IBM SPSS Modeler 18.2.1 — węzły źródłowe, procesowe i wyników

Uwaga

Przed skorzystaniem z niniejszych informacji oraz produktu, którego one dotyczą, należy zapoznać się z informacjami zamieszczonymi w sekcji “Uwagi” na stronie 395.

Informacje o produkcie

Niniejsze wydanie publikacji dotyczy wersji 18, wydania 2, modyfikacji 0 produktu IBM SPSS Modeler oraz wszystkich następnych wydań i modyfikacji do czasu, aż w kolejnym wydaniu publikacji zostanie zawarta informacja o stosownej zmianie.

Spis treści

Przedmowa

IBM® SPSS Modeler to oferowane przez IBM zaawansowane środowisko eksploracji danych. SPSS Modeler pomaga przedsiębiorstwom i instytucjom w rozwijaniu relacji z klientami i obywatelami w oparciu o pogłębioną interpretację dostępnych danych. Organizacje korzystają z wiedzy uzyskanej dzięki programowi SPSS Modeler w bardzo szerokim spektrum zastosowań, m.in. do zatrzymywania najbardziej wartościowych klientów, określania możliwości sprzedaży wiązanej, przyciągania nowych klientów, wykrywania oszustw, ograniczania ryzyka i podnoszenia jakości usług publicznych.

Interfejs graficzny produktu SPSS Modeler zachęca użytkowników, aby wykorzystywali specjalistyczną wiedzę, dzięki której możliwe będzie opracowanie bardziej wydajnych modeli predykcyjnych i skrócenie czasu potrzebnego do uzyskania rozwiązania. SPSS Modeler oferuje wiele technik modelowania, takich jak predykcja, klasyfikacja, segmentacja i algorytmy do wykrywania związków. Po utworzeniu modeli program IBM SPSS Modeler Solution Publisher umożliwia udostępnienie ich osobom podejmującym decyzje w całym przedsiębiorstwie lub zapisanie w bazie danych.

Informacje o programie IBM Business Analytics

Oprogramowanie IBM Business Analytics dostarcza kompletne, spójne i dokładne informacje, na których mogą polegać osoby decyzyjne chcące polepszyć wyniki biznesowe. Wszechstronne portfolio obejmujące moduły: analiza biznesowa, analiza prognostyczna, zarządzanie wynikami i strategiami finansowymi oraz aplikacje analityczne, zapewnia jasny, natychmiastowy i pozwalający na podjęcie działań wgląd w bieżące wyniki oraz daje możliwość przewidywania przyszłych wyników. W połączeniu z licznymi rozwiązaniami branżowymi, sprawdzonymi praktykami i profesjonalnymi usługami, organizacje o różnych rozmiarach mogą wspomagać najwyższą produktywność, w sposób pewny zautomatyzować decyzje i uzyskać lepsze wyniki.

Oprogramowanie IBM SPSS Predictive Analytics będące częścią tego portfolio wspomaga organizacje w zakresie przewidywania przyszłych zdarzeń oraz proaktywnie wpływać na ten wgląd z korzyścią dla wyników finansowych. Klienci komercyjni, rządowi i uczelnie na całym świecie polegają na technologii IBM SPSS zapewniającej przewagę konkurencyjną, dzięki której przyciągają, zatrzymują i pozyskują nowych klientów, walcząc z nieuczciwością i ograniczając ryzyko. Wdrażając oprogramowanie IBM SPSS do swojej codziennej działalności, organizacje stają się przewidującymi przedsiębiorstwami, zdolnymi do zarządzania i automatyzacji decyzji w celu realizacji celów biznesowych i osiągnięcia mierzalnej przewagi konkurencyjnej. W celu uzyskania dalszych informacji lub skontaktowania się z przedstawicielem, proszę wejść na stronę http://www.ibm.com/spss.

Wsparcie techniczne

Wsparcie techniczne jest dostępne w celu zapewnienia klientom obsługi technicznej. Klienci mogą się kontaktować z działem Wsparcia technicznego w celu uzyskania pomocy dotyczącej korzystania z produktów IBM lub pomocy w instalacji dla jednego z obsługiwanych środowisk sprzętowych. Aby skontaktować się z działem Wsparcia technicznego, wejdź na stronę internetową IBM pod adresem http://www.ibm.com/support. W przypadku prośby o pomoc, należy przygotować swoje dane identyfikacyjne, dane swojej organizacji, a także dane dotyczące usług wsparcia.

Rozdział 1. O programie IBM SPSS Modeler

IBM SPSS Modeler to zestaw narzędzi do eksploracji danych. Produkt umożliwia szybkie opracowywanie modeli predykcyjnych przy wykorzystaniu wiedzy specjalistycznej i stosowanie tych modeli w procesach biznesowych jako wsparcia przy podejmowaniu decyzji. Rozwiązania zawarte w oprogramowaniu IBM SPSS Modeler zapewniają możliwość wykorzystywania branżowego modelu CRISP-DM i pozwalają na obsługę całego procesu eksploracji danych: od pozyskiwania danych do uzyskiwania lepszych wyników biznesowych.

Oprogramowanie IBM SPSS Modeler umożliwia korzystanie z wielu metod modelowania opartych na sztucznej inteligencji, uczeniu maszynowym i statystykach. Metody dostępne na palecie Modelowanie pozwalają na ekstrahowanie nowych informacji z danych i tworzenie modeli predykcyjnych. Każda metoda ma określone mocne strony i jest dostosowana do rozwiązywania określonych problemów.

Oprogramowanie SPSS Modeler można zakupić jako produkt samodzielny lub jako program kliencki używany wraz z oprogramowaniem SPSS Modeler Server. Dostępnych jest wiele opcji dodatkowych, które przedstawiono w kolejnych rozdziałach. Aby uzyskać więcej informacji, patrz https://www.ibm.com/analytics/us/en/technology/spss/.

Produkty IBM SPSS Modeler

Rodzina produktów IBM SPSS Modeler i towarzyszącego im oprogramowania składa się z elementów przedstawionych poniżej.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console (jest częścią produktu IBM SPSS Deployment Manager)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Modeler Server adapters for IBM SPSS Collaboration and Deployment Services

IBM SPSS Modeler

Oprogramowanie SPSS Modeler to w pełni funkcjonalna wersja produktu instalowana i uruchamiana na komputerze osobistym. Oprogramowanie SPSS Modeler można uruchomić lokalnie jako produkt samodzielny lub korzystać z niego w trybie rozproszonym wraz z serwerem IBM SPSS Modeler Server. Tego typu rozwiązanie zapewnia zwiększenie wydajności obsługi dużych zbiorów danych.

Dzięki oprogramowaniu SPSS Modeler można szybko tworzyć dokładne modele predykcyjne, stosując intuicyjne metody niewymagające umiejętności programowania. Unikatowy interfejs graficzny pozwala na wizualizowanie procedur eksploracji danych. Zaawansowane metody opracowywania analiz dostępne w programie umożliwiają określanie wcześniej niezauważalnych wzorców i trendów zawartych w danych. Użytkownik może modelować wyniki i poznawać czynniki wpływające na ich wartości. W ten sposób można wykorzystywać nowe szanse biznesowe i obniżać ryzyko.

Dostępne są dwie edycje oprogramowania SPSS Modeler: SPSS Modeler Professional oraz SPSS Modeler Premium. Więcej informacji można znaleźć w temacie “Wydania programu IBM SPSS Modeler” na stronie 2.

IBM SPSS Modeler Server

Oprogramowanie SPSS Modeler działa w oparciu o architekturę klient-serwer, w której żądania wymagające zaangażowania dużych zasobów kierowane są do zaawansowanego oprogramowania serwerowego. Takie rozwiązanie umożliwia bardziej wydajną obsługę dużych zbiorów danych.

SPSS Modeler Server to produkt wymagający dodatkowej licencji, działający stale na serwerze w trybie analizy rozproszonej. Współpracuje on z co najmniej jedną instalacją oprogramowania IBM SPSS Modeler. W ten sposób oprogramowanie SPSS Modeler Server poprawia wydajność podczas obsługi dużych zbiorów danych, ponieważ operacje wymagające dużej mocy obliczeniowej można wykonywać na serwerze bez potrzeby pobierania danych na komputer kliencki. Oprogramowanie IBM SPSS Modeler Server optymalizuje również obsługę SQL i funkcje modelowania wewnątrz bazy danych, co dodatkowo zwiększa wydajność działania i sprzyja automatyzacji pracy.

IBM SPSS Modeler Administration Console

Oprogramowanie Modeler Administration Console to graficzny interfejs użytkownika służący do obsługi wielu opcji konfiguracji SPSS Modeler Server, które można dostosować również za pomocą pliku opcji. Konsola udostępniona w aplikacji IBM SPSS Deployment Manager pozwala na monitorowanie i konfigurowanie instalacji SPSS Modeler Server. Konsola jest dostępna bez dodatkowych opłat dla aktualnych użytkowników SPSS Modeler Server. Aplikację można zainstalować tylko na komputerach z systemem Windows, jednak administrować można serwerem zainstalowanym na dowolnej obsługiwanej platformie.

IBM SPSS Modeler Batch

Eksploracja danych jest zazwyczaj procesem interaktywnym, jednak oprogramowanie SPSS Modeler można też uruchomić z poziomu wiersza komend i zrezygnować z używania graficznego interfejsu użytkownika. Niekiedy użytkownik wykonuje długotrwałe lub powtarzalne zadania, które mogą być realizowane bez nadzoru. Oprogramowanie SPSS Modeler Batch to specjalna wersja produktu pozwalająca na wykonywanie wszystkich funkcji analitycznych SPSS Modeler bez potrzeby używania standardowego interfejsu użytkownika. Oprogramowanie SPSS Modeler Server jest wymagane do korzystania z aplikacji SPSS Modeler Batch.

IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher umożliwia tworzenie spakowanych wersji strumieni programu SPSS Modeler, które można uruchamiać za pomocą zewnętrznych środowisk wykonawczych lub osadzać w aplikacji zewnętrznej. W ten sposób można publikować i wdrażać pełne strumienie SPSS Modeler w celu używania ich w środowiskach, w których nie zainstalowano programu SPSS Modeler. SPSS Modeler Solution Publisher jest dystrybuowany jako część produktu IBM SPSS Collaboration and Deployment Services - Scoring, który do działania wymaga oddzielnej licencji. Wraz z licencją użytkownik otrzymuje oprogramowanie SPSS Modeler Solution Publisher Runtime umożliwiające uruchamianie opublikowanych strumieni.

Więcej informacji na temat programu SPSS Modeler Solution Publisher znajduje się w dokumentacji produktu IBM SPSS Collaboration and Deployment Services. W Centrum wiedzy IBM SPSS Collaboration and Deployment Services dostępne są sekcje "IBM SPSS Modeler Solution Publisher" oraz "IBM SPSS Analytics Toolkit".

IBM SPSS Modeler Server Adapters for IBM SPSS Collaboration and Deployment Services

Dostępnych jest wiele adapterów dla IBM SPSS Collaboration and Deployment Services, które umożliwiają współpracę programów SPSS Modeler i SPSS Modeler Server z repozytorium IBM SPSS Collaboration and Deployment Services. Dzięki temu strumień SPSS Modeler wdrożony w repozytorium można udostępnić wielu użytkownikom lub uzyskać do niego dostęp z poziomu uproszczonej aplikacji klienckiej IBM SPSS Modeler Advantage. Adapter należy zainstalować na systemie hostującym repozytorium.

Wydania programu IBM SPSS Modeler

Dostępne są następujące wydania oprogramowania SPSS Modeler.

SPSS Modeler Professional

Oprogramowanie SPSS Modeler Professional zapewnia wszystkie narzędzia wymagane do obsługi większości typów danych ustrukturyzowanych, takich jak np. zachowania i interakcje śledzone w systemach CRM, dane demograficzne, zachowania zakupowe i dane sprzedażowe.

SPSS Modeler Premium

Oprogramowanie SPSS Modeler Premium wymaga oddzielnej licencji. Dzięki temu rozwiązaniu oprogramowanie SPSS Modeler Professional może obsługiwać wyspecjalizowane dane oraz nieustrukturyzowane dane tekstowe. SPSS Modeler Premium obejmuje IBM SPSS Modeler Text Analytics:

Program **IBM SPSS Modeler Text Analytics** korzysta z zaawansowanych rozwiązań lingwistycznych oraz przetwarzania języka naturalnego w celu szybkiego przetwarzania różnego rodzaju nieustrukturyzowanych danych tekstowych, wyodrębniania i porządkowania kluczowych pojęć oraz grupowania tych pojęć w kategorie. Wyodrębnione pojęcia i kategorie można łączyć z istniejącymi danymi ustrukturyzowanymi, takimi jak dane demograficzne, a następnie stosować w celu modelowania, korzystając z produktu IBM SPSS Modeler i zawartego w nim pełnego pakietu narzędzi do eksploracji danych, aby w rezultacie takiego połączenia podejmować lepsze decyzje przy zmniejszonej ilości zakłóceń.

IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription oferuje te same funkcje analiz predykcyjnych, co tradycyjny klient IBM SPSS Modeler. Użytkownicy edycji Subscription mogą regularnie pobierać aktualizacje produktu.

Dokumentacja

Dokumentacja jest dostępna w programie SPSS Modeler z poziomu menu Pomoc. Spowoduje to otwarcie internetowego Centrum Wiedzy, które jest zawsze dostępne poza produktem.

Pełna dokumentacja dla każdego produktu (w tym instrukcje instalacji) jest również dostępna w formacie PDF, w osobnym skompresowanym folderze, jako część materiałów do pobrania z produktem. Najnowsze dokumenty PDF można także pobrać z Internetu pod adresem http://www.ibm.com/support/docview.wss?uid=ibm10874788.

Dokumentacja SPSS Modeler Professional

Pakiet dokumentacji produktu SPSS Modeler Professional (bez instrukcji instalacyjnych) zawiera następujące publikacje.

- **IBM SPSS Modeler — podręcznik użytkownika.** Ogólne wprowadzenie do obsługi oprogramowania SPSS Modeler, w tym opisy procedur tworzenia strumieni danych, obsługi braków danych, tworzenia wyrażeń CLEM pracy z projektami i raportami, a także przygotowywania strumieni do wdrożenia w IBM SPSS Collaboration and Deployment Services lub IBM SPSS Modeler Advantage.
- **IBM SPSS Modeler — węzły źródłowe, procesowe i wyników.** Opisy wszystkich węzłów używanych do odczytywania, przetwarzania i tworzenia wynikowych postaci danych w różnych formatach. Czyli wszystkich węzłów poza węzłami modelowania.
- **IBM SPSS Modeler — węzły modelowania.** Opisy wszystkich węzłów używanych do tworzenia modeli eksploracji danych. Oprogramowanie IBM SPSS Modeler umożliwia korzystanie z wielu metod modelowania opartych na sztucznej inteligencji, uczeniu maszynowym i statystykach.
- **IBM SPSS Modeler — podręcznik zastosowań.** Przykłady zawarte w niniejszym przewodniku stanowią skrócone informacje związane z konkretnymi metodami i technikami modelowania. Wersja elektroniczna tego podręcznika jest również dostępna z poziomu menu Pomoc. Więcej informacji można znaleźć w temacie “Przykłady aplikacji” na stronie 4.
- **IBM SPSS Modeler — podręcznik tworzenia skryptów w języku Python i automatyzacji.** Informacje na temat automatyzacji działania systemu za pomocą skryptów Python wraz z właściwościami służącymi do obsługi węzłów i strumieni.
- **IBM SPSS Modeler — podręcznik wdrażania.** Informacje na temat uruchamiania strumieni IBM SPSS Modeler przedstawione w postaci krokowych operacji wykonywanych podczas przetwarzania zadań w programie IBM SPSS Deployment Manager.
- **IBM SPSS Modeler CLEF Developer's Guide.** Z oprogramowaniem CLEF można zintegrować inne programy pozwalające na przetwarzanie danych lub obsługę algorytmów modelujących w postaci węzłów w IBM SPSS Modeler.

- **IBM SPSS Modeler — podręcznik eksploracji w bazie danych.** Informacje na temat wydajnego wykorzystywania bazy danych w celu zwiększenia wydajności i zakresu funkcji analitycznych za pomocą algorytmów innych firm.
- **IBM SPSS Modeler Server — podręcznik administracji i wydajności.** Informacje na temat konfiguracji i funkcji administracyjnych w oprogramowaniu IBM SPSS Modeler Server.
- **IBM SPSS Deployment Manager — Podręcznik użytkownika.** Informacje dotyczące korzystania z interfejsu użytkownika konsoli administracyjnej zawartej w aplikacji Deployment Manager podczas monitorowania i konfigurowania serwera IBM SPSS Modeler Server.
- **IBM SPSS Modeler — podręcznik CRISP-DM.** Szczegółowy podręcznik metodologii CRISP-DM w kontekście eksploracji danych za pomocą oprogramowania SPSS Modeler.
- **IBM SPSS Modeler Batch — podręcznik użytkownika.** Pełny podręcznik obsługi oprogramowania IBM SPSS Modeler w trybie wsadowym obejmujący szczegółowe informacje na temat pracy w trybie wsadowym i korzystania z argumentów z poziomu wiersza komend. Ten podręcznik jest dostępny tylko w formacie PDF.

Dokumentacja SPSS Modeler Premium

Pakiet dokumentacji produktu SPSS Modeler Premium (bez instrukcji instalacyjnych) zawiera następujące publikacje.

- **SPSS Modeler Text Analytics — podręcznik użytkownika.** Informacje na temat używania analiz tekstu za pomocą oprogramowania SPSS Modeler obejmują procedury dotyczące węzłów eksploracji tekstu, interaktywnego pulpitu roboczego, szablonów oraz innych zasobów.

Przykłady aplikacji

Podczas gdy narzędzia do eksploracji danych w programie SPSS Modeler mogą pomóc w rozwiązaniu szeregu problemów biznesowych i organizacyjnych, przykłady aplikacji udostępniają krótkie, ukierunkowane wprowadzenia do konkretnych metod i technik modelowania. Używane tutaj zestawy danych są znacznie mniejsze niż ogromne składnice danych zarządzane przez programy do eksploracji danych, lecz używane koncepcje i metody są skalowalne odpowiednio do potrzeb rzeczywistych aplikacji.

Dostęp do przykładów można uzyskać, klikając opcję **Przykłady aplikacji** w menu Pomoc programu SPSS Modeler.

Pliki danych i przykładowe strumienie są instalowane w folderze Dema, w katalogu instalacyjnym produktu. Aby uzyskać więcej informacji, zobacz “Folder Demos”.

Przykłady modelowania w bazach danych. Przykłady zamieszczono w publikacji *IBM SPSS Modeler — podręcznik eksploracji w bazie danych*.

Przykłady skryptów. Przykłady zamieszczono w publikacji *IBM SPSS Modeler — podręcznik tworzenia skryptów w języku Python i automatyzacji*.

Folder Demos

Pliki danych i przykładowe strumienie używane z przykładami do aplikacji są instalowane w folderze Demos wewnątrz katalogu instalacyjnego produktu (na przykład: C:\Program Files\IBM\SPSS\Modeler\<version>\Demos). Dostęp do tego folderu można także uzyskać z grupy programów IBM SPSS Modeler w menu Start systemu Windows lub klikając opcję Dema na liście ostatnich katalogów w oknie dialogowym **Plik > Otwórz strumień**.

Monitorowanie wykorzystania licencji

Podczas pracy z produktem SPSS Modeler wykorzystanie licencji jest monitorowanie i regularnie rejestrowane. Metryka wykorzystania licencji nosi nazwę *AUTHORIZED_USER* (użytkownik autoryzowany) lub *CONCURRENT_USER* (użytkownik pracujący jednocześnie), a typ rejestrowanej metryki zależy od typu licencji na produkt SPSS Modeler, którą posiada użytkownik.

Generowane pliki dzienników mogą być przetwarzane przez program IBM License Metric Tool, z którego uzyskać można raporty o wykorzystaniu licencji.

Pliki dzienników wykorzystania licencji są tworzone w tym samym katalogu, w którym zapisywane są dzienniki klienta SPSS Modeler (domyślnie %ALLUSERSPROFILE%/IBM/SPSS/Modeler/<wersja>/log).

Rozdział 2. Węzły źródeł

Przegląd

Węzły źródłowe umożliwiają importowanie danych zapisanych w wielu formatach, w tym jako pliki płaskie, IBM SPSS Statistics (.sav), SAS, Microsoft Excel oraz relacyjne bazy danych zgodne z ODBC. Można również wygenerować dane syntetyczne, korzystając z węzła danych użytkownika.

Paleta Źródła zawiera następujące węzły:

Źródło Analytic Server umożliwia uruchamianie strumienia w systemie Hadoop Distributed File System (HDFS). Informacje zawarte w źródle danych Analytic Server pochodzą z różnych obszarów, takich jak pliki tekstowe i bazy danych. Więcej informacji można znaleźć w temacie “Węzeł źródłowy Analytic Server” na stronie 13.

Węzeł Baza danych umożliwia importowanie danych z różnych innych pakietów za pośrednictwem ODBC (Open Database Connectivity), np. Microsoft SQL Server, Db2, Oracle inne. Więcej informacji można znaleźć w temacie “Węzeł źródłowy bazy danych” na stronie 17.

Węzeł Plik zmiennych odczytuje dane z plików tekstowych — czyli z plików, których rekordy zawierają stałą liczbę zmiennych, ale różnią się liczbą znaków. Ten węzeł jest również przydatny w przypadku zmiennych z tekstem nagłówka o ustalonej długości i niektórych typów adnotacji. Więcej informacji można znaleźć w temacie “Węzeł Plik zmiennych” na stronie 26.

Węzeł Plik kolumnowy importuje dane z plików tekstowych ze stałymi polami — czyli z plików, w których pola nie są separowane, ale rozpoczynają się w tym samym miejscu i mają stałą długość. Dane wygenerowane maszynowe lub pochodzące ze starszych wersji często są zapisywane w formacie ze stałymi polami. Więcej informacji można znaleźć w temacie “Węzeł Plik kolumnowy” na stronie 30.

Węzeł Plik Statistics odczytuje dane z pliku w formacie *.sav* lub *.zsav* używanym przez program IBM SPSS Statistics, jak również pliki pamięci podręcznej zapisane w programie IBM SPSS Modeler, które również używają tego samego formatu.

Węzeł Data Collection importuje dane z ankiet w różnych formatach używanych w oprogramowaniu do badań rynku zgodnym z modelem danych Data Collection. Aby możliwe było korzystanie z tego węzła, konieczne jest zainstalowanie programu Data Collection Developer Library. Więcej informacji można znaleźć w temacie “Węzeł Data Collection” na stronie 33.

Węzeł źródłowy IBM Cognos importuje dane z baz danych Cognos Analytics.

Węzeł źródłowy IBM Cognos TM1 importuje dane z baz danych Cognos TM1.

Węzeł Plik SAS importuje dane SAS do programu IBM SPSS Modeler. Więcej informacji można znaleźć w temacie “Węzeł źródłowy SAS” na stronie 43.

Węzeł Plik Excel importuje dane z programu Microsoft Excel w pliku w formacie .xlsx. Źródło danych ODBC nie jest wymagane. Więcej informacji można znaleźć w temacie “Węzeł źródłowy programu Excel” na stronie 44.

Węzeł źródłowy XML importuje dane w formacie XML do strumienia. Można zaimportować jeden plik lub wszystkie pliki z katalogu. Można opcjonalnie określić plik schematu, z którego odczytywana będzie struktura XML.

Węzeł Dane niestandardowe udostępnia prosty sposób na utworzenie danych syntetycznych — od podstaw lub poprzez zmianę istniejących danych. Jest to przydatne na przykład podczas tworzenia testowego zbioru danych do modelowania. Więcej informacji można znaleźć w temacie “Węzeł Dane niestandardowe” na stronie 47.

Węzeł Symulacje Generowanie zapewnia łatwy sposób na wygenerowanie danych objętych symulacją — od podstaw, korzystając z rozkładów statystycznych określonych przez użytkownika lub automatycznie, korzystając z rozkładów uzyskanych po uruchomieniu węzła Symulacje Dopasowanie dla istniejących danych historycznych. Jest to przydatne, kiedy ma zostać przeprowadzona ocena wyniku modelu predykcyjnego przy braku pewności dla danych wejściowych modelu.

Węzeł Widok danych może być używany do uzyskiwania dostępu do źródeł danych zdefiniowanych w widokach danych analitycznych IBM SPSS Collaboration and Deployment Services. Widok danych analitycznych definiuje standardowy interfejs i tworzy powiązania wielu fizycznych źródeł danych z tym interfejsem. Więcej informacji można znaleźć w temacie “Węzeł widoku danych” na stronie 65.

Węzeł źródłowy Dane geoprzestrzenne umożliwia przeniesienie danych z mapy lub danych przestrzennych do sesji eksploracji danych. Więcej informacji można znaleźć w temacie “Węzeł źródłowy Dane geoprzestrzenne” na stronie 66.

Węzeł źródłowy JSON importuje dane z pliku JSON. Więcej informacji można znaleźć w “Węzeł źródłowy JSON” na stronie 67.

Aby rozpocząć strumień, należy dodać węzeł źródłowy do obszaru roboczego strumienia. Następnie należy dwukrotnie kliknąć węzeł, aby otworzyć jego okno dialogowe. Różne karty w oknie dialogowym umożliwiają odczytywanie danych, wyświetlanie pól i wartości oraz ustawianie różnych opcji, takich jak filtry, typy danych, role zmiennych i sprawdzanie braków danych.

Ustawienia składowania i formatowania zmiennej

Opcje na karcie Dane dla węzłów Plik kolumnowy, Plik zmiennych, Źródło XML i Dane niestandardowe umożliwiają określenie typu składowania dla zmiennych w przypadku ich importowania lub tworzenia w programie IBM SPSS Modeler. W przypadku węzłów Plik kolumnowy, Plik zmiennych i Dane niestandardowe można również określić formatowanie zmiennej oraz inne metadane.

W przypadku danych odczytanych z innych źródeł składowanie jest ustalane automatycznie, ale można je zmienić za pomocą funkcji przekształcenia, np. to_integer, w węźle wypełniania lub wyliczania.

Zmienna Kolumna **Zmienna** umożliwia wyświetlanie i wybieranie zmiennych z bieżącego zbioru danych.

Nadpisz Zaznaczenie pola wyboru w kolumnie **Nadpisz** spowoduje aktywowanie opcji w kolumnach **Składowanie** i **Format wejściowy**.

Składowanie danych

Składowanie to sposób przechowywania danych w zmiennej. Przykładowo w zmiennej zawierającej wartości 1 i 0 składowane są dane w postaci liczb całkowitych. Różni się to od poziomu pomiaru, który opisuje użycie danych i nie wpływa na składowanie. Można na przykład ustawić poziom pomiaru dla zmiennej całkowitej zawierającej wartości 1 i 0 jako *Flaga*. Zwykle oznacza to, że 1 = *Prawda*, a 0 = *Fałsz*. Składowanie musi być określone w źródle, natomiast poziom pomiaru można zmienić za pomocą węzła Typy w dowolnym miejscu w strumieniu. Więcej informacji można znaleźć w temacie “Poziomy pomiaru” na stronie 139.

Dostępne typy składowania to:

- **Łańcuch** Używany w przypadku zmiennych zawierających dane nienumeryczne, zwane również danymi alfanumerycznymi. Łańcuch może zawierać dowolną sekwencję znaków, np. *fred*, *Klasa 2* lub *1234*. Należy pamiętać, że liczby użyte w łańcuchach nie mogą być wykorzystywane do obliczeń.
- **Liczba całkowita** Zmienna, której wartości są liczbami całkowitymi.
- **Liczba rzeczywista** Wartości są liczbami, które mogą mieć miejsca dziesiętne (bez ograniczenia do liczb całkowitych). Format wyświetlania jest określany w oknie dialogowym Właściwości strumienia i może być zastąpiony dla pojedynczych zmiennych w węźle Typy (karta Format).
- **Data** Wartości daty określone w standardowym formacie, takie jak rok, miesiąc i dzień (np. 2007-09-26). Konkretny format jest określany w oknie dialogowym Właściwości strumienia.
- **Czas** Czas mierzony jako czas trwania. Przykładowo połączenie z serwisem trwające 1 godzinę, 26 minut i 38 sekund może być zapisane jako 01:26:38, w zależności od obowiązującego formatu czasu określonego w oknie dialogowym Właściwości strumienia.
- **Znacznik czasu** Wartości składające się z daty i czasu, na przykład 2007–09–26 09:04:00, ponownie w zależności od obowiązujących formatów daty i czasu określonych w oknie dialogowym Właściwości strumienia. Należy pamiętać, że konieczne może być ujęcie wartości znacznika czasu w podwójnym cudzysłowie, aby były interpretowane jako pojedyncza wartość, a nie jako osobne wartości daty i czasu. (Dotyczy to na przykład sytuacji, kiedy wartości są wprowadzane w węźle Dane niestandardowe).

- **Lista** Zmienna składowania Lista wprowadzona w programie SPSS Modeler, wersja 17, wraz z nowymi poziomami pomiaru Geoprzestrzenny i Przedziałowy, zawiera wiele wartości dla pojedynczego rekordu. Dostępne są wersje list dla wszystkich pozostałych typów składowania.

Tabela 1. Lista ikon typów składowania

Ikona	Typ składowania
[]	Lista łańcuchów
[]	Lista liczb całkowitych
[]	Lista liczb rzeczywistych
[]	Lista godzin
[]	Lista dat
[]	Lista znaczników czasu
[[]]	Lista o głębokości większej niż zero

Dodatkowo do użycia z poziomem pomiaru Przedziałowy dostępne są wersje listy następujących poziomów pomiaru.

Tabela 2. Lista ikon poziomu pomiaru

Ikona	Poziom pomiaru
[]	Lista ilościowych
[]	Lista jakościowych
[]	Lista flag
[]	Lista nominalnych
[]	Lista porządkowych

Listy mogą być importowane do programu SPSS Modeler w jednym z trzech węzłów źródłowych (Analytic Server, Geoprzestrzenny lub Plik zmiennych) lub mogą być utworzone na podstawie strumieni za pośrednictwem węzłów działania zmiennych Wyliczanie lub Wypełnianie.

Więcej informacji na temat list i ich interakcji z poziomami pomiaru Przedziałowy i Geoprzestrzenny zawiera sekcja “Składowanie listy i powiązane poziomy pomiaru” na stronie 12.

Przekształcanie typów składowania. Można przekształcić typ składowania dla zmiennej, korzystając z różnych funkcji przekształcania, takich jak to_string i to_integer, dostępnych w węźle wypełniania. Więcej informacji można znaleźć w temacie “Przekształcanie sposobu składowania za pomocą węzła Wypełnianie” na stronie 162. Funkcje przekształcania (i inne funkcje wymagające określonego typu danych wejściowych, np. wartość daty lub czasu) są uzależnione od bieżących formatów określonych w oknie dialogowym Właściwości strumienia. Na przykład, aby wykonać przekształcenie zmiennej łańcuchowej o wartościach *Sty 2018*, *Lut 2018* itd. do postaci składowania daty, jako domyślny format daty strumienia należy wybrać **MIE RRRR**. Funkcje przekształcania są również dostępne z węzła wyliczeń i umożliwiają tymczasowe przekształcenie podczas wyliczania. Węzła wyliczeń można także użyć do wykonywania innych działań, takich jak rekodowanie zmiennych łańcuchowych przez wartości jakościowe. Więcej informacji można znaleźć w temacie “Rekodowanie wartości za pomocą węzła wyliczeń” na stronie 161.

Wczytywanie danych mieszanych. Należy zwrócić uwagę, że podczas wczytywania zmiennych z liczbowym typem składowania (liczby całkowite, rzeczywiste, czas, znacznik czasu lub data) wszelkie wartości nieliczbowe są zamieniane na null lub braki systemowe. Wynika to z faktu, że w odróżnieniu od niektórych aplikacji produkt IBM SPSS Modeler nie zezwala na przechowywanie w zmiennej danych różnego typu. Aby uniknąć takiej sytuacji, wszelkie zmienne z danymi mieszanymi należy wczytywać jako łańcuchy, zmieniając w razie potrzeby typ składowania w węźle źródłowym lub aplikacji zewnętrznej.

Format wejściowy zmiennej (tylko węzły Plik kolumnowy, Plik zmiennych i Dane niestandardowe)

W przypadku wszystkich typów składowania, z wyjątkiem łańcucha i liczb całkowitych, można określić opcje formatowania dla wybranej zmiennej, wybierając je z listy rozwijanej. Przykładowo podczas scalania danych z różnych lokalizacji konieczne może być określenie kropki (.) jako separatora dziesiętnego dla jednej zmiennej, podczas gdy druga będzie wymagała przecinka.

Opcje wejściowe określone w węźle źródłowym zastępują opcje formatowania określone w oknie dialogowym właściwości strumienia; jednak nie pozostają później w strumieniu. Ich zadaniem jest poprawne przeanalizowanie danych wejściowych w oparciu o wiedzę na temat danych. Określone formaty są używane jako wskazówka do analizowania danych podczas ich wczytywania do programu IBM SPSS Modeler, nie do określania sposobu ich formatowania po wczytaniu do IBM SPSS Modeler. Aby określić formatowanie dla zmiennej w innym miejscu strumienia, należy użyć karty Format w węźle typu. Więcej informacji można znaleźć w temacie “Karta ustawień formatu zmiennej” na stronie 150.

Opcje różnią się w zależności od typu składowania. Na przykład dla typu składowania liczb rzeczywistych jako separator dziesiętny można wybrać opcje **Kropka (.)** lub **Przecinek (,)**. W przypadku zmiennych znacznika czasu po wybraniu opcji **Określ** z listy rozwijanej otwierane jest osobne okno dialogowe. Więcej informacji można znaleźć w temacie “Ustawianie opcji formatu zmiennej” na stronie 151.

Dla wszystkich typów składowania można również wybrać opcję **Jak dla strumienia**, aby podczas importu używać wartości domyślnych strumienia. Ustawienia strumienia są określane w oknie dialogowym właściwości strumienia.

Opcje dodatkowe

Korzystając z karty Dane, można określić kilka dodatkowych opcji:

- Aby wyświetlić ustawienia składowania dla danych, które nie są już połączone za pośrednictwem bieżącego węzła (na przykład dane uczące), należy wybrać opcję **Widok ustawień niewykorzystanych zmiennych**. Wcześniejsze zmienne można skasować, klikając przycisk **Wyczyść**.
- Korzystając z tego okna dialogowego, w dowolnym czasie można kliknąć przycisk **Odśwież**, aby ponownie załadować zmienne ze źródła danych. Jest to przydatne w przypadku zmiany połączeń danych na węzeł źródłowy lub podczas pracy na różnych kartach okna dialogowego.

Składowanie listy i powiązane poziomy pomiaru

Zmienna składowania Lista wprowadzona w programie SPSS Modeler, wersja 17, do pracy z nowymi poziomami pomiaru Geoprzestrzenny i Przedziałowy, zawiera wiele wartości dla pojedynczego rekordu. Listy ujmuje się w nawiasy kwadratowe ([]). Przykłady listy: [1,2,4,16] i ["abc", "def"].

Listę można zaimportować do programu SPSS Modeler w jednym z trzech węzłów źródłowych (Analytic Server, Geoprzestrzenny lub Plik zmiennych), które tworzone są na podstawie strumieni za pośrednictwem węzłów działania zmiennych Wyliczanie lub Wypełnianie lub wygenerowane za pośrednictwem węzła Łączenie w przypadku korzystania z metody łączenia Warunek z rangowaniem.

Dla list określana jest głębokość; na przykład prosta lista z elementami ujętymi w pojedyncze nawiasy kwadratowe, w formacie [1,3], jest rejestrowana w programie IBM SPSS Modeler jako lista o głębokości zero. Oprócz prostych list o głębokości zero można korzystać z list zagnieżdżonych, w których każda wartość sama stanowi listę.

Głębokość list zagnieżdżonych jest uzależniona od powiązanego poziomu pomiaru. Jeśli poziom nie jest określony, nie ma ustawionego limitu głębokości, w przypadku poziomu przedziałowego głębokość wynosi zero, a w przypadku poziomu geoprzestrzennego głębokość musi wynosić od zera do dwóch włącznie, co zależy od liczby zagnieżdżonych elementów.

Dla list o głębokości zero można ustawić poziom pomiaru geoprzestrzenny lub przedziałowy. Oba te poziomy są nadrzędnymi poziomami pomiaru, a informacje o poziomie podrzędnym pomiaru można wprowadzić w oknie dialogowym Wartości. Przedziałowy poziom podrzędny pomiaru określa poziom pomiaru elementów znajdujących się na liście. Wszystkie poziomy pomiaru (z wyjątkiem tych bez określonego typu i geoprzestrzennego) są dostępne jako poziomy podrzędne dla przedziałów. Geoprzestrzenny poziom pomiaru obejmuje sześć poziomów podrzędnych: Punkt, Łańcuch, Wielokąt, Multipunkt, Multiłańcuch i Multiwielokąt; więcej informacji zawiera temat “Geoprzestrzenne podpoziomy pomiarów” na stronie 141.

Uwaga: Przedziałowy poziom pomiaru może być stosowany tylko w przypadku list o głębokości 0, geoprzestrzenny poziom pomiaru może być stosowany tylko w przypadku list o maksymalnej głębokości wynoszącej 2, a poziom pomiaru bez określonego typu może być stosowany dla list o dowolnej głębokości.

Poniższy przykład przedstawia różnicę pomiędzy listą o głębokości zero a listą zagnieżdżoną za pośrednictwem struktury geoprzestrzennych podrzędnych poziomów pomiaru Punkt i Łańcuch:

- Geoprzestrzenny podrzędny poziom pomiaru Punkt ma głębokość zmiennej wynoszącą zero:
  [1,3] dwie współrzędne
  [1,3,-1] trzy współrzędne
- Geoprzestrzenny podrzędny poziom pomiaru Łańcuch ma głębokość zmiennej wynoszącą jeden:
  [[1,3], [5,0]] dwie współrzędne
  [[1,3,-1], [5,0,8]] trzy współrzędne

Zmienna Punkt (o głębokości zero) jest zwykłą listą, na której każda wartość składa się z dwóch lub trzech współrzędnych. Zmienna Łańcuch (o głębokości jeden) jest listą punktów, a każdy punkt składa się z dodatkowej serii wartości listy.

Więcej informacji na temat tworzenia listy zawiera temat “Wyliczanie zmiennej listy lub geoprzestrzennej” na stronie 158.

Nieobsługiwane znaki sterujące

Niektóre procesy w programie SPSS Modeler nie mogą obsługiwać danych, które zawierają różne znaki sterujące. Jeśli dane zawierają takie znaki, może zostać wyświetlony następujący komunikat o błędzie:

```
W wartościach zmiennej {0} napotkano nieobsługiwane znaki sterujące
```

Nieobsługiwane znaki to: od 0x0 do 0x3F włącznie oraz 0x7F; znaki tabulacji (0x9(\t)), nowego wiersza (0xA(\n)) i powrotu karetki (0xD(\r)) nie powodują problemów.

Jeśli wyświetlony zostanie komunikat o błędzie dotyczący nieobsługiwanych znaków, w strumieniu, za węzłem źródłowym, należy użyć węzła wypełniania oraz wyrażenia CLEM **stripctrlchars**, aby zastąpić te znaki.

Węzeł źródłowy Analytic Server

Źródło Analytic Server umożliwia uruchamianie strumienia w systemie Hadoop Distributed File System (HDFS). Informacje zawarte w źródle danych Analytic Server mogą pochodzić z różnych obszarów, takich jak:

- Pliki tekstowe w systemie HDFS
- Bazy danych
- HCatalog

Zwykle strumień z węzłem źródłowym Analytic Server jest wykonywany w systemie HDFS; jeśli jednak strumień zawiera węzeł, który nie może być wykonywany w systemie HDFS, wówczas maksymalna możliwa część strumienia zostanie przekazana do źródła Analytic Server, a następnie program SPSS Modeler Server podejmie próbę przetworzenia pozostałej części strumienia. W przypadku bardzo dużych zbiorów danych konieczne będzie wykonanie podpróby; na przykład, zastępując węzeł próby strumieniem.

Aby użyć własnego połączenia Analytic Server zamiast połączenia domyślnego zdefiniowanego przez administratora, należy usunąć zaznaczenie pola **Użyj domyślnego serwera analitycznego** i wybrać własne połączenie. Szczegółowe informacje na temat konfigurowania kilku połączeń Analytic Server zawiera temat Łączenie się z serwerem Analytic Server.

Źródło danych. Zakładając, że użytkownik lub administrator SPSS Modeler Server nawiązał połączenie, należy wybrać źródło danych zawierające dane, jakie mają zostać użyte. Źródło danych zawiera zmienne i metadane powiązane ze źródłem. Aby wyświetlić listę dostępnych źródeł danych, należy kliknąć przycisk **Wybierz**. Więcej informacji można znaleźć w temacie “Wybieranie źródła danych”.

Jeśli konieczne jest utworzenie nowego źródła danych lub przeprowadzenie edycji istniejącego, należy kliknąć opcję **Uruchom edytora źródła danych...**.

Należy zauważyć, że użycie kilku połączeń Analytic Server może być przydatne podczas sterowania przepływem danych. Przykładowo, w przypadku użycia węzła źródłowego i węzła eksportu Analytic Server użytkownik może chcieć użyć różnych połączeń Analytic Server w różnych gałęziach strumienia, tak aby po uruchomieniu poszczególnych gałęzi korzystały one z własnego serwera Analytic Server i aby żadne dane nie były przekazywane do serwera IBM SPSS Modeler Server. Należy pamiętać, że jeśli gałąź zawiera więcej niż jedno połączenie Analytic Server, dane będą pobierane z serwerów Analytic Server na serwer IBM SPSS Modeler Server. Więcej informacji, w tym informacje o ograniczeniach, można znaleźć w sekcji Właściwości strumienia Analytic Server.

Wybieranie źródła danych

Tabela Źródła danych zawiera listę dostępnych źródeł danych. Należy wybrać źródło, jakie ma zostać użyte, i kliknąć przycisk **OK**.

Kliknięcie opcji **Pokaż właściciela** pozwala wyświetlić właściciela źródła danych.

Opcja **Filtruj wg** umożliwia filtrowanie listy źródeł danych na podstawie **słowa kluczowego**, w czasie którego sprawdzane są kryteria filtrowania w odniesieniu do nazwy źródła danych oraz jego opisu lub **właściciela**. Jako kryteria filtrowania można wprowadzić kombinację znaków łańcuchowych, numerycznych lub wieloznacznych, opisanych poniżej. W każdym łańcuchu wielkość liter jest rozróżniana. Kliknięcie przycisku **Odśwież** spowoduje zaktualizowanie tabeli Źródła danych.

_ Podkreślenie może reprezentować każdy pojedynczy znak w łańcuchu wyszukiwania.

% Znak procentu może reprezentować dowolną sekwencję w łańcuchu wyszukiwania, składającą się z zera lub większej liczby znaków.

Poprawianie danych uwierzytelniających

Jeśli dane uwierzytelniające do uzyskiwania dostępu do programu Analytic Server różnią się od danych uwierzytelniających programu SPSS Modeler Server, podczas uruchamiania strumienia w programie Analytic Server konieczne będzie wprowadzenie danych uwierzytelniających Analytic Server. Jeśli użytkownik nie zna swoich danych uwierzytelniających, powinien skontaktować się z administratorem serwera.

Obsługiwane węzły

Wiele węzłów SPSS Modeler można wykonywać w systemie HDFS, jednak w przypadku niektórych ich wykonywanie może przebiegać inaczej, a niektóre obecnie nie są obsługiwane. W tym temacie omówiono szczegółowo obsługę na bieżącym poziomie.

Uwagi ogólne

- Niektóre znaki, które są zwykle do zaakceptowania w nazwie zmiennej programu Modeler ujętej w cudzysłów, nie będą akceptowane w programie Analytic Server.
- Aby strumień programu Modeler mógł być uruchomiony w programie Analytic Server, musi się rozpoczynać od co najmniej jednego węzła źródłowego Analytic Server i kończyć pojedynczym węzłem modelowania lub węzłem eksportu Analytic Server.
- Zaleca się, aby ustawić składowanie przewidywanych zmiennych ilościowych jako liczby rzeczywiste, a nie jako liczby całkowite. Modele oceniające dla ciągłych zmiennych przewidywanych zawsze zapisują w plikach danych wynikowych wartości rzeczywiste, natomiast model danych wynikowych do wykonywania ocen korzysta ze składowania zmiennej przewidywanej. Dlatego, jeśli dla przewidywanej zmiennej ilościowej zastosowane jest składowanie z użyciem liczby całkowitej, wystąpi brak zgodności ocen pomiędzy zapisanymi wartościami i modelem danych, co spowoduje błędy podczas próby odczytania danych poddanych ocenie.
- Jeśli zmienna jest geoprzestrzenna, w przypadku @OFFSET funkcja nie jest obsługiwana.

Źródło

- Strumień, który nie rozpoczyna się od węzła źródłowego Analytic Server, zostanie uruchomiony lokalnie.

Operacje na rekordach

Wszystkie operacje związane z rekordami są obsługiwane, z wyjątkiem szeregów czasowych i węzłów siatki czasoprzestrzeni. Poniżej przedstawiono dalsze uwagi dotyczące funkcji obsługiwanych węzłów.

Selekcja

- Obsługuje ten sam zestaw funkcji co Węzeł wyliczenia.

Losowanie

- Próbkowanie na poziomie bloku nie jest obsługiwane.
- Metody złożonego losowania nie są obsługiwane.
- N pierwszych prób dla Odrzuć próbę nie jest obsługiwanych.
- N pierwszych prób dla N>20000 nie jest obsługiwanych.
- Próbkowanie 1-w-n nie jest obsługiwane, jeśli opcja Maksymalna wielkość próby nie jest ustawiona.
- Próbkowanie 1-w-n nie jest obsługiwane, jeśli N * Maksymalna wielkość próby > 20000.
- Próbkowanie losowo % ze wszystkich na poziomie bloku nie jest obsługiwane.
- Losowo % ze wszystkich aktualnie obsługuje dostarczanie wartości początkowej.

Agregacja

- Klucze ciągłe nie są obsługiwane. Jeśli ponownie wykorzystywany jest istniejący strumień, który skonfigurowano w celu sortowania danych, i ustawienie to zostanie następnie użyte w węźle Agregacja, należy zmienić strumień, tak aby usunąć węzeł sortowania.

- Statystyki porządkowe (mediana, 1. kwartyl, 3. kwartyl) są wyliczane w przybliżeniu, a ich obsługę umożliwia karta Optymalizacja.

Sortowanie

- Karta Optymalizacja nie jest obsługiwana.

W środowisku rozproszonym dostępna jest ograniczona liczba operacji, które zachowują kolejność rekordu ustaloną w węźle sortowania.

- Ustawienie węzła sortowania, po którym następuje węzeł eksportu, tworzy posortowane źródło danych.
- Ustawienie węzła sortowania przed węzłem Losowanie z ustawieniem próbkowania rekordu na **Pierwsze** zwróci *N* pierwszych rekordów.

Ogólnie węzeł sortowania powinien znajdować się jak najbliżej operacji wymagających posortowanych rekordów.

Łączenie

- Łączenie wg kolejności nie jest obsługiwane.
- Karta Optymalizacja nie jest obsługiwana.
- Operacje łączenia przebiegają relatywnie powoli. Jeśli w systemie HDFS jest wolne miejsce, znacznie szybsze będzie połączenie źródeł danych jeden raz i używanie połączonego źródła w kolejnych strumieniach niż łączenie źródeł danych w każdym strumieniu.

Transformacje R

Polecenia R w węźle powinny zawierać operacje wykonywane dla jednego rekordu w jednym czasie.

Operacje na zmiennych

Wszystkie operacje na zmiennych są obsługiwane, z wyjątkiem węzłów Anonimizacja, Transpozycja, Przedziały czasowe i Historia. Poniżej przedstawiono dalsze uwagi dotyczące funkcji obsługiwanych węzłów.

Auto Przygotowanie

- Uczenie węzła nie jest obsługiwane. Zastosowanie transformacji w wyuczonym węźle Auto Przygotowanie dla nowych danych jest obsługiwane.

Wyliczanie

- Wszystkie funkcje węzła Wyliczanie są obsługiwane, z wyjątkiem funkcji sekwencji.
- Wyliczanie nowej zmiennej jako liczebności zasadniczo jest operacją sekwencyjną i dlatego nie jest obsługiwane.
- Zmienne podziału nie mogą być wyliczane w tym samym strumieniu, w którym zostały użyte do podziału; konieczne będzie utworzenie dwóch strumieni: jednego, który będzie dzielił zmienną podziału i drugiego, w którym ta zmienna będzie użyta do podziału.

Wypełnianie

- Obsługuje ten sam zestaw funkcji co Węzeł wyliczenia.

Kategoryzacja

Następujące funkcje nie są obsługiwane.

- Kategoryzacja optymalna
- Rangi
- N-tyle -> Tworzenie N-tyli: Suma wartości
- N-tyle -> Wiązania: Pozostaw w bieżącej i Przydziel losowo
- N-tyle ->N użytkownika: gdy wartości przekraczają 100, a dowolna wartość N, gdy 100% N nie jest równe zero.

Analiza RFM

- Opcja Pozostaw w bieżącej dla wiązań nie jest obsługiwana. RFM — oceny aktualności, częstości i kwoty nie zawsze będą zgodne z tymi wyliczonymi przez program Modeler na podstawie tych samych danych. Zakresy wyników będą takie same , ale przypisania ocen (numery przedziałów) mogą różnić się o jeden.

Wykresy

Wszystkie węzły wykresów są obsługiwane.

Modelowanie

Obsługiwane są następujące węzły modelowania: Szeregi czasowe, TCM, Izotoniczna-AS, Rozszerzenie Model, Drzewo-AS, Drzewo C&R, Quest, CHAID, Liniowy, Liniowy-AS, Sieć neuronowa, GLE, LSVM, Dwustopniowa-AS, Drzewa losowe, STP, Reguły asocjacyjne, XGBoost-AS, Las losowy i K-średnie-AS. Poniżej przedstawiono dalsze uwagi dotyczące funkcji tych węzłów.

Liniowy

Podczas budowania modeli w oparciu o duże zbiory danych zwykle dokonywana jest zmiana celu na Bardzo duże zbiory danych lub określane są podziały.

- Kontynuacja uczenia dla istniejących modeli PSM nie jest obsługiwana.
- Cel, jakim jest budowanie modelu standardowego, jest zalecany tylko w przypadku, gdy definiowane są zmienne podziału, dzięki czemu liczba rekordów w każdym podziale nie jest zbyt duża, przy czym definicja określenia „zbyt duża” zależy od potęgi poszczególnych węzłów w klastrze Hadoop. Należy jednak zachować ostrożność, aby podziały nie były zdefiniowane jako zbyt małe, ponieważ powstanie zbyt wiele rekordów do zbudowania modelu.
- Cel, jakim jest wartość logiczna, nie jest obsługiwany.
- Cel, jakim jest agregacja bootstrap, nie jest obsługiwany.
- Użycie celu Bardzo duże zbiory danych nie jest zalecane w przypadku kilku rekordów; jego użycie spowoduje, że często model nie zostanie zbudowany lub zbudowany model będzie uszkodzony.
- Automatyczne przygotowanie danych nie jest obsługiwane. Może to powodować problemy podczas próby zbudowania modelu na podstawie danych zawierających wiele braków danych; zwykle są one wprowadzane jako część operacji automatycznego przygotowania danych. Aby uniknąć tego problemu, należy użyć modelu drzewa lub sieci neuronowej w ustawieniach zaawansowanych w celu podstawienia wybranych braków danych.
- Statystyka dokładności nie jest obliczana dla modeli rozdzielonych.

Sieci neuronowe

Podczas budowania modeli w oparciu o duże zbiory danych zwykle dokonywana jest zmiana celu na Bardzo duże zbiory danych lub określane są podziały.

- Kontynuacja uczenia istniejących modeli standardowych lub PSM nie jest obsługiwana.
- Cel, jakim jest budowanie modelu standardowego, jest zalecany tylko w przypadku, gdy definiowane są zmienne podziału, dzięki czemu liczba rekordów w każdym podziale nie jest zbyt duża, przy czym definicja określenia „zbyt duża” zależy od potęgi poszczególnych węzłów w klastrze Hadoop. Należy jednak zachować ostrożność, aby podziały nie były zdefiniowane jako zbyt małe, ponieważ powstanie zbyt wiele rekordów do zbudowania modelu.
- Cel, jakim jest wartość logiczna, nie jest obsługiwany.
- Cel, jakim jest agregacja bootstrap, nie jest obsługiwany.
- Użycie celu Bardzo duże zbiory danych nie jest zalecane w przypadku kilku rekordów; jego użycie spowoduje, że często model nie zostanie zbudowany lub zbudowany model będzie uszkodzony.
- Jeśli w danych występuje wiele braków danych, w celu podstawienia braków danych należy użyć ustawień zaawansowanych.
- Statystyka dokładności nie jest obliczana dla modeli rozdzielonych.

C&R Tree, CHAID i Quest

Podczas budowania modeli w oparciu o duże zbiory danych zwykle dokonywana jest zmiana celu na Bardzo duże zbiory danych lub określane są podziały.

- Kontynuacja uczenia dla istniejących modeli PSM nie jest obsługiwana.

- Cel, jakim jest budowanie modelu standardowego, jest zalecany tylko w przypadku, gdy definiowane są zmienne podziału, dzięki czemu liczba rekordów w każdym podziale nie jest zbyt duża, przy czym definicja określenia „zbyt duża” zależy od potęgi poszczególnych węzłów w klastrze Hadoop. Należy jednak zachować ostrożność, aby podziały nie były zdefiniowane jako zbyt małe, ponieważ powstanie zbyt wiele rekordów do zbudowania modelu.
- Cel, jakim jest wartość logiczna, nie jest obsługiwany.
- Cel, jakim jest agregacja bootstrap, nie jest obsługiwany.
- Użycie celu Bardzo duże zbiory danych nie jest zalecane w przypadku kilku rekordów; jego użycie spowoduje, że często model nie zostanie zbudowany lub zbudowany model będzie uszkodzony.
- Sesje interaktywne nie są obsługiwane.
- Statystyka dokładności nie jest obliczana dla modeli rozdzielonych.
- Jeśli obecna jest zmienna podziału, modele drzew budowane lokalnie w programie Modeler nieco różnią się od modeli drzew budowanych przez program Analytic Server, w wyniku czego uzyskiwane są różne oceny. Algorytmy używane w obu przypadkach są poprawne; algorytmy stosowane przez Analytic Server są po prostu nowsze. Ze względu na fakt, że algorytmy drzew wykazują tendencję do tworzenia wielu reguł heurystycznych, różnice pomiędzy tymi dwoma komponentami są normalne.

Ocena modelu

Wszystkie modele obsługiwane do modelowania są również obsługiwane w przypadku oceniania. Ponadto w przypadku oceniania obsługiwane są również tworzone lokalnie modele użytkowe dla następujących węzłów: C&RT, Quest, CHAID, liniowy i sieć neuronowa (niezależnie od tego, czy jest to model standardowy, wzmocniony spakowany, czy bardzo duży zbiór danych), regresja, C5.0, logistyczny, Genlin, GLMM, Coksa, SVM, sieć Bayesa, dwustopniowy, KNN, lista decyzyjna, dyskryminacyjny, samonauczanie, wykrywanie anomalii, Apriori, Carma, K-średnie, Kohonena, R i eksploracja tekstu.

- Surowe lub skorygowane skłonności nie będą oceniane. W ramach obejścia ten sam skutek można uzyskać poprzez ręczne obliczenie surowej skłonności za pomocą węzła wyliczania, stosując następujące wyrażenie: if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif.

R Polecenia R w modelu użytkowym powinny zawierać operacje wykonywane dla jednego rekordu w jednym czasie.

Wynik Węzły Macierz, Analiza, Audyt danych, Transformacja, Globalne, Statystyki, Średnie i Tabela są obsługiwane. Poniżej przedstawiono dalsze uwagi dotyczące funkcji obsługiwanych węzłów.

Audyt danych

Węzeł Audyt danych nie może utworzyć dominanty dla zmiennych ilościowych.

Średnie

Węzeł Średnie nie może utworzyć błędu standardowego lub przedziału ufności 95%.

Tabela

Węzeł Tabela jest obsługiwany poprzez zapisywanie tymczasowego źródła danych Analytic Server zawierającego wyniki z wcześniejszych operacji. Następnie węzeł Tabela przechodzi kolejno przez zawartość źródła danych.

Eksport

Strumień może rozpoczynać się od węzła źródłowego Analytic Server i kończyć węzłem eksportu innym niż węzeł eksportu Analytic Server, ale dane zostaną przeniesione z systemu HDFS do SPSS Modeler Server i ostatecznie do lokalizacji eksportu.

Węzeł źródłowy bazy danych

Węzeł źródłowy Baza danych umożliwia importowanie danych z różnych innych pakietów za pośrednictwem ODBC (Open Database Connectivity), np. Microsoft SQL Server, Db2, Oracle i inne.

Aby odczytać lub zapisać dane w bazie danych, należy mieć zainstalowane źródło danych ODBC, które jest skonfigurowane dla odpowiedniej bazy danych z uprawnieniami odczytu i zapisu zgodnie z potrzebami. IBM SPSS

Data Access Pack obejmuje zestaw sterowników ODBC, których można użyć w tym celu, a sterowniki są dostępne na stronie WWW z materiałami do pobrania. W przypadku pytań dotyczących tworzenia lub określania uprawnień dla źródeł danych ODBC należy skontaktować się z administratorem bazy danych.

Obsługiwane sterowniki ODBC

W celu uzyskania najnowszych informacji na temat obsługiwanych i przetestowanych pod kątem współpracy z produktem IBM SPSS Modeler baz danych i sterowników ODBC należy zapoznać się z tabelami kompatybilności i odwiedzić korporacyjną witrynę wsparcia pod adresem http://www.ibm.com/support.

Gdzie zainstalować sterowniki

Uwaga: Sterowniki ODBC należy zainstalować i skonfigurować na każdym komputerze, gdzie może występować przetwarzanie.

- Jeśli program IBM SPSS Modeler działa w trybie lokalnym (samodzielnym), sterowniki należy zainstalować na komputerze lokalnym.
- Jeśli program IBM SPSS Modeler jest uruchomiony w trybie analizy rozproszonej dla zdalnego serwera IBM SPSS Modeler Server, sterowniki powinny być zainstalowane na komputerze, na którym zainstalowany jest serwer IBM SPSS Modeler Server. Dla serwerów IBM SPSS Modeler Server z systemem UNIX zobacz również punkt „Konfigurowanie sterowników ODBC w systemach UNIX” w dalszej części tej sekcji.
- Jeśli wymagany jest dostęp do tych samych źródeł danych zarówno z programu IBM SPSS Modeler, jak i serwera IBM SPSS Modeler Server, sterowniki ODBC należy zainstalować na obu komputerach.
- Jeśli program IBM SPSS Modeler jest uruchamiany za pomocą usług terminalowych, sterowniki ODBC należy zainstalować na serwerze usług terminalowych, na którym zainstalowany jest program IBM SPSS Modeler.

Dostęp do danych z bazy danych

Aby uzyskać dostęp do danych z bazy danych, należy wykonać następujące kroki.

- Zainstaluj sterownik ODBC i skonfiguruj źródło danych dla bazy danych, jaka będzie używana.
- W oknie dialogowym Baza danych nawiąż połączenie z bazą danych za pośrednictwem trybu tabeli lub trybu zapytania SQL.
- Wybierz tabele z bazy danych.
- Korzystając z kart w oknie dialogowym Baza danych można zmienić typy użycia oraz filtrować zmienne danych.

Więcej szczegółów na temat kolejnych kroków zamieszczono w tematach zawartych w powiązanej dokumentacji.

Uwaga: Po wywołaniu zapisanych procedur (SP) bazy danych z programu SPSS Modeler wyświetlana jest pojedyncza zmienna wynikowa o nazwie RowsAffected, a nie oczekiwany wynik z zapisanych procedur. Taka sytuacja ma miejsce, jeśli ODBC nie zwraca wystarczających informacji, aby określić model danych wynikowych dla zapisanych procedur. Program SPSS Modeler tylko w ograniczony sposób obsługuje zapisane procedury (SP), które zwracają wynik, dlatego zaleca się, aby zamiast korzystać z zapisanych procedur, wyodrębnić funkcję SELECT z SP i wykonać jedną z następujących czynności.

- Należy utworzyć widok na podstawie funkcji SELECT i wybrać ten widok w węźle źródłowym bazie danych
- Funkcji SELECT można użyć bezpośrednio w węźle źródłowym bazy danych.

Ustawianie opcji węzła bazy danych

Opcje na karcie Dane w oknie dialogowym węzła źródłowego bazy danych umożliwiają uzyskanie dostępu do bazy danych i odczytywanie danych z wybranej tabeli.

Tryb. Wybranie opcji **Tabela** umożliwia połączenie z tabelą za pośrednictwem elementów sterujących w oknie dialogowym.

Wybranie opcji **Zapytanie SQL** pozwala utworzyć zapytanie dla wybranej poniżej bazy danych z użyciem zapytanie SQL. Więcej informacji można znaleźć w temacie “Tworzenie zapytań dla bazy danych” na stronie 25.

Źródło danych. W przypadku obu trybów, Tabela i Zapytanie SQL, można wprowadzić nazwę w polu Źródło danych lub wybrać z listy rozwijanej opcję **Dodaj nowe połączenie z bazą danych**.

Do nawiązania połączenia z bazą danych i wybrania tabeli za pośrednictwem okna dialogowego używane są następujące opcje:

Nazwa tabeli. Jeśli nazwa tabeli, do której użytkownik zamierza uzyskać dostęp jest znana, należy ją wpisać w polu Nazwa tabeli. W przeciwnym razie należy kliknąć przycisk **Wybierz**, aby otworzyć okno dialogowe z listą dostępnych tabel.

Ujmij w cudzysłów nazwy tabeli i kolumny. Należy określić, czy podczas wysyłania zapytań do bazy danych nazwy tabeli i kolumn mają być ujęte w cudzysłów (jeśli na przykład zawierają spacje lub znaki interpunkcyjne).

- Opcja **W razie potrzeby** będzie powodowała ujęcie w cudzysłów nazw tabeli i kolumn *wyłącznie*, jeśli będą one zawierały znaki niestandardowe. Znaki niestandardowe to znaki niezgodne z formatem ASCII, spacje i wszystkie znaki niealfanumeryczne oprócz kropki (.).
- Opcję **Zawsze** należy wybrać, aby *wszystkie* nazwy tabeli i kolumn były ujmowane w cudzysłów.
- Opcję **Nigdy** należy wybrać, aby nazwy tabeli i kolumn *nigdy* nie były ujmowane w cudzysłów.

Usuń wiodące i końcowe spacje. Tę opcję należy wybrać, aby odrzucać spacje wiodące i końcowe w łańcuchach.

Uwaga: Porównania między łańcuchami, które nie korzystają ze wstępnego generowania kodu SQL, mogą zwracać odmienne wyniki w przypadku, gdy istnieją spacje końcowe.

Odczytywanie pustych łańcuchów Oracle. W przypadku odczytu z lub zapisu do bazy danych Oracle należy pamiętać, że inaczej niż w przypadku IBM SPSS Modeler i inaczej niż w przypadku większości innych baz danych, Oracle traktuje i przechowuje puste wartości strumienia jako równoważne wartościom null. Oznacza to, że te same dane wyodrębnione z bazy danych Oracle mogą zachowywać się w odmienny sposób w przypadku wyodrębnienia z pliku lub innej bazy danych, zaś dane mogą zwracać odmienne wyniki.

Dodawanie połączenia z bazą danych

Aby otworzyć bazę danych, najpierw należy wybrać źródło danych, a którym ma zostać nawiązane połączenie. Na karcie Dane należy wybrać z listy rozwijanej Źródło danych opcję **Dodaj nowe połączenie z bazą danych**.

Spowoduje to otwarcie okna dialogowego Połączenia z bazą danych.

Uwaga: Inną metodą otwarcia tego okna dialogowego jest wybranie z menu głównego opcji: **Narzędzia > Bazy danych...**

Źródła danych. Zawiera listę dostępnych źródeł danych. Jeśli wybrana baza danych nie jest widoczna, należy listę przewinąć w dół. Po wybraniu źródła danych i wprowadzeniu haseł, należy kliknąć przycisk **Połącz**. Kliknięcie przycisku **Odśwież** spowoduje zaktualizowanie listy.

Nazwa użytkownika i hasło. Jeśli źródło danych jest zabezpieczone hasłem, należy wprowadzić nazwę użytkownika i powiązane z nią hasło.

Dane uwierzytelniające. Jeśli skonfigurowano dane uwierzytelniające w programie IBM SPSS Collaboration and Deployment Services, można wybrać tę opcję, aby wybrać je w repozytorium. Nazwa użytkownika i hasło z danych uwierzytelniających muszą być zgodne z nazwą użytkownika i hasłem wymaganymi do uzyskania dostępu do bazy danych.

Połączenia. Wyświetla aktualnie podłączone bazy danych.

- **Domyślny.** Opcjonalnie można wybrać jedno połączenie jako domyślne. W ten sposób węzły źródła lub eksportu bazy danych będą zawierały to połączenie jako ustawienie predefiniowane dla źródła danych; ustawienie to można jednak w razie potrzeby edytować.
- **Zapisz.** Opcjonalnie można wybrać jedno lub kilka połączeń, jakie można będzie ponownie wyświetlić w kolejnych sesjach.
- **Źródło danych.** Łańcuchy połączenia dla aktualnie podłączonych baz danych.
- **Wstępne ustawienie.** Wskazuje (za pomocą symbolu gwiazdki *), czy wprowadzone zostały wstępne ustawienia wartości dla danego połączenia z bazą danych. Aby wprowadzić wartości wstępnych ustawień, należy kliknąć kolumnę w wierszu odpowiadającym połączeniu z bazą danych i wybrać opcję Określ z listy. Więcej informacji można znaleźć w temacie “Określanie wartości wstępnych ustawień dla połączenia z bazą danych” na stronie 22.

Aby usunąć połączenia, należy wybrać połączenie z listy i kliknąć przycisk **Usuń**.

Po wyborze należy kliknąć przycisk **OK**.

Aby odczytać lub zapisać dane w bazie danych, należy mieć zainstalowane źródło danych ODBC, które jest skonfigurowane dla odpowiedniej bazy danych z uprawnieniami odczytu i zapisu zgodnie z potrzebami. IBM SPSS Data Access Pack obejmuje zestaw sterowników ODBC, których można użyć w tym celu, a sterowniki są dostępne na stronie WWW z materiałami do pobrania. W przypadku pytań dotyczących tworzenia lub określania uprawnień dla źródeł danych ODBC należy skontaktować się z administratorem bazy danych.

Obsługiwane sterowniki ODBC

W celu uzyskania najnowszych informacji na temat obsługiwanych i przetestowanych pod kątem współpracy z produktem IBM SPSS Modeler baz danych i sterowników ODBC należy zapoznać się z tabelami kompatybilności i odwiedzić korporacyjną witrynę wsparcia pod adresem http://www.ibm.com/support.

Gdzie zainstalować sterowniki

Uwaga: Sterowniki ODBC należy zainstalować i skonfigurować na każdym komputerze, gdzie może występować przetwarzanie.

- Jeśli program IBM SPSS Modeler działa w trybie lokalnym (samodzielnym), sterowniki należy zainstalować na komputerze lokalnym.
- Jeśli program IBM SPSS Modeler jest uruchomiony w trybie analizy rozproszonej dla zdalnego serwera IBM SPSS Modeler Server, sterowniki powinny być zainstalowane na komputerze, na którym zainstalowany jest serwer IBM SPSS Modeler Server. Dla serwerów IBM SPSS Modeler Server z systemem UNIX zobacz również punkt „Konfigurowanie sterowników ODBC w systemach UNIX” w dalszej części tej sekcji.
- Jeśli wymagany jest dostęp do tych samych źródeł danych zarówno z programu IBM SPSS Modeler, jak i serwera IBM SPSS Modeler Server, sterowniki ODBC należy zainstalować na obu komputerach.
- Jeśli program IBM SPSS Modeler jest uruchamiany za pomocą usług terminalowych, sterowniki ODBC należy zainstalować na serwerze usług terminalowych, na którym zainstalowany jest program IBM SPSS Modeler.

Konfigurowanie sterowników ODBC w systemach UNIX

Domyślnie narzędzie DataDirect Driver Manager nie jest skonfigurowane dla serwera IBM SPSS Modeler Server w systemach UNIX. Aby skonfigurować system UNIX w celu załadowania programu DataDirect Driver Manager, należy wprowadzić następujące komendy:

```
cd <modeler_server_install_directory>/bin
rm -f libspssodbc.so
ln -s libspssodbc_datadirect.so libspssodbc.so
```

Powoduje to usunięcie domyślnego łącza i tworzy łącze do programu DataDirect Driver Manager.

Uwaga: Opakowanie sterownika UTF16 jest wymagane do użycia sterownika SAP HANA lub IBM Db2 CLI dla niektórych baz danych. DashDB wymaga sterownika IBM Db2 CLI. Aby utworzyć łącze dla opakowania sterownika UTF16, zamiast powyższego rozwiązania wprowadź następujące komendy:

```
rm -f libspssodbc.so
ln -s libspssodbc_datadirect_utf16.so libspssodbc.so
```

Aby skonfigurować program SPSS Modeler Server:

1. Skonfiguruj skrypt uruchamiania serwera SPSS Modeler Server modelersrv.sh, aby określić położenie pliku pakietu IBM SPSS Data Access Pack odbc.sh, dodając następujący wiersz w pliku modelersrv.sh:

   ```
   . /<pathtoSDAPinstall>/odbc.sh
   ```

   Gdzie <pathtoSDAPinstall> to pełna ścieżka do instalacji produktu IBM SPSS Data Access Pack.
2. Ponownie uruchom system SPSS Modeler Server.

Dodatkowo, tylko dla platform SAP HANA i IBM Db2, dodaj następującą definicję parametrów do DSN w pliku odbc.ini, aby uniknąć przepełnienia bufora podczas połączenia:

```
DriverUnicodeType=1
```

Uwaga: Opakowanie libspssodbc_datadirect_utf16.so jest również kompatybilne z innymi obsługiwanymi sterownikami ODBC serwera SPSS Modeler Server.

Potencjalne problemy z bazą danych

W zależności od używanej bazy danych istnieją potencjalne problemy, o których należy pamiętać.

IBM Db2

Podczas podjęcia próby zapisania w pamięci podręcznej węzła w strumieniu, który odczytuje dane z bazy danych Db2, może zostać wyświetlony następujący komunikat:

```
Nie można znaleźć domy⌂ślnego obszaru tabel o wielkoś⌂ci strony co najmniej 4096,
 do używania którego jest uprawniony ID autoryzowanego użytkownika TEST.
```

Aby skonfigurować bazę Db2, tak aby zapisywanie w pamięci podręcznej w bazie danych działało poprawnie w programie SPSS Modeler, administrator bazy danych powinien utworzyć obszar danych tymczasowy dla użytkownika i nadać uprawnienia dostępu do tego obszaru tabel odpowiednim kontom Db2.

W nowym obszarze tabel zalecamy zastosowanie wielkości strony 32768, ponieważ zwiększy to limit liczby zmiennych, które można pomyślnie zapisać w pamięci podręcznej.

IBM Db2 for z/OS

- Ocenianie podzbioru algorytmów z włączoną ufnością za pośrednictwem wygenerowanego kodu SQL może spowodować wystąpienie błędu lub wyjątku. Problem ten jest specyficzny dla Db2 for z/OS; aby go usunąć, należy użyć składnika SPSS Modeler Server Scoring Adapter for Db2 on z/OS.
- Podczas uruchamiania strumieni w Db2 for z/OS mogą wystąpić błędy bazy danych, jeśli aktywowano limit czasu bezczynności połączeń bazy danych i limit ten jest zbyt niski. W programie Db2 for z/OS, wersja 8, ustawienie domyślne zostało zmienione z braku określonego limitu czasu na 2 minuty. Rozwiązaniem jest zwiększenie wartości parametru systemu Db2 IDLE THREAD TIMEOUT (IDTHTOIN) lub zresetowanie wartości do 0.

Oracle

Po uruchomieniu strumienia, który zawiera węzeł agregacji wartości zwrócone dla 1. i 3. kwartyla po przekazaniu kodu SQL do bazy danych Oracle mogą różnić się od wartości zwróconych w trybie rodzimym.

Określanie wartości wstępnych ustawień dla połączenia z bazą danych

W niektórych bazach danych można określić pewne ustawienia domyślne dla połączenia z bazą danych. Wszystkie te ustawienia mają zastosowanie do eksportu do bazy danych.

Poniżej przedstawiono typy baz danych obsługujących tę funkcję.

- Wydania produktu SQL Server Enterprise i Developer. Więcej informacji można znaleźć w temacie "Ustawienia dla bazy danych SQL Server".
- Wydania produktu Oracle Enterprise lub Personal. Więcej informacji można znaleźć w temacie "Ustawienia dla bazy danych Oracle".
- IBM Db2 for z/OS i Teradata łączą się z bazą danych lub schematem w podobny sposób. Więcej informacji można znaleźć w temacie "ustawienia dla baz danych IBM Db2 for z/OS, IBM Db2 LUW i Teradata" na stronie 23.

Jeśli nawiązane jest połączenie z bazą danych lub schematem, który nie obsługuje tej funkcji, wyświetlony zostanie komunikat **Nie można skonfigurować wstępnych ustawień dla tego połączenia z bazą danych**.

Ustawienia dla bazy danych SQL Server

Ustawienia te są wyświetlane dla wydań produktu SQL Server Enterprise i Developer.

Użyj kompresji. Po zaznaczeniu tej opcji tworzone są tabele do eksportu z zastosowaniem kompresji.

Kompresja dla. Należy wybrać poziom kompresji.

- **Wiersz.** Umożliwia kompresję na poziomie wiersza (np. odpowiednik zapisu CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW); w języku SQL).
- **Strona.** Umożliwia kompresję na poziomie strony (np. CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE); in SQL).

Ustawienia dla bazy danych Oracle

Ustawienia dla bazy danych Oracle — opcja podstawowa

Te ustawienia są wyświetlane w przypadku wydań produktu Oracle Enterprise lub Personal dla opcji podstawowej.

Użyj kompresji. Po zaznaczeniu tej opcji tworzone są tabele do eksportu z zastosowaniem kompresji.

Kompresja dla. Należy wybrać poziom kompresji.

- **Domyślny.** Umożliwia zastosowanie domyślnej kompresji (np. zapis CREATE TABLE MYTABLE(...) COMPRESS; w języku SQL). W takim przypadku uzyskany zostanie efekt taki sam, jak po zastosowaniu opcji **Podstawowe**.
- **Podstawowe.** Umożliwia zastosowanie podstawowej kompresji (np. zapis CREATE TABLE MYTABLE(...) COMPRESS BASIC; w języku SQL).

Ustawienia dla bazy danych Oracle — opcja zaawansowana

Te ustawienia są wyświetlane w przypadku wydań produktu Oracle Enterprise lub Personal Advanced dla opcji zaawansowanej.

Użyj kompresji. Po zaznaczeniu tej opcji tworzone są tabele do eksportu z zastosowaniem kompresji.

Kompresja dla. Należy wybrać poziom kompresji.

- **Domyślny.** Umożliwia zastosowanie domyślnej kompresji (np. zapis CREATE TABLE MYTABLE(...) COMPRESS; w języku SQL). W takim przypadku uzyskany zostanie efekt taki sam, jak po zastosowaniu opcji **Podstawowe**.
- **Podstawowe.** Umożliwia zastosowanie podstawowej kompresji (np. zapis CREATE TABLE MYTABLE(...) COMPRESS BASIC; w języku SQL).

- **OLTP.** Umożliwia zastosowanie kompresji OLTP (np. zapis CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP; w języku SQL).
- **Zapytanie Niska/Wysoka.** (Tylko serwery Exadata) Umożliwia zastosowanie hybrydowej kompresji kolumnowej dla zapytania (np. zapis CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW; lub CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH; w języku SQL). Kompresja zapytania jest przydatna w środowiskach zajmujących się magazynowaniem danych; WYSOKA oznacza wyższy współczynnik kompresji niż NISKA.
- **Archiwizacja Niska/Wysoka.** (Tylko serwery Exadata) Umożliwia zastosowanie hybrydowej kompresji kolumnowej dla archiwum (np. zapis CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW; lub CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH; w języku SQL). Kompresja dla archiwum jest przydatna do kompresji danych, które będą składowane przez długi okres; WYSOKA oznacza wyższy współczynnik kompresji niż NISKA.

ustawienia dla baz danych IBM Db2 for z/OS, IBM Db2 LUW i Teradata

Po określeniu wstępnych ustawień dla baz danych IBM Db2 for z/OS, IBM Db2 LUW lub Teradata zostanie wyświetlony monit o wybranie następujących opcji:

Użyj bazy danych Server Scoring Adapter lub **Użyj schematu Server Scoring Adapter**. Po wybraniu aktywowana jest opcja **Baza danych Server Scoring Adapter** lub **Schemat Server Scoring Adapter**.

Baza danych Server Scoring Adapter lub **Schemat Server Scoring Adapter** Z listy rozwijanej należy wybrać odpowiednie połączenie.

Ponadto w przypadku bazy danych Teradata można również ustawić szczegóły kategoryzowania zapytania w celu udostępniają dodatkowych metadanych dotyczących elementów, takich jak zarządzanie obciążeniem, segregowanie, identyfikowanie i rozwiązywanie zapytań oraz śledzenie użycia bazy danych.

Sprawdzenie pisowni kategoryzowania zapytania. Tę opcję należy wybrać, jeśli kategoryzowanie zapytania będzie ustawione raz dla całego czasu przez jaki nawiązane jest połączenie z bazą danych Teradata (**Dla sesji**) lub jeśli będzie ono ustawiane za przy każdym uruchomieniu strumienia (**Dla transakcji**).

Uwaga: Jeśli kategoryzowanie zapytań będzie ustawiane w strumieniu, po skopiowaniu strumienia na inny komputer kategoryzowanie zostanie utracone. Aby tego uniknąć, można użyć skryptów w celu uruchomienia strumienia oraz słowa kluczowego *querybanding*w skrypcie, aby zastosować odpowiednie ustawienia.

Wymagane uprawnienia do bazy danych

Aby funkcje bazy danych SPSS Modeler działały poprawnie, należy nadać uprawnienia dostępu do następujących elementów, przypisując je do wszystkich używanych identyfikatorów użytkowników:

Db2 LUW

SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSCAT.TABLESPACES

SYSCAT.SCHEMATA

Db2/z SYSIBM.SYSDUMMY1

SYSIBM.SYSFOREIGNKEYS

SYSIBM.SYSINDEXES

SYSIBM.SYSKEYCOLUSE

SYSIBM.SYSKEYS

SYSIBM.SYSPARMS

SYSIBM.SYSRELS

SYSIBM.SYSROUTINES

SYSIBM.SYSROUTINES_SRC

SYSIBM.SYSSYNONYMS

SYSIBM.SYSTABCONST

SYSIBM.SYSTABCONSTPKC

SYSIBM.SYSTABLES

SYSIBM.SYSTRIGGERS

SYSIBM.SYSVIEWDEP

SYSIBM.SYSVIEWS

SYSIBM.SYSDUMMYU

SYSIBM.SYSPACKSTMT

Teradata

DBC.Functions

DBC.USERS

Wybór tabeli bazy danych

Po nawiązaniu połączenia ze źródłem danych można wybrać zmienne do zaimportowania z określonej tabeli lub widoku. Korzystając z kart Dane w oknie dialogowym Baza danych, można wprowadzić nazwę tabeli w polu Nazwa tabeli lub kliknąć przycisk **Wybierz**, aby otworzyć okno dialogowe Wybierz tabelę/widok z listą dostępnych tabel i widoków.

Pokaż właściciela tabeli. Tę opcję należy wybrać, jeśli źródło danych wymaga, aby przed uzyskaniem dostępu do tabeli wybrać właściciela tabeli. Zaznaczenie tej opcji należy usunąć w przypadku źródeł danych, dla których taki wymóg nie istnieje.

Uwaga: Bazy danych SAS i Oracle zwykle wymagają wskazania właściciela tabeli.

Tabele/widoki. Należy wybrać tabelę lub widok do zaimportowania.

Pokaż. Wyświetla listę kolumn w źródle danych, z którym obecnie nawiązane jest połączenie. Należy kliknąć jedną z następujących opcji, aby dostosować widok dostępnych tabel:

- Kliknięcie przycisku **Tabele użytkownika** umożliwia wyświetlenie zwykłych tabel baz danych utworzonych przez użytkowników bazy danych.
- Kliknięcie przycisku **Tabele systemowe** pozwala wyświetlić tabele baz danych, których właścicielem jest system (na przykład tabele udostępniające informacje na temat bazy danych, takie jak szczegóły indeksów). Ta opcja może być używana do wyświetlania kart używanych w bazach danych programu Excel. (Należy pamiętać, że dostępny jest również osobny węzeł źródłowy programu Excel. Więcej informacji można znaleźć w temacie “Węzeł źródłowy programu Excel” na stronie 44.
- Kliknięcie przycisku **Widoki** pozwala wyświetlać tabele wirtualne w oparciu o zapytanie obejmujące co najmniej jedną zwykłą tabelę.
- Opcja **Synonimy** umożliwia wyświetlenie synonimów utworzonych w bazie danych dla istniejących tabel.

Filtry Nazwa/Właściciel. Te filtry umożliwiają filtrowanie listy wyświetlanych tabel według nazwy lub właściciela. Na przykład typ SYS spowoduje wyświetlenie tylko tabel tego właściciela. W przypadku wyszukiwania z użyciem symboli wieloznacznych podkreślenie (_) może reprezentować dowolny pojedynczy znak, a znak procentu (%) może reprezentować dowolną sekwencję dla zera lub większej liczby znaków.

Ustaw jako domyślne. Umożliwia zapisanie bieżących ustawień jako ustawienia domyślne dla bieżącego użytkownika. Ustawienia te zostaną przywrócone w przyszłości, kiedy użytkownik otworzy nowe okno dialogowe selektora tabel *dla tej samej nazwy źródła danych i tych samych danych logowania*.

Tworzenie zapytań dla bazy danych

Po nawiązaniu połączenia ze źródłem danych można zaimportować zmienne za pośrednictwem zapytań SQL. Z głównego okna dialogowego należy wybrać opcję **Zapytanie SQL** jako tryb połączenia. Spowoduje to dodanie edytora zapytań do okna dialogowego. Korzystając z edytora zapytań można utworzyć jedno lub więcej zapytań SQL, których zestaw wynikowy zostanie wczytany do strumienia danych.

Jeśli określanych jest kilka zapytań SQL, należy je rozdzielić średnikami (;) i upewnić się, że nie została wybrana instrukcja wyboru wielokrotnego.

Aby anulować i zamknąć okno edytora zapytań, należy jako tryb połączenia wybrać opcję **Tabela**.

W zapytaniu SQL można uwzględnić parametry strumienia SPSS Modeler (typ zmiennej definiowanej przez użytkownika). Więcej informacji można znaleźć w temacie “Korzystanie z parametrów strumienia w zapytaniu SQL” na stronie 26.

Załaduj zapytanie. Należy kliknąć tę opcję, aby otworzyć przeglądarkę plików, która umożliwi załadowanie wcześniej zapisanego zapytania.

Zapisz zapytanie. Kliknięcie ten opcji spowoduje otwarcie okna dialogowego Zapisz zapytanie, którego można użyć do zapisania bieżącego zapytania.

Domyślne ustawienia importu. Tę opcję należy kliknąć, aby zaimportować na przykład instrukcję SQL SELECT utworzoną automatycznie za pośrednictwem tabeli i kolumn wybranych w oknie dialogowym.

Wyczyść. Pozwala wyczyścić zawartość obszaru roboczego. Tej opcji należy użyć, aby zacząć od nowa.

Rozdziel tekst. Opcja domyślna **Nigdy** oznacza, że zapytanie zostanie wysłane do bazy danych w całości. Alternatywnie można wybrać opcję **W razie potrzeby**; wówczas program SPSS Modeler podejmie próbę przeanalizowania zapytania i określenia, czy zawiera ono instrukcje SQL, które powinny zostać wysłane do bazy danych jedna po drugiej.

Korzystanie z parametrów strumienia w zapytaniu SQL

Podczas tworzenia zapytań SQL do importowania zmiennych można uwzględnić parametry strumienia SPSS Modeler, które były wcześniej zdefiniowane. Obsługiwane są wszystkie typy parametrów strumienia.

W poniższej tabeli przedstawiono, w jaki sposób niektóre przykłady parametrów strumienia będę interpretowane w zapytaniu SQL.

Tabela 3. Przykłady parametrów strumienia

Nazwa parametru strumienia (przykładowa)	Składowanie	Wartość parametru strumienia	Interpretacja
PString	Łańcuch	ss	'ss'
PInt	Liczba całkowita	5	5
PReal	Liczba rzeczywista	5,5	5,5
PTime	Czas	23:05:01	t{'23:05:01'}
PDate	Data	2011-03-02	d{'2011-03-02'}
PTimeStamp	Znacznik czasu	2011-03-02 23:05:01	ts{'2011-03-02 23:05:01'}
PColumn	Nieznane	IntValue	IntValue

W zapytaniu SQL parametr strumienia określany jest w taki sam sposób, jak w wyrażeniu CLEM, a mianowicie: '$P-<nazwa_parametru>', gdzie <nazwa_parametru> oznacza nazwę, jaka została zdefiniowana dla parametru strumienia.

W przypadku odniesienia do zmiennej typ składowania musi być zdefiniowany jako nieznany, a wartość parametru musi być w razie potrzeby ujęta w cudzysłów. Zatem, korzystając z przykładów przedstawionych w tabeli, po wprowadzeniu zapytania SQL:

```
select "IntValue" from Table1 where "IntValue" < '$P-PInt';
```

będzie ono interpretowane jako:

```
select "IntValue" from Table1 where "IntValue" < 5;
```

Jeśli odniesienie do zmiennej IntValue (Wartość wewnętrzna) zostanie utworzone za pośrednictwem parametru PColumn, aby uzyskać taki sam wynik, konieczne będzie określenie zapytania w następujący sposób:

```
select "IntValue" from Table1 where "'$P-PColumn'" < '$P-PInt';
```

Węzeł Plik zmiennych

Węzły Plik zmiennych umożliwiają odczytywanie danych z plików tekstowych z danymi swobodnymi (pliki, których rekordy zawierają stałą liczbę zmiennych, ale różną liczbę znaków), znanych również jako pliki tekstowe w formacie separowanym. Węzeł tego typu jest również przydatny w przypadku zmiennych z tekstem nagłówka o ustalonej długości i niektórych typów adnotacji. Rekordy są odczytywane kolejno i przekazywane w strumieniu do czasu, aż cały plik zostanie odczytany.

Uwaga dotycząca odczytu danych geoprzestrzennych

Jeśli węzeł zawiera dane geoprzestrzenne i został utworzony w wyniku eksportu z pliku płaskiego, konieczne jest wykonanie kilku dodatkowych kroków w celu skonfigurowania metadanych geoprzestrzennych. Aby uzyskać więcej informacji, zobacz “Importowanie danych geoprzestrzennych do węzła Plik zmiennych” na stronie 29.

Uwagi dotyczące odczytu danych pliku tekstowego w formacie separowanym

- Rekordy muszą być separowane znakiem nowego wiersza wstawionym na końcu każdego wiersza. Znak nowego wiersza nie może być używany w żadnym innym celu (na przykład w nazwie zmiennej lub wartości). Spacje wiodące i końcowe powinny zostać usunięte, aby zachować odstęp, ale nie jest to konieczne. Opcjonalnie spacje te mogą zostać usunięte przez węzeł.
- Zmienne muszą być separowane przecinkami lub innym znakiem, który powinien być używany tylko jako separator, czyli nie powinien występować w nazwach zmiennych lub wartościach. Jeśli nie jest to możliwe, wówczas wszystkie zmienne tekstowe mogą zostać ujęte w podwójny cudzysłów, o ile żadna nazwa zmiennej lub wartość tekstowa nie zawiera znaku podwójnego cudzysłowu. Jeśli nazwy zmiennych lub wartości zawierają znaki podwójnego cudzysłowu, wówczas alternatywnie pola tekstowe mogą zostać ujęte w znaki pojedynczego cudzysłowu (apostrof), o ile znaki pojedynczego cudzysłowu nie są używane w innych miejscach w wartościach. Jeśli nie można użyć ani pojedynczego, ani podwójnego cudzysłowu, wówczas wartości tekstowe należy poprawić, usuwając lub zastępując znak separatora lub pojedynczy/podwójny cudzysłów.
- W każdym wierszu, z uwzględnieniem wiersza nagłówka, powinna znajdować się taka sama liczba zmiennych.
- W pierwszym wierszu powinny znajdować się nazwy zmiennych. W przeciwnym razie należy usunąć zaznaczenie pola **Odczytaj nazwy zmiennych z pliku**, aby nadać każdej zmiennej ogólną nazwę, taką jak Field1, Field2 (Zmienna1, Zmienna2) itd.
- Drugi wiersz musi zawierać pierwszy rekord danych. Nie może zawierać pustych wierszy ani komentarzy.
- Wartości liczbowe nie mogą zawierać separatora tysięcy ani symbolu grupowania — na przykład 3,000.00 bez przecinka. Separatora dziesiętnego (kropka w USA lub w Wielkiej Brytanii) należy używać tylko wówczas, gdy ma to zastosowanie.
- Wartości daty i godziny powinny być zapisywane w jednym z formatów rozpoznawanych w oknie dialogowym opcji strumienia, np. DD/MM/RRRR lub GG:MM:SS. Wszystkie zmienne daty i czasu w pliku powinny być zapisane w tym samym formacie, a w przypadku zmiennych, która zawierają datę, wszystkie wartości tej zmiennej muszą mieć taki sam format.

Określanie opcji dla węzła Plik zmiennych

Opcje ustawiane są na karcie Plik w oknie dialogowym węzła Plik zmiennych.

Plik Należy określić nazwę pliku. Można wpisać nazwę pliku lub kliknąć przycisk wielokropka (**...**), aby wybrać plik. Po wybraniu pliku wyświetlana jest jego ścieżka, a zawartość jest wyświetlana bez separatorów w panelu poniżej.

Przykładowy tekst wyświetlany ze źródła danych można skopiować i wkleić do następujących elementów sterujących: znaki komentarza na końcu wiersza oraz separatory określone przez użytkownika. Do kopiowania i wklejania można użyć kombinacji klawiszy Ctrl-C i Ctrl-V.

Odczytaj nazwy zmiennych z pliku Ta opcja jest wybrana domyślnie; po jej zaznaczeniu pierwszy wiersz w pliku danych jest traktowany jako etykieta kolumny. Jeśli pierwszy wiersz nie jest nagłówkiem, należy usunąć zaznaczenie, aby automatycznie każdej zmiennej nadawana była ogólna nazwa, np. *Field1, Field2* (Zmienna1, Zmienna 2) (dla każdej zmiennej w zbiorze danych).

Określ liczbę zmiennych. Należy określić liczbę zmiennych w każdym rekordzie. Liczba zmiennych może zostać wykryta automatycznie, o ile rekordy są zakończone nową linią. Liczbę można również ustawić ręcznie.

Pomiń znaki nagłówka. Należy określić, ile znaków ma zostać zignorowanych na początku pierwszego rekordu.

Znaki komentarza do końca wiersza. Należy określić znaki, takie jak # lub !, aby wskazać adnotacje w danych. Jeśli jeden z tych znaków znajdzie się w pliku danych, zignorowane zostaną wszystkie wpisy do znaku nowej linii, bez tego znaku.

Usuń wiodące i końcowe spacje. Tę opcję należy wybrać, aby odrzucać spacje wiodące i końcowe w łańcuchach podczas importowania.

Uwaga: Porównania między łańcuchami, które nie korzystają ze wstępnego generowania kodu SQL, mogą zwracać odmienne wyniki w przypadku, gdy istnieją spacje końcowe.

Nieprawidłowe znaki. Należy wybrać opcję **Odrzuć**, aby usunąć niepoprawne znaki z źródła danych. Opcję **Zamień na** należy wybrać, aby zastąpić niepoprawne znaki określonym symbolem (tylko jeden znak). Znaki niepoprawne to znaki null lub dowolne znaki, które nie istnieją w określonej metodzie kodowania.

Kodowanie. Określa typ używanej metody kodowania tekstu. Można wybrać domyślne ustawienie systemowe, domyślne ustawienie strumienia lub UTF-8.

- Domyślne ustawienia systemowe są określone w Panelu sterowania systemu Windows (lub w przypadku trybu rozproszonego — na serwerze).
- Ustawienie domyślne strumienia jest określane w oknie dialogowym Właściwości strumienia.

Separator dziesiętny Należy wybrać typ separatora dziesiętnego, jaki jest używany w źródle danych. **Jak dla strumienia** oznacza znak, który został wybrany na karcie Opcje w oknie dialogowym właściwości strumienia. W przeciwnym razie należy wybrać opcję **Kropka (.)** lub **Przecinek (,)**, aby wszystkie dane z tego okna dialogowego były odczytywane z użyciem wybranego znaku jako separatora dziesiętnego.

Separatorem linii jest znak nowego wiersza Tę opcję należy wybrać, aby jako separatora linii użyć znaku nowego wiersza, a nie separatora zmiennej. Ta opcja może być na przykład przydatna, jeśli w wierszu znajduje się zbyt duża liczba separatorów, co powoduje zawijanie linii. Należy pamiętać, że po zaznaczeniu tej opcji na liście separatorów nie będzie można wybrać opcji **Nowy wiersz**.

Uwaga: Po wybraniu tej opcji wszystkie puste wartości na końcu wierszy danych zostaną usunięte.

Separatory. Korzystając z pól wyboru tego elementu sterującego, można określić, które znaki, takie jak przecinek (,), będą definiowały granice zmiennej w pliku. Można również określić więcej niż jeden separator, np. ", |", o ile w rekordach używanych jest kilka separatorów. Domyślnym separatorem jest przecinek.

Uwaga: Jeśli przecinek jest również zdefiniowany jako separator dziesiętny, ustawienia domyślne nie będą miały tutaj zastosowania. Jeśli przecinek jest ustawiony jako separator zmiennych oraz jako separator dziesiętny, należy z listy separatorów wybrać opcję **Inne**. Następnie należy ręcznie wprowadzić przecinek w polu.

Zaznaczenie opcji **Dopuszczaj wiele pustych separatorów** pozwala traktować wiele sąsiadujących pustych znaków separacji jako pojedynczy separator. Przykładowo, jeśli po wartości danych występują cztery spacje, a następnie kolejna wartość danych, wówczas ta grupa będzie traktowana jako dwie zmienne a nie jako pięć.

Wiersze do przeskanowania dla kolumny i typu Należy określić, ile wierszy i kolumn ma zostać przeskanowanych w celu wyszukania danych odpowiedniego typu.

Automatyczne rozpoznawanie dat i godzin Aby program IBM SPSS Modeler mógł automatycznie rozpoznawać wpisy danych, takie jak daty lub godziny, należy zaznaczyć to pole wyboru. Przykład: oznacza to, że wpis, taki jak 07-11-1965, zostanie zidentyfikowany jako data, a 02:35:58 jako godzina; jednak niejednoznaczne wpisy, takie jak 07111965 lub 023558, będą wyświetlane jako liczby całkowite, ponieważ pomiędzy cyframi nie ma żadnych separatorów.

Uwaga: Aby w przypadku korzystania z plików danych z wcześniejszych wersji programu IBM SPSS Modeler uniknąć potencjalnych problemów z danymi, to pole jest domyślnie wyłączone, informując, że zapis został dokonany w wersjach wcześniejszych niż 13.

Traktuj nawiasy kwadratowe jako listy Po zaznaczeniu tego pola wyboru dane ujęte w nawiasy kwadratowe są traktowane jako pojedyncza wartość, nawet jeśli zawartość stanowią znaki separatora, takie jak przecinki i znaki podwójnego cudzysłowu. Przykładowo, może to dotyczyć danych geoprzestrzennych dwu- lub trzywymiarowych, w których współrzędne ujęte w nawiasy kwadratowe są przetwarzane jak pojedynczy element. Więcej informacji zawiera temat “Importowanie danych geoprzestrzennych do węzła Plik zmiennych” na stronie 29.

Cudzysłowy. Korzystając z list rozwijanych można określić, w jaki sposób pojedyncze i podwójne cudzysłowy będą traktowane podczas importu. Można wybrać opcję **Odrzuć** dla wszystkich cudzysłowów, **Dołącz jako tekst** poprzez umieszczenie ich w wartości zmiennej lub **Złącz w pary i odrzuć**, aby utworzyć pary cudzysłowów i usunąć je. Jeśli znak cudzysłowu nie zostanie dopasowany, użytkownik otrzyma komunikat o błędzie. Opcje **Odrzuć** i **Złącz w pary i odrzuć** powodują zachowanie wartości zmiennej (bez cudzysłowów) w postaci łańcucha.

Uwaga: Po użyciu opcji **Złącz w pary i odrzuć** spacje są zachowywane. Użycie opcji **Odrzuć** powoduje usunięcie spacji wiodących i końcowych wewnątrz cudzysłowów i poza nimi (np. dla zapisu **' " ab c" , "d ef " , " gh i " '** uzyskany zostanie następujący efekt: **'ab c, d ef, gh i'**). W przypadku użycia opcji **Dołącz jako tekst** cudzysłowy są traktowane jako zwykłe znaki, dlatego spacje wiodące i końcowe będą usuwane.

Korzystając z tego okna dialogowego, w dowolnym czasie można kliknąć przycisk **Odśwież**, aby ponownie załadować zmienne ze źródła danych. Jest to przydatne w przypadku zmiany połączeń danych na węzeł źródłowy lub podczas pracy na różnych kartach okna dialogowego.

Importowanie danych geoprzestrzennych do węzła Plik zmiennych

Jeśli węzeł zawiera dane geoprzestrzenne, został utworzony w wyniku eksportu z pliku płaskiego i jest używany w tym strumieniu, w którym został utworzony, zachowa on dane geoprzestrzenne i nie jest konieczna żadna dalsza konfiguracja.

Jeśli jednak węzeł został wyeksportowany i jest używany w innym strumieniu, dane geoprzestrzenne listy są automatycznie przekształcane na format łańcuchowy; konieczne jest wykonanie kilku dodatkowych kroków, aby przywrócić typ składowania Lista oraz powiązane metadane geoprzestrzenne.

Aby uzyskać więcej informacji na temat list, patrz “Składowanie listy i powiązane poziomy pomiaru” na stronie 12.

Więcej informacji na temat szczegółowych ustawień metadanych geoprzestrzennych zawiera temat “Geoprzestrzenne podpoziomy pomiarów” na stronie 141.

Aby skonfigurować metadane geoprzestrzenne, należy wykonać następujące kroki.

1. Na karcie Plik w węźle pliku zmiennych należy zaznaczyć pole wyboru **Traktuj nawiasy kwadratowe jako listy**. Zaznaczenie tego pola wyboru oznacza, że dane ujęte w nawiasy kwadratowe są traktowane jako pojedyncza wartość, nawet jeśli zawartość stanowią znaki separatora, takie jak przecinki i znaki podwójnego cudzysłowu. Jeśli to pole wyboru nie zostanie zaznaczone, dane będą odczytywane jako typ składowania łańcuchowego; wszystkie przecinki w zmiennej będą przetwarzane jak separatory, a struktura danych będzie interpretowana niepoprawnie.
2. Jeśli dane zawierają pojedyncze lub podwójne cudzysłowy, należy wybrać opcję **Złącz w pary i odrzuć** odpowiednio w polach **Apostrofy** i **Cudzysłowy**.
3. Na karcie Dane węzła pliku zmiennych w przypadku zmiennych danych geoprzestrzennych należy zaznaczyć pole wyboru **Nadpisz** i zmienić typ **składowania** z łańcucha na listę.
4. Domyślnie typ **składowania** Lista jest ustawiony jako *Lista liczb rzeczywistych*, a podstawowy typ składowania Wartość zmiennej listy jest ustawiony jako *Liczba rzeczywista*. Aby zmienić podstawowy typ składowania Wartość na Głębokość, należy kliknąć opcję **Określ...**, aby wyświetlić podokno dialogowe Składowanie.
5. W podoknie dialogowym Składowanie można zmodyfikować następujące ustawienia:
   - **Składowanie** Określa ogólny typ składowania dla zmiennej danych. Domyślnie typ składowania jest ustawiony jako Lista; jednak na liście rozwijanej dostępne są wszystkie pozostałe typy składowania (Łańcuch, Liczba całkowita, Liczba rzeczywista, Data, Czas i Znacznik czasu). Jeśli wybrany zostanie jakikolwiek inny typ składowania niż Lista, opcje **Składowanie wartości** i **Głębokość** będą niedostępne.
   - **Składowanie wartości** Określa typy składowania elementów na liście, w odróżnieniu od zmiennej jako całości. Podczas importowania zmiennych geoprzestrzennych jedyne dostępne typy to Liczba rzeczywista i Liczba całkowita; ustawieniem domyślnym jest Liczba rzeczywista.
   - **Głębokość** Określa głębokość zmiennej listy. Wymagana głębokość zależy od typu zmiennej geoprzestrzennej i powinna spełniać następujące kryteria:
     – Punkt – 0

– Łańcuch – 1
– Wielokąt – 1
– Multipunkt – 1
– Multiłańcuch – 2
– Multiwielokąt – 2

Uwaga: Użytkownik musi znać typ zmiennej geoprzestrzennej przekształcanej z powrotem na listę oraz wymaganą dla tej zmiennej głębokość. Jeśli te informacje zostaną ustawione niepoprawnie, użycie zmiennej będzie niemożliwe.

6. Na karcie Typy w węźle Plik zmiennych w przypadku zmiennej danych geoprzestrzennych należy sprawdzić, czy komórka **Poziom pomiaru** zawiera poprawny poziom pomiaru. Aby zmienić poziom, w komórce **Poziom pomiaru** należy kliknąć opcję **Określ...**, co spowoduje wyświetlenie okna dialogowego Wartości.
7. W oknie dialogowym Wartości dla listy wyświetlane są wartości **Poziom pomiaru**, **Składowanie** i **Głębokość**. Należy zaznaczyć opcję **Określ wartości i etykiety** i z listy rozwijanej **Typ** wybrać poprawny typ dla opcji **Poziom pomiaru**. W zależności od wybranej wartości **Typ** może zostać wyświetlony monit o podanie dodatkowych szczegółów, takich jak dane reprezentujące 2 lub 3 wymiary oraz jaki układ współrzędnych jest używany.

Węzeł Plik kolumnowy

Węzły Plik kolumnowy umożliwiają importowanie danych z plików tekstowych ze stałymi zmiennymi (pliki, w których zmienne nie są separowane, ale rozpoczynają się w tym samym miejscu i mają stałą długość). Dane wygenerowane maszynowe lub pochodzące ze starszych wersji często są zapisywane w formacie ze stałymi polami. Korzystając z karty Plik węzła Plik kolumnowy, można w prosty sposób określić pozycję i długość kolumn w danych.

Ustawianie opcji dla węzła Plik kolumnowy

Karta Plik węzła Plik kolumnowy umożliwia wprowadzenie danych do programu IBM SPSS Modeler i określenie położenia kolumn i długości rekordów. Korzystając z panelu podglądu danych w środkowej części okna dialogowego można klikając dodać strzałki, aby określić punkty przerwania pomiędzy zmiennymi.

Plik. Należy określić nazwę pliku. Można wprowadzić nazwę pliku lub kliknąć przycisk wielokropka (**...**) w celu wybrania pliku. Po wybraniu pliku wyświetlana jest jego ścieżka, a jego zawartość wraz z separatorami jest wyświetlana w poniższym panelu.

Panelu podglądu danych można użyć do określenia położenia i długości kolumny. Linijka w górnej części okna podglądu ułatwia zmierzenie długości zmiennych i określenie punktu przerwania pomiędzy nimi. Linie punktów przerwania można określić, klikając w obszarze linijki ponad zmiennymi. Punkty przerwania można przesuwać przeciągając je; można je również odrzucać, przeciągając je poza obszar podglądu danych.

- Każda linia punktu przerwania automatycznie dodaje nową zmienną do zmiennych w poniższej tabeli.
- Pozycje początkowe wskazane przez strzałki są automatycznie dodawane do kolumny początkowej w poniższej tabeli.

Jeden rekord w jednym wierszu. Tę opcję należy wybrać, aby pominąć znak nowej linii na końcu każdego rekordu.

Pomiń wiersze nagłówka. Należy określić, ile wierszy ma zostać zignorowanych na początku pierwszego rekordu. Jest to przydatne do ignorowania nagłówków kolumn.

Długość rekordu. Należy określić liczbę znaków w każdym rekordzie.

Zmienna. Wszystkie zmienne zdefiniowane dla tego pliku danych są tutaj wyświetlane. Możliwe są dwa sposoby zdefiniowania zmiennych:

- Należy określić zmienne interaktywnie za pośrednictwem panelu podglądu danych powyżej.

- Należy określić zmienne ręcznie, dodając puste wiersze zmiennych w poniższej tabeli. Kliknięcie przycisku po prawej stronie panelu zmiennych umożliwia dodanie nowych zmiennych. Następnie w pustej zmiennej można wprowadzić nazwę zmiennej, pozycję początkową oraz długość. Te opcje spowoduję automatyczne dodanie strzałek do panelu podglądu danych, który można w prosty sposób skorygować.

Aby usunąć wcześniej zdefiniowane zmienne, należy wybrać zmienną z listy i kliknąć czerwony przycisk usuwania.

Początek. Należy określić the pozycję pierwszego znaku w zmiennej. Przykładowo, jeśli druga zmienna w rekordzie rozpoczyna się na szóstym znaku, w punkcie początkowym należy prowadzić wartość 16.

Długość. Należy określić, ile znaków znajduje się w najdłuższej wartości każdej zmiennej. Pozwala to określić punkt odcięcia dla kolejnej zmiennej.

Usuń wiodące i końcowe spacje. Tę opcję należy wybrać, aby odrzucić spacje wiodące i końcowe w łańcuchach podczas importowania.

Uwaga: Porównania między łańcuchami, które nie korzystają ze wstępnego generowania kodu SQL, mogą zwracać odmienne wyniki w przypadku, gdy istnieją spacje końcowe.

Nieprawidłowe znaki. Należy wybrać opcję **Odrzuć**, aby usunąć niepoprawne znaki z danych wejściowych. Opcję **Zamień na** należy wybrać, aby zastąpić niepoprawne znaki określonym symbolem (tylko jeden znak). Znaki niepoprawne to znaki null (0) lub dowolne znaki, które nie istnieją w bieżącym kodowaniu.

Kodowanie. Określa typ używanej metody kodowania tekstu. Można wybrać domyślne ustawienie systemowe, domyślne ustawienie strumienia lub UTF-8.

- Domyślne ustawienia systemowe są określone w Panelu sterowania systemu Windows (lub w przypadku trybu rozproszonego — na serwerze).
- Ustawienie domyślne strumienia jest określane w oknie dialogowym Właściwości strumienia.

Separator dziesiętny. Należy wybrać typ separatora dziesiętnego używanego w źródle danych. **Jak dla strumienia** oznacza znak wybrany na karcie Opcje w oknie dialogowym właściwości strumienia. W przeciwnym razie należy wybrać opcję **Kropka (.)** lub **Przecinek (,)**, aby wszystkie dane z tego okna dialogowego były odczytywane z użyciem wybranego znaku jako separatora dziesiętnego.

Automatyczne rozpoznawanie dat i godzin. Aby program IBM SPSS Modeler mógł automatycznie rozpoznawać wpisy danych, takie jak daty lub godziny, należy zaznaczyć to pole wyboru. Przykład: oznacza to, że wpis, taki jak 07-11-1965, zostanie zidentyfikowany jako data, a 02:35:58 jako godzina; jednak niejednoznaczne wpisy, takie jak 07111965 lub 023558, będą wyświetlane jako liczby całkowite, ponieważ pomiędzy cyframi nie ma żadnych separatorów.

Uwaga: Aby w przypadku korzystania z plików danych z wcześniejszych wersji programu IBM SPSS Modeler, to pole jest domyślnie wyłączone, informując, że zapis został wykonany w wersjach wcześniejszych niż 13.

Wiersze do przeskanowania dla typu. Należy określić, ile wierszy ma zostać przeskanowanych w celu wyszukania danych określonego typu.

Korzystając z tego okna dialogowego, w dowolnym czasie można kliknąć przycisk **Odśwież**, aby ponownie załadować zmienne ze źródła danych. Jest to przydatne w przypadku zmiany połączeń danych na węzeł źródłowy lub podczas pracy na różnych kartach okna dialogowego.

Węzeł Plik Statistics

Węzeł Plik Statistics umożliwia odczyt danych bezpośrednio z zapisanego pliku IBM SPSS Statistics (*.sav* lub *.zsav*). Ten format zastępuje teraz plik pamięci podręcznej z wcześniejszych wersji programu IBM SPSS Modeler. Aby zaimportować zapisany plik pamięci podręcznej, należy użyć węzła Plik IBM SPSS Statistics.

Importuj Plik. Należy określić nazwę pliku. Można wpisać nazwę pliku lub kliknąć przycisk wielokropka (**...**), aby wybrać plik. Po wybraniu pliku zostanie wyświetlona jego ścieżka.

Plik jest zaszyfrowany hasłem. Należy zaznaczyć to pole, jeśli wiadomo, że plik jest zabezpieczony hasłem; po wyświetleniu monitu należy wprowadzić **Hasło**. Jeśli plik jest zabezpieczony hasłem i nie zostanie ono wprowadzone, przy próbie przejścia do innej karty, odświeżenia danych, wyświetlenia podglądu zawartości węzła lub próbie wykonania strumienia zawierającego węzeł zostanie wyświetlone ostrzeżenie.

Uwaga: Pliki zabezpieczone hasłem można otwierać tylko w programie IBM SPSS Modeler w wersji 16 lub wyższej.

Nazwy zmiennych. Należy wybrać metodę obsługi nazw zmiennych i etykiet podczas importowania z pliku *.sav* lub *.zsav* programu IBM SPSS Statistics. Wybrane tutaj metadane do uwzględnienia będą dostępne podczas całej pracy w programie IBM SPSS Modeler i można je wyeksportować ponownie w celu użycia z narzędziem IBM SPSS Statistics.

- **Odczytaj nazwy i etykiety.** To pole należy zaznaczyć, aby w programie IBM SPSS Modeler odczytywane były nazwy i etykiety zmiennych. Domyślnie ta opcja jest zaznaczona i w węźle typu wyświetlane są nazwy zmiennych. Etykiety mogą być wyświetlane na wykresach, w przeglądarkach modeli oraz w innego typu wynikach, w zależności od opcji określonych w oknie dialogowym właściwości strumienia. Domyślnie opcja wyświetlania etykiet w wynikach jest wyłączona.
- **Odczytaj etykiety jako nazwy.** Tę opcję należy wybrać, aby odczytywać opisowe etykiety zmiennej z pliku *.sav* lub *.zsav* programu IBM SPSS Statistics zamiast krótkich nazw zmiennych i używać tych etykiet jako nazwy zmiennych w programie IBM SPSS Modeler.

Wartości. Należy wybrać metodę obsługi wartości i etykiet podczas importowania z pliku *.sav* lub *.zsav* programu IBM SPSS Statistics. Wybrane tutaj metadane do uwzględnienia będą dostępne podczas całej pracy w programie IBM SPSS Modeler i można je wyeksportować ponownie w celu użycia z narzędziem IBM SPSS Statistics.

- **Odczytaj dane i etykiety.** Tę opcję należy wybrać, aby odczytywać rzeczywiste wartości i wartości etykiet w programie IBM SPSS Modeler. Domyślnie ta opcja jest zaznaczona i wartości są wyświetlane w węźle typu. Etykiety wartości mogą być wyświetlane w konstruktorze wyrażeń, na wykresach, w przeglądarkach modeli oraz w innego typu wynikach, w zależności od opcji określonych w oknie dialogowym właściwości strumienia.
- **Odczytaj etykiety jako dane.** Tę opcję należy wybrać, jeśli zamiast kodów numerycznych lub symbolicznych używanych do reprezentowania wartości używane były etykiety wartości z pliku *.sav* lub *.zsav*. Na przykład, po wybraniu tej opcji dla danych zawierających zmienną płci, których wartości 1 i 2 w rzeczywistości oznaczają odpowiednio *male* (mężczyzna) i *female* (kobieta), nastąpi przekształcenie zmiennej na łańcuch i zaimportowanie zmiennych *male* i *female* jako wartości rzeczywiste.

  Istotne jest, aby przed zaznaczeniem tej opcji w danych programu IBM SPSS Statistics uwzględnić braki danych. Przykładowo, jeśli zmienna numeryczna używa etykiet tylko dla braków danych (0 = *No Answer* (Brak odpowiedzi), –99 = *Unknown* (Nieznane)), wówczas zaznaczenie powyższej opcji spowoduje zaimportowanie tylko etykiet wartości *No Answer* i *Unknown* i przekształcenie zmiennej na łańcuch. W takich przypadkach należy importować same wartości i ustawić braki danych w węźle typu.

Użyj informacji o formacie zmiennej w celu wymuszenia typu składowania. Jeśli to pole nie jest zaznaczone, wartości zmiennych sformatowane w pliku *.sav* jako liczby całkowite (tj. zmienne określone jako F*n*.0 w widoku zmiennych w programie IBM SPSS Statistics) są importowane z użyciem składowania jako liczba całkowita. Wszystkie pozostałe wartości zmiennych, z wyjątkiem łańcuchów, są importowane jako liczby rzeczywiste.

Jeśli to pole jest zaznaczone (ustawienie domyślne), wszystkie wartości zmiennych, z wyjątkiem łańcuchów, są importowane jako liczby rzeczywiste, niezależnie od tego, czy w pliku *.sav* zostały sformatowane jaki liczby całkowite.

Zestawy wielokrotnych odpowiedzi. Wszystkie zestawy wielokrotnych odpowiedzi zdefiniowane w pliku IBM SPSS Statistics zostaną automatycznie zachowane po zaimportowaniu pliku. Zestawy wielokrotnych odpowiedzi można wyświetlać i edytować w dowolnym węźle po wybraniu zakładki Filtr. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi” na stronie 153.

Węzeł Data Collection

Węzły źródłowe Data Collection importują dane sondażowe korzystając z narzędzia Survey Reporter Developer Kit, które jest udostępniane z produktem Data Collection. Ten format odróżnia *obserwacje* — rzeczywiste odpowiedzi na pytania zgromadzone w czasie ankiety — od *metadanych*, które opisują sposób gromadzenia i rozmieszczania obserwacji. Metadane składają się z informacji, takich jak teksty pytań, nazwy i opisy zmiennych, definicje zmiennych wielokrotnych odpowiedzi, tłumaczenia różnych tekstów oraz definicje struktury obserwacji.

Uwaga: Ten węzeł wymaga narzędzia Survey Reporter Developer Kit, które jest udostępniane razem z produktem Data Collection. Poza zainstalowaniem narzędzia Developer Kit nie jest wymagana żadna dodatkowa konfiguracja.

Komentarze

- Dane sondażowe są odczytywane z formatu VDATA (płaski, tabelaryczny) lub z źródeł w formacie hierarchicznym HDATA, o ile zawierają źródło metadanych.
- Typy są instalowane automatycznie na podstawie informacji z metadanych.
- Podczas importowania danych sondażowych do programu SPSS Modeler pytania są renderowane jako zmienne, wraz z rekordem dla każdego respondenta.

Opcje pliku importu Data Collection

Na karcie Plik w węźle Data Collection można określić opcje dotyczące metadanych oraz obserwacji, jakie mają zostać zaimportowane.

Ustawienia metadanych

Uwaga: Aby wyświetlić pełną listę dostępnych typów plików dostawcy, należy zainstalować program Survey Reporter Developer Kit, dostępny wraz z oprogramowaniem Data Collection.

Dostawca metadanych. Dane sondażowe można zaimportować w różnych formatach obsługiwanych przez program Data Collection Survey Reporter Developer Kit. Poniżej przedstawiono dostępne typy dostawców:

- **DataCollectionMDD.** Odczytuje metadane z pliku definicji kwestionariusza (*.mdd*). Jest to standardowy format modelu danych Data Collection.
- **ADO Database.** Odczytuje obserwacje i metadane z plików ADO. Należy określić nazwę i lokalizację pliku *.adoinfo*, który zawiera metadane. Nazwa wewnętrzna komponentu (DSC) to *mrADODsc*.
- **In2data Database.** Odczytuje obserwacje i metadane z bazy danych In2data. Nazwa wewnętrzna komponentu (DSC) to *mrI2dDsc*.
- **Data Collection Log File.** Odczytuje metadane ze standardowego pliku dziennika Data Collection. Zwykle pliki dzienników mają rozszerzenie *.tmp*. Jednak niektóre pliki dzienników mają inne rozszerzenie nazwy pliku. W razie konieczności można zmienić nazwę pliku, tak aby miała rozszerzenie *.tmp*. Nazwa wewnętrzna komponentu (DSC) to *mrLogDsc*.
- **Quancept Definitions File.** Przekształca metadane na skrypt Quancept. Należy podać nazwę pliku *.qdi* Quancept. Nazwa wewnętrzna komponentu (DSC) to *mrQdiDrsDsc*.
- **Quanvert Database.** Odczytuje obserwacje i metadane z bazy danych Quanvert. Należy określić nazwę i lokalizację pliku *.qvinfo* lub *.pkd*. Nazwa wewnętrzna komponentu (DSC) to *mrQvDsc*.
- **Data Collection Participation Database.** Odczytuje tabele przykładowe i przebiegu historii oraz tworzy pochodne zmienne jakościowe odpowiadające kolumnom w tabelach. Nazwa wewnętrzna komponentu (DSC) to *mrSampleReportingMDSC*.
- **Statistics File.** Odczytuje obserwacje i metadane z pliku IBM SPSS Statistics *.sav*. Zapisuje obserwacje w pliku IBM SPSS Statistics *.sav*, dzięki czemu możliwa jest analiza w programie IBM SPSS Statistics. Zapisuje metadane z pliku IBM SPSS Statistics *.sav* w pliku *.mdd*. Nazwa wewnętrzna komponentu (DSC) to *mrSavDsc*.
- **Surveycraft File.** Odczytuje obserwacje i metadane z bazy danych SurveyCraft. Należy określić nazwę pliku *.vq* SurveyCraft. Nazwa wewnętrzna komponentu (DSC) to *mrSCDsc*.

- **Data Collection Scripting File.** Odczytuje metadane z pliku mrScriptMetadata. Zwykle pliki te mają rozszerzenie *.mdd* lub *.dms*. Nazwa wewnętrzna komponentu (DSC) to *mrScriptMDSC*.
- **Triple-S XML File.** Odczytuje metadane z pliku Triple-S file w formacie XML. Nazwa wewnętrzna komponentu (DSC) to *mrTripleSDsc*.

Właściwości metadanych. Opcjonalnie można wybrać opcję **Właściwości**, aby określić wersję ankiety do zaimportowania oraz język, kontekst i typ etykiety, jakie mają zostać użyte. Więcej informacji można znaleźć w temacie "Właściwości metadanych importu Data Collection" na stronie 35.

Ustawienia danych z obserwacji

Uwaga: Aby wyświetlić pełną listę dostępnych typów plików dostawcy, należy zainstalować program Survey Reporter Developer Kit, dostępny wraz z oprogramowaniem Data Collection.

Pobierz ustawienia danych z obserwacji. Podczas odczytywania metadanych tylko z plików *.mdd* należy kliknąć przycisk **Pobierz ustawienia danych z obserwacji**, aby określić, jakie źródła obserwacji są powiązane z wybranymi metadanymi oraz konkretne ustawienia wymagane do uzyskania dostępu do danego źródła. Ta opcja jest dostępna tylko dla plików *.mdd*.

Dostawca danych z obserwacji. Obsługiwane są następujące typy dostawców:

- **ADO Database.** Odczytuje obserwacje, korzystając z interfejsu Microsoft ADO. Należy wybrać opcję OLE-DB UDL dla typu obserwacji i wprowadzić łańcuch połączenia w polu UDL danych z obserwacji. Więcej informacji można znaleźć w temacie "Łańcuch połączenia z bazą danych" na stronie 36. Nazwa wewnętrzna komponentu (DSC) to *mrADODsc*.
- **Delimited Text File (Excel).** Odczytuje dane z pliku rozdzielanego przecinkami (.CSV), takiego jaki może być uzyskany z programu Excel. Nazwa wewnętrzna to *mrCsvDsc*.
- **Data Collection Data File.** Odczytuje dane z pliku natywnego formatu danych Data Collection. Nazwa wewnętrzna to *mrDataFileDsc*.
- **In2data Database.** Odczytuje obserwacje i metadane z pliku bazy danych In2data (*.i2d*). Nazwa wewnętrzna to *mrI2dDsc*.
- **Data Collection Log File.** Odczytuje obserwacje ze standardowego pliku dziennika Data Collection. Zwykle pliki dzienników mają rozszerzenie *.tmp*. Jednak niektóre pliki dzienników mają inne rozszerzenie nazwy pliku. W razie konieczności można zmienić nazwę pliku, tak aby miała rozszerzenie *.tmp*. Nazwa wewnętrzna to *mrLogDsc*.
- **Quantum Data File.** Odczytuje dane z pliku ASCII w formacie Quantum (*.dat*). Nazwa wewnętrzna to *mrPunchDsc*.
- **Quancept Data File.** Odczytuje obserwacje z pliku *.drs*, *.drz* lub *.dru* komponentu Quancept. Nazwa wewnętrzna to *mrQdiDrsDsc*.
- **Quanvert Database.** Odczytuje obserwacje z pliku *qvinfo* lub *.pkd* komponentu Quanvert. Nazwa wewnętrzna to *mrQvDsc*.
- **Data Collection Database (MS SQL Server).** Odczytuje obserwacje do relacyjnej bazy danych Microsoft SQL Server. Więcej informacji można znaleźć w temacie "Łańcuch połączenia z bazą danych" na stronie 36. Nazwa wewnętrzna to *mrRdbDsc2*.
- **Statistics File.** Odczytuje obserwacje z pliku IBM SPSS Statistics *.sav*. Nazwa wewnętrzna to *mrSavDsc*.
- **Surveycraft File.** Odczytuje obserwacje z pliku *.qdt* komponentu SurveyCraft. Pliki *.vq* i *.qdt* muszą znajdować się w tym samym katalogu, a użytkownik musi mieć prawo do odczytu i zapisu w obu plikach. Nie jest to domyślny sposób, w jaki są tworzone przy użyciu komponentu SurveyCraft, dlatego jeden z tych plików musi zostać przeniesiony, aby zaimportować dane SurveyCraft. Nazwa wewnętrzna to *mrScDsc*.
- **Triple-S Data File.** Odczytuje obserwacje z pliku danych Triple-S, w formacie o ustalonej długości lub rozdzielonym przecinkami. Nazwa wewnętrzna to *mr TripleDsc*.
- **Data Collection XML.** Odczytuje obserwacje z pliku danych XML Data Collection. Zwykle ten format może być używany do przeniesienia obserwacji z jednej lokalizacji do drugiej. Nazwa wewnętrzna to *mrXmlDsc*.

Typ obserwacji. Określa, czy obserwacje są odczytywane z pliku, folderu, OLE-DB UDL lub ODBC DSN i odpowiednio aktualizuje opcje w oknie dialogowym. Poprawne opcje będą zależały od typu dostawcy. W przypadku dostawców baz danych można określić opcje dla połączenia OLE-DB lub ODBC. Więcej informacji można znaleźć w temacie “Łańcuch połączenia z bazą danych” na stronie 36.

Projekt danych z obserwacji. Podczas odczytu obserwacji z bazy danych Data Collection można wprowadzić nazwę projektu. Dla pozostałych typów obserwacji to ustawienie powinno pozostać puste.

Import zmiennej

Importuj zmienne systemowe. Określa, czy zmienne systemowe są importowane, z uwzględnieniem zmiennych, które wskazują status wywiadu (w toku, ukończono, data ukończenia itd.). Można wybrać opcje **Brak**, **Wszystkie** lub **Wspólne**.

Importuj zmienne z kodami pytań otwartych. Steruje importem zmiennych reprezentujących kody używane do otwartych odpowiedzi „Inne” dla zmiennych jakościowych.

Importuj zmienne określające pliki źródłowe elementów. Steruje importem zmiennych, które zawierają nazwy plików obrazów zeskanowanych odpowiedzi.

Importuj zmienne wielokrotnych odpowiedzi jako. Zmienne wielokrotnych odpowiedzi można zaimportować jako wiele zmiennych typu flaga (zestaw wielokrotnych dychotomii); jest to domyślna metoda dla nowych strumieni. Strumienie utworzone w programie IBM SPSS Modeler w wersji wcześniejszej niż 12.0 importowały wielokrotne odpowiedzi do pojedynczej zmiennej, z wartościami rozdzielonymi przecinkami. Starsza metoda jest nadal używana, aby możliwe było uruchomienie istniejących strumieni tak, jak poprzednio; zalecane jednak jest zaktualizowanie starszych strumieni, aby możliwe było użycie nowej metody. Więcej informacji można znaleźć w temacie “Importowanie zestawów wielokrotnych odpowiedzi” na stronie 36.

Właściwości metadanych importu Data Collection

Podczas importowania danych sondażowych Data Collection w oknie dialogowym Właściwości metadanych można określić wersję ankiety do zaimportowania, a także język, kontekst i typ etykiety, jakie mają zostać użyte. Należy pamiętać, że język, kontekst i typ etykiety można importować pojedynczo.

Wersja. Każda wersja ankiety może być traktowana jako obraz stanu metadanych użytych do zgromadzenia konkretnego zbioru obserwacji. Jeśli w kwestionariuszu zostaną wprowadzone zmiany, utworzonych może być kilka jego wersji. Można zaimportować najnowszą wersję, wszystkie wersje lub konkretną wersję.

- **Wszystkie wersje.** Tę opcję należy zaznaczyć, aby użyć kombinacji (nadzbiór) wszystkich dostępnych wersji. (Czasami używana jest nazwa superwersja). Jeśli pomiędzy wersjami występuje konflikt, pierwszeństwo ma najnowsza wersja. Przykładowo, jeśli etykieta kategorii różni się pomiędzy wersjami, użyty zostanie tekst z najnowszej wersji.
- **Najnowsza wersja.** Tę opcję należy wybrać, aby użyć najnowszej wersji.
- **Określ wersję.** Tę opcję należy wybrać, aby użyć konkretnej wersji ankiety.

Wybranie wszystkich wersji jest na przykład przydatne w celu wyeksportowania obserwacji dla więcej niż jednej wersji, a w definicjach zmiennej i kategorii wprowadzono zmiany, co oznacza, że obserwacje zgromadzone z użyciem jednej wersji nie są poprawne w innej wersji. Zaznaczenie wszystkich wersji, dla których mają zostać wyeksportowane obserwacje oznacza, że można wyeksportować obserwacje zgromadzone przy użyciu różnych wersji w tym samym czasie, bez generowania błędów związanych z poprawnością wynikających z różnic pomiędzy wersjami. Jednak w zależności od zmian pomiędzy wersjami niektóre błędy związane z poprawnością nadal mogą być generowane.

Język. Pytania i powiązany z nimi tekst można zapisać w metadanych w wielu językach. Można dla ankiety użyć języka domyślnego lub określić konkretny język. Jeśli pozycja jest niedostępna w określonym języku, używany jest wersja w języku domyślnym.

Kontekst. Należy wybrać kontekst użytkownika, jaki będzie używany. Kontekst użytkownika decyduje tym, które teksty są wyświetlane. Przykładowo można zaznaczyć opcję **Pytanie**, aby wyświetlać teksty pytań lub **Analiza**, aby wyświetlać krótsze teksty, odpowiednie do wyświetlania w czasie analizy danych.

Typ etykiety. Wyświetla listę etykiet, jakie zostały zdefiniowane. Domyślnie jest to **label** (etykieta); jest ona używana w tekstach pytań w kontekście użytkownika Pytanie oraz w opisach zmiennych w kontekście użytkownika Analiza. Pozostałe typy etykiet można zdefiniować dla instrukcji, opisów itp.

Łańcuch połączenia z bazą danych

Podczas korzystania z węzła Data Collection do importowania obserwacji z bazy danych za pośrednictwem komponentu OLE-DB lub ODBC należy wybrać opcję **Edycja** na karcie Plik, aby uzyskać dostęp do okna dialogowego łańcucha połączenia, w którym można dostosować łańcuch przekazany do dostawcy w celu dostosowania połączenia.

Ustawienia zaawansowane

Korzystając z węzła Data Collection do importowania obserwacji z bazy danych, która wymaga jawnego logowania, należy wybrać opcję **Zaawansowane**, aby wprowadzić identyfikator użytkownika i hasło do uzyskania dostępu do źródła danych.

Importowanie zestawów wielokrotnych odpowiedzi

Zmienne wielokrotnych odpowiedzi można zaimportować z programu Data Collection jako zestawy wielokrotnych dychotomii, ze osobną zmienną flagi dla każdej możliwej wartości. Przykładowo, jeśli respondenci są proszeni o wybranie z listy muzeów, które odwiedzili, zestaw będzie zawierał osobną zmienną flagi dla każdego muzeum z listy.

Po zaimportowaniu danych można dodać lub edytować zestawy wielokrotnych odpowiedzi z dowolnego węzła, który zawiera kartę Filtrowanie. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi” na stronie 153.

Importowanie wielokrotnych odpowiedzi do pojedynczej zmiennej (dot. strumieni utworzonych we wcześniejszych wersjach)

W starszych wersjach produktu SPSS Modeler wielokrotne odpowiedzi nie były importowane w opisanych powyżej sposób — były importowane do pojedynczej zmiennej, a wartości były rozdzielane przecinkami. Ta metoda jest nadal dostępna, aby umożliwić obsługę istniejących strumieni, jednak zalecane jest zaktualizowanie tego typu strumieni, tak aby możliwe było zastosowanie nowej metody.

Uwag dotyczące importu kolumny Data Collection

Kolumny z danych Data Collection są odczytywane w programie SPSS Modeler w sposób, którego podsumowanie przedstawiono w poniższej tabeli.

Tabela 4. Podsumowanie importu kolumny Data Collection

Typ kolumny Data Collection	Przechowywanie SPSS Modeler	Poziom pomiaru
Flaga logiczna (tak/nie)	Łańcuch	Flaga (wartości 0 i 1)
Jakościowy	Łańcuch	Nominalny
Data lub znacznik czasu	Znacznik czasu	Ilościowy
Podwójnej precyzji (wartość zmiennopozycyjna w określonym zakresie)	Liczba rzeczywista	Ilościowy
Całkowita długa (wartość całkowita w określonym zakresie)	Liczba całkowita	Ilościowy
Tekst (dowolny opis)	Łańcuch	Nieokreślony

Tabela 4. Podsumowanie importu kolumny Data Collection (kontynuacja)

Typ kolumny Data Collection	Przechowywanie SPSS Modeler	Poziom pomiaru
Poziom (określa siatki lub pętle w pytaniu)	Nie występuje w przypadku formatu VDATA i nie jest importowany do programu SPSS Modeler	
Obiekt (dane binarne, takie jak kopia napisanego tekstu lub nagrania głosowego)	Zaimportowanie do programu SPSS Modeler nie jest możliwe.	
Brak (typ nieznany)	Zaimportowanie do programu SPSS Modeler nie jest możliwe.	
Kolumna Respondent.Serial (tworzy powiązanie unikalnego identyfikatora z każdym respondentem)	Liczba całkowita	Nieokreślony

Aby uniknąć możliwych niespójności pomiędzy etykietami wartości odczytywanymi z metadanych a rzeczywistymi wartościami, wszystkie metadane wartości są przekształcane na małe litery. Przykładowo, etykieta wartości *E1720_years* (E1720_lata) zostanie przekształcona na *e1720_years* (e1720_lata).

węzeł źródłowy IBM Cognos

Węzeł źródłowy IBM Cognos umożliwia przeniesienie danych z bazy danych Cognos lub pojedynczych raportów listy do sesji eksploracji danych. W ten sposób można połączyć funkcje analizy biznesowej Cognos z możliwościami analizy predykcyjnej programu IBM SPSS Modeler. Możliwe jest importowanie danych relacyjnych, zamodelowanych wymiarowo (DMR) oraz danych OLAP.

Korzystając z połączenia z serwerem Cognos, najpierw należy wybrać lokalizację, z której dane lub raporty mają zostać zaimportowane. Lokalizacja obejmuje model Cognos oraz wszystkie foldery, zapytania, raporty, widoki, skróty, adresy URL i definicje zadań powiązane z danym modelem. Model Cognos definiuje reguły biznesowe, opisy danych, relacje danych, wymiary i hierarchie biznesowe oraz inne zadania administracyjne.

W przypadku importowania danych wybierane są obiekty, jakie mają zostać zaimportowane z wybranego pakietu. Do obiektów możliwych do zaimportowania należą obiekty zapytań (które reprezentują tabele baz danych) lub pojedyncze elementy zapytań (które reprezentują kolumny tabeli). Więcej informacji można znaleźć w “Ikony obiektów Cognos”.

Jeśli dla pakietu zdefiniowano filtry, można zaimportować jeden lub kilka filtrów. Jeśli importowany filtr jest powiązany z importowanymi danymi, filtr ten jest stosowany przed zaimportowaniem danych. Dane do zaimportowania muszą być zapisane w formacie UTF-8.

W przypadku importowania raportu należy wybrać pakiet lub folder w pakiecie zawierający co najmniej jeden raport. Następnie należy wybrać pojedynczy raport, jaki ma zostać zaimportowany. Importować można tylko pojedyncze raporty listy; nie ma możliwości obsługi wielu list.

Jeśli zostały zdefiniowane parametry dla obiektu danych lub dla raportu, przed zaimportowaniem obiektu lub raportu można określić wartości dla tych parametrów.

Uwaga: Węzeł źródłowy Cognos obsługuje jedynie pakiety CQM Cognos. Pakiety DQM nie są obsługiwane.

Ikony obiektów Cognos

Obiekty różnego typu, które można importować z bazy danych Cognos Analytics, są reprezentowane przez różne ikony, które przedstawiono w poniższej tabeli.

Tabela 5. Ikony obiektów Cognos

Ikona	Obiekt
	Pakiet
	Przestrzeń nazw
	Obiekt zapytania
	Pozycja zapytania
	Wymiar miary
	Miara
	Wymiar
	Hierarchia poziomów
	Poziom
	Filtrowanie
	Raport
	Samodzielne obliczenie

Importowanie danych Cognos

Aby zaimportować dane z bazy danych IBM Cognos Analytics (obsługiwana jest wersja 11 lub nowsza), na karcie Dane w oknie dialogowym IBM Cognos należy ustawić opcję **Tryb** na wartość **Dane**.

Connection. Kliknięcie opcji **Edytuj** umożliwia wyświetlenie okna dialogowego, w którym można zdefiniować szczegóły nowego połączenia Cognos, za pośrednictwem którego możliwe będzie importowanie danych lub raportów. Jeśli użytkownik jest już zalogowany na serwerze Cognos za pośrednictwem programu IBM SPSS Modeler, może również edytować szczegóły dotyczące bieżącego połączenia. Więcej informacji można znaleźć w “Połączenia Cognos” na stronie 40.

Lokalizacja. Po nawiązaniu połączenia z serwerem Cognos należy kliknąć opcję **Edytuj** obok tego pola, aby wyświetlić listę dostępnych pakietów w celu zaimportowania ich zawartości. Więcej informacji można znaleźć w “Wybór lokalizacji Cognos” na stronie 40.

Zawartość. Wyświetla nazwę wybranego pakietu razem z przestrzeniami nazw powiązanymi z danym pakietem. Dwukrotne kliknięcie przestrzeni nazw pozwala wyświetlić obiekty, jakie można zaimportować. Obiekty różnego typu są oznaczane różnymi ikonami. Więcej informacji można znaleźć w “Ikony obiektów Cognos” na stronie 37.

Aby wybrać obiekt do zaimportowania, należy zaznaczyć obiekt i kliknąć znajdującą się wyżej strzałkę w prawo (jedną z dwóch), aby przenieść obiekt do panelu **Zmienne do zaimportowania**. Zaznaczenie obiektu zapytania spowoduje zaimportowanie wszystkich jego elementów zapytań. Dwukrotne kliknięcie obiektu zapytania powoduje jego rozwinięcie, dzięki czemu można wybrać jeden lub więcej pojedynczych elementów zapytań. Aby dokonać wielokrotnego wyboru, można użyć metody Ctrl+kliknięcie (wybór pojedynczych elementów), Shift+kliknięcie (wybór bloku elementów) oraz Ctrl+A (wybór wszystkich elementów).

Aby wybrać filtr do zastosowania (o ile w pakiecie zdefiniowano filtry), należy przejść do filtru w panelu zawartości, wybrać filtr i kliknąć niższą z dwóch strzałek w prawo, aby przenieść filtr do panelu **Filtry do zastosowania**. Aby dokonać wielokrotnego wyboru, można użyć metody Ctrl+kliknięcie (wybór pojedynczych filtrów) oraz Shift+kliknięcie (wybór bloku filtrów).

Zmienne do zaimportowania. Wyświetla listę obiektów bazy danych, jakie zostały wybrane do zaimportowania do programu IBM SPSS Modeler w celu przetworzenia. Jeśli konkretny obiekt nie jest już potrzebny, należy go zaznaczyć i kliknąć strzałkę w lewo, aby przywrócić go do panelu **Zawartość**. Wielokrotnego wyboru można dokonać w taki sam sposób, jak w przypadku opcji **Zawartość**.

Filtry do zastosowania. Wyświetla wszystkie filtry, jakie zostały wybrane do zastosowania w danych przed ich zaimportowaniem. Jeśli konkretny filtr nie jest już potrzebny, należy go zaznaczyć i kliknąć strzałkę w lewo, aby przywrócić go do panelu **Zawartość**. Wielokrotnego wyboru można dokonać w taki sam sposób, jak w przypadku opcji **Zawartość**.

Parametry. Jeśli przycisk ten jest włączony, zaznaczony obiekt ma zdefiniowane parametry. Przed zaimportowaniem danych możesz użyć parametrów do dopasowania danych (na przykład przeprowadzić sparametryzowanych obliczeń). Jeśli parametry są zdefiniowane, ale nie mają ustalonej wartości domyślnej, przycisk wyświetla trójkąt ostrzegawczy. Kliknięcie przycisku umożliwia wyświetlenie parametrów i opcjonalnie przeprowadzenie ich edycji. Jeśli przycisk jest nieaktywny, dla raportu nie zdefiniowano żadnych parametrów.

Przed importem wykonaj agregację danych. To pole wyboru należy zaznaczyć, aby zaimportować dane zagregowane, a nie dane surowe.

Importowanie raportów Cognos

Aby zaimportować predefiniowany raport z bazy danych IBM Cognos, na karcie Dane w oknie dialogowym IBM Cognos należy ustawić opcję **Tryb** na wartość **Raport**. Importować można tylko pojedyncze raporty listy; nie ma możliwości obsługi wielu list.

Connection. Kliknięcie opcji **Edytuj** umożliwia wyświetlenie okna dialogowego, w którym można zdefiniować szczegóły nowego połączenia Cognos, za pośrednictwem którego możliwe będzie importowanie danych lub raportów. Jeśli użytkownik jest już zalogowany na serwerze Cognos za pośrednictwem programu IBM SPSS Modeler, może również edytować szczegóły dotyczące bieżącego połączenia. Więcej informacji można znaleźć w “Połączenia Cognos” na stronie 40.

Lokalizacja. Po nawiązaniu połączenia z serwerem Cognos należy kliknąć opcję **Edytuj** obok tego pola, aby wyświetlić listę dostępnych pakietów w celu zaimportowania ich zawartości. Więcej informacji można znaleźć w “Wybór lokalizacji Cognos” na stronie 40.

Zawartość. Wyświetla nazwę wybranego pakietu lub folderu, jaki zawiera raporty. Należy przejść do odpowiedniego raportu, zaznaczyć go i kliknąć strzałkę w prawo, aby przenieść raport do pola **Raport do zaimportowania**.

Raport do zaimportowania. Wskazuje raport, jaki został wybrany do zaimportowania do programu IBM SPSS Modeler. Jeśli raport nie jest już potrzebny, należy go zaznaczyć i kliknąć strzałkę w lewo, aby przywrócić raport do panelu **Zawartość** lub wprowadzić inny raport do tego pola.

Parametry. Jeśli przycisk ten jest włączony, zaznaczony obiekt ma zdefiniowane parametry. W celu dokonania poprawek przed zaimportowaniem raportu można użyć parametrów (na przykład określić datę początkową i końcową dla danych raportu). Jeśli parametry są zdefiniowane, ale nie mają ustalonej wartości domyślnej, przycisk wyświetla trójkąt ostrzegawczy. Kliknięcie przycisku umożliwia wyświetlenie parametrów i opcjonalnie przeprowadzenie ich edycji. Jeśli przycisk jest nieaktywny, dla raportu nie zdefiniowano żadnych parametrów.

Połączenia Cognos

W oknie dialogowym Cognos Connections można wybrać serwer Cognos Analytics (obsługiwana jest wersja 11 lub nowsza), z którego obiekty bazy danych będą importowane lub eksportowane.

Adres URL serwera cognos Należy wpisać adres URL serwera Cognos Analytics, z którego dane będą importowane lub eksportowane. Jest to wartość właściwości środowiska „zewnętrznego dyspozytora URI" konfiguracji IBM Cognos na serwerze Cognos. W razie braku pewności, który adres URL ma zostać użyty, należy skontaktować się z administratorem systemu Cognos.

Tryb Należy wybrać opcję **Ustaw dane uwierzytelniające**, aby zalogować się, używając konkretnej przestrzeni nazw, nazwy użytkownika i hasła (na przykład jako administrator). Opcję **Użyj połączenia anonimowego** należy wybrać, aby zalogować się bez danych uwierzytelniających; w takim przypadku nie są wypełniane żadne inne pola.

Alternatywnie, jeśli dostępne są dane uwierzytelniające IBM Cognos, które zostały zapisane w repozytorium IBM SPSS Collaboration and Deployment Services, można użyć tych danych uwierzytelniających zamiast wpisywania nazwy i hasła użytkownika lub tworzenia połączenia anonimowego. Aby użyć istniejących danych uwierzytelniających, należy wybrać opcję **Zapisane dane uwierzytelniające** i wprowadzić wartość w polu **Nazwa danych uwierzytelniających** lub ją wyszukać.

Przestrzeń nazw Cognos jest modelowana na podstawie domeny w programie IBM SPSS Collaboration and Deployment Services.

ID przestrzeni nazw Należy określić dostawcę zabezpieczeń dla uwierzytelniania Cognos, który jest używany do zalogowania się na serwerze. Dostawca autoryzacji jest używany do definiowania i zarządzania użytkownikami, grupami i rolami oraz do kontroli procesu autoryzacji. Należy pamiętać, że jest to identyfikator przestrzeni nazw, a nie nazwa przestrzeni nazw (nie zawsze identyfikator jest taki sam jak nazwa).

Nazwa użytkownika Należy wprowadzić nazwę użytkownika Cognos, która jest używana do zalogowania się na serwerze.

Hasło Należy wprowadzić hasło, które jest powiązane z określoną nazwą użytkownika.

Zapisz jako domyślne Kliknięcie tego przycisku pozwala zapisać ustawienia jako domyślne, aby uniknąć konieczności ponownego wstawiania ich po każdym otwarciu węzła.

Wybór lokalizacji Cognos

Okno dialogowe Określ lokalizację pozwala na wybranie pakietu Cognos, z którego można zaimportować dane, lub pakietu bądź folderu, z którego można zaimportować raporty.

Foldery publiczne. W przypadku importowania danych ta opcja pozwala wyświetlić pakiety i foldery dostępne dla wybranego serwera. Należy wybrać pakiet, jaki ma zostać użyty, i kliknąć przycisk **OK**. Dla węzła źródłowego Cognos można wybrać tylko jeden pakiet.

W przypadku importowania raportów wyświetlane są foldery i pakiety zawierające raporty dostępne z wybranego serwera. Należy wybrać pakiet lub folder raportów i kliknąć przycisk **OK**. Dla węzła źródłowego Cognos można wybrać tylko jeden pakiet lub folder raportów, ale foldery raportów mogą zawierać inne foldery raportów, a także pojedyncze raporty.

Określanie parametrów dla danych lub raportów

Jeśli dla obiektu danych lub dla raportu w programie Cognos Analytics zostały zdefiniowane parametry, przed zaimportowaniem obiektu lub raportu można określić wartości dla tych parametrów. Przykładem parametrów dla raportu mogą być daty początku i końca dla zawartości raportu.

Nazwa. Nazwa parametru, jak określona w bazie danych Cognos.

Typ. Opis parametru.

Wartość. Wartość, która ma zostać przypisana parametrowi. Aby wprowadzić lub edytować wartość, kliknij dwukrotnie jej komórkę w tabeli. Wartości nie są w tym miejscu sprawdzane, dzięki czemu wszystkie nieprawidłowe wartości są wykrywane w czasie wykonywania.

Automatycznie usuń niepoprawne parametry z tabeli. Ta opcja zostaje wybrana domyślnie i spowoduje usunięcie wszystkich niepoprawnych parametrów znalezionych w obiekcie danych lub w raporcie.

Węzeł źródłowy IBM Cognos TM1

Węzeł źródłowy IBM Cognos TM1 umożliwia wprowadzenie danych Cognos TM1 do sesji eksploracji danych. W ten sposób można połączyć funkcje Cognos do planowania w przedsiębiorstwie z możliwościami analizy predykcyjnej programu IBM SPSS Modeler. Istnieje możliwość zaimportowania spłaszczonej wersji danych wielowymiarowej kostki OLAP.

Uwaga: Użytkownik TM1 musi mieć następujące uprawnienia: prawo do zapisu kostek, prawdo do odczytu wymiarów oraz prawo do zapisu elementów wymiaru. Ponadto do zaimportowania i wyeksportowania danych Cognos TM1 za pośrednictwem programu SPSS Modeler wymagana jest instalacja IBM Cognos TM1 10.2, pakiet poprawek 3. Istniejące strumienie, które utworzono na podstawie poprzednich wersji, będą nadal działać.

Dla tego węzła nie są wymagane dane uwierzytelniające administratora. Jeśli jednak nadal używana jest starsza wersja niż węzeł TM1 17.1, dane uwierzytelniające administratora są nadal wymagane.

SPSS Modeler umożliwia współpracę z Cognos TM1 wyłącznie w trybach IntegratedSecurityMode 1, 4 i 5.

Przed zaimportowaniem danych konieczne jest ich zmodyfikowanie w programie TM1. Dane do zaimportowania muszą być zapisane w formacie UTF-8.

Po nawiązaniu połączenia z hostem administracyjnym IBM Cognos TM1 najpierw należy wybrać serwer TM1, z którego dane będą importowane; na serwerze dostępna jest co najmniej jedna kostka TM1. Następnie można wybrać odpowiednią kostkę, a w kostce wybrać kolumny i wiersze do zaimportowania.

Uwaga: Przed użyciem węzła źródłowego lub eksportu TM1 w programie SPSS Modeler należy zweryfikować niektóre ustawienia w pliku tm1s.cfg; jest to plik konfiguracji serwera TM1 znajdujący się w katalogu głównym serwera TM1.

- HTTPPortNumber — należy ustawić poprawny numer portu; zwykle jest to 1-65535. Należy pamiętać, że nie jest to numer portu, który wcześniej podawany był dla połączenia w węźle; jest to używany przez TM1 port wewnętrzny, który został domyślnie wyłączony. W razie konieczności należy skontaktować się z administratorem TM1, aby potwierdzić poprawność ustawienia dla tego portu.
- UseSSL — jeśli ta opcja zostanie ustawiona na *True* (Prawda), jako protokół transportu użyty zostanie protokół HTTPS. W tym przypadku należy zaimportować certyfikat TM1 do środowiska JRE serwera SPSS Modeler Server.

Importowanie danych IBM Cognos TM1

Aby zaimportować dane z bazy danych IBM Cognos TM1, na karcie Dane w oknie dialogowym IBM Cognos TM1 należy wybrać odpowiedniego hosta administracyjnego TM1 oraz powiązane serwer, kostkę i szczegóły danych.

Uwaga: Przed zaimportowaniem danych należy przeprowadzić odpowiednie przetwarzanie z użyciem TM1, aby dane były w formacie rozpoznawanym przez program IBM SPSS Modeler. Obejmuje to filtrowanie danych za pomocą edytora podzbiorów w celu uzyskania informacji na temat odpowiedniej wielkości i odpowiedniego kształtu dla importu.

Należy pamiętać, że wartości zerowe (0) zaimportowane z programu TM1 będą traktowane jako wartości „null" (TM1 nie odróżnia wartości pustych od zerowych). Należy również pamiętać, że dane nienumeryczne (lub metadane) z *wymiarów stałych* mogą być importowane do programu IBM SPSS Modeler. Jednak importowanie *miar* nienumerycznych nie jest obecnie obsługiwane.

Host administracyjny Należy wpisać adres URL hosta administracyjnego, z zainstalowanym serwerem TM1, z którym ma zostać nawiązane połączenie. Host administracyjny jest zdefiniowany jako pojedynczy adres URL dla wszystkich serwerów TM1. Za pomocą tego adresu URL można wykryć wszystkie zainstalowane i uruchomione w danym środowisku serwery IBM Cognos TM1 oraz uzyskać do nich dostęp.

TM1 Server. Po nawiązaniu połączenia z hostem administracyjnym należy wybrać serwer, który zawiera dane do zaimportowania, i kliknąć przycisk **Login**. Jeśli wcześniej nie zostało nawiązane połączenie z tym serwerem, zostanie wyświetlony monit o wprowadzenie danych w polach **Nawa użytkownika** i **Hasło**; alternatywnie, można wyszukać wcześniej wprowadzone dane logowania, zapisane jako **Zapisane dane uwierzytelniające**.

Wybierz widok kostki TM1 do zaimportowania. Wyświetla nazwy kostek na serwerze TM1, do których można zaimportować dane. Dwukrotne kliknięcie kostki umożliwia wyświetlenie danych, jakie można zaimportować.

Uwaga:

Do programu IBM SPSS Modeler można zaimportować tylko kostki z wymiarem.

Jeśli dla elementu w kostce TM1 zdefiniowano alias (na przykład jeśli wartość 23277 ma alias Sales), to zostanie zaimportowana wartość, a nie alias.

Aby wybrać dane do zaimportowania, należy wybrać widok i kliknąć strzałkę w prawo, aby przenieść je do panelu **Widok do zaimportowania**. Jeśli wymagany widok jest niewidoczny, należy kliknąć dwukrotnie kostkę, aby rozwinąć jej listę widoków. Można wybrać widok publiczny lub prywatny.

Wymiary wierszowe. Wyświetla nazwę wymiaru wierszowego w danych, jakie zostały wybrane do zaimportowania. Należy przewinąć listę poziomów i wybrać odpowiedni.

Wymiary kolumnowe. Wyświetla nazwę wymiaru kolumnowego w danych, jakie zostały wybrane do zaimportowania. Należy przewinąć listę poziomów i wybrać odpowiedni.

Wymiary kontekstowe. Tylko wyświetlanie. Pokazuje wymiary kontekstowe, które odnoszą się do wybranych kolumn i wierszy.

Węzeł źródłowy TWC

Węzeł źródłowy TWC importuje dane pogodowe z firmy The Weather Company, IBM Business. Można ich użyć do uzyskania historycznych i prognozowanych danych pogodowych dla określonej lokalizacji. Pozwoli to na opracowywanie rozwiązań biznesowych zależnych od pogody, co pozwoli na łatwiejsze podejmowanie decyzji dzięki zastosowaniu najbardziej dokładnych i precyzyjnych danych pogodowych.

Węzeł ten umożliwia wprowadzenie danych dotyczących pogody, takich jak latitude, longitude, time, day_ind (wskazuje dzień lub noc), temp, dewpt (punkt rosy), rh (wilgotność względna), temperatura feels_like, heat_index, wc (temperatura odczuwalna), wx_phrase (duże zachmurzenie, częściowe zachmurzenie itp), pressure, clds (chmury), vis (widoczność), wspd (prędkość wiatru), gust, wdir (kierunek wiatru), wdir_cardinal (NW, NNW, N itp), uv_index (indeks UV) oraz uv_desc (niski, wysoki itp.).

Węzeł źródłowy TWC używa następujących interfejsów API:

- TWC Historical Observations Airport (http://goo.gl/DplOKj) do pozyskiwania historycznych danych pogodowych
- TWC Hourly Forecast (http://goo.gl/IJhhvZ) do pozyskiwania prognoz

Lokalizacja

Szerokość geograficzna. Należy wprowadzić wartość szerokości geograficznej lokalizacji, dla której mają zostać pozyskane dane pogodowe, w formacie [-90.0~90.0].

Długość geograficzne. Należy wprowadzić wartość długości geograficznej lokalizacji, dla której mają zostać pozyskane dane pogodowe, w formacie [-180.0~180.0].

Różne

Klucz licencyjny. Klucz licencyjny jest wymagany. Należy wprowadzić klucz licencyjny uzyskany z The Weather Company. Jeśli klucz nie jest dostępny, należy skontaktować się z administratorem lub przedstawicielem IBM.

Zamiast wydawać klucz wszystkim użytkownikom, administrator może określić klucz w nowym pliku config.cfg na serwerze IBM SPSS Modeler Server; w takim przypadku to pole może pozostać puste. Jeśli klucz zostanie określony w obu lokalizacjach, klucz z tego okna dialogowego ma pierwszeństwo. Uwaga do administratorów: aby dodać klucz licencyjny na serwerze, należy utworzyć nowy plik o nazwie config.cfg o zawartości LicenseKey=<LICENCEKEY> (gdzie <LICENSEKEY> oznacza klucz licencyjny) w lokalizacji <ModelerServerInstallation>\ext\bin\pasw.twcdata.

Jednostki. Należy wybrać jednostkę miary, jaka będzie użyta: **Angielskie**, **Metryczne** lub **Hybrydowe**. Ustawieniem domyślnym jest **Metryczne**.

Format czasu

UTC. Wybierz format czasu UTC, jeśli importujesz historyczne dane pogodowe i nie chcesz, aby SPSS Modeler uzyskiwał dostęp do interfejsu TWC Hourly Forecast API. W przypadku wybrania tej opcji klucz licencyjny musi zapewniać dostęp wyłącznie do interfejsu TWC Historical Observations Airport API.

Lokalne. Wybierz lokalny format czasu, jeśli SPSS Modeler ma uzyskiwać dostęp do interfejsu TWC Hourly Forecast API w celu przeliczenia czasu z UTC na lokalny. W przypadku wybrania tej opcji klucz licencyjny musi zapewniać dostęp do obu interfejsów API TWC.

Typ danych

Historyczne. Aby zaimportować historyczne dane pogodowe, należy wybrać opcję **Historyczne**, a następnie określić datę początkową i końcową w formacie RRRRMMDD (na przykład 20120101 dla 1 stycznia 2012 roku).

Prognoza. Aby zaimportować prognozowane dane pogodowe, należy wybrać opcję **Prognoza**, the następnie określić godzinę prognozy.

Węzeł źródłowy SAS

Ta funkcja jest dostępna w programach SPSS Modeler Professional i SPSS Modeler Premium.

Węzeł źródłowy SAS umożliwia udostępnienie danych SAS dla sesji eksploracji danych. Można zaimportować cztery typy plików:

- SAS dla systemu Windows/OS2 (*.sd2*)
- SAS dla systemu UNIX (*.ssd*)
- Pliki transportowe SAS (*.tpt*)
- SAS wersja 7/8/9 (*.sas7bdat*)

Podczas importowania danych wszystkie zmienne są zachowywane i żaden typ zmiennej nie ulega zmianie. Wybierane są wszystkie obserwacje.

Ustawianie opcji dla węzła źródłowego SAS

Importuj. Należy wybrać typ pliku SAS do transportu. Można wybrać: **SAS dla Windows/OS2 (.sd2)**, **SAS dla UNIX (.SSD)**, **Pliki transportowe SAS (.tpt)** lub **SAS wersja 7/8/9 (.sas7bdat)**.

Importuj plik. Należy określić nazwę pliku. Można wprowadzić nazwę pliku lub kliknąć przycisk wielokropka (**...**), aby wybrać lokalizację pliku.

Członek. Należy wybrać członka do zaimportowania z pliku transportowego SAS wybranego powyżej. Można prowadzić nazwę członka lub kliknąć opcję **Wybierz**, aby przeglądać wszystkich członków w pliku.

Odczytaj formaty użytkownika z pliku danych SAS. Tę opcję należy wybrać, aby odczytać formaty użytkownika. W plikach SAS zapisywane są dane i formaty danych (takie jak etykiety zmiennych) w różnych plikach. Najczęściej importowane będą również formaty. Jednak w przypadku dużego zbioru danych można usunąć zaznaczenie tej opcji, aby zaoszczędzić pamięć.

Plik formatu. Jeśli plik formatu jest wymagany, to pole tekstowe jest aktywne. Można wprowadzić nazwę pliku lub kliknąć przycisk wielokropka (**...**), aby wybrać lokalizację pliku.

Nazwy zmiennych. Należy wybrać metodę obsługi nazw zmiennych oraz etykiet podczas importowania z pliku SAS. Wybrane tutaj metadane do uwzględnienia zostaną zachowane podczas całej pracy w programie IBM SPSS Modeler i można je wyeksportować ponownie w celu użycia w systemie SAS.

- **Odczytaj nazwy i etykiety.** To pole należy zaznaczyć, aby w programie IBM SPSS Modeler odczytywane były nazwy i etykiety zmiennych. Domyślnie ta opcja jest zaznaczona i w węźle typu wyświetlane są nazwy zmiennych. Etykiety mogą być wyświetlane w konstruktorze wyrażeń, na wykresach, w przeglądarkach modeli oraz w innego typu wynikach, w zależności od opcji określonych w oknie dialogowym właściwości strumienia.
- **Odczytaj etykiety jako nazwy.** Tę opcję należy wybrać, aby odczytywać opisowe etykiety zmiennej z pliku SAS zamiast krótkich nazw zmiennych i używać tych etykiet jako nazwy zmiennych w programie IBM SPSS Modeler.

Węzeł źródłowy programu Excel

Węzeł źródłowy programu Excel umożliwia importowanie danych z programu Microsoft Excel w formacie pliku .xlsx.

Typ pliku. Należy wybrać typ pliku Excel, jaki ma zostać zaimportowany.

Importuj Plik. Określa nazwę i lokalizację pliku arkusza do zaimportowania.

Użyj nazwanego zakresu komórek. Umożliwia określenie nazwanego zakresu komórek, zgodnie z definicją w arkuszu Excel. Należy kliknąć przycisk wielokropka (**...**), aby dokonać wyboru z listy dostępnych zakresów. Jeśli użyty zostanie nazwany zakres, inne ustawienia arkusza i zakresu danych nie będą miały już zastosowania i wyniku tego zostaną wyłączone.

Wybierz arkusz. Określa arkusz do zaimportowania według indeksu lub według nazwy.

- **Według indeksu.** Należy określić wartość indeksu dla arkusza, jaki ma zostać zaimportowany, rozpoczynając od wartości 0 dla pierwszego arkusza, 1 dla drugiego itd.

- **Według nazwy.** Należy określić nazwę arkusza, jaki ma zostać zaimportowany. Należy kliknąć przycisk wielokropka (**...**). Pozwoli to wybrać dostępny arkusz z listy.

Zakres w arkuszu. Dane można importować, rozpoczynając od pierwszego niepustego wiersza lub określonego zakresu komórek.

- **Początek zakresu od pierwszego niepustego wiersza.** Lokalizuje pierwszą niepustą komórkę i używa jej jako górny lewy róg zakresu danych.
- **Określony zakres komórek.** Umożliwia wybranie określonego zakresu komórek według wiersza i kolumny. Na przykład, aby określić zakres w programie Excel A1:D5, można wprowadzić A1 w pierwszym polu i D5 w drugim (lub alternatywnie R1C1 i R5C4). Zwrócone zostaną wszystkie wiersze w określonym zakresie, również wiersze puste.

Na pustych wierszach. Jeśli więcej niż jeden wiersz jest pusty, można wybrać opcję **Przerwij odczyt** lub opcję **Zwróć puste wiersze**, aby kontynuować odczyt wszystkich danych do końca arkusza, wraz z pustymi wierszami.

Pierwszy wiersz zawiera nazwy zmiennych. Oznacza, że pierwszy wiersz w określonym zakresie powinien być używany jako nazwy zmiennych (kolumn). Jeśli ta opcja nie zostanie zaznaczona, nazwy zmiennych będą generowane automatycznie.

Składowanie zmiennej i poziom pomiaru

Podczas odczytywania wartości z programu Excel zmienne z numerycznym typem składowania są odczytywane domyślnie z typem pomiaru *Ilościowy*, z zmienne łańcuchowe są odczytywane jako typ *Nominalny*. Istnieje możliwość ręcznej zmiany poziomu pomiaru (ilościowy a nominalny) na karcie Typ, ale typ składowania jest określany automatycznie (choć w razie konieczności można go zmienić za pomocą funkcji przekształcenia, takiej jak to_integer w węźle wypełniania lub wyliczeń). Więcej informacji można znaleźć w temacie “Ustawienia składowania i formatowania zmiennej” na stronie 9.

Domyślnie zmienne z mieszanymi wartościami numerycznymi i łańcuchowymi są odczytywane jako liczby, co oznacza, że dowolne wartości łańcuchowe zostaną w programie IBM SPSS Modeler ustawione na wartość null (systemowe braki danych). Wynika to z faktu, że w odróżnieniu od programu Excel, program IBM SPSS Modeler nie zezwala na występowanie w zmiennej różnych typów składowania. Aby tego uniknąć, można ręcznie ustawić format komórki na Tekst w arkuszu Excel, co spowoduje, że wszystkie wartości (w tym liczbowe) będą odczytywane jako łańcuchy.

Węzeł źródłowy XML

Ta funkcja jest dostępna w programach SPSS Modeler Professional i SPSS Modeler Premium.

Węzeł źródłowy XML umożliwia importowanie danych z pliku w formacie XML do strumienia IBM SPSS Modeler. XML to standardowy węzeł wymiany danych i wiele organizacji wybiera właśnie ten format. Przykładowo, urząd skarbowy może chcieć przeanalizować dane przesłane online, które są w formacie XML (patrz http://www.w3.org/standards/xml/).

Zaimportowanie danych XML do strumienia IBM SPSS Modeler umożliwi wykonanie wielu funkcji analiz predykcyjnych, korzystając ze źródła. Dane XML są analizowane w formacie tabelarycznym, w którym kolumny odpowiadają różnym poziomom zagnieżdżenia elementów XML i atrybutów. Elementy XML są wyświetlane w formacie XPath (patrz http://www.w3.org/TR/xpath20/).

Ważne: Węzeł źródłowy XML nie uwzględnia deklaracji przestrzeni nazw. Tak więc na przykład pliki XML nie mogą zawierać dwukropka (:) w znaczniku name. Jeśli będą go zawierały, podczas wykonywania wystąpią błędy dotyczące niepoprawnych znaków.

Czytaj z jednego pliku. Domyślnie program SPSS Modeler odczytuje pojedynczy plik, który jest określany w polu **Źródło danych XML**.

Czytaj pliki XML z katalogu. Należy wybrać tę opcję, aby odczytywać wszystkie pliki XML z konkretnego katalogu. W wyświetlonym polu **Katalog** należy określić lokalizację. Zaznaczenie pola wyboru **Uwzględnij podkatalogi** pozwoli dodatkowo odczytywać pliki XML ze wszystkich podkatalogów wybranego katalogu.

Źródło danych XML. Należy wpisać pełną ścieżkę i nazwę pliku źródłowego XML, jaki ma zostać zaimportowany lub użyć przycisku Przeglądaj, aby znaleźć plik.

Schemat XML. (Opcjonalnie) Należy podać ścieżkę i nazwę pliku XSD lub DTD, z którego odczytywana będzie struktura XML, lub użyć przycisku Przeglądaj, aby znaleźć ten plik. Jeśli to pole pozostanie puste, struktura będzie odczytywana z źródłowego pliku XML. Plik XSD lub DTD może zawierać więcej niż jeden element główny (root). W takim przypadku po skoncentrowaniu się na innej zmiennej wyświetlane jest okno dialogowe, w którym można wybrać element główny, jaki ma zostać użyty. Więcej informacji można znaleźć w temacie "Wybór z wielu elementów głównych (root)".

Uwaga: Wskaźniki XSD są ignorowane przez program SPSS Modeler

Struktura XML. Drzewo hierarchiczne przedstawiające strukturę pliku źródłowego XML (lub schematu, jeśli został określony polu **Schemat XML**). Aby zdefiniować granicę rekordu, należy wybrać element i kliknąć przycisk strzałki w prawo, aby skopiować element do pola **Rekordy**.

Pokaż atrybuty. Wyświetla lub ukrywa atrybuty elementów XML w polu **Struktura XML**.

Rekordy (XPath). Wyświetla składnię XPath dla elementu skopiowanego z pola struktury XML. Ten element jest wyróżniony w strukturze XML i definiuje granicę rekordu. Za każdym razem, kiedy ten element zostanie napotkany z pliku źródłowym, tworzony jest nowy rekord. Jeśli to pole jest puste, pierwszy element podrzędny elementu głównego będzie ustalał granicę rekordu.

Czytaj wszystkie dane. Domyślnie, wszystkie dane z pliku źródłowego są wczytywane do strumienia.

Określ czytane dane. Tę opcję należy wybrać, aby zaimportować pojedyncze elementy i/lub atrybuty. Po wybraniu tej opcji aktywowana jest tabela Zmienne, w której można określić dane do zaimportowania.

Pola. W tej tabeli wyświetlane są elementy i atrybuty wybrane do zaimportowania, o ile wybrano opcję **Określ czytane dane**. Można wpisać składnię XPath elementu lub atrybutu bezpośrednio w kolumnie XPath lub wybrać element lub atrybut w strukturze XML i kliknąć przycisk strzałki w prawo, aby skopiować element do tabeli. Aby skopiować wszystkie elementy podrzędne i atrybuty elementu, należy wybrać element w strukturze XML i kliknąć przycisk podwójnej strzałki.

- **XPath.** Składnia XPath elementów do zaimportowania.
- **Lokalizacja.** Lokalizacja elementów do zaimportowania w strukturze XML. Opcja **Ustalona ścieżka** wyświetla ścieżkę elementu w odniesieniu do elementu wyróżnionego w strukturze XML (lub pierwszy element podrzędny pod elementem głównym, jeśli żaden element nie jest wyróżniony). **Dowolna lokalizacja** oznacza element o danej nazwie znajdujący się w dowolnej lokalizacji w strukturze XML. Opcja **Użytkownika** jest wyświetlana, jeśli lokalizacja jest wpisywana bezpośrednio w kolumnie XPath.

Wybór z wielu elementów głównych (root)

Prawidłowo utworzony plik XML może mieć tylko jeden element główny, natomiast pliki XSD lub DTD mogą zawierać kilka elementów głównych. Jeśli jeden z elementów głównych jest zgodny z elementem w pliku źródłowym XML, wówczas użyty zostanie ten element główny; w przeciwnym razie należy wybrać element główny.

Wybierz element główny (root). Należy wybrać element główny, jaki będzie używany. Domyślnie, jest to pierwszy element główny w strukturze XSD lub DTD.

Usuwanie niepotrzebnych spacji z danych źródłowych XML

Znaki podziału wiersza w danych źródłowych XML mogą być zaimplementowane przez kombinację znaków [CR][LF]. W niektórych przypadkach znaki podziału wiersza mogą występować w środku łańcucha tekstowego, na przykład:

```
<description>An in-depth look at creating applications[CR][LF]
with XML.</description>
```

Te znaki podziału wiersza mogą być niewidoczne w przypadku otwarcia pliku za pomocą niektórych aplikacji, takich jak przeglądarka WWW. Jednak jeśli dane są wczytywane do strumienia za pośrednictwem węzła źródłowego XML, znaki podziału wiersza są przekształcane na szereg znaków spacji.

Można to poprawić, używając węzła wypełniania w celu usunięcia niepotrzebnych spacji:

Tutaj zamieszczono przykład, w jaki sposób można to zrobić:

1. Dołącz węzeł wypełniania do węzła źródłowego XML.
2. Otwórz węzeł wypełniania i użyj selektora zmiennych, aby wybrać zmienną z niepotrzebnymi spacjami.
3. Ustaw wartość **Zastąp** na **W oparciu o warunek**, a opcję **Warunek** na **prawda**.
4. W polu **Zamień na** wprowadź replace(" ","",@FIELD) i kliknij przycisk OK.
5. Dołącz węzeł tabeli do węzła wypełnienia i uruchom strumień.

W wynikach węzła Tabela tekst będzie teraz wyświetlany bez dodatkowych spacji.

Węzeł Dane niestandardowe

Węzeł Dane niestandardowe udostępnia prosty sposób na utworzenie danych syntetycznych — od podstaw lub poprzez zmianę istniejących danych. Jest to przydatne na przykład podczas tworzenia testowego zbioru danych do modelowania.

Tworzenie danych od podstaw

Węzeł Dane niestandardowe jest dostępny z palety Źródła i można go dodać bezpośrednio do obszaru roboczego strumienia.

1. Kliknij zakładkę **Źródła** w palecie węzłów.
2. Przeciągnij i upuść lub kliknij dwukrotnie węzeł Dane niestandardowe, aby dodać go do obszaru roboczego strumienia.
3. Kliknij dwukrotnie, aby otworzyć jego okno dialogowe i określić zmienne i wartości.

Uwaga: Węzły danych niestandardowych wybrane z palety Źródła będą puste, bez żadnych informacji na temat zmiennych i danych. Umożliwia to utworzenie danych syntetycznych całkowicie od podstaw.

Generowanie danych z istniejącego źródła danych

Można również wygenerować węzeł danych użytkownika z dowolnego węzła w strumieniu, który nie jest węzłem końcowym:

1. Zdecyduj, w którym punkcie strumienia węzeł ma zostać zastąpiony.
2. Kliknij prawym przyciskiem myszy węzeł, z którego dane zostaną przekazane do węzła danych użytkownika, i wybierz z menu opcję **Generuj węzeł danych użytkownika**.
3. Węzeł danych użytkownika zostanie wyświetlony we wszystkich procesach w dół strumienia z nim powiązanych, zastępując istniejący węzeł w punkcie wskazanym w strumieniu danych. Po wygenerowaniu węzeł dziedziczy z metadanych wszystkie informacje na temat struktury danych i typu zmiennych (o ile są dostępne).

Uwaga: Jeśli dane nie były uruchomione we wszystkich węzłach w strumieniu, wówczas węzły nie są w pełni określone, co oznacza, że dane dotyczące składowania i wartości danych mogą być niedostępne podczas zastępowania węzła danych niestandardowych.

Ustawianie opcji dla węzła Dane niestandardowe

Okno dialogowe węzła danych użytkownika zawiera kilka narzędzi, jakich można używać do wprowadzania wartości i definiowania struktury danych syntetycznych. Dla wygenerowanego węzła w tabeli na karcie Dane znajdują się nazwy zmiennych z oryginalnego źródła danych. Dla węzła dodanego z palety Źródła tabela jest pusta. Korzystając z opcji tabeli, można wykonać następujące zadania:

- Dodać nowe zmienne za pomocą przycisku Dodaj nową zmienną po prawej stronie tabeli.
- Zmienić nazwy istniejących zmiennych.
- Określić typ składowania danych dla poszczególnych zmiennych.
- Określić wartości.
- Zmienić kolejność wyświetlania zmiennych.

Wprowadzanie danych

Dla każdej zmiennej można określić wartości lub wstawić wartości z oryginalnego zbioru danych, używając przycisku wyboru wartości po prawej stronie tabeli. Więcej informacji na temat określania wartości można znaleźć w opisach reguł zamieszczonych poniżej. Zmienna może pozostać pusta — takie zmienne są wypełniane przez systemową wartość null ($null$).

Aby określić wartości łańcucha, należy je wpisać w kolumnie Wartości, rozdzielając je spacjami:

```
Fred Ethel Martin
```

Łańcuchy mogą zawierać spacje ujęte w podwójne cudzysłowy:

```
"Bill Smith" "Fred Martin" "Jack Jones"
```

W przypadku zmiennych numerycznych można wprowadzić wiele wartości w taki sam sposób (wymienić je, wstawiając między nimi spacje):

```
10 12 14 16 18 20
```

Lub można określić taki sam szereg liczb, ustawiając jego ograniczenia (10, 20) oraz liczbę kroków między nimi (2). Używając tej metody, należy wpisać:

```
10,20,2
```

Te dwie metody można połączyć, osadzając jedną w drugiej, np.

```
1 5 7 10,20,2 21 23
```

Taki zapis da następujące wartości:

```
1 5 7 10 12 14 16 18 20 21 23
```

Wartości daty i czasu można wprowadzić z użyciem bieżącego domyślnego formatu, jaki został wybrany w oknie dialogowym właściwości strumienia, przykładowo:

```
11:04:00 11:05:00 11:06:00
2007-03-14 2007-03-15 2007-03-16
```

W przypadku wartości znacznika czasu, które zawierają składnik daty i czasu, konieczne jest użycie cudzysłowów:

```
"2007-03-14 11:04:00" "2007-03-14 11:05:00" "2007-03-14 11:06:00"
```

Dodatkowe informacje można znaleźć w komentarzach dotyczących składowania danych zamieszczonych poniżej.

Generuj dane. Umożliwia określenie sposobu generowania rekordów po uruchomieniu strumienia.

- **Wszystkie kombinacje wartości**. Generuje rekordy zawierające wszystkie możliwe kombinacje wartości zmiennych, dlatego każda zmienna będzie wyświetlana w kilku rekordach. Niekiedy może to spowodować wygenerowanie zbyt dużej ilości danych, dlatego często ten węzeł może być poprzedzony węzłem próby.
- **W kolejności wprowadzania**. Generuje rekordy w kolejności, w jakiej wprowadzane były wartości zmiennych danych. Każda wartość zmiennej będzie wyświetlana tylko w jednym rekordzie. Łączna liczba rekordów jest równa największej liczbie wartości dla jednej zmiennej. Jeśli w zmiennych jest mniej wartości niż największa liczba, wstawiane są wartości niezdefiniowane ($null$).

Przykład

Przykładowo, przedstawione poniżej wpisy spowodują wygenerowanie rekordów przedstawionych w dwóch poniższych przykładowych tabelach.

- **Wiek.** 30,60,10
- **Ciśnienie krwi.** NISKIE
- **Cholesterol.** W NORMIE WYSOKI
- **Lek.** (puste)

Tabela 6. Generowanie zbioru zmiennych — ustawienie Wszystkie kombinacje wartości

Wiek	CIŚNIENIE KRWI	Cholesterol	Lek
30	NISKIE	W NORMIE	$null$
30	NISKIE	WYSOKI	$null$
40	NISKIE	W NORMIE	$null$
40	NISKIE	WYSOKI	$null$
50	NISKIE	W NORMIE	$null$
50	NISKIE	WYSOKI	$null$
60	NISKIE	W NORMIE	$null$
60	NISKIE	WYSOKI	$null$

Tabela 7. Generowanie zbioru zmiennych — ustawienie W kolejności wprowadzania

Wiek	CIŚNIENIE KRWI	Cholesterol	Lek
30	NISKIE	W NORMIE	$null$
40	$null$	WYSOKI	$null$
50	$null$	$null$	$null$
60	$null$	$null$	$null$

Składowanie danych

Składowanie to sposób przechowywania danych w zmiennej. Przykładowo w zmiennej zawierającej wartości 1 i 0 składowane są dane w postaci liczb całkowitych. Różni się to od poziomu pomiaru, który opisuje użycie danych i nie wpływa na składowanie. Można na przykład ustawić poziom pomiaru dla zmiennej całkowitej zawierającej wartości 1 i 0 jako *Flaga*. Zwykle oznacza to, że 1 = *Prawda*, a 0 = *Fałsz*. Składowanie musi być określone w źródle, natomiast poziom pomiaru można zmienić za pomocą węzła Typy w dowolnym miejscu w strumieniu. Więcej informacji można znaleźć w temacie “Poziomy pomiaru” na stronie 139.

Dostępne typy składowania to:

- **Łańcuch** Używany w przypadku zmiennych zawierających dane nienumeryczne, zwane również danymi alfanumerycznymi. Łańcuch może zawierać dowolną sekwencję znaków, np. *fred*, *Klasa 2* lub *1234*. Należy pamiętać, że liczby użyte w łańcuchach nie mogą być wykorzystywane do obliczeń.

- **Liczba całkowita** Zmienna, której wartości są liczbami całkowitymi.
- **Liczba rzeczywista** Wartości są liczbami, które mogą mieć miejsca dziesiętne (bez ograniczenia do liczb całkowitych). Format wyświetlania jest określany w oknie dialogowym Właściwości strumienia i może być zastąpiony dla pojedynczych zmiennych w węźle Typy (karta Format).
- **Data** Wartości daty określone w standardowym formacie, takie jak rok, miesiąc i dzień (np. 2007-09-26). Konkretny format jest określany w oknie dialogowym Właściwości strumienia.
- **Czas** Czas mierzony jako czas trwania. Przykładowo połączenie z serwisem trwające 1 godzinę, 26 minut i 38 sekund może być zapisane jako 01:26:38, w zależności od obowiązującego formatu czasu określonego w oknie dialogowym Właściwości strumienia.
- **Znacznik czasu** Wartości składające się z daty i czasu, na przykład 2007–09–26 09:04:00, ponownie w zależności od obowiązujących formatów daty i czasu określonych w oknie dialogowym Właściwości strumienia. Należy pamiętać, że konieczne może być ujęcie wartości znacznika czasu w podwójnym cudzysłowie, aby były interpretowane jako pojedyncza wartość, a nie jako osobne wartości daty i czasu. (Dotyczy to na przykład sytuacji, kiedy wartości są wprowadzane w węźle Dane niestandardowe).
- **Lista** Zmienna składowania Lista wprowadzona w programie SPSS Modeler, wersja 17, wraz z nowymi poziomami pomiaru Geoprzestrzenny i Przedziałowy, zawiera wiele wartości dla pojedynczego rekordu. Dostępne są wersje list dla wszystkich pozostałych typów składowania.

Tabela 8. Lista ikon typów składowania

Ikona	Typ składowania
[A]	Lista łańcuchów
[ ]	Lista liczb całkowitych
[ ]	Lista liczb rzeczywistych
[ ]	Lista godzin
[ ]	Lista dat
[ ]	Lista znaczników czasu
[[]]	Lista o głębokości większej niż zero

Dodatkowo do użycia z poziomem pomiaru Przedziałowy dostępne są wersje listy następujących poziomów pomiaru.

Tabela 9. Lista ikon poziomu pomiaru

Ikona	Poziom pomiaru
[ ]	Lista ilościowych
[ ]	Lista jakościowych

Tabela 9. Lista ikon poziomu pomiaru (kontynuacja)

Ikona	Poziom pomiaru
[ ]	Lista flag
[ ]	Lista nominalnych
[ ]	Lista porządkowych

Listy mogą być importowane do programu SPSS Modeler w jednym z trzech węzłów źródłowych (Analytic Server, Geoprzestrzenny lub Plik zmiennych) lub mogą być utworzone na podstawie strumieni za pośrednictwem węzłów działania zmiennych Wyliczanie lub Wypełnianie.

Więcej informacji na temat list i ich interakcji z poziomami pomiaru Przedziałowy i Geoprzestrzenny zawiera sekcja “Składowanie listy i powiązane poziomy pomiaru” na stronie 12.

Przekształcanie typów składowania. Można przekształcić typ składowania dla zmiennej, korzystając z różnych funkcji przekształcania, takich jak to_string i to_integer, dostępnych w węźle wypełniania. Więcej informacji można znaleźć w temacie “Przekształcanie sposobu składowania za pomocą węzła Wypełnianie” na stronie 162. Funkcje przekształcania (i inne funkcje wymagające określonego typu danych wejściowych, np. wartość daty lub czasu) są uzależnione od bieżących formatów określonych w oknie dialogowym Właściwości strumienia. Na przykład, aby wykonać przekształcenie zmiennej łańcuchowej o wartościach *Sty 2018*, *Lut 2018* itd. do postaci składowania daty, jako domyślny format daty strumienia należy wybrać **MIE RRRR**. Funkcje przekształcania są również dostępne z węzła wyliczeń i umożliwiają tymczasowe przekształcenie podczas wyliczania. Węzła wyliczeń można także użyć do wykonywania innych działań, takich jak rekodowanie zmiennych łańcuchowych przez wartości jakościowe. Więcej informacji można znaleźć w temacie “Rekodowanie wartości za pomocą węzła wyliczeń” na stronie 161.

Wczytywanie danych mieszanych. Należy zwrócić uwagę, że podczas wczytywania zmiennych z liczbowym typem składowania (liczby całkowite, rzeczywiste, czas, znacznik czasu lub data) wszelkie wartości nieliczbowe są zamieniane na null lub braki systemowe. Wynika to z faktu, że w odróżnieniu od niektórych aplikacji produkt IBM SPSS Modeler nie zezwala na przechowywanie w zmiennej danych różnego typu. Aby uniknąć takiej sytuacji, wszelkie zmienne z danymi mieszanymi należy wczytywać jako łańcuchy, zmieniając w razie potrzeby typ składowania w węźle źródłowym lub aplikacji zewnętrznej.

Uwaga: Wygenerowane węzły danych użytkownika mogą już zawierać informacje na temat składowania pobrane z węzła źródłowego, o ile został określony. Węzeł, który nie jest określony, nie zawiera informacji na temat typu składowania lub użycia.

Reguły dotyczące określania wartości

W przypadku zmiennych symbolicznych pomiędzy wieloma wartościami należy wstawić spacje, na przykład:

```
WYSOKIE ŚREDNIE NISKIE
```

W przypadku zmiennych numerycznych można wprowadzić wiele wartości w taki sam sposób (wymienić je, wstawiając między nimi spacje):

```
10 12 14 16 18 20
```

Lub można określić taki sam szereg liczb, ustawiając jego ograniczenia (10, 20) oraz liczbę kroków między nimi (2). Używając tej metody, należy wpisać:

```
10,20,2
```

Te dwie metody można połączyć, osadzając jedną w drugiej, np.

1 5 7 10,20,2 21 23

Taki zapis da następujące wartości:

1 5 7 10 12 14 16 18 20 21 23

Węzeł Symulacje Generowanie

Węzeł Symulacje Generowanie zapewnia łatwy sposób na wygenerowanie danych objętych symulacją — bez danych historycznych, korzystając z rozkładów statystycznych określonych przez użytkownika lub automatycznie, korzystając z rozkładów uzyskanych po uruchomieniu węzła Symulacje Dopasowanie dla istniejących danych historycznych. Generowanie symulowanych danych jest przydatne, kiedy ma zostać przeprowadzona ocena wyniku modelu predykcyjnego przy braku pewności dla danych wejściowych modelu.

Tworzenie danych bez danych historycznych

Węzeł Symulacje Generowanie jest dostępny z palety Źródła i można go dodać bezpośrednio do obszaru roboczego strumienia.

1. Kliknij zakładkę **Źródła** w palecie węzłów.
2. Przeciągnij i upuść lub kliknij dwukrotnie węzeł Symulacje Generowanie, aby dodać go do obszaru roboczego strumienia.
3. Kliknij dwukrotnie, aby otworzyć okno dialogowe i określić zmienne, typy składowania, rozkłady statystyczne i parametry rozkładu.

Uwaga: Węzły Symulacje Generowanie wybrane z palety Źródła będą puste, bez żadnych informacji na temat zmiennych i rozkładu. Umożliwia to tworzenie całkowicie symulowanych danych bez danych historycznych.

Generowanie symulowanych danych z użyciem istniejących danych historycznych

Węzeł Symulacje Generowanie może być również utworzony poprzez wykonanie węzła końcowego Symulacje Dopasowanie:

1. Kliknij prawym przyciskiem myszy węzeł Symulacje Dopasowanie i wybierz z menu opcję **Uruchom**.
2. Węzeł Symulacje Generowanie zostaje wyświetlony w obszarze roboczym strumienia z łączem aktualizacji do węzła Symulacje Dopasowanie.
3. Po wygenerowaniu węzeł Symulacje Generowanie dziedziczy informacje na temat wszystkich zmiennych, typów składowania i rozkładu statystycznego z węzła Symulacje Dopasowanie.

Definiowanie łącza aktualizacji do węzła Symulacje Dopasowanie

Można utworzyć połączenie pomiędzy węzłem Symulacje Generowanie a węzłem Symulacje Dopasowanie. Jest to przydatne, aby zaktualizować co najmniej jedna zmienną na podstawie informacji dotyczących najlepszego dopasowania rozkładu, ustalonego w oparciu o dopasowanie do danych historycznych.

1. Kliknij węzeł Symulacje Generowanie prawym przyciskiem myszy.
2. Z menu wybierz opcję **Zdefiniuj łącze aktualizacji**. Kursor zmieni się na kursor łącza.
3. Kliknij kolejny węzeł. Jeśli ten węzeł jest węzłem Symulacje Dopasowanie, tworzone jest łącze. Jeśli ten węzeł jest węzłem Symulacje Dopasowanie, nie jest tworzone żadne łącze, a kursor ponownie zmienia się na zwykły.

Jeśli zmienne w węźle Symulacje Generowanie różnią się od zmiennych w węźle Symulacje Dopasowanie, wyświetlany jest komunikat z informacją o różnicy.

Jeśli węzeł Symulacje Dopasowanie jest używany aktualizowania powiązanego węzła Symulacje Generowanie, wynik zależy od tego, czy w obu węzłach znajdują się takie same zmienne oraz czy zmienne te są odblokowane w węźle Symulacje Generowanie. Wyniki aktualizacji węzła Symulacje Dopasowanie są wyświetlane w następującej tabeli.

Tabela 10. Wyniki aktualizacji węzła Symulacje Dopasowanie

	Zmienna w węźle	**dopasowania symulacji**
Zmienna w węźle Symulacje Generowanie	**Dostępne**	**Braki**
Dostępna i odblokowana.	Zmienna jest nadpisywana.	Zmienna jest usuwana.
Brak.	Zmienna jest dodawana.	Brak zmiany.
Dostępna i zablokowana.	Rozkład zmiennej nie jest nadpisywany. Aktualizowane są informacje w oknie dialogowym Szczegóły dopasowania oraz korelacje.	Zmienna nie jest nadpisywana. Korelacje są ustawiane na wartość zero.
Zaznaczane jest pole wyboru **Nie czyść wartości min. i maks. podczas ponownego dopasowania**.	Zmienna jest nadpisywana, oprócz wartości	w kolumnie Minimum i Maksimum.
Zaznaczane jest pole wyboru **Nie obliczaj ponownie korelacji podczas ponownego dopasowania**.	Jeśli zmienna nie jest zablokowana, zostaje nadpisana.	Korelacje nie są nadpisywane.

Usuwanie łącza aktualizacji do węzła Symulacje Dopasowanie

Istnieje możliwość usunięcia łącza pomiędzy węzłem Symulacje Generowanie a węzłem Symulacje Dopasowanie; w tym celu należy wykonać następujące kroki:

1. Kliknij węzeł Symulacje Generowanie prawym przyciskiem myszy.
2. Z menu wybierz opcję **Usuń łącze aktualizacji**. Łącze zostaje usunięte.

Ustawianie opcji dla węzła Symulacje Generowanie

Opcje na karcie Dane w oknie dialogowym węzła Symulacje Generowanie umożliwiają wykonanie następujących operacji:

- Wyświetlanie, określanie i edytowanie informacji o rozkładzie statystycznym dla zmiennych.
- Wyświetlanie, określanie i edytowanie korelacji pomiędzy zmiennymi.
- Określanie liczby iteracji i obserwacji, jakie będą objęte symulacją.

Wybierz element. Umożliwia przełączanie pomiędzy trzema widokami węzła Symulacje Generowanie: Zmienne symulowane, Korelacje i Opcje zaawansowane.

Widok Zmienne symulowane

Jeśli węzeł Symulacje Generowanie został wygenerowany lub zaktualizowany na podstawie węzła Symulacje Dopasowanie z użyciem danych historycznych, w widoku Zmienne symulowane można wyświetlić i edytować informacje o rozkładzie statystycznym dla każdej zmiennej. Na kartę **Typy** węzła Symulacje Generowanie z węzła Symulacje Dopasowanie kopiowane są następujące informacje na temat każdej zmiennej:

- Poziom pomiaru
- Wartości
- Braki
- Sprawdź
- Rola

Jeśli nie są dostępne żadne dane historyczne, można zdefiniować zmienne i określić ich rozkłady, wybierając typ składowania oraz wybierając typ rozkładu i wprowadzając wymagane parametry. Wygenerowanie danych w taki sposób oznacza, że informacje na temat poziomu pomiaru poszczególnych zmiennych nie będą dostępne, dopóki dane nie zostaną określone, na przykład na karcie **Typy** lub w węźle typu.

Widok Zmienne symulowane zawiera kilka narzędzi, których można użyć do wykonania następujących zadań:

- Dodawanie i usuwanie zmiennych.
- Zmienić kolejność wyświetlania zmiennych.
- Określanie typu składowania dla każdej zmiennej.
- Określanie rozkładu statystycznego dla każdej zmiennej.
- Określanie wartości parametrów rozkładu statystycznego każdej zmiennej.

Zmienne symulowane. Ta tabela zawiera jeden pusty wiersz, jeśli węzeł Symulacje Generowanie został dodany do obszaru roboczego strumienia z palety Źródła. Jeśli ten wiersz jest edytowany, nowy pusty wiersz jest dodawany u dołu tabeli. Jeśli węzeł Symulacje Generowanie został utworzony z węzła Symulacje Dopasowanie, ta tabela będzie zawierała jeden wiersz dla każdej zmiennej z danych historycznych. Dodatkowe wiersze można dodać do tabeli, klikając ikonę **Dodaj nową zmienną**.

Tabela Zmienne symulowane składa się z następujących kolumn:

- **Zmienna.** Zawiera nazwy zmiennych. Nazwy zmiennych mogą być edytowane poprzez wpisanie ich w komórkach.
- **Składowanie.** Komórki w tej kolumnie zawierają listę rozwijaną typów składowania. Dostępne typy składowania to: **Łańcuch**, **Liczba całkowita**, **Liczba rzeczywista**, **Czas**, **Data** i **Znacznik czasu**. Wybrany typ składowania określa, które rozkłady będą dostępne w kolumnie rozkładu. Jeśli węzeł Symulacje Generowanie został utworzony na podstawie węzła Symulacje Dopasowanie, typ składowania jest kopiowany z węzła Symulacje Dopasowanie.

  Uwaga: W przypadku zmiennych z typem składowania data/czas należy określić parametry rozkładu jako liczby całkowite. Przykładowo, aby określić średnią datę jako 1 stycznia 1970, należy użyć liczby całkowitej 0. Oznaczona liczba całkowita reprezentuje liczbę sekund po (lub przed) północą 1 stycznia 1970.
- **Status.** Ikony w kolumnie Status określają status dopasowania dla każdej zmiennej.

Żaden rozkład nie został określony dla zmiennej lub brakuje co najmniej jednego parametru rozkładu. W celu uruchomienia symulacji należy określić rozkład dla tej zmiennej i wprowadzić poprawne wartości parametrów.

Zmienna jest ustawiona jako najlepiej dopasowany rozkład. **Uwaga:** Ta ikona może być wyświetlana tylko w przypadku utworzenia węzła Symulacje Generowanie na podstawie węzła Symulacje Dopasowanie.

Najlepiej dopasowany rozkład został zastąpiony alternatywnym rozkładem z podokna dialogowego Szczegóły dopasowania. Więcej informacji można znaleźć w temacie “Szczegóły dopasowania” na stronie 59.

Rozkład został ręcznie określony lub był edytowany i może zawierać parametr określony na więcej niż jednym poziomie.

- **Zablokowane.** Zablokowanie symulowanej zmiennej, poprzez zaznaczenie pola wyboru w kolumnie z ikoną blokady, wyklucza zmienną z automatycznego aktualizowania za pośrednictwem połączonego węzła Symulacje

Dopasowanie. Najczęściej jest to przydatne w przypadku ręcznego określania rozkładu, kiedy użytkownik chce mieć pewność, że nie będzie miało na niego wpływu automatyczne dopasowanie rozkładu podczas wykonywania połączonego węzła Symulacje Dopasowanie.

- **Rozkład.** Komórki w tej kolumnie zawierają listę rozwijaną rozkładów statystycznych. Wybrany typ składowania określa, które rozkłady będą dostępne w tej kolumnie dla danej zmiennej. Więcej informacji można znaleźć w temacie “Rozkłady” na stronie 61.

  Uwaga: Nie można określić rozkładu stałego dla każdej zmiennej. Jeśli każda zmienna w wygenerowanych danych ma być stała, można po węźle danych użytkownika należy wstawić węzeł zrównoważenia.
- **Parametry.** W tej kolumnie wyświetlane są parametry rozkładu powiązane z każdym dopasowanym rozkładem. W przypadku kilku wartości parametru są one rozdzielane przecinkami. Określenie wielu wartości dla parametru powoduje wygenerowanie wielu iteracji dla symulacji. Więcej informacji można znaleźć w temacie “Iteracje” na stronie 61. Jeśli brakuje parametrów, jest to wskazywane przez ikonę wyświetlaną w kolumnie Status. Aby określić wartości dla parametrów, należy kliknąć tę kolumnę w wierszu odpowiadającym danej zmiennej i wybrać opcję **Określ** z listy. Spowoduje to otwarcie podokna dialogowego Określ parametry. Więcej informacji można znaleźć w temacie “Określanie parametrów” na stronie 60. Ta kolumna jest wyłączona, jeśli w kolumnie rozkładu wybrano Empiryczny.
- **Minimum, Maksimum.** W tej kolumnie w przypadku niektórych rozkładów można określić wartość minimalną i/lub wartość maksymalną dla danych poddawanych symulacji. Dane objęte symulacją mniejsze niż określona wartość minimalna i większe od określonej wartości maksymalnej będą odrzucone, nawet jeśli będą poprawne dla określonego rozkładu. Aby określić wartości minimalne i maksymalne, należy kliknąć tę kolumnę w wierszu odpowiadającym danej zmiennej i wybrać opcję **Określ** z listy. Spowoduje to otwarcie podokna dialogowego Określ parametry. Więcej informacji można znaleźć w temacie “Określanie parametrów” na stronie 60. Ta kolumna jest wyłączona, jeśli w kolumnie rozkładu wybrano Empiryczny.

Użyj najlepszego dopasowania. Ta opcja jest włączona, jeśli węzeł Symulacje Generowanie został utworzony automatycznie z węzła Symulacje Dopasowanie z użyciem danych historycznych, a w tabeli Zmienne symulowane wybrano pojedynczy wiersz. Zastępuje informacje dla zmiennej w wybranym wierszu informacjami o najlepszym dopasowaniu dla tej zmiennej. Jeśli informacje w wybranym wierszu były edytowane, naciśnięcie tego przycisku spowoduje ich przywrócenie do danych o najlepszym dopasowaniu rozkładu ustalonych na podstawie węzła Symulacje Dopasowanie.

Szczegóły dopasowania. Ta opcja jest włączona tylko w przypadku automatycznego utworzenia węzła Symulacje Generowanie z węzła Symulacje Dopasowanie. Otwiera podokno dialogowe Szczegóły dopasowania. Więcej informacji można znaleźć w temacie “Szczegóły dopasowania” na stronie 59.

Korzystając z ikon w widoku Zmienne symulowane można wykonać kilka przydatnych zadań. Ikony te zostały opisane w poniższej tabeli.

Tabela 11. Ikony w widoku Zmienne symulowane.

Ikona	Podpowiedź	Opis
	**Edytuj parametry rozkładu**	Ta ikona jest aktywna tylko po wybranie w tabeli Zmienne symulowane pojedynczego wiersza. Otwiera podokno dialogowe Określ parametry dla wybranego wiersza. Więcej informacji można znaleźć w temacie “Określanie parametrów” na stronie 60.
	**Dodaj nową zmienną**	Ta ikona jest aktywna tylko po wybranie w tabeli Zmienne symulowane pojedynczego wiersza. Dodaje nowy pusty wiersz u dołu tabeli Zmienne symulowane.

Tabela 11. Ikony w widoku Zmienne symulowane (kontynuacja).

Ikona	Podpowiedź	Opis
	**Utwórz wiele kopii**	Ta ikona jest aktywna tylko po wybranie w tabeli Zmienne symulowane pojedynczego wiersza. Otwiera podokno dialogowe Klonuj zmienną. Więcej informacji można znaleźć w temacie "Klonowanie zmiennych" na stronie 58.
	**Usuń wybraną zmienną**	Usuwa wybrany wiersz z tabeli Zmienne symulowane.
	**Przenieś na początek**	Ta ikona jest aktywna tylko w przypadku, gdy wybrany wiersz nie jest najwyższym wierszem w tabeli Zmienne symulowane. Przesuwa wybrany wiersz na samą górę tabeli Zmienne symulowane. Ta czynność wpływa na kolejność zmiennych w danych poddawanych symulacji.
	**Przenieś w górę**	Ta ikona jest aktywna tylko w przypadku, gdy wybrany wiersz nie jest najwyższym wierszem w tabeli Zmienne symulowane.. Przesuwa wybrany wiersz w tabeli Zmienne symulowane w górę o jedną pozycję. Ta czynność wpływa na kolejność zmiennych w danych poddawanych symulacji.
	**Przenieś w dół**	Ta ikona jest aktywna tylko w przypadku, gdy wybrany wiersz nie jest najniższym wierszem w tabeli Zmienne symulowane. Przesuwa wybrany wiersz w tabeli Zmienne symulowane w dół o jedną pozycję. Ta czynność wpływa na kolejność zmiennych w danych poddawanych symulacji.
	**Przenieś na koniec**	Ta ikona jest aktywna tylko w przypadku, gdy wybrany wiersz nie jest najniższym wierszem w tabeli Zmienne symulowane. Przesuwa wybrany wiersz na sam dół tabeli Zmienne symulowane. Ta czynność wpływa na kolejność zmiennych w danych poddawanych symulacji.

Nie czyść wartości min. i maks. podczas ponownego dopasowania. Po wybraniu tej opcji wartości minimalne i maksymalne nie są nadpisywane po zaktualizowaniu rozkładów w wyniku wykonania połączonego węzła Symulacje Dopasowanie.

Widok Korelacja

Zmienne wejściowe dla modeli predykcyjnych często są również nazywane zmiennymi skorelowanymi — na przykład, wysokość i waga. Korelacje pomiędzy zmiennymi, które zostaną poddane symulacji, muszą być wyjaśnione, tak aby symulowane wartości zachowały te korelacje.

Jeśli węzeł Symulacje Generowanie został utworzony lub zaktualizowany na podstawie węzła Symulacje Dopasowanie z użyciem danych historycznych, w widoku Korelacje można wyświetlać i edytować obliczone korelacje pomiędzy parami zmiennych. Jeśli dane historyczne nie są dostępne, można określić korelacje ręcznie w oparciu o wiedzę na temat korelacji zmiennych.

Uwaga: Przed wygenerowaniem jakichkolwiek danych macierz korelacji jest automatycznie sprawdzana, aby ustalić, czy jest ona częściowo ostateczna i może zostać odwrócona. Macierz może zostać odwrócona, jeśli jej kolumny są liniowo niezależne. Jeśli macierzy korelacji nie można odwrócić, zostanie ona automatycznie skorygowana, tak aby możliwe było jej odwrócenie.

Informacje na temat korelacji mogą być wyświetlane w formacie macierzy lub listy.

Macierz korelacji. Wyświetla korelacje pomiędzy parami zmiennych w postaci macierzy. Nazw zmiennych są wyświetlane w kolejności alfabetycznej w dół po lewej stronie i w górnej macierzy. Edytować można tylko komórki pod przekątną; wprowadzane muszą być wartości od -1,000 do 1,000 włącznie. Komórki powyżej przekątnej są aktualizowane po przejściu z komórki stanowiącej lustrzane odbicie danej komórki poniżej przekątnej do innej komórki; wówczas w obu komórkach wyświetlana jest ta sama wartość. Komórki na przekątnej są zawsze nieaktywne, a ich korelacja zawsze wynosi 1,000. Domyślna wartość dla wszystkich pozostałych komórek wynosi 0,000. Wartość 0,000 oznacza, że pomiędzy powiązaną parą zmiennych nie ma żadnej korelacji. W macierzy uwzględniane są tylko zmienne ilościowe i porządkowe. Zmienne nominalne, jakościowe i typu flaga przypisane do stałego rozkładu nie są wyświetlane w tabeli.

Lista korelacji. Wyświetla korelacje pomiędzy parami zmiennych w tabeli. Każdy wiersz tabeli przedstawia korelacje pomiędzy parą zmiennych. Nie można dodawać ani edytować wierszy. Kolumny z nagłówkami Zmienna 1 i Zmienna 2 zawierają nazwy zmiennych, które można edytować. Kolumna Korelacja zawiera korelacje, które można edytować; należy wprowadzać wartości od -1,000 do 1,000 włącznie. Domyślna wartość dla wszystkich komórek wynosi 0,000. Na liście uwzględniane są tylko zmienne ilościowe i porządkowe. Zmienne nominalne, jakościowe i typu flaga przypisane do stałego rozkładu nie są wyświetlane na liście.

Resetuj korelacje. Otwiera okno dialogowe Resetuj korelacje. Jeśli dane historyczne są dostępne, można wybrać jedną z trzech opcji:

- **Dopasowane.** Zastępuje bieżące korelacje tymi, które zostały obliczone z użyciem danych historycznych.
- **Zera.** Zastępuje bieżące korelacje zerami.
- **Anuluj.** Zamyka okno dialogowe. Korelacje nie ulegają zmianie.

Jeśli dane historyczne są niedostępne, ale dokonano zmian w korelacjach, można zastąpić bieżące korelacje zerami lub anulować operację.

Pokaż jako. Wybranie opcji **Tabela** umożliwia wyświetlenie korelacji w postaci macierzy. Wybranie opcji **Lista** umożliwia wyświetlenie korelacji w postaci listy.

Nie obliczaj ponownie korelacji podczas ponownego dopasowania. Tę opcję należy wybrać, aby ręcznie określić korelacje i uniemożliwić ich nadpisanie po automatycznym dopasowaniu rozkładów za pośrednictwem węzła Symulacje Dopasowanie i danych historycznych.

Użyj dopasowanej wielodzielczej tabeli kontyngencji dla wejściowych zmiennych z rozkładem kategorialnym. Domyślnie, wszystkie zmienne z rozkładem jakościowym są uwzględniane w tabeli kontyngencji (lub wielowymiarowej tabeli kontyngencji w zależności od liczby zmiennych z rozkładem jakościowym). Tabela kontyngencji jest tworzona, podobnie jak korelacje, podczas wykonywania węzła Symulacje Dopasowanie. Tabeli kontyngencji nie można wyświetlać. Po wybraniu tej opcji zmienne z rozkładem jakościowym są poddawane symulacji z użyciem rzeczywistych wartości procentowych z tabeli kontyngencji. Oznacza to, że każde powiązanie pomiędzy zmiennymi nominalnymi jest tworzone ponownie w nowych, poddawanych symulacji danych. Po usunięciu zaznaczenia tej opcji zmienne z rozkładem jakościowym są poddawane symulacji z użyciem oczekiwanych wartości procentowych z tabeli kontyngencji. Po zmodyfikowaniu zmiennej zostaje ona usunięta z tabeli kontyngencji.

Widok Opcje zaawansowane

Liczba obserwacji do symulacji. Wyświetla opcje dla określania liczby obserwacji do uwzględnienia w symulacji oraz sposobu nadawania nazw poszczególnym iteracjom.

- **Maksymalna liczba obserwacji.** Określa maksymalną liczbę obserwacji danych symulowanych oraz powiązanych wartości zmiennych przewidywanych, które mają zostać wygenerowane. Domyślna wartość to 100 000, wartość minimalna wynosi 1000, a wartość maksymalna wynosi 2 147 483 647.
- **Iteracje.** Liczba jest obliczana automatycznie i nie można jej edytować. Iteracja jest tworzona automatycznie za każdym razem, kiedy dla parametru rozkładu określonych jest kilka wartości.
- **Wiersze ogółem.** Ta opcja jest aktywna, jeśli liczba iteracji jest większa niż 1. Liczba jest obliczana automatycznie przy użyciu przedstawionego równania i nie można jej edytować.
- **Utwórz zmienną iteracyjną.** Ta opcja jest aktywna, jeśli liczba iteracji jest większa niż 1. Po jej wybraniu aktywowane jest pole **Nazwa**. Więcej informacji można znaleźć w temacie “Iteracje” na stronie 61.
- **Nazwa.** Ta opcja jest aktywna, jeśli zaznaczono pole wyboru **Utwórz zmienną iteracyjną**, a liczba iteracji jest większa niż 1. Nazwę zmiennej iteracyjnej można edytować, wpisując ją w polu. Więcej informacji można znaleźć w temacie “Iteracje” na stronie 61.

Wartość początkowa. Ustawienie wartości początkowej generatora liczb losowych umożliwia powielenie symulacji.

- **Replikacja wyników.** Po wybraniu tej opcji aktywowany jest przycisk **Utwórz** i pole **Wartość początkowa generatora liczb losowych**.
- **Wartość początkowa.** Ta opcja jest aktywna po zaznaczeniu pola wyboru **Replikuj wyniki**. W tym polu można określić liczbę całkowitą, która będzie używana jako wartość początkowa generatora liczb losowych. Wartość domyślna to 629111597.
- **Utwórz.** Ta opcja jest aktywna po zaznaczeniu pola wyboru **Replikuj wyniki**. Tworzy w polu **Wartość początkowa generatora liczb losowych** pseudolosową liczbę całkowitą z przedziału od 1 do 999999999 włącznie.

Klonowanie zmiennych

W oknie dialogowym Klonuj zmienną można określić, ile kopii wybranej zmiennej ma zostać utworzonych oraz podać nazwy każdej kopii. Przydatne jest posiadanie kilka kopii zmiennych podczas badania złożonych skutków, na przykład procentów lub wskaźników wzrostu w kolejnych okresach.

Na pasku tytułu w oknie dialogowym znajduje się nazwa wybranej zmiennej.

Liczba kopii do utworzenia. Zawiera liczbę kopii zmiennej do utworzenia. Kliknięcie strzałek umożliwia wybranie liczby kopii do utworzenia. Minimalna liczba kopii to 1, a maksymalna to 512. Liczba kopii jest początkowo ustawiona na 10.

Kopiuj znaki przyrostka. Zawiera znaki, które są dodawane na końcu nazwy zmiennej dla każdej kopii. Znaki te oddzielają nazwę zmiennej od numeru kopii. Znaki przyrostka mogą być edytowane poprzez wpisanie ich w tym polu. Ta zmienna może pozostać pusta; w takim przypadku pomiędzy nazwą zmiennej a numerem kopii nie będzie żadnych znaków. Domyślnym znakiem jest podkreślenie.

Początkowa liczba kopii. Zawiera numer przyrostka dla pierwszej kopii. Kliknięcie strzałek umożliwia wybór początkowej liczby kopii. Minimalna początkowa liczba kopii wynosi 1, a maksymalna 1000. Domyślnie początkowa liczba kopii jest ustawiona na 1.

Krok kopiowania numeru. Zawiera wielkość przyrostu dla liczby przyrostków. Kliknięcie strzałek umożliwia wybranie wielkości przyrostu. Minimalny przyrost wynosi 1, a maksymalny 255. Początkowo wielkość przyrostu jest ustawiona na 1.

Pola. Zawiera podgląd nazw zmiennych dla kopii, który jest aktualizowany w przypadku edycji dowolnej zmiennej w oknie dialogowym Klonuj zmienną. Ten tekst jest generowany automatycznie i nie można go edytować.

OK. Generuje wszystkie kopie zgodnie z opcjami wybranymi w oknie dialogowym. Kopie są dodawane do tabeli Zmienne symulowane w oknie dialogowym węzła Symulacje Generowanie, bezpośrednio pod wierszem, który zawiera skopiowaną zmienną.

Anuluj. Zamyka okno dialogowe. Wszelkie zmiany, jakie zostały wprowadzone, zostają odrzucone.

Szczegóły dopasowania

Okno dialogowe Szczegóły dopasowania jest dostępne tylko po utworzeniu węzła Symulacje Generowanie lub zaktualizowaniu go przez wykonanie węzła Symulacje Dopasowanie. Wyświetla wyniki automatycznego dopasowania rozkładu dla wybranej zmiennej. Rozkłady są uporządkowane według dobroci dopasowania, a najlepiej pasujący rozkład wymieniony jest jako pierwszy. W tym oknie dialogowym można wykonać następujące zadania:

- Zbadać rozkłady dopasowane do danych historycznych.
- Wybrać jeden z dopasowanych rozkładów.

Zmienna. Zawiera nazwę wybranej zmiennej. Tekstu nie można edytować.

Traktuj jako. Wyświetla typ pomiaru wybranej zmiennej. Dane są pobierane z tabeli Zmienne symulowane dostępnej w oknie dialogowym węzła Symulacje Generowanie. Typ pomiaru można zmienić, klikając strzałkę i wybierając typ pomiaru z listy rozwijanej. Dostępne są trzy opcje: **Ilościowy**, **Nominalny** i **Porządkowy**.

Rozkłady. Tabela Rozkłady wyświetla wszystkie rozkłady odpowiednie dla danego typu pomiaru. Rozkłady, które zostały dopasowane do danych historycznych są uporządkowane według dobroci dopasowania, w kolejności od najlepszego do najgorszego dopasowania. Dobroć dopasowania jest określana na podstawie statystyki dopasowania wybieranej w węźle Symulacje Dopasowanie. Rozkłady, które nie zostały dopasowane do danych historycznych, są wyświetlane w tabeli w porządku alfabetycznym pod rozkładami, które zostały dopasowane.

Tabela Rozkład zawiera następujące kolumny:

- **Wykorzystanie.** Wybrany przycisk opcji wskazuje, który rozkład jest obecnie wybrany dla zmiennej. Można zastąpić najlepiej dopasowany rozkład, wybierając przycisk opcji żądanego rozkładu w kolumnie Użycie. Wybranie przycisku opcji w kolumnie Użycie powoduje również wyświetlenie wykresu rozkładu nałożonego na histogram (lub wykres słupkowy) danych historycznych dla wybranej zmiennej. Jednorazowo można wybrać tylko jeden rozkład.
- **Rozkład.** Zawiera nazwę rozkładu. Tej kolumny nie można edytować.
- **Statystyki dopasowania.** Zawiera obliczone statystyki dopasowania dla rozkładu. Tej kolumny nie można edytować. Zawartość komórki zależy od typu pomiaru zmiennej:
  - **Ilościowa.** Zawiera wyniki testów Andersona-Darlinga i Kołmogorowa-Smirnowa. Pokazane są także powiązane z testami wartości p. Statystyka dopasowania wybrana jako kryterium dopasowania w węźle Symulacje Dopasowanie wyświetlana jest jako pierwsza i jest używana do uporządkowania rozkładów. Statystyki Andersona-Darlinga są wyświetlane jako: A=*aval* P=*pval*. Statystyki Kołmogorowa-Smirnowa są wyświetlane jako K=*kval* P=*pval*. Jeśli statystyki nie można obliczyć, zamiast liczby wyświetlana jest kropka.
  - **Nominalny i Porządkowy.** Zawiera wyniki testu chi-kwadrat. Wartość p powiązana z testem również jest wyświetlana. Statystyki są wyświetlane jako Chi-Sq=*val* P=*pval*. Jeśli rozkład nie został dopasowany, wyświetlany jest komunikat Niedopasowane. Jeśli rozkład nie może być dopasowany matematycznie, wyświetlany jest komunikat Nie można dopasować.

  Uwaga: Ta komórka jest zawsze pusta dla rozkładu empirycznego.
- **Parametry.** Zawiera parametry rozkładu powiązane z każdym dopasowanym rozkładem. Parametry są wyświetlane jako *nazwa_parametru* = *wartość_parametru*, przy czym parametry są rozdzielane pojedynczą spacją. Dla rozkładu jakościowego kategorii nazwy parametrów są kategoriami, a wartości parametrów to powiązane prawdopodobieństwa. Jeśli rozkład nie został dopasowany do danych historycznych, komórka jest pusta. Tej kolumny nie można edytować.

Miniatura histogramu. Wyświetla wykres wybranego rozkładu nałożony na histogram danych historycznych dla wybranej zmiennej.

Miniatura rozkładu. Wyświetla objaśnienie i ilustrację wybranego rozkładu.

OK. Zamyka okno dialogowe i aktualizuje wartości w kolumnach Poziom, Rozkład, Parametry oraz Minimum i Maksimum tabeli Zmienne symulowane dla wybranej zmiennej na podstawie informacji z wybranego rozkładu. Ikona w kolumnie Status również jest aktualizowana, tak aby wskazywała, czy wybrany rozkład jest rozkładem z najlepszym dopasowaniem do danych.

Anuluj. Zamyka okno dialogowe. Wszelkie zmiany, jakie zostały wprowadzone, zostają odrzucone.

Określanie parametrów

W oknie dialogowym Określ parametry można ręcznie określić wartości parametrów dla rozkładu wybranej zmiennej. Można również wybrać inny rozkład dla wybranej zmiennej.

Okno dialogowe Określ parametry można otworzyć na trzy sposoby:

- Należy dwukrotnie kliknąć nazwę zmiennej w tabeli Zmienne symulowane w oknie dialogowym węzła Symulacje Generowanie.
- Należy kliknąć kolumnę Parametry lub Minimum, Maksimum w tabeli Zmienne symulowane i wybrać z listy opcję **Określ**.
- W tabeli Zmienne symulowane należy zaznaczyć wiersz, a następnie kliknąć ikonę **Edytuj parametry rozkładu**.

Zmienna. Zawiera nazwę wybranej zmiennej. Tekstu nie można edytować.

Rozkład. Zawiera rozkład dla wybranej zmiennej. Dane są pobierane z tabeli Zmienne symulowane. Rozkład można zmienić, klikając strzałkę i wybierając rozkład z listy rozwijanej. Dostępne rozkłady zależą od typu składowania wybranej zmiennej.

Boki. Ta opcja jest dostępna po wybraniu rozkładu rzutu kostką w polu **Rozkład**. Klikając strzałki można wybrać liczbę boków lub kategorii,, na jakie zmienna zostanie podzielona. Minimalna liczba boków to dwa, a maksymalna to 20. Liczba boków jest początkowo ustawiona na 6.

Parametry rozkładu. Tabela Parametry rozkładu zawiera jeden wiersz dla każdego parametru wybranego rozkładu.

Uwaga: W rozkładzie używany jest parametr wskaźnika, w którym parametr skali $\alpha = k$, a odwrotny parametr skali $\beta = 1/\theta$.
Tabela składa się z dwóch kolumn:

- **Parametr.** Zawiera nazwy parametrów. Tej kolumny nie można edytować.
- **Wartości.** Zawiera wartości parametrów. Jeśli węzeł Symulacje Generowanie został utworzony lub zaktualizowany na podstawie węzła Symulacje Dopasowanie, komórki w tej kolumnie zawierają wartości parametrów, które zostały określone poprzez dopasowanie rozkładu do danych historycznych. Jeśli węzeł Symulacje Generowanie został dodany do obszaru roboczego strumienia w palety Węzły źródeł, komórki w tej kolumnie są puste. Wartości mogą być edytowane poprzez wpisanie ich w komórkach. Więcej informacji parametrach wymaganych przez poszczególne rozkłady oraz akceptowanych wartości parametrów można znaleźć w temacie “Rozkłady” na stronie 61.

  W przypadku kilku wartości parametrów należy je rozdzielić przecinkami. Określenie wielu wartości dla parametru spowoduje wygenerowanie wielu iteracji dla symulacji. Można określić wiele wartości dla jednego parametru.

  Uwaga: W przypadku zmiennych z typem składowania data/czas należy określić parametry rozkładu jako liczby całkowite. Przykładowo, aby określić średnią datę jako 1 stycznia 1970, należy użyć liczby całkowitej 0.

Uwaga: Po wybraniu rozkładu rzutu kostką tabela parametrów rozkładu będzie się nieco różnić. Tabela zawiera jeden wiersz dla każdego boku (lub kategorii). Dostępne są kolumny Wartość i Prawdopodobieństwo. W kolumnie Wartość dostępne są etykiety dla każdej kategorii. Wartości domyślne dla etykiet są liczbami całkowitymi 1-N, gdzie N oznacza liczbę boków. Etykiety mogą być edytowane poprzez wpisanie ich w komórkach. W komórkach można wprowadzić dowolne wartości. Aby użyć wartości, która nie jest liczbą, typ składowania zmiennej daty należy zmienić na łańcuch, o ile jeszcze nie został ustawiony jako łańcuch. Kolumna Prawdopodobieństwo zawiera prawdopodobieństwo dla każdej kategorii. Prawdopodobieństw nie można edytować i są one obliczane jako 1/N.

Podgląd. Przedstawia wykres próbny rozkładu na podstawie określonych parametrów. Jeśli dla jednego parametru określono dwie lub więcej wartości, wyświetlane są wykresy próbne dla każdej wartości parametru. Jeśli dla wybranej zmiennej dostępne są dane historyczne, wykres rozkładu zostaje nałożony na histogram danych historycznych.

Ustawienia opcjonalne. Opcje te umożliwiają określenie wartości minimalnej i/lub wartości maksymalnej dla danych poddawanych symulacji. Dane objęte symulacją mniejsze niż określona wartość minimalna i większe od określonej wartości maksymalnej będą odrzucone, nawet jeśli będą poprawne dla określonego rozkładu.

- **Określ minimum.** Tę opcję należy wybrać, aby aktywować pole **Odrzuć wartości poniżej**. To pole wyboru jest nieaktywne, jeśli wybrano rozkład empiryczny.
- **Odrzuć wartości poniżej.** Aktywne po wybraniu opcji **Określ minimum**. Należy wprowadzić wartość minimalną dla danych objętych symulacją. Każda wartość objęta symulacją, która będzie mniejsza od tej wartości, zostanie odrzucona.
- **Określ maksimum.** Tę opcję należy wybrać, aby aktywować pole **Odrzuć wartości powyżej**. To pole wyboru jest nieaktywne, jeśli wybrano rozkład empiryczny.
- **Odrzuć wartości powyżej.** Aktywne po wybraniu opcji **Określ maksimum**. Należy wprowadzić wartość maksymalną dla danych objętych symulacją. Każda wartość objęta symulacją, która będzie większa od tej wartości, zostanie odrzucona.

OK. Zamyka okno dialogowe i aktualizuje wartości w kolumnach Rozkład, Parametry oraz Minimum, Maksimum w tabeli Zmienne symulowane dla wybranej zmiennej. Ikona w kolumnie Status również zostaje zaktualizowana, tak aby odpowiadała wybranemu rozkładowi.

Anuluj. Zamyka okno dialogowe. Wszelkie zmiany, jakie zostały wprowadzone, zostają odrzucone.

Iteracje

Jeśli dla zmiennej stałej lub parametru rozkładu określono więcej niż jedną wartość, generowany jest niezależny zestaw symulowanych obserwacji — a tym samym osobną symulację — dla każdej określonej wartości. Dzięki temu możliwe jest zbadanie skutków różnic zmiennych lub parametrów. Każdy zbiór symulowanych obserwacji jest określany jako *iteracja*. W danych poddawanych symulacji iteracje są zestawiane.

Jeśli pole wyboru **Utwórz zmienną iteracyjną** w widoku Opcje zaawansowane w oknie dialogowym węzła Symulacje Generowanie jest zaznaczone, zmienna iteracji zostaje dodana do danych poddawanych symulacji jako zmienna nominalna z numerycznym typem składowania. Nazwę tej zmiennej można edytować, wpisując ją w polu **Nazwa** w widoku opcji zaawansowanych. Pole zawiera etykietę wskazującą, do której iteracji należą poszczególne obserwacje objęte symulacją. Forma etykiet zależy od typu iteracji:

- **Iteracja zmiennej stałej.** Etykieta jest nazwą zmiennej, po której następuje znak równości, a za nim wartość zmiennej dla tej iteracji, czyli:

  nazwa_zmiennej = *wartość_zmiennej*
- **Iteracja parametru rozkładu.** Etykieta jest nazwą zmiennej, po której następuje dwukropek, po nim nazwa parametru poddawanego iteracji, następnie znak równości i wartość parametru dla tej iteracji, czyli:

  nazwa_zmiennej:*nazwa_parametru* = *wartość_parametru*
- **Iteracja parametru rozkładu dla rozkładu jakościowego lub rozstępu.** Etykieta jest nazwą zmiennej, po której następuje przecinek, następnie słowo „Iteration” (Iteracja) oraz numer iteracji, czyli:

  nazwa_zmiennej: Iteration *numer_iteracji*

Rozkłady

Istnieje możliwość ręcznego określenia rozkładu prawdopodobieństwa dla dowolnej zmiennej; w tym celu należy otworzyć okno dialogowe Określ parametry dla danej zmiennej, wybrać odpowiedni rozkład z listy **Rozkład** i wprowadzić parametry rozkładu w tabeli **Parametry rozkładu**. Poniżej przedstawiono uwagi na temat konkretnych rozkładów:

- **Jakościowy.** Rozkład jakościowy opisuje zmienną wejściową, która ma ustaloną liczbę wartości zwanych kategoriami. Każda kategoria posiada powiązane prawdopodobieństwo przypisane w taki sposób, by suma wszystkich prawdopodobieństw dla wszystkich kategorii wynosi jeden.

Uwaga: Jeśli określone zostanie prawdopodobieństwo dla tych kategorii, którego suma nie będzie wynosiła 1, użytkownik otrzyma ostrzeżenie.

- **Dwumianowy ujemny – niepowodzenia.** Opisuje rozkład liczby niepowodzeń w sekwencji prób, zanim osiągnięto określoną liczbę sukcesów. Parametr *Threshold* (Wartość graniczna) jest określoną liczbą sukcesów, a parametr *Probability* (Prawdopodobieństwo) to prawdopodobieństwo sukcesu dla podanej próby.
- **Dwumianowy ujemny – próby.** Opisuje rozkład liczby wymaganych prób przed zaobserwowaniem określonej liczby sukcesów. Parametr *Threshold* (Wartość graniczna) jest określoną liczbą sukcesów, a parametr *Probability* (Prawdopodobieństwo) to prawdopodobieństwo sukcesu dla podanej próby.
- **Zakres.** Ten rozkład składa się ze zbioru przedziałów, dla których przydzielone zostało prawdopodobieństwo w taki sposób, by suma wartości prawdopodobieństwa dla wszystkich przedziałów wynosiła 1. Wartości z podanego przedziału są rysowane z wykorzystaniem rozkładu jednostajnego zdefiniowanego dla tego przedziału. Przedziały określa się, wprowadzając wartość minimalną, wartość maksymalną oraz powiązane z nim prawdopodobieństwo.

  Na przykład: przewiduje się, że koszt surowca ma 40% szans na znalezienie się w przedziale od 10 USD do 15 USD za jednostkę i 60% szans na znalezienie się w przedziale od 15 USD do 20 USD za jednostkę. Koszt zostałby zamodelowany za pomocą rozkładu Przedziału składającego się z dwóch przedziałów [10 - 15] i [15 - 20] z ustawieniem prawdopodobieństwa powiązanego z pierwszym przedziałem jako 0,4 oraz prawdopodobieństwa skojarzonego z drugim przedziałem jako 0,6. Przedziały nie muszą być przedziałami sąsiadującymi ze sobą, a nawet mogą na siebie zachodzić. Na przykład: można określić przedziały: 10 - 15 USD i 20 - 25 USD lub 10 - 15 USD i 13 - 16 USD.
- **Rozkład Weibulla.** Parametr *Location* (Lokalizacja) jest opcjonalnym parametrem lokalizacji określającym, w którym miejscu znajduje się początek rozkładu.

W poniższej tabeli przedstawiono rozkłady dostępne dla niestandardowego dopasowywania rozkładu oraz możliwe do zaakceptowania wartości dla parametrów. Niektóre z tych rozkładów są dostępne dla niestandardowego dopasowywania do określonych typów składowania, nawet jeśli nie są dopasowane automatycznie do tych typów składowania w węźle Symulacje Dopasowanie.

Tabela 12. Rozkłady dostępne dla niestandardowego dopasowania

Rozkład	Typ składowania obsługiwany przez niestandardowe dopasowanie	Parametry	Limity parametrów	Uwagi
Bernoulliego	Liczba całkowita, rzeczywista, data/czas	Prawdopodobieństwo	$0 \leq$ *Prawdop.* ≤ 1	
Beta	Liczba całkowita, rzeczywista, data/czas	Kształt 1 Kształt 2 Minimum Maksimum	≥ 0 ≥ 0 $<$ *Maksimum* $>$ *Minimum*	Wartości minimalne i maksymalne są opcjonalne.
Dwumianowy	Liczba całkowita, rzeczywista, data/czas	Liczba prób (n) Prawdopodobieństwo Minimum Maksimum	$>$ 0, liczba całkowita $0 \leq$ *Prawdop.* ≤ 1 $<$ *Maksimum* $>$ *Minimum*	Liczba prób musi być liczbą całkowitą. Wartości minimalne i maksymalne są opcjonalne.
Jakościowy	Liczba całkowita, rzeczywista, data/czas, łańcuch	Nazwa kategorii (lub etykieta)	$0 \leq$ *Wartość* ≤ 1	Wartość jest prawdopodobieństwem kategorii. Suma wartości musi wynosić 1, w przeciwnym razie generowane jest ostrzeżenie.

Tabela 12. Rozkłady dostępne dla niestandardowego dopasowania (kontynuacja)

Rozkład	Typ składowania obsługiwany przez niestandardowe dopasowanie	Parametry	Limity parametrów	Uwagi
Rzutu kostką	Liczba całkowita, łańcuch	Boki	$2 \leq boki \leq 20$	Prawdopodobieństwo każdej kategorii (bok) jest obliczane jako 1/N, gdzie N jest liczbą boków. Tych prawdopodobieństw nie można edytować.
Empiryczny	Liczba całkowita, rzeczywista, data/czas			Nie można edytować rozkładu empirycznego ani wybierać go jako typ. Rozkład empiryczny jest dostępny tylko w przypadku danych historycznych.
Wykładniczy	Liczba całkowita, rzeczywista, data/czas	Skala Minimum Maksimum	> 0 < *Maksimum* > *Minimum*	Wartości minimalne i maksymalne są opcjonalne.
Stała	Liczba całkowita, rzeczywista, data/czas, łańcuch	Wartość		Nie można określić rozkładu stałego dla każdej zmiennej. Jeśli każda zmienna w wygenerowanych danych ma być stała, można po węźle danych użytkownika należy wstawić węzeł zrównoważenia.
Gamma	Liczba całkowita, rzeczywista, data/czas	Kształt Skala Minimum Maksimum	≥ 0 ≥ 0 < *Maksimum* > *Minimum*	Wartości minimalne i maksymalne są opcjonalne. W rozkładzie używany jest parametr wskaźnika, w którym parametr skali $\alpha = k$, a odwrotny parametr skali $\beta = 1/\theta$.
Lognormalny	Liczba całkowita, rzeczywista, data/czas	Kształt 1 Kształt 2 Minimum Maksimum	≥ 0 ≥ 0 < *Maksimum* > *Minimum*	Wartości minimalne i maksymalne są opcjonalne.
Ujemny dwumianowy - niepowodzenia	Liczba całkowita, rzeczywista, data/czas	Wartość graniczna Prawdopodobieństwo Minimum Maksimum	≥ 0 $0 \leq Prawdop. \leq 1$ < *Maksimum* > *Minimum*	Wartości minimalne i maksymalne są opcjonalne.
Ujemny dwumianowy - próby	Liczba całkowita, rzeczywista, data/czas	Wartość graniczna Prawdopodobieństwo Minimum Maksimum	≥ 0 $0 \leq Prawdop. \leq 1$ < *Maksimum* > *Minimum*	Wartości minimalne i maksymalne są opcjonalne.

Tabela 12. Rozkłady dostępne dla niestandardowego dopasowania (kontynuacja)

Rozkład	Typ składowania obsługiwany przez niestandardowe dopasowanie	Parametry	Limity parametrów	Uwagi
Normalny	Liczba całkowita, rzeczywista, data/czas	Średnia Odchylenie standardowe Minimum Maksimum	≥ 0 > 0 < *Maksimum* > *Minimum*	Wartości minimalne i maksymalne są opcjonalne.
Poissona	Liczba całkowita, rzeczywista, data/czas	Średnia Minimum Maksimum	≥ 0 < *Maksimum* > *Minimum*	Wartości minimalne i maksymalne są opcjonalne.
Przedział	Liczba całkowita, rzeczywista, data/czas	Początek(X) Koniec(X) Prawdopodobieństwo(X)	0 ≤ *Wartość* ≤ 1	X to indeks każdego przedziału. Suma wartości prawdopodobieństwa musi wynosić 1.
Trójkątny	Liczba całkowita, rzeczywista, data/czas	Dominanta Minimum Maksimum	*Minimum* ≤ *Wartość* ≤ *Maksimum* < *Maksimum* > *Minimum*	
Jednostajny	Liczba całkowita, rzeczywista, data/czas	Minimum Maksimum	< *Maksimum* > *Minimum*	
Weibulla	Liczba całkowita, rzeczywista, data/czas	Wskaźnik Skala Lokalizacja Minimum Maksimum	> 0 > 0 ≥ 0 < *Maksimum* > *Minimum*	Lokalizacja, maksimum i minimum są opcjonalne.

Węzeł Rozszerzenie Import

Dzięki węzłowi Rozszerzenie Eksport można wykonywać skrypty R lub Python for Spark służące do importu danych.

Węzeł Rozszerzenie Import — karta Polecenia

Wybierz język poleceń: **R** albo **Python for Spark**. Następnie wprowadź lub wklej swój skrypt służący do importowania danych. Gdy polecenia będą gotowe, można kliknąć przycisk **Wykonaj**, aby wykonać węzeł Rozszerzenie Import.

Węzeł Rozszerzenie Import — karta Wynik z konsoli

Karta **Wynik z konsoli** zawiera wszelkie wyniki odbierane podczas wykonywania skryptu w języku R lub Python for Spark na karcie Polecenia (na przykład, jeśli używany jest skrypt R, to na tej karcie wyświetlane są wyniki odbierane z konsoli R podczas wykonywania skryptu z pola **Polecenia R** na karcie **Polecenie**). Wyniki te mogą zawierać komunikaty o błędach lub ostrzeżenia generowane podczas wykonywania skryptu w języku R lub Python. Wyniki można wykorzystać przede wszystkim do debugowania skryptu. Karta **Wynik z konsoli** zawiera także skrypt z pola **Polecenia R** lub **Polecenia Python**.

Po każdym wykonaniu skryptu Rozszerzenie Import zawartość karty **Wynik z konsoli** jest nadpisywana wynikami z konsoli R lub środowiska Python for Spark. Wyników nie można edytować.

Filtrowanie lub zmiana nazw zmiennych

Istnieje możliwość zmiany nazwy lub wykluczenia zmiennych w dowolnym punkcie strumienia. Przykładowo, pracownik naukowo-badawczy może nie być zainteresowany poziomem potasu (dane na poziomie zmiennej) u

pacjentów (dane na poziomie rekordu); dlatego może odfiltrować zmienną K (potas). Można to zrobić, używając osobnego węzła filtrowania lub karty Filtrowanie w węźle źródłowym lub wynikowym. Działanie jest takie samo, niezależnie od węzła wybranego do uzyskania dostępu.

- Za pośrednictwem węzłów źródłowych, takich jak Plik zmiennych, Plik kolumnowy, Plik Statistics, XML lub Importowanie przez rozszerzenie, można zmienić nazwę lub odfiltrować zmienne podczas odczytywania danych w programie IBM SPSS Modeler.
- Korzystając z węzła filtrowania można zmienić nazwy lub odfiltrować zmienne w dowolnym punkcie strumienia.
- Węzły Eksport Statistics, Transformacja Statistics, Model Statistics i Wynik Statistics umożliwiają zmianę nazwy lub filtrowanie zmiennych, tak aby były zgodne ze standardami nadawania nazw w programie IBM SPSS Statistics. Więcej informacji można znaleźć w temacie “Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics” na stronie 364.
- Karta Filtrowanie w dowolnym z powyższych węzłów umożliwia zdefiniowanie lub edytowanie zestawów wielokrotnych odpowiedzi. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi” na stronie 153.
- Węzła Filtrowanie można użyć do mapowania zmiennych z jednego źródła na inne.

Węzeł widoku danych

Węzeł widoku danych umożliwia uwzględnienie danych zdefiniowanych w widoku danych analitycznych IBM SPSS Collaboration and Deployment Services w strumieniu. Widok danych analitycznych definiuje strukturę sposobu uzyskiwania dostępu do danych, która opisuje jednostki użyte w modelach predykcyjnych oraz reguły biznesowe. Widok wiąże źródło danych z fizycznymi źródłami danych w celu przeprowadzenia analizy.

Analiza predykcyjna wymaga, aby dane były rozmieszczone w tabelach, w których każdy wiersz odpowiada jednostce, dla której wykonywane są predykcje. Każda kolumna w tabeli reprezentuje możliwy do zmierzenia atrybut jednostki. Niektóre atrybuty można wyliczyć poprzez zagregowanie wartości dla innego atrybutu. Przykładowo, wiersze w tabeli mogą reprezentować klientów, a w kolumnach znajdą się dane, takie jak nazwisko, płeć klienta, kod pocztowy i liczba zakupów dokonanych przez klienta w ostatnim roku na kwotę powyżej 500 USD. Wartości w ostatniej kolumnie są wyznaczane na podstawie historii zamówień klienta, która zwykle jest zapisywana w jednej lub kliku powiązanych tabelach.

Proces analizy predykcyjnej obejmuje użycie różnych zbiorów danych w całym cyklu życia modelu. Podczas początkowego wdrożenia modelu predykcyjnego wykorzystywane są dane historyczne, dla których często dostępne są znane wyniki dotyczące przewidywanego zdarzenia. Aby ocenić skuteczność i dokładność modelu, przeprowadzana jest walidacja modelu kandydackiego w oparciu o inne dane. Po przeprowadzeniu walidacji modelu jest on wdrażany do środowiska produkcyjnego w celu wygenerowania ocen dla wielu jednostek w procesie wsadowym lub dla pojedynczych jednostek w procesie wykonywanym w czasie rzeczywistym. Jeśli model ten zostanie połączony z regułami biznesowymi w procesie zarządzania decyzjami, w oparciu o dane objęte symulacją można przeprowadzić walidację wyników takiego połączenia. Jednak pomimo tego, że na różnych etapach procesu wdrożenia modelu dane różnią się, każdy zbiór danych musi udostępniać taki sam zbiór atrybutów dla modelu. Zbiór atrybutów pozostaje stały; dane poddawane analizie ulegają zmianie.

Widok danych analitycznych obejmuje następujące składniki, które odpowiadają specjalnym potrzebom analizy predykcyjnej:

- Schemat widoku danych lub model danych, który definiuje interfejs logiczny uzyskiwania dostępu do danych jako zbiór atrybutów rozmieszczonych w powiązanych tabelach. Atrybuty w modelu mogą być wyliczane na podstawie innych atrybutów.
- Co najmniej jeden plan dostępu do danych, który udostępnia atrybuty modelu danych z wartościami fizycznymi. Dane dostępne dla modelu są dobierane poprzez określenie, który plan dostępu do danych będzie aktywny dla konkretnej aplikacji.

Ważne: Aby użyć węzła widoku danych, w danej lokalizacji należy najpierw zainstalować i skonfigurować program IBM SPSS Collaboration and Deployment Services Repository. Widok danych analitycznych, do którego odwołuje się węzeł, zwykle jest tworzony i zapisywany za pośrednictwem programu IBM SPSS Deployment Manager.

Ustawienie opcji dla węzła Widok danych

Opcje na karcie **Dane** w węźle Widok danych umożliwiają określenie ustawień danych dla widoku danych analitycznych wybranego w programie IBM SPSS Collaboration and Deployment Services Repository.

Widok danych analitycznych. Należy kliknąć przycisk wielokropka (**...**). Pozwala to wybrać widok danych analitycznych. Jeśli obecnie połączenie z serwerem nie jest nawiązane, należy określić adres URL serwera w oknie dialogowym Repozytorium: serwer, kliknąć przycisk **OK** i określić dane uwierzytelniające połączenia w oknie dialogowym Repozytorium: dane uwierzytelniające. Aby uzyskać więcej informacji na temat logowania do repozytorium i pobierania obiektów, należy zapoznać się z publikacją IBM SPSS Modeler — podręcznik użytkownika.

Nazwa tabeli. Należy wybrać tabelę z modelu danych w widoku danych analitycznych. Każda tabela w modelu danych reprezentuje koncepcję lub jednostkę biorącą udział w procesie analizy predykcyjnej. Zmienne tabel odpowiadają atrybutom jednostek reprezentowanym przez tabele. Przykładowo podczas analizowania zamówień klienta model danych może zawierać tabelę dla klientów oraz tabelę dla zamówień. Tabela klientów może zawierać atrybuty dla identyfikatora, wieku, płci, stanu cywilnego oraz kraju zamieszkania klienta. W tabeli zamówień mogą znaleźć się atrybuty identyfikatora zamówienia, liczby pozycji w zamówieniu, całkowitego kosztu oraz identyfikatora klienta, który złożył zamówienie. Atrybut identyfikatora klienta może być użyty do powiązania klientów w tabeli klientów z ich zamówieniami w tabeli zamówień.

Plan dostępu do danych. Należy wybrać plan dostępu do danych z widoku danych analitycznych. Plan dostępu do danych umożliwia powiązanie tabel modelu danych w widoku danych analitycznych z fizycznymi źródłami danych. Widok danych analitycznych zwykle obejmuje kilka planów dostępu do danych. Po zmianie używanego planu dostępu do danych zmieniane są dane używane w strumieniu. Przykładowo, jeśli widok danych analitycznych zawiera plan dostępu do danych przeznaczony do uczenia modelu oraz plan dostępu do danych przeznaczony do testowania modelu, można przełączyć dane uczące na dane testujące, zmieniając używany plan dostępu do danych.

Atrybuty opcjonalne. Jeśli konkretny atrybut nie jest wymagany przez aplikację korzystającą z widoku danych analitycznych, można oznaczyć ten atrybut jako opcjonalny. W przeciwieństwie do atrybutów wymaganych atrybuty opcjonalne mogą zawierać wartości null. Konieczne może być dostosowanie aplikacji, tak aby obejmowała obsługę wartości null dla atrybutów opcjonalnych. Przykładowo, w czasie wywoływania reguły biznesowej utworzonej w programie IBM Operational Decision Manager program IBM Analytical Decision Management tworzy zapytania dla obsługi reguł w celu ustalenia, które wartości wejściowe są wymagane. Jeśli rekord, który ma zostać poddany ocenie, zawiera wartość null dla dowolnej zmiennej wymaganej w ramach obsługi reguł, reguła nie zostanie wywołana, a do zmiennych wynikowych dla reguły zostaną wprowadzone wartości domyślne. Jeśli opcjonalna zmienna zawiera wartość null, reguła jest wywoływana. W celu kontrolowania przetwarzania reguła może sprawdzać, czy wartości null są obecne.

Aby określić atrybuty jako opcjonalne, należy kliknąć opcję **Atrybuty opcjonalne** i wybrać właściwe atrybuty.

Uwzględnij dane XML w polu. Tę opcję należy wybrać, aby utworzyć zmienną zawierającą możliwy do wykonywania model obiektu danych XML dla każdego wiersza danych. Te informacje są wymagane, jeśli dane będą używane w programie IBM Operational Decision Manager. Należy określić nazwę dla nowej zmiennej.

Węzeł źródłowy Dane geoprzestrzenne

Węzeł źródłowy Dane geoprzestrzenne umożliwia przeniesienie danych z mapy lub danych przestrzennych do sesji eksploracji danych. Dane można zaimportować w jeden z dwóch sposobów:

- W postaci pliku kształtu (.shp)
- Poprzez połączenie serwera ESRI zawierającego hierarchiczny system plików, który obejmuje pliki map.

Uwaga: Można łączyć się tylko z usługami map publicznych.

Predykcje przestrzenno-czasowych modeli predykcyjny (STP) mogą obejmować mapy lub elementy przestrzenne. Aby uzyskać więcej informacji na temat tych modeli, należy zapoznać się z tematem „Węzeł modelowania Predykcja

przestrzenno-czasowa (STP)" w sekcji dotyczącej modeli szeregów czasowych w publikacji na temat węzłów modelowania programu Modeler (ModelerModelingNodes.pdf).

Ustawianie opcji dla węzła źródłowego danych geoprzestrzennych

Typ źródła danych Istnieje możliwość zaimportowania danych z **pliku kształtu** (.shp) lub poprzez nawiązanie połączenia z **usługą map**.

Jeśli używana jest opcja **Plik kształtu**, można wprowadzić nazwę i ścieżkę pliku lub przejść do wybranego pliku. Plik musi znajdować się w katalogu lokalnym lub musi być dostępny za pośrednictwem zmapowanego napędu; dostęp do pliku można uzyskać, korzystając ze ścieżki UNC.

Uwaga: Dane kształtu wymagają zarówno pliku .shp, jak i pliku .dbf. Te dwa pliki muszą mieć taką samą nazwę i muszą znajdować się w tym samym folderze. Plik .dbf jest importowany automatycznie po wybraniu pliku .shp. Ponadto, dostępny może być plik .prj, który określa układ współrzędnych dla danych kształtu.

Jeśli używana jest **Usługa map**, należy wprowadzić adres URL do usługi i kliknąć przycisk **Połącz**. Po nawiązaniu połączenia z usługą warstwy dla tej usługi są wyświetlane w dolnej części okna dialogowego w strukturze drzewa panelu **Dostępne mapy**; należy rozwinąć drzewo i wybrać wymaganą warstwę.

Uwaga: Można łączyć się tylko z usługami map publicznych.

Automatyczna definicja danych geoprzestrzennych

Domyślnie, program SPSS Modeler automatycznie definiuje, o ile to możliwe, zmienne danych geoprzestrzennych w węźle źródłowym z poprawnymi metadanymi. Metadane mogą zawierać informacje na temat poziomu pomiaru zmiennej geoprzestrzennej (np. Punkt lub Wielokąt) oraz układu współrzędnych, jaki będzie użyty dla zmiennych, w tym szczegóły, takie jak punkt początkowy (lub na przykład, szerokość geograficzna 0, długość geograficzna 0) oraz jednostki miary. Więcej informacji o poziomach pomiaru zawiera temat "Geoprzestrzenne podpoziomy pomiarów" na stronie 141.

Pliki .shp i .dbf, które tworzą plik kształtu, zawierają wspólną zmienną identyfikującą, która jest używana jako klucz. Przykładowo, plik .shp może zawierać kraje, a zmienna nazwy kraju jest używana jako identyfikator, natomiast plik .dbf może zawierać informacje na temat tych krajów wraz z nazwą kraju, która również jest używana jako identyfikator.

Uwaga: Jeśli układ współrzędnych różni się od domyślnego układu współrzędnych programu SPSS Modeler konieczne może być ponowne rzutowanie danych, aby możliwe było użycie odpowiedniego układu współrzędnych. Aby uzyskać więcej informacji, zobacz "Węzeł Zmiana rzutowania" na stronie 186.

Węzeł źródłowy JSON

Użyj węzła źródłowego JSON, aby zaimportować dane z pliku JSON do strumienia SPSS Modeler, używając kodowania UTF-8. Dane w pliku mogą mieć postać *obiektu*, *tablicy* albo *wartości*. Węzeł źródłowy JSON obsługuje tylko odczyt *tablicy* obiektów, a obiekt nie może być zagnieżdżony.

Przykładowe dane JSON:

```
[
 {
  "After": 122762,
  "Promotion": 1467,
  "Cost": 23.99,
  "Class": "Confection",
  "Before": 114957
 },
 {
  "After": 137097,
  "Promotion": 1745,
```

```
  "Cost": 79.29,
  "Class": "Drink",
  "Before": 123378
 }
]
```

Gdy program SPSS Modeler odczytuje dane z pliku JSON, wykonuje następujące przekształcenia.

Tabela 13. Przekształcenia typów składowania danych JSON

Wartość JSON	Typ składowania danych w programie SPSS Modeler
string	Łańcuch
number(int)	Liczba całkowita
number(real)	Liczba rzeczywista
true	1(Integer)
false	0(Integer)
null	Brakujące wartości

W oknie dialogowym węzła źródłowego JSON dostępne są następujące opcje.

Źródło danych JSON. Wybierz plik JSON do zaimportowania.

Format łańcucha JSON. Określ format łańcucha JSON. Wybierz opcję **Rekordy**, jeśli plik JSON jest zbiorem par nazwa-wartość. Węzeł źródłowy JSON zaimportuje nazwy jako nazwy zmiennych w programie SPSS Modeler. Lub wybierz opcję **Wartości**, jeśli dane JSON używają tylko wartości (bez nazw).

Wspólne zakładki węzłów źródłowych

Poniżej przedstawiono opcje, jaki można określić dla wszystkich węzłów źródłowych po kliknięciu odpowiedniej zakładki:

- **Zakładka Dane.** Służy do zmiany domyślnego typu składowania.
- **Zakładka Filtr.** Służy do eliminacji lub zmiany nazwy zmiennych danych. Ta zakładka oferuje taką samą funkcjonalność jak węzeł filtrowania. Więcej informacji można znaleźć w temacie “Ustawianie opcji filtrowania” na stronie 152.
- **Karta Typy.** Służy do ustawiania poziomów pomiaru. Ta zakładka oferuje taką samą funkcjonalność jak węzeł typu.
- **Zakładka Adnotacje.** Ta karta jest używana dla wszystkich węzłów i oferuje opcje zmiany nazwy węzłów, obsługi podpowiedzi użytkownika i zapisywania długich powiadomień.

Ustawianie poziomów pomiaru w węźle źródłowym

Właściwości zmiennej można określić w węźle źródłowym lub w osobnym węźle typu. Działanie jest podobne w przypadku obu węzłów. Dostępne są następujące właściwości:

- **Zmienna** Należy dwukrotnie kliknąć dowolną nazwę zmiennej, aby określić etykiety wartości i zmiennej dla danych w programie IBM SPSS Modeler. Przykładowo można tutaj wyświetlić lub zmodyfikować metadane zmiennej zaimportowane z programu IBM SPSS Statistics. Podobnie, można utworzyć nowe etykiety dla zmiennych i ich wartości. Określone tutaj etykiety są wyświetlane w programie IBM SPSS Modeler w zależności od opcji wybranych w oknie dialogowym właściwości strumienia.
- **Poziom pomiaru** Jest to poziom pomiaru używany do opisania charakterystyk danych w określonej zmiennej. Jeśli wszystkie szczegóły zmiennej są znane, jest ona nazywana **w pełni określoną**. Aby uzyskać więcej informacji, zobacz “Poziomy pomiaru” na stronie 139.

  Uwaga: Poziom pomiaru zmiennej różni się od typu składowania, który wskazuje, że dane są zapisywane jako łańcuchy, liczby całkowite, liczby rzeczywiste, daty, godziny, znaczniki czasu lub listy.

- **Wartości** Ta kolumna umożliwia określenie opcji do odczytywania wartości danych ze zbiory danych lub użycie opcji **Określ** w celu określenia poziomów pomiaru i wartości w osobnym oknie dialogowym. Można również wybrać opcję przepuszczenia zmiennych bez odczytywania ich wartości. Aby uzyskać więcej informacji, zobacz “Wartości danych” na stronie 143.

  Uwaga: Nie można poprawiać wartości komórek w tej kolumnie, jeśli odpowiadająca jej pozycja **Zmienna** zawiera listę.
- **Braki** Służy do określania sposoby traktowania braków danych dla zmiennej. Aby uzyskać więcej informacji, zobacz “Definiowanie braków danych” na stronie 148.

  Uwaga: Nie można poprawiać wartości komórek w tej kolumnie, jeśli odpowiadająca jej pozycja **Zmienna** zawiera listę.
- **Sprawdź** W tej kolumnie można ustawić opcje, dzięki którym wartości zmiennej będą zgodne z określonymi wartościami lub zakresami. Aby uzyskać więcej informacji, zobacz “Sprawdzanie wartości typu” na stronie 148.

  Uwaga: Nie można poprawiać wartości komórek w tej kolumnie, jeśli odpowiadająca jej pozycja **Zmienna** zawiera listę.
- **Rola** Służy do przekazania węzłom modelowania informacji, czy zmienne będą **wejściowe** (zmienne predykcyjne) czy **przewidywane** (zmienne przewidywane), na potrzeby procesu uczenia maszynowego. Dostępne są również role **Łącznie** i **Brak** oraz **Partycja**, która wskazuje zmienną użytą do podziału rekordów na osobne próby na potrzeby uczenia, testowania i walidacji. Wartość **Podział** określa, że dla każdej możliwej wartości zmiennej tworzone będą osobne modele. Aby uzyskać więcej informacji, zobacz “Ustawianie roli zmiennej” na stronie 149.

Więcej informacji można znaleźć w temacie “Węzeł Typy” na stronie 137.

Kiedy przeprowadzić określanie w węźle źródłowym

Istnieją dwa sposoby uzyskania wiedzy na temat składowania danych wartości dla zmiennych. *Określanie* może może mieć miejsce w węźle źródłowym, po wprowadzeniu danych do programu IBM SPSS Modeler po raz pierwszy lub po wstawieniu węzła typu do strumienia danych.

Określanie w węźle źródłowym jest przydatne, jeśli:

- Zbiór danych jest mały.
- Użytkownik planuje wyznaczenie nowych zmiennych za pomocą konstruktora wyrażeń (określanie sprawia, że wartości zmiennych są dostępne za pośrednictwem konstruktora wyrażeń).

Ogólnie, jeśli zbiór danych nie jest bardzo duży i jeśli w dalszej części strumienia zmienne nie będą dodawane, najwygodniejszą metodą jest określenie w węźle źródłowym.

Uwaga: W przypadku eksportu danych w węźle eksportu do bazy danych wymagane jest pełne określenie danych.

Filtrowanie zmiennych z węzła źródłowego

Karta Filtrowanie w oknie dialogowym węzła źródłowego umożliwia wykluczenie zmiennych z operacji w dalszej części strumienia w oparciu o wstępna sprawdzenie danych. Jest na przykład przydatna, jeśli w danych znajdują się zmienne zduplikowane lub jeśli użytkownik jest już wystarczająco zaznajomiony z danymi, aby wykluczać nieodpowiednie zmienne. Alternatywnie można dodać osobny węzeł Filtrowanie w dalszej części strumienia. Działanie jest podobne w obu przypadkach. Więcej informacji można znaleźć w temacie “Ustawianie opcji filtrowania” na stronie 152.

Rozdział 3. Węzły operacji związanych z rekordami

Przegląd operacji związanych z rekordami

Węzły operacji związanych z rekordami umożliwiają wprowadzanie zmian w danych na poziomie rekordu. Operacje te są ważne na etapie **Zrozumienie danych** oraz na etapie **Przygotowanie danych** w procesie eksploracji danych, ponieważ umożliwiają dostosowanie danych do konkretnych potrzeb biznesowych.

Przykładowo, w oparciu o wyniki audytu danych przeprowadzonego za pośrednictwem węzła Audyt danych (paleta wyników) można podjąć decyzję, czy rekordy dotyczące zakupów danego klienta w ciągu ostatnich trzech miesięcy mają zostać połączone. Za pośrednictwem węzła Łączenie można połączyć rekordy w oparciu o wartości zmiennej kluczowej, takiej jak *Customer ID* (Id. klienta). Lub można odkryć, że nie można połączyć dany zawierającej informacje na temat liczby odsłon strony WWW, jeśli liczba rekordów przekracza milion. Korzystając z węzła Losowanie można wybrać podzbiór danych, jakie będą używane podczas modelowania.

Paleta operacji związanych z rekordami zawiera następujące węzły:

Węzeł Selekcja wybiera lub odrzuca podzbiór rekordów ze strumienia danych na podstawie określonego warunku. Na przykład można wybrać rekordy należące do konkretnego regionu sprzedaży.

Węzeł Losowanie wybiera podzbiór rekordów. Obsługiwanych jest wiele typów prób, w tym próby: warstwowana, zgrupowana i nielosowa (strukturalna). Próbkowanie może być przydatne do zwiększenia wydajności oraz podczas wyboru grup powiązanych rekordów lub transakcji do analizy.

Węzeł Zrównoważenie poprawia dysproporcje w zbiorze danych, tak aby spełniał określone warunki. Dyrektywa równoważenia koryguje proporcje rekordów, w których warunek został spełniony, na podstawie określonego czynnika.

Węzeł Agregacja zastępuje sekwencję rekordów wejściowych zsumowanymi, zagregowanymi rekordami wyjściowymi.

Węzeł Agregacja RFM (Recency — Aktualność, Frequency — Częstość, Monetary — Kwota) umożliwia dostęp do historycznych danych transakcyjnych klienta, usunięcie danych nieużywanych i połączenie wszystkich pozostałych danych transakcyjnych w jednym wierszu, który informuje, kiedy ostatnio dokonywana była transakcja, ilu transakcji dokonano oraz jaka jest łączna kwota tych transakcji.

Węzeł Sortowanie sortuje rekordy w kolejności rosnącej lub malejącej na podstawie wartości jednej lub większej liczby zmiennych.

Węzeł Łączenie na podstawie wielu rekordów wejściowych tworzy pojedynczy rekord wyjściowy zawierający niektóre lub wszystkie zmienne wejściowe. Jest przydatny podczas scalania danych z różnych źródeł, takich jak dane wewnętrzne klienta oraz dane demograficzne osób, które dokonały zakupu.

Węzeł Dołączanie łączy zestawy rekordów. Jest to przydatne do łączenia zbiorów danych z podobnymi strukturami zawierającymi inne dane.

Węzeł Powtórzenia usuwa zduplikowane rekordy, przekazując pierwszy odmienny rekord do strumienia danych lub odrzucając pierwszy rekord i przekazując do strumienia danych wszystkie duplikaty.

Węzeł Szeregi czasowe buduje i ocenia modele szeregów czasowych w jednym kroku. Tego węzła z danymi można używać w środowisku lokalnym lub rozproszonym; w środowisku rozproszonym można wykorzystać moc programu IBM SPSS Analytic Server

Algorytm Spectral Clustering© przy wykorzystaniu kilku wektorów własnych rzutuje dane na przestrzeń o mniejszej liczbie wymiarów. Następnie w nowej przestrzeni stosowany jest algorytm grupowania metodą k-średnich w celu podzielenia danych na skupienia. Algorytm ten działa dość szybko w przypadku małych rekordów z wieloma zmiennymi, ale charakteryzuje się dużym kosztem obliczeniowym w przypadku dużych zbiorów danych. Węzeł Grupowanie spektralne w produkcie SPSS Modeler eksponuje podstawowe funkcje i często używane parametry biblioteki Spectral Clustering. Węzeł jest zaimplementowany w języku Python.

Siatka czasoprzestrzeni stanowi rozszerzenie geokodowanych lokalizacji przestrzennych. Mówiąc dokładniej, siatka czasoprzestrzeni to łańcuch alfanumeryczny reprezentujący obszar czasu i przestrzeni o regularnych kształtach.

Węzeł Szeregi czasowe buduje i ocenia modele szeregów czasowych w jednym kroku.

Węzeł optymalizacji CPLEX zapewnia możliwość korzystania z zaawansowanej optymalizacji matematycznej (CPLEX) za pośrednictwem pliku modelu OPL (Optimization Programming Language). Ta funkcja jest dostępna w produkcie IBM Analytical Decision Management, jednak teraz węzła CPLEX można również używać w programie SPSS Modeler bez IBM Analytical Decision Management.

Więcej informacji na temat optymalizacji CPLEX i OPL zawiera dokumentacja IBM Analytical Decision Management https://www.ibm.com/support/knowledgecenter/SS6A3P_18.0.0/configurableapps/knowledge_center/product_landing.html.

Wiele węzłów w palecie operacji związanych z rekordami wymaga użycia wrażenia CLEM. Jeśli użytkownik jest zaznajomiony z wyrażeniami CLEM, może wpisać wyrażenie w tym polu. Jednaj wszystkie pola wyrażenia zawierają przycisk otwierający konstruktora wyrażeń CLEM, który ułatwia automatyczne utworzenie wyrażeń.

Rysunek 1. Przycisk konstruktora wyrażeń

Węzeł selekcji

Węzły selekcji umożliwiają wybór lub odrzucenie podzbioru rekordów ze strumienia danych w oparciu o określony warunek, taki jak BP (ciśnienie krwi) = "HIGH" (Wysokie).

Dominanta. Określa, czy rekordy spełniające dany warunek będą uwzględniane w strumieniu danych, czy będą z niego wykluczane.

- **Uwzględnij.** Tę opcję należy wybrać, aby uwzględnić rekordy spełniające określony warunek wyboru.
- **Odrzuć.** Tę opcję należy wybrać, aby wykluczyć rekordy spełniające określony warunek wyboru.

Warunek. Wyświetla warunek wyboru, jaki będzie używany do przetestowania każdego rekordu, określany za pośrednictwem wyrażenia CLEM. Można wprowadzić wyrażenie w oknie lub użyć konstruktora wyrażeń, klikając przycisk kalkulatora (Konstruktor wyrażeń) po prawej stronie okna.

Jeśli wybrana zostanie opcja odrzucenia rekordów na podstawie warunku, takiego jak:

```
(var1='value1' and var2='value2'),
```

gdzie var1 to war1, value1 to wartość1, var2 to war2, value2 to wartość2, węzeł selekcji domyślnie odrzuci również rekordy z wartościami null dla wszystkich wybranych zmiennych. Aby tego uniknąć, należy zastąpić następujący warunek oryginalnym:

```
and not(@NULL(var1) and @NULL(var2))
```

Węzły selekcji są również używane do wyboru proporcji rekordów. Zwykle dla tej operacji używany będzie inny węzeł, węzeł próby. Jednak jeśli warunek, jaki ma zostać użyty, jest bardziej złożony niż udostępnione parametry, można utworzyć własny warunek, używając węzła selekcji. Przykładowo można utworzyć następujący warunek:

```
BP = "HIGH" and random(10) <= 4
```

gdzie BP to ciśnienie krwi, HIGH — Wysokie a random(10) to losowanie liczb z przedziału od 0 do 9. Spowoduje to wybranie około 40% rekordów przedstawiających wysokie ciśnienie krwi i przepuszczenie tych rekordów w dół strumienia w celu przeprowadzenia dalszej analizy.

Węzeł próby

Węzły próby umożliwiają wybór podzbioru rekordów na potrzeby analizy lub określenie proporcji rekordów do odrzucenia. Obsługiwanych jest wiele typów prób, w tym próby: warstwowana, zgrupowana i nielosowa (strukturalna). Dobór próby może być stosowany z kliku przyczyn:

- W celu zwiększenia wydajności poprzez oszacowanie modeli na podzbiorze danych. Modele oszacowane na podstawie próby często są tak samo dokładne, jak te wyznaczone z pełnego zbioru danych; mogą być również bardziej dokładne, jeśli zwiększona wydajność pozwoli na eksperymentowanie z różnymi metodami, na które w innych okolicznościach nie zwrócono by uwagi.
- W celu wybrania grup powiązanych rekordów lub transakcji do analizy, tak jak w przypadku wyboru wszystkich towarów z koszyka zakupów w sklepie online (lub w koszyku w supermarkecie) lub właściwości w określonym sąsiedztwie.
- W celu zidentyfikowania jednostek lub obserwacji dla kontroli losowej pod kątem zapewnienia jakości, zapobiegania oszustwom lub ze względów bezpieczeństwa.

Uwaga: Jeśli dane mają zostać po prostu podzielone na próby testujące i uczące w celu przeprowadzenia walidacji, można użyć węzła podziału na podzbiory. Więcej informacji można znaleźć w temacie “Węzeł Partycja” na stronie 176.

Typy prób

Próby zespołowe. Raczej grupy prób niż pojedyncze jednostki. Załóżmy na przykład, że dostępny jest plik danych zawierający po jednym rekordzie na ucznia. Jeśli przeprowadzone zostanie grupowanie według szkoły i wielkość próby będzie wynosiła 50%, wówczas wybranych zostanie 50% szkół, a z każdej wybranej szkoły zostaną wybrani wszyscy uczniowie. Uczniowie ze szkół, które nie zostaną wybrane, będą odrzuceni. Średnio można oczekiwać, że wybranych zostanie około 50% uczniów, jednak ze względu na różną wielkość szkół, wartość procentowa może nie być dokładna. Podobnie można pogrupować towary w koszyku zakupów według identyfikatora transakcji, aby mieć pewność, że wszystkie towary dla wybranych transakcji zostaną zachowane. Przykład pogrupowania właściwości według miast zawiera strumień przykładowy *complexsample_property.str*.

Próby warstwowe. Próby wybierane są niezależnie w niepokrywających się podgrupach populacji lub warstwy. Przykładowo można zapewnić dobór próby wśród mężczyzn i kobiet w jednakowych proporcjach lub zapewnić, że reprezentowany będzie każdy region lub każda grupa socjoekonomiczna w populacji miejskiej. Można również określić inną wielkość próby dla każdej warstwy (na przykład, jeśli przypuszcza się, że jedna grupa była niedostatecznie reprezentowana w oryginalnych danych). Przykład podziału na warstwy właściwości według hrabstwa zawiera strumień przykładowy *complexsample_property.str*.

Systematyczny dobór próby lub dobór próby co N-ty rekord. Jeśli wybór w sposób losowy jest trudny do uzyskania, jednostki można podzielić na próby w sposób systematyczny (wg stałego przedziału) lub sekwencyjny.

Wagi losowania. Wagi losowania są automatycznie wyliczane podczas losowania próby złożonej i w przybliżeniu odpowiadają „częstości”, z jaką każda próbkowana jednostka reprezentuje oryginalne dane. Dlatego, suma wag w próbie powinna szacować wielkość oryginalnych danych.

Operat losowania

Operat losowania definiuje potencjalne źródło obserwacji, jakie będą uwzględnione w próbie lub badaniu. W niektórych przypadkach możliwe jest zidentyfikowanie każdego pojedynczego elementu populacji i uwzględnienie każdego z nich w tej samej próbie — na przykład, jeśli dobór próby dotyczy towarów opuszczających linię produkcyjną. Znacznie częściej uzyskanie dostępu do każdej możliwej obserwacji nie będzie możliwe. Na przykład, nie można mieć pewności, kto będzie głosował w wyborach, dopóki wybory się nie zakończą. W takim przypadku można użyć rejestru wyborczego jako operatu losowania, nawet jeśli niektóre zarejestrowane osoby nie zagłosują, a inne zagłosują pomimo tego, że nie znalazły się na listach w chwili sprawdzania rejestru. Osoby, które nie zostały

uwzględnione przez operat losowania, nie mają szans na ujęcie w próbie. Pytanie, jakie należy zadać w odniesieniu do każdej obserwacji w czasie rzeczywistym, to: czy operat losowania jest wystarczająco zbliżony do charakteru populacji, która jest poddawana ocenie.

Opcje węzła próby

Można wybrać metodę **prostą** lub **złożoną**, odpowiednio do wymagań.

Opcje prostego doboru próby

Metoda prosta umożliwia wybranie losowej wartości procentowej rekordów, wybranie rekordów ciągłych lub wybranie co *n-tego* rekordu.

Dominanta. Tę opcję należy wybrać, aby przepuścić (uwzględnić) lub odrzucić (wykluczyć) rekordy dla następujących trybów:

- **Uwzględnij próbę.** Uwzględnia wybrane rekordy w strumieniu danych i odrzuca wszystkie pozostałe. Przykładowo, jeśli wybrany zostanie tryb **Uwzględnij próbę**, a opcja **Co N-ty rekord** zostanie ustawiona na wartość 5, wówczas uwzględniany będzie każdy co piąty rekord, co utworzy zbiór danych, który będzie w przybliżeniu stanowił jedną piątą oryginalnej wielkości. Jest to tryb domyślny dla doboru próby z danych i jedyny tryb, jakiego można użyć w przypadku metody złożonej.
- **Odrzuć próbę.** Wyklucza wybrane rekordy i uwzględnia wszystkie pozostałe. Przykładowo, jeśli tryb zostanie ustawiony na **Odrzuć próbę**, a opcja **Co N-ty rekord** zostanie ustawiona na wartość 5, wówczas każdy co piąty rekord zostanie odrzucony. Ten tryb jest dostępny tylko w przypadku metody prostej.

Przykład. Metodę doboru próby można wybrać, ustawiając jedną z następujących opcji:

- **Pierwsza.** Tę opcję należy wybrać, aby zastosować dobór próby z danych w bezpośrednim sąsiedztwie. Przykładowo, jeśli maksymalna wielkość próby jest ustawiona na 10000, wówczas wybranych zostanie pierwszych 10 000 rekordów.
- **Co N-ty rekord.** Tę opcję należy wybrać, aby dobór próby z danych odbywał się poprzez przepuszczenie lub odrzucenie każdego *n*-tego rekordu. Na przykład, jeśli wartość *n* jest ustawiona na 5, wówczas wybrany będzie każdy co piąty rekord.
- **Losowo % ze wszystkich.** Tę opcję należy wybrać, aby próba dobierana była jako losowa wartość procentowa z danych. Na przykład, jeśli wartość procentowa zostanie ustawiona na 20, wówczas 20% danych zostanie przepuszczonych do strumienia danych lub odrzuconych, w zależności od wybranego trybu. Do określenia wartości procentowej doboru próby należy użyć zmiennej. Można również określić wartość początkową, używając elementu sterującego **Ustaw wartość początkową generatora liczb losowych**.

  Użyj próbkowania na poziomie bloku (tylko w bazach danych). Ta opcja jest włączona tylko po wybraniu losowego procentowego doboru próby podczas wykonywania eksploracji w bazie danych w bazie Oracle lub IBM Db2. W takiej sytuacji próbkowanie na poziomie może być bardziej skuteczne.

  Uwaga: Po każdym uruchomieniu doboru próby przy tych samych ustawieniach próby losowej uzyskiwana liczba zwracanych wierszy nie jest dokładnie taka sama. Dzieje się tak dlatego, ponieważ dla każdego rekordu wejściowego istnieje prawdopodobieństwo N/100, że zostanie on uwzględniony w próbie (gdzie N to wartość **Losowo % ze wszystkich** określona w węźle), a prawdopodobieństwa są niezależne; dlatego wyniki nie wynoszą dokładnie N%.

Maksymalna wielkość próby. Określa maksymalną liczbę rekordów do uwzględnienia w próbie. Ta opcja jest nadmiarowa i dlatego jest wyłączona, jeśli wybrane są ustawienia **Pierwsza** i **Uwzględnij**. Należy również pamiętać, że użycie tej opcji w połączeniu z opcją **Losowo % ze wszystkich** może uniemożliwić wybranie niektórych rekordów. Przykładowo, jeśli dostępnych jest 10 milionów rekordów w zbiorze danych i wybranych zostanie 50% rekordów z maksymalną wielkością próby wynoszącą trzy miliony rekordów, wówczas wybranych zostanie 50% z pierwszych sześciu milionów rekordów, a pozostałe cztery miliony rekordów nie mają szansy na wybranie. Aby uniknąć tego ograniczenia, należy wybrać **złożoną** metodę doboru próby i zażądać próby losowej z trzech milionów rekordów bez określania zmiennej grupowania lub warstwującej.

Opcje złożonego doboru próby

Opcje próby złożonej umożliwiają dokładniejszą kontrolę próby, z uwzględnieniem prób zgrupowanych, warstwowych i ważonych wraz z innymi opcji.

Zespoły i warstwy. Umożliwia określenie w razie potrzeby zmiennych grupowania, warstwujących i ważących zmiennych wejściowych. Więcej informacji można znaleźć w temacie "Okno Zespoły i warstwy — ustawienia".

Typ próby.

- **Losowa.** Umożliwia wybranie grup lub rekordów losowo dla każdej warstwy.
- **Systematyczna.** Rekordy wybierane są według stałego przedziału. Ta opcja działa podobnie jak metoda *Co N-ty rekord*, z wyjątkiem pozycji pierwszego rekordu, która zmienia się w zależności od wartości początkowej. Wartość *n* jest określana automatycznie w oparciu o wielkość próby lub jej proporcje.

Jednostki próby. Jako podstawowe jednostki próby można wybrać proporcje lub liczebności.

Wielkość próby. Wielkość próby można określić na kilka sposobów:

- **Stała.** Umożliwia określenie ogólnej wielkości próby jako liczebność lub proporcję.
- **Użytkownika.** Umożliwia określenie wielkości próby dla każdej podgrupy lub warstwy. Ta opcja jest dostępna tylko po określeniu zmiennej warstwowania w podoknie dialogowym Zespoły i warstwy.
- **Zmienne.** Umożliwia użytkownikowi pobranie zmiennej, która definiuje wielkość próby dla każdej podgrupy lub warstwy. Ta zmienna powinna mieć taką samą wartość dla każdego rekordu z konkretnej warstwy; przykładowo, jeśli próba jest warstwowana według hrabstwa, wówczas wszystkie rekordy, dla których *county = Surrey* (hrabstwo = Surrey) muszą mieć taką samą wartość. Zmienna musi być numeryczna, a jej wartości muszą być zgodne z wybranymi jednostkami próby. W przypadku proporcji wartości powinny być większe od 0 i mniejsze niż 1; dla liczebności wartość minimalna wynosi 1.

Minimalna próba w warstwie. Określa minimalna liczbę rekordów (lub minimalną liczbę grup, jeśli wybrano zmienną grupowania).

Minimalna próba w warstwie. Określa maksymalną liczbę rekordów lub grup. Jeśli ta opcja zostanie wybrana bez określenia zmiennej grupowania lub warstwowania, wybrana zostanie próba losowa lub systematyczna o określonej wielkości.

Ustaw wartość początkową generatora liczb losowych. W przypadku próbkowania lub dzielenia na podzbiory rekordów w oparciu o losową wartość procentową opcja ta pozwala na zduplikowanie tych samych wyników w innej sesji. Określenie wartości początkowej używanej przez generator liczb pseudolosowych zapewni, że podczas każdego wykonywania węzła przypisywane będą te same rekordy. Wprowadź żądaną wartość początkową generatora lub kliknij przycisk **Utwórz**, aby automatycznie wygenerować wartość losową. Jeśli nie wybrano tej opcji, przy każdej próbie wykonania węzła wygenerowana zostanie inna próba.

Uwaga: Jeśli używana jest opcja **Ustaw wartość początkową generatora liczb losowych** w przypadku rekordów odczytanych z bazy danych, przed przeprowadzeniem próby konieczne może być sortowanie węzła, aby po każdym wykonaniu węzła uzyskany wynik był taki sam. Wynika to z faktu, że wartość początkowa generatora liczb losowych zależy od kolejności rekordów, która w relacyjnej bazie danych nie musi pozostawać jednakowa. Więcej informacji można znaleźć w temacie "Węzeł Sortowanie" na stronie 84.

Okno Zespoły i warstwy — ustawienia

Okno dialogowe Zespoły i warstwy umożliwia wybór zmiennych grupowania, warstwowania i zmiennych wagi podczas losowania próby złożonej.

Grupy. Określa zmienną jakościową użytą w rekordach grupowania. Rekordy są próbkowane na podstawie przynależności do grupy, przy czym niektóre grupy mogą być uwzględniane, a inne nie. Jeśli jednak rekord z danej grupy jest uwzględniony, wszystkie są uwzględniane. Przykładowo podczas analizowania powiązań produktów w

koszykach zakupów należy pogrupować towary według identyfikatorów transakcji, aby upewnić się, że uwzględnione będą wszystkie towary dla wybranych transakcji. Zamiast próbkowania rekordów — które zniszczyłoby informacje jej na temat tego, które towary zostały sprzedane razem — można wykonać próbkowanie transakcji, tak aby upewnić się, czy wszystkie rekordy dla wybranych transakcji zostały zachowane.

Warstwy według. Określa zmienną jakościową używaną do warstwowania rekordów, tak aby próby były wybierane niezależnie w niepokrywających się podgrupach populacji lub warstwy. Jeśli na przykład wybrana zostanie 50% próba warstwowana według płci, wówczas wybrane zostaną dwie 50% próby: jedna dla mężczyzn, a druga dla kobiet. Przykładowo, warstwą mogą być grupy socjoekonomiczne, kategorie stanowisk, grupy wiekowe lub grupy etniczne, które zapewnią odpowiednie wielkości próby to dla interesujących podgrup. Jeśli w oryginalnym zbiorze danych jest trzy razy więcej kobiet niż mężczyzn, współczynnik ten zostanie zachowany poprze próbkowanie osobno dla każdej grupy. Można również określić wiele zmiennych warstwowania (na przykład, próbkowanie linii produktów w regionach i odwrotnie).

Uwaga: Jeśli podział na warstwy jest dokonywany według zmiennej, która zawiera braki danych (wartości null lub systemowe braki danych, puste łańcuchy, białe znaki i puste lub zdefiniowane przez użytkownika braki danych), wówczas nie można określić niestandardowej wielkości próby dla warstwy. Aby użyć niestandardowych wielkości prób podczas podziału na warstwy według zmiennej zawierającej braki danych lub puste wartości, wówczas należy je wypełnić we wcześniejszej części strumienia.

Użyj wagi wejściowej. Określa zmienną użytą do ważenia rekordów przed rozpoczęciem próbkowania. Przykładowo, jeśli zmienna ważąca ma wartości z przedziału od 1 do 5, wówczas dla rekordów z wagą 5 prawdopodobieństwo wybrania jest pięć razy większe. Wartości tej zmiennej zostaną zastąpione przez końcowe wagi wynikowe wygenerowane przez węzeł (patrz kolejny akapit).

Nowa waga wyjściowa. Określa nazwę zmiennej, w której wagi końcowe są zapisywane, jeśli nie określono żadnej wejściowej zmiennej ważącej. (Jeśli wejściowa zmienna ważąca została określona, jej wartości są zastępowane przez wagi końcowe, jak opisano powyżej, i nie jest tworzona żadna osobna wynikowa zmienna ważąca). Wartości wag wynikowych określają liczbę rekordów reprezentowanych przez poszczególne próbkowane rekordy w oryginalnych danych. Suma wartości wagi daje oszacowanie wielkości próby. Przykładowo, jeśli dostępna jest 10% próba, waga wynikowa będzie wynosiła 10 dla wszystkich rekordów, co oznacza, że każdy próbkowany rekord reprezentuje w przybliżeniu dziesięć rekordów w oryginalnych danych. W próbie warstwowej lub ważonej wartości wag wynikowych mogą różnić się w oparciu o proporcję próby dla każdej warstwy.

Komentarze

- Losowanie zespołowe jest przydatne, jeśli nie można uzyskać pełnej listy populacji, z jakiej ma zostać pobrana próba, ale można uzyskać kompletne listy dla określonych grup lub zespołów. Jest również używany, jeśli próba losowa spowoduje utworzenie listy obiektów badanych, z którymi kontakt byłby niepraktyczny. Przykładowo, łatwiejPrzykładowo, łatwiej byłoby odwiedzić wszystkich rolników w jednym hrabstwie niż określonych rolników znajdujących się w różnych częściach kraju.
- Można określić zmienne losowania zespołowego oraz próby warstwowej, aby pobrać próbę w grupach niezależnie dla każdej warstwy. Przykładowo można pobrać próbę dla wartości właściwości podzielonych na warstwy według hrabstwa i grupy według miast w każdym hrabstwie. Dzięki temu niezależna próba miast zostanie pobrana z każdego hrabstwa. Niektóre miasta zostaną uwzględnione, inne nie, ale dla każdego miasta, które zostanie uwzględnione, wszystkie właściwości odnoszące się do miasta zostaną uwzględnione.
- Aby wybrać losową próbę jednostek z każdej grupy, można połączyć w łańcuch dwa węzły próby. Na przykład można najpierw pobrać próbę dla gmin podzielonych na warstwy według hrabstwa, jak opisano powyżej. Następnie dołączyć drugi węzeł próby i wybrać *town* (miasto) jako zmienną warstwującą, co umożliwi pobranie próby dla proporcjonalnej liczby rekordów z każdej gminy.
- W sytuacji, kiedy do jednoznacznego określenia grup konieczne jest zastosowanie kombinacji zmiennych, można wygenerować nową zmienną, używając węzła wyliczeń. Przykładowo, jeśli kilka sklepów korzysta z tego samego systemu numeracji transakcji, można wyznaczyć nową zmienną, która połączy identyfikatory sklepów i transakcji.

Wielkości prób dla warstw

Podczas losowania próby warstwowej domyślną opcją jest dobór próby składającej się z takiej samej proporcji rekordów lub grup z każdej warstwy. Jeśli na przykład jedna grupa przewyższa liczebnie inną według współczynnika 3, zazwyczaj uzasadnione jest zachowanie tego samego współczynnika w próbie. Jeśli jednak nie jest to konieczne, można określić wielkość próby osobno dla każdej warstwy.

Okno dialogowe Wielkości prób dla warstw wyświetla listę poszczególnych wartości zmiennej warstwowania, umożliwiając zastąpienie wartości domyślnych dla warstwy. Jeśli wybrano wiele zmiennych warstwowania, na liście wyświetlana jest każda możliwa kombinacja wartości, dzięki czemu można na przykład określić wielkość każdej grupy etnicznej w mieście lub każde miasto w danym hrabstwie. Wielkości są określane jako proporcje lub liczebności, zgodnie z bieżącymi ustawianiami w węźle próby.

Aby określić wielkości prób dla warstw

1. W węźle próby zaznacz opcję **Złożone**, a następnie wybierz co najmniej jedną zmienną warstwowania. Więcej informacji można znaleźć w temacie “Okno Zespoły i warstwy — ustawienia” na stronie 76.
2. Wybierz opcję **Użytkownika**, a następnie **Określ wielkość**.
3. W oknie dialogowym Wielkości prób dla warstw kliknij przycisk **Odczytaj wartości** u dołu po lewej stronie, aby wypełnić pola. W razie konieczności można określić wartości we wcześniejszym węźle źródłowym lub w węźle typu. Więcej informacji można znaleźć w temacie “Co to jest określanie?” na stronie 143.
4. Kliknij dowolny wiersz, aby zastąpić domyślną wielkość warstwy.

Uwagi do wielkości próby

Niestandardowe (użytkownika) wielkości próby mogą być przydatne, jeśli dla różnych warstw istnieją różne wariancje, przykładowo w celu utworzenia wielkości prób proporcjonalnych do odchylenia standardowego. (Jeśli obserwacje w danej warstwie są bardziej zróżnicowane, podczas doboru próby należy wybrać ich więcej, aby uzyskana próba była reprezentatywna). Lub jeśli warstwa jest mała, można użyć większej proporcji próby, aby upewnić się, że uwzględniona zostanie minimalna liczba obserwacji.

Uwaga: Jeśli podział na warstwy jest dokonywany według zmiennej, która zawiera braki danych (wartości null lub systemowe braki danych, puste łańcuchy, białe znaki i puste lub zdefiniowane przez użytkownika braki danych), wówczas nie można określić niestandardowej wielkości próby dla warstwy. Aby użyć niestandardowych wielkości prób podczas podziału na warstwy według zmiennej zawierającej braki danych lub puste wartości, wówczas należy je wypełnić we wcześniejszej części strumienia.

Węzeł ważenia

Węzły ważenia umożliwiają poprawę dysproporcji w zbiorach danych, tak aby były zgodne z określonymi kryteriami testowymi. Załóżmy na przykład, że zbiór danych zawiera tylko dwie wartości — *low* (niskie) lub *high* (wysokie) — przy czym 90% obserwacji ma wartości *low*, a tylko 10% *high*. Wiele technik modelowania ma problemy z tak odchylonymi danymi, ponieważ mają tendencję do uczenia tylko wyniku *low* i ignorowania wyniku *high*, ponieważ występuje on rzadziej. Jeśli dane są dobrze równoważone i liczba wyników *low* i *high* jest w przybliżeniu taka sama, modele mają większą szansę na znalezienie wzorców, które rozróżnią te dwie grupy. W takim przypadku węzeł ważenia pomaga w utworzeniu dyrektywy równoważenia, która zredukuje liczbę obserwacji z wynikiem *low*.

Równoważenie jest wykonywane poprzez zduplikowanie, a następnie odrzucenie rekordów na podstawie określonych warunków. Rekordy, dla których nie ma żadnych warunków, zawsze są przepuszczane. Ponieważ ten proces działa poprzez zduplikowanie i/lub odrzucenie rekordów, oryginalna kolejność danych w operacjach w dalszej części strumienia zostaje utracona. Wartości zależne od kolejności należy wyliczyć przed dodaniem do strumienia danych węzła ważenia.

Uwaga: Węzły ważenia mogą być generowane automatycznie z rozkładów i histogramów. Przykładowo można zrównoważyć dane, aby pokazać równe proporcje we wszystkich kategoriach zmiennej jakościowej, w sposób przedstawiony na wykresie rozkładu.

Przykład. Podczas budowania strumienia RFM w celu zidentyfikowania ostatnich klientów, którzy pozytywnie odpowiedzieli na poprzednie kampanie marketingowe, dział marketingu firmy zajmującej się sprzedażą używa węzła ważenia w celu zrównoważenia różnic pomiędzy odpowiedziami prawda i fałsz w danych.

Ustawianie opcji dla węzła ważenia

Dyrektywy równoważenia rekordów. Wyświetla listę bieżących dyrektyw równoważenia. Każda dyrektywa obejmuje czynnik i warunek, który informuje oprogramowanie, aby „zwiększyć proporcję rekordów za pośrednictwem określonych czynników, dla których warunek jest prawdziwy". Czynnik niższy niż 1,0 oznacza, że proporcja wskazanych rekordów zostanie zmniejszona. Przykładowo, aby zmniejszyć liczbę rekordów, w których lek Y jest lekiem zastosowanym w leczeniu, można utworzyć dyrektywę równoważenia z czynnikiem wynoszącym 0,7 i warunkiem Drug = "drugY" (Lek = "lekY"). Ta dyrektywa oznacza, że liczba rekordów, w których lek Y jest lekiem zastosowanym w leczeniu, zostanie zmniejszona do 70% dla wszystkich operacji w dalszej części strumienia.

Uwaga: Czynniki równoważenia umożliwiające redukcję mogą być określane z dokładnością do czterech miejsc dziesiętnych. Czynniki ustawione na wartość niższą niż 0,0001 spowodują wystąpienie błędu, ponieważ wyniki nie zostaną poprawnie obliczone.

- **Utwórz warunki**, klikając przycisk po prawej stronie pola tekstowego. Spowoduje to wstawienie pustego wiersza umożliwiającego dodanie nowych warunków. Aby utworzyć wyrażenie CLEM dla warunku, należy kliknąć przycisk konstruktora wyrażeń.
- **Usuń dyrektywy** za pomocą czerwonego przycisku usuwania.
- **Sortowanie dyrektyw** — jest możliwe za pomocą strzałek w górę i w dół.

Równoważ tylko dane uczące. Jeśli w strumieniu występuje zmienna dzieląca na podzbiory, ta opcja spowoduje zrównoważenie danych tylko w podzbiorze uczącym. Może to być szczególnie przydatne w przypadku generowania skorygowanych ocen skłonności, które wymagają niezrównoważonych podzbiorów uczących lub walidacyjnych. Jeśli w strumieniu nie ma żadnej zmiennej dzielącej na podzbiory (lub określono wiele zmiennych dzielących na podzbiory), wówczas ta opcja jest ignorowana i równoważone są wszystkie dane.

Węzeł Agregacja

Agregacja to zadanie przygotowywania danych często stosowane do zmniejszenia wielkości zbioru danych. Przed wykonaniem agregacji należy wyczyścić dane, koncentrując się szczególnie na brakach danych. Po przeprowadzeniu agregacji potencjalnie przydatne informacje na temat braków danych mogą zostać utracone.

Węzeł Agregacja umożliwia zastąpienie sekwencji wprowadzanych rekordów podsumowaniem, czyli zagregowanymi rekordami wynikowymi. Przykładowo, dostępny może być zbiór rekordów danych wejściowych dotyczących sprzedaży, takich jak te pokazane w tabeli.

Tabela 14. Przykład rekordu danych wejściowych dot. sprzedaży

Age	Sex	Region	Branch	Sales
23	P	W	8	4
45	P	W	16	4
37	P	W	8	5
30	P	W	5	7
44	P	N	4	9
25	P	N	2	11
29	F	W	16	6
41	F	N	4	8
23	F	N	6	2
45	F	N	4	5
33	F	N	6	10

Rekordy te można zagregować, przyjmując jako zmienne kluczowe *Sex* (Płeć) i *Region*. Następnie można przeprowadzić agregację według zmiennej *Age* (Wiek), tryb **Średnia** oraz *Sales* (Sprzedaż), tryb **Suma**. Po wybraniu opcji **Dołącz liczebność rekordów w zmiennej** w oknie dialogowym Agregacja zagregowany wynik będzie wyświetlany w sposób przedstawiony w poniższej tabeli.

Tabela 15. Przykład zagregowanego rekordu

Age (mean)	Sex	Region	Sales (sum)	Record Count
35,5	F	N	25	4
29	F	W	6	1
34,5	P	N	20	2
33,75	P	W	20	4

Na podstawie tych danych można się na przykład dowiedzieć, że średni wiek czterech kobiet (F) z działu sprzedaży w regionie północnym (N) wynosi 35,5, a suma łącznie ich sprzedaży to 25 jednostek.

Uwaga: Zmienne, takie jak *Branch* (Gałąź) są automatycznie odrzucane, jeśli tryb agregacji nie zostanie określony.

Ustawianie opcji dla węzła Agregacja

W węźle agregacji można określić następujące wartości.

- Co najmniej jedną zmienną kluczową stanowiącą kategorię dla agregacji
- Co najmniej jedną zmienną agregacji, dla której obliczane są wartości agregacji
- Co najmniej jeden tryb agregacji (typ agregacji) dla wyniku dla każdej zmiennej agregacji

Można również określić domyślne tryby agregacji, jakie będą stosowane w przypadku nowych zmiennych oraz użyć wyrażeń (podobnych do formuł) w celu utworzenia kategorii agregacji.

Należy pamiętać, że w celu zwiększenia wydajności korzystne może być włączenie przetwarzania równoległego dla operacji agregacji.

Zmienne grupujące. Wyświetla listę zmiennych, jakie mogą zostać użyte jako kategorie dla agregacji. Zmiennymi kluczowymi mogą być zmienne ilościowe (numeryczne) oraz zmienne jakościowe. Jeśli wybrana zostanie więcej niż jedna zmienna kluczowa, wartości zostaną połączone, tworząc wartość kluczową dla agregowania rekordów. Dla każdej unikalnej zmiennej kluczowej wygenerowany zostanie jeden zagregowany rekord. Na przykład, jeśli Sex (Płeć) i Region są zmiennymi kluczowymi, dla każdej unikatowej kombinacji zmiennych M (Mężczyzna) i F (Kobieta) z regionami N (Północny) i S (Południowy) — cztery unikatowe kombinacje — utworzony zostanie zagregowany rekord. Aby dodać zmienną kluczową, należy użyć przycisku Selektor zmiennych, który znajduje się po prawej stronie okna.

Pozostała część okna dialogowego jest podzielona na dwa główne obszary — **Agregacje podstawowe** i **Wyrażenia agregujące**.

Agregacje podstawowe

Agregowane zmienne. Wyświetla listę zmiennych, których wartości zostaną poddane agregacji, oraz wybrane tryby agregacji. Aby dodać zmienne do listy, należy użyć przycisku Selektor zmiennych, który znajduje się po prawej stronie. Dostępne są następujące tryby agregacji.

Uwaga: Niektóre tryby nie mają zastosowania w przypadku zmiennych nienumerycznych (np. **Suma** dla zmiennej data/czas). Tryby, których nie można użyć dla wybranej zmiennej agregacji, są wyłączone.

- **Suma.** Należy wybrać tę opcję, aby zwrócone zostały zsumowane wartości dla każdej kombinacji zmiennej kluczowej. Suma to łączne wartości wszystkich obserwacji bez braków danych.
- **Średnia.** Należy wybrać tę opcję, aby zwrócone zostały wartości średnie dla każdej kombinacji zmiennej kluczowej. Średnia jest miarą tendencji centralnej i jest średnią arytmetyczną (suma podzielona przez liczbę obserwacji).
- **Min.** Należy wybrać tę opcję, aby zwrócone zostały wartości minimalne dla każdej kombinacji zmiennej kluczowej.
- **Maks.** Należy wybrać tę opcję, aby zwrócone zostały wartości maksymalne dla każdej kombinacji zmiennej kluczowej.
- **OdchStd.** Należy wybrać tę opcję, aby zwrócone zostało odchylenie standardowe dla każdej kombinacji zmiennej kluczowej. Odchylenie standardowe jest miarą rozproszenia wokół średniej i jest równe pierwiastkowi kwadratowemu miary wariancji.
- **Mediana.** Należy wybrać tę opcję, aby zwrócone zostały wartości mediany dla każdej kombinacji zmiennej kluczowej. W przeciwieństwie do średniej, na którą wpływ może mieć nawet kilka ekstremalnie wysokich lub niskich wartości, mediana jest miarą tendencji centralnej niewrażliwą na wartości odstające. Jest również znana jako 50. percentyl lub 2. kwartyl.
- **Liczebności.** Należy wybrać tę opcję, aby zwrócone zostały wartości liczności inne niż null dla każdej kombinacji zmiennej kluczowej.
- **Wariancja.** Należy wybrać tę opcję, aby zwrócone zostały wartości wariancji dla każdej kombinacji zmiennej kluczowej. Wariancja jest miarą rozproszenia wokół średniej, równą sumie podniesionych do kwadratu odchyleń od średniej, podzielonej przez liczbę obserwacji minus jeden.
- **1. Kwartyl.** Należy wybrać tę opcję, aby zwrócone zostały wartości 1. kwartyla (25. percentyl) dla każdej kombinacji zmiennej kluczowej.
- **3. Kwartyl.** Należy wybrać tę opcję, aby zwrócone zostały wartości 3. kwartyla (75. percentyl) dla każdej kombinacji zmiennej kluczowej.

Uwaga: Podczas uruchamiania strumienia zawierającego węzeł agregacji wartości zwracane dla 1. i 3. kwartyla po przekazaniu kodu SQL do bazy danych Oracle mogą różnić się od wartości zwróconych w trybie rodzimym.

Ustawienia domyślnych obliczeń. Należy określić domyślny tryb agregacji, jaki będzie używany w przypadku nowych zmiennych. Jeśli często używana jest taka sama agregacja, należy tutaj wybrać do najmniej jeden tryb i za pomocą przycisku Zastosuj do wszystkich po prawej stronie zastosować wybrane tryby do wszystkich wymienionych powyżej zmiennych.

Nowe rozszerzenie nazwy zmiennej. Tę opcję należy wybrać, aby dodać przyrostek lub przedrostek, taki jak „**1** ” lub „new” (**nowa**), do zduplikowanych zagregowanych zmiennych. Na przykład, jeśli wybrana zostanie opcja dodania przyrostka, a jako rozszerzenie ustawiona zostanie wartość **1**, wynikiem agregacji wartości minimalnych dla zmiennej **Age** (Wiek) będzie zmienna o nazwie Age_Min_1. Uwaga: Rozszerzenia agregacji, takie jak _Min (Minimum) lub Max_ (Maksimum), będą automatycznie dodawane do nowej zmiennej, wskazując typ wykonanej agregacji. Aby wskazać preferowany styl rozszerzenia, należy wybrać opcję **Przyrostek** lub **Przedrostek**.

Dołącz liczebność rekordów w zmiennej. Należy wybrać tę opcję, aby domyślnie w każdym rekordzie wynikowym o nazwie Record_Count (Liczebność rekordu) uwzględnić dodatkową zmienną. To pole wskazuje, ile rekordów wejściowych zostało zagregowanych w celu utworzenia poszczególnych rekordów agregacji. Należy utworzyć niestandardową nazwę dla tej zmiennej, wpisując ją w polu edycji.

Uwaga: Systemowe wartości null są wykluczane podczas obliczania agregacji, ale są uwzględniane w liczebności rekordu. Puste wartości natomiast są uwzględniane zarówno w agregacji, jak i w liczebności rekordu. Aby wykluczyć puste wartości, można użyć węzła wypełniania w celu zastąpienia pustych wartości wartościami null. Puste wartości można również usunąć za pomocą węzła Selekcja.

Wyrażenia agregujące

Wyrażenia są podobne do formuł budowanych z wartości, nazw zmiennych, operatorów i funkcji. W przeciwieństwie do funkcji, które działają dla jednego rekordu naraz, wyrażenia agregujące działają w przypadku grupy, zestawu lub zbioru rekordów.

Uwaga: Wyrażenia agregujące można również utworzyć, jeśli strumień będzie obejmował połączenie z bazą danych (za pośrednictwem węzła Źródło bazy danych).

Nowe wyrażenia są tworzone jako zmienne pochodne; aby utworzyć wyrażenie, należy użyć funkcji *Agregaty bazy danych*, które są dostępne za pośrednictwem konstruktora wyrażeń.

Więcej informacji na temat konstruktora wyrażeń zawiera publikacja IBM SPSS Modeler — podręcznik użytkownika (ModelerUsersGuide.pdf).

Należy pamiętać, że istnieje połączenie pomiędzy opcją **Zmienne grupujące** i utworzonymi wyrażeniami agregującymi, ponieważ wyrażenia agregujące są grupowane na podstawie zmiennej kluczowej.

Poprawne wyrażenia agregujące to takie, które służą do oceny wyników agregacji; poniżej przedstawiono kilka przykładów poprawnych wyrażeń agregujących oraz reguły, które nimi sterują:

- Funkcje skalarne umożliwiają połączenie wielu funkcji agregujących w celu utworzenia jednego wyniku agregacji. Na przykład:

  ```
  max(C01) - min(C01)
  ```

- Funkcja agregująca może działać na wynikach wielu funkcji skalarnych. Na przykład:

  ```
  sum (C01*C01)
  ```

Ustawienia optymalizacji agregacji

W węźle optymalizacji można określić następujące wartości.

Zmienne grupujące są posortowane. Należy wybrać tę opcję, jeśli wiadomo, że wszystkie rekordy zawierające takie same wartości kluczowe są w wartościach wejściowych pogrupowane (na przykład, jeśli wartości wejściowe są posortowane według zmiennych kluczowych). Dzięki temu można zwiększyć wydajność.

Zezwalaj na przybliżanie mediany i kwartyli. Statystyki porządkowe (mediana, 1. kwartyl i 3. kwartyl) obecnie nie są obsługiwane podczas przetwarzania danych w programie Analytic Server. Jeśli używany jest program Analytic Server, można zaznaczyć to pole wyboru, aby używać przybliżonych wartości dla tych statystyk zamiast wartości obliczonych poprzez kategoryzację danych i obliczyć dla nich oszacowanie na podstawie rozkładu w przedziałach. Domyślnie ta opcja nie jest zaznaczona.

Liczba przedziałów. Ta opcja jest dostępna wyłącznie po zaznaczeniu pola wyboru **Zezwalaj na przybliżanie mediany i kwartyli**. Należy wybrać liczbę przedziałów, jaka będzie użyta podczas oszacowania statystyki; liczba przedziałów wpływa na wartość **Błąd maksymalny %**. Domyślnie liczba przedziałów jest ustawiona na 1000, co odpowiada maksymalnemu błędowi wynoszącemu 0,1 procent zakresu.

Węzeł Agregacja RFM

Węzeł agregacji RFM (Recency — Aktualność, Frequency — Częstość, Monetary — Kwota) umożliwia dostęp do historycznych danych transakcyjnych klienta, usunięcie danych nieużywanych i połączenie wszystkich pozostałych danych transakcyjnych w jednym wierszu, używając ich unikatowego identyfikatora jako klucza, który informuje, kiedy ostatnio dokonywana była transakcja (aktualność), ile transakcji dokonano (częstość) oraz jaka jest łączna kwota tych transakcji (kwota).

Przed wykonaniem agregacji należy poświęcić nieco czasu na wyczyszczenie danych, koncentrując się szczególnie na brakach danych.

Po zidentyfikowaniu i przekształceniu danych za pośrednictwem węzła agregacji RFM można użyć węzła analizy RFM w celu przeprowadzenia dalszej analizy. Więcej informacji można znaleźć w temacie “Węzeł Analiza RFM” na stronie 173.

Należy pamiętać, że po uruchomieniu danych w węźle agregacji RFM nie będą one zawierały żądnych wartości przewidywanych; dlatego zanim możliwe będzie użycie tych danych jako wartości wejściowych na potrzeby dalszej

analizy predykcyjnej z użyciem węzłów modelowania, takich jak C5.0 lub CHAID, konieczne będzie połączenie ich z innymi danymi klienta (np. poprzez dopasowanie identyfikatorów klienta). Więcej informacji można znaleźć w temacie “Węzeł Łączenie” na stronie 85.

Węzły Agregacja RFM i Analiza RFM w programie IBM SPSS Modeler są skonfigurowane, tak aby korzystały z niezależnej kategoryzacji; oznacza to, że rangowanie i kategoryzacja danych są przeprowadzane dla każdej miary wartości aktualności, częstości i kwoty, bez względu na ich wartości lub pozostałe dwie miary.

Ustawianie opcji dla węzła Agregacja RFM

Karta Ustawienia węzła agregacji RFM zawiera następujące pola.

Oblicz okres od Należy określić datę, od której obliczana będzie aktualność transakcji. Może to być **Ustalona data** wprowadzana przez użytkownika lub **Dzisiejsza data** ustawiana przez system. Wartość **Dzisiejsza data** jest wprowadzana domyślnie i jest automatycznie aktualizowana w czasie wykonywania węzła.

Uwaga: Wyświetlana wartość **Ustalona data** może różnić się dla różnych lokalizacji. Przykładowo, jeśli wartość 2007-8-10 jest zapisana w strumieniu jako Fri Aug 10 00:00:00 CST 2007, oznacza to godzinę i datę w strefie czasowej 'UTC+8'. Jednak w strefie czasowej 'UTC-8' będzie wyświetlana jako Thu Aug 9 12:00:00 EDT 2007.

Wartości identyfikatorów są posortowane Jeśli dane zostały wstępnie posortowane tak, że wszystkie rekordy o tym samym identyfikatorze są wyświetlane razem w strumieniu danych, należy wybrać tę opcję w celu przyspieszenia przetwarzania. Jeśli dane nie zostały wstępnie posortowane (lub nie ma co do tego pewności), należy pozostawić tę opcję niezaznaczoną. Węzeł posortuje dane automatycznie.

Identyfikator To pole należy zaznaczyć, aby było używane podczas identyfikacji klientów i ich transakcji. Aby wyświetlić zmienne, jakie można wybrać, należy użyć przycisku Selektor zmiennych, który znajduje się po prawej stronie.

Data Należy wybrać zmienną daty, aby na jej podstawie obliczana była aktualność. Aby wyświetlić zmienne, jakie można wybrać, należy użyć przycisku Selektor zmiennych, który znajduje się po prawej stronie.

Należy pamiętać, że w celu użycia jako wartości wejściowej wymagana jest zmienna z typem składowania data lub znacznik czasu w odpowiednim formacie. Na przykład, jeśli dostępna jest zmienna łańcuchowa z wartościami *Jan 2007* (Sty 2007), *Feb 2007* (Lut 2007) itd., można ją przekształcić na zmienną daty, korzystając z węzła wypełniania i funkcji to_date(). Więcej informacji można znaleźć w temacie “Przekształcanie sposobu składowania za pomocą węzła Wypełnianie” na stronie 162.

Wartość Należy wybrać zmienną, jaka będzie używana do obliczenia całkowitej wartości kwoty dla transakcji klienta. Aby wyświetlić zmienne, jakie można wybrać, należy użyć przycisku Selektor zmiennych, który znajduje się po prawej stronie. *Uwaga*: Musi to być wartość liczbowa.

Nowe rozszerzenie nazwy zmiennej Należy wybrać tę opcję, aby dołączyć przyrostek lub przedrostek, taki jak „12_month” (12_mies.), dla nowo wygenerowanych zmiennych aktualności, częstości i kwoty. Aby wskazać preferowany styl rozszerzenia, należy wybrać opcję **Przyrostek** lub **Przedrostek**. Może to być na przykład przydatne podczas badania kilku okresów.

Usuń rekordy z wartością poniżej Jeśli jest to wymagane, można określić wartość minimalną; wówczas szczegóły transakcji poniżej tej kwoty nie będą uwzględniane podczas obliczania sum RFM. Jednostki tej wartości odnoszą się do wybranej zmiennej **Wartość**.

Uwzględnij tylko najnowsze transakcje W przypadku przeprowadzania analizy dużej bazy danych można wybrać, aby używane były tylko najnowsze rekordy. Można wybrać, aby używane były tylko dane zapisane po konkretnej dacie lub w ostatnim okresie:

- **Data transakcji po** Należy określić datę transakcji, po której rekordy będą uwzględniane w analizie.

- **Transakcje w ostatnich** Należy określić liczbę i typ okresów (dni, tygodnie, miesiące lub lata) wstecz od daty **Oblicz okres od**, dla których rekordy będą uwzględniane w analizie.

Zapisz dane drugiej najnowszej transakcji Aby poznać datę drugiej najnowszej transakcji dla każdego klienta, należy zaznaczyć to pole. Ponadto można również zaznaczyć pole **Zapisz dane trzeciej najnowszej transakcji**. Może to na przykład ułatwić zidentyfikowanie klientów, którzy dokonywali wielu transakcji jakiś czas temu, a ostatnio dokonali tylko jednej.

Węzeł Sortowanie

Węzły Sortowanie umożliwiają sortowanie rekordów rosnąco lub malejąco na podstawie wartości co najmniej jednej zmiennej. Przykładowo węzły Sortowanie są często stosowane do wyświetlania i wybierania rekordów z najbardziej powszechnymi wartościami danych. Zwykle dane najpierw agregowane są za pomocą węzła Agregacja, a następnie za pomocą węzła Sortowanie zagregowane dane są sortowane wg liczebności rekordów w kolejności malejącej. Wyświetlenie tych wyników w tabeli umożliwi eksplorację danych i podejmowanie decyzji, takich jak wybór rekordów dla 10 najlepszych klientów.

Karta Ustawienia węzła sortowania zawiera następujące pola.

Sortuj według. W tabeli wyświetlane są wszystkie zmienne wybrane jako klucze sortowania. Zmienna kluczowa działa najlepiej podczas sortowania, jeśli jest wartością liczbową.

- **Dodawanie zmiennych** — należy użyć przycisku Selektor zmiennych, który znajduje się po prawej stronie.
- **Wybór porządku** — należy kliknąć strzałkę **Rosnąco** lub **Malejąco** w kolumnie *Porządek* w tabeli.
- **Usuwanie zmiennych** — w tym celu należy użyć czerwonego przycisku usuwania.
- **Sortowanie dyrektyw** — jest możliwe za pomocą strzałek w górę i w dół.

Domyślny porządek sortowania. Można wybrać **Rosnąco** lub **Malejąco**, określając w tym sposób domyślny porządek sortowania podczas dodawania nowych zmiennych.

Uwaga: Węzeł Sortowanie nie ma zastosowania, jeśli poniżej strumienia modelu występuje węzeł Powtórzenia. Informacje na temat węzła Powtórzenia zawiera temat “Węzeł Powtórzenia” na stronie 93.

Ustawienia optymalizacji sortowania

W przypadku pracy z danymi, które zostały już posortowane według określonych zmiennych kluczowych, można określić, które zmienne są już posortowane, zezwalając systemowi na bardziej efektywne posortowanie pozostałych danych. Na przykład dane mają być posortowane według zmiennych *Age* (Wiek) (malejąco) i *Drug* (Lek) (rosnąco), ale wiadomo, że zostały już posortowane według zmiennej *Age* (malejąco).

Dane są wstępnie posortowane. Określa, że dane są już posortowane według co najmniej jednej zmiennej.

Określ porządek istniejącego sortowania. Należy określić zmienne, które zostały już posortowane. Korzystając z okna dialogowego Wybierz zmienne, można dodać zmienne do listy. W kolumnie *Porządek* należy określić, czy dana zmienna jest posortowana w porządku rosnącym czy malejącym. Jeśli określanych jest wiele zmiennych, należy upewnić się, czy wyświetlane są we właściwym porządku. Strzałki po prawej stronie listy pozwalają ustawić zmienne w poprawnym porządku. Jeśli istniejący porządek sortowania zostanie błędnie określony, po uruchomieniu strumienia wyświetlony zostanie błąd, wskazujący numer rekordu, w którym sortowanie jest niezgodne z dokonanym wyborem.

Uwaga: Szybkość sortowania może zostać zwiększona przez aktywowanie przetwarzania równoległego.

Węzeł Łączenie

Funkcją węzła Łączenie jest utworzenie z kilku rekordów wejściowych pojedynczego rekordu wynikowego zawierającego wszystkie lub niektóre zmienne wejściowe. Jest to pomocna operacja, jeśli użytkownik chce połączyć dane z różnych źródeł, na przykład dane wewnętrzne klienta i dane demograficzne osób, które dokonały zakupu. Dane można połączyć w następujący sposób.

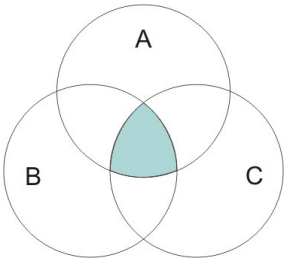
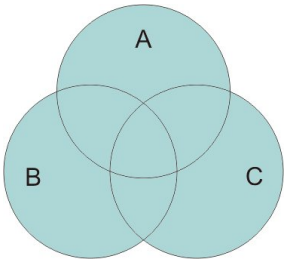
- Łączenie wg **porządku** łączy odpowiednie rekordy ze wszystkich źródeł w kolejności ich wstawiania aż do wyczerpania najmniejszego źródła danych. Istotne jest, aby podczas korzystania z tej opcji dane były posortowane za pomocą węzła Sortowanie.
- Łączenie przy użyciu zmiennej **kluczowej**, takiej jak *Customer ID* (Id. klienta), pozwala określić, w jaki sposób dopasować rekordy z jednego źródła danych z rekordami z innych źródeł. Możliwe jest utworzenie kilku typów złączeń, w tym złączenie wewnętrzne, pełne złączenie zewnętrzne, częściowe złączenie zewnętrzne oraz anty złączenie. Więcej informacji można znaleźć w temacie “Typy złączeń”.
- Łączenie według **warunku** pozwala określić warunek, jaki musi zostać spełniony, aby nastąpiło połączenie. Warunek można określić bezpośrednio w węźle lub utworzyć go za pośrednictwem konstruktora wyrażeń.
- Łączenie według **warunku z rangowaniem** jest lewostronnym złączeniem zewnętrznym, w którym określane są warunek, jaki musi zostać spełniony, aby nastąpiło połączenie, oraz wyrażenie rangujące, które sortuje w kolejności rosnącej. Jest to metoda najczęściej stosowana do łączenia danych geoprzestrzennych; warunek można określić bezpośrednio w węźle lub utworzyć go za pośrednictwem konstruktora wyrażeń.

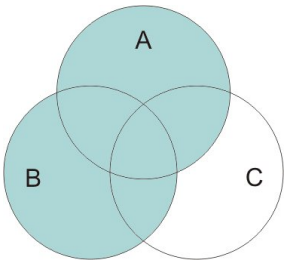
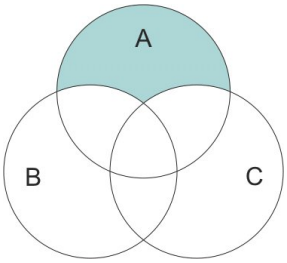
Typy złączeń

Jeśli do łączenia danych używana jest zmienna kluczowa, należy zastanowić się, które rekordy będą uwzględniane, a które zostaną wykluczone. Istnieją różne złączenia, a ich szczegółowe omówienie znajduje się poniżej.

Dwa podstawowe typy złączeń to złączenia wewnętrzne i złączenia zewnętrzne. Te metody są często stosowane do łączenia tabel z powiązanych zbiorów danych na podstawie wspólnych wartości zmiennej kluczowej, takiej jak *Customer ID* (Id. klienta). Złączenia wewnętrzne umożliwiają czyste łączenie, a wynikiem jest zbiór danych, który zawiera tylko kompletne rekordy. Złączenia zewnętrzne również obejmują kompletne rekordy z połączonych danych, ale umożliwiają również uwzględnianie unikatowych danych z co najmniej jednej tabeli wejściowej.

Poniżej zamieszczono dokładniejszy opis możliwych typów złączeń.

	**Złączenie wewnętrzne** obejmuje tylko rekordy, w których wartość zmiennej wejściowej jest taka sama dla wszystkich tabel wejściowych. Oznacza to, że niedopasowane rekordy nie będą uwzględniane w wynikowym zbiorze danych.
	**Pełne złączenie zewnętrzne** obejmuje wszystkie rekordy, dopasowane i niedopasowane, z tabel wejściowych. Lewostronne i prawostronne złączenia zewnętrzne są określane jako częściowe złączenia zewnętrzne i zostały opisane poniżej.

	**Częściowe złączenie zewnętrzne** obejmuje wszystkie rekordy dopasowane przy użyciu zmiennej kluczowej, jak również rekordy niedopasowane z określonych tabel. (Lub, innymi słowy, wszystkie rekordy z określonych tabel i tylko rekordy dopasowane z pozostałych). Tabele (takie jak przedstawione tutaj A i B) można wybrać w celu uwzględnienia w złączeniu zewnętrznym za pomocą przycisku Wybierz na karcie Łączenie. Złączenia częściowe są również nazywane lewostronnymi lub prawostronnymi złączeniami zewnętrznymi, jeśli łączenie przeprowadzane jest tylko pomiędzy dwiema tabelami. Ponieważ program IBM SPSS Modeler umożliwia łączenie więcej niż dwóch tabel, używane jest określenie częściowe złączenie zewnętrzne.
	**Anty złączenie** obejmuje tylko niedopasowane rekordy dla pierwszej tabeli wejściowej (tabela A na rysunku). Tego typu złączenie stanowi przeciwieństwo złączenia wewnętrznego i nie uwzględnia kompletnych rekordów w wynikowym zbiorze danych.

Przykładowo, jeśli w jednym zbiorze danych znajdują się informacje na temat gospodarstw rolnych, a w drugim roszczenia ubezpieczeniowe powiązane z gospodarstwami rolnymi, można dopasować rekordy z pierwszego źródła do rekordów z drugiego źródła, korzystając z opcji łączenia.

Aby ustalić, czy klient z próby obejmującej gospodarstwo rolne zgłosił roszczenie ubezpieczeniowe, należy użyć opcji złączenia wewnętrznego, co spowoduje zwrócenie listy przedstawiającej wszystkie dopasowane identyfikatory z dwóch prób.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalue
1	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0...
2	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0...
3	id620	name620	north	880.000	74.000	6.000	426988.000	rapeseed	arable_de	118885.0...

Rysunek 2. Przykładowy wynik dla metody złączenia wewnętrznego

Użycie opcji pełnego złączenia zewnętrznego spowoduje zwrócenie rekordów dopasowanych i niedopasowanych z tabel wejściowych. Systemowy brak danych ($null$) będzie stosowany w przypadku wszystkich niekompletnych wartości.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop	claimtype	claimvalu
1	id601	$null$	$null$	$null$	$null$	$null$	$null$	$null$	decomm...	74703.1(
2	id602	name602	north	1780.000	42.000	9.000	734118.000	maize	$null$	$nul
3	id604	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes	decomm...	281082.0
4	id605	name605	north	1700.000	46.000	8.000	621148.000	wheat	decomm...	122006.0
5	id606	$null$	$null$	$null$	$null$	$null$	$null$	$null$	arable_de	122135.0

Rysunek 3. Przykładowy wynik dla metody pełnego złączenia zewnętrznego

Częściowe złączenie zewnętrzne obejmuje wszystkie rekordy dopasowane przy użyciu zmiennej kluczowej, jak również rekordy niedopasowane z określonych tabel. W tabeli wyświetlane są wszystkie rekordy dopasowane według zmiennej identyfikacyjnej, jak również rekordy dopasowane z pierwszego zbioru danych.

	id	claimtype	claimvalue	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	$null$	$null$	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id604	decomm...	281082.0...	name604	southwest	1860.000	103.0...	3.000	625251.000	potatoes
3	id605	decomm...	122006.0...	name605	north	1700.000	46.000	8.000	621148.000	wheat
4	id607	$null$	$null$	name607	southeast	1820.000	29.000	6.000	211605.000	maize
5	id608	$null$	$null$	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
6	id609	$null$	$null$	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
7	id615	$null$	$null$	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
8	id618	$null$	$null$	name618	southeast	1180.000	98.000	3.000	368646.000	maize

Rysunek 4. Przykładowy wynik dla metody częściowego złączenia zewnętrznego

Jeśli używana jest opcja anty złączenia, w tabeli zwrócone zostaną tylko niedopasowane rekordy dla pierwszej tabeli wejściowej.

	id	name	region	farmsize	rainfall	landquality	farmincome	maincrop
1	id602	name602	north	1780.000	42.000	9.000	734118.000	maize
2	id607	name607	southeast	1820.000	29.000	6.000	211605.000	maize
3	id608	name608	southeast	1640.000	108.0...	7.000	1167040.0...	maize
4	id609	name609	southwest	1600.000	101.0...	5.000	756755.000	wheat
5	id615	name615	midlands	920.000	86.000	6.000	442554.000	potatoes
6	id618	name618	southeast	1180.000	98.000	3.000	368646.000	maize
7	id619	name619	north	840.000	64.000	8.000	457552.000	potatoes

Rysunek 5. Przykładowy wynik dla metody anty złączenia

Określanie metody łączenia oraz kluczy

Karta Łączenie węzła łączenia zawiera następujące pola.

Metoda łączenia Umożliwia wybranie metody, jaka będzie używana do łączenia rekordów. Wybranie opcji **Klucze** lub **Warunek** aktywuje dolną część okna dialogowego.

- **Kolejność** Łączy rekordy według porządku, np. łączony jest *n*-ty rekord z każdego rekordu wejściowego w celu utworzenia *n*-tego rekordu wynikowego. Jeśli w rekordzie nie będzie już żadnego pasującego rekordu wejściowego, rekordy wynikowe nie zostaną utworzone. Oznacza to, że liczba utworzonych rekordów jest liczbą rekordów w najmniejszym zbiorze danych.
- **Klucze** Używa zmiennej kluczowej, takiej jak *Transaction ID* (Id. transakcji), do połączenia rekordów z taką samą wartością w zmiennej kluczowej. Jest to odpowiednik złączenia równościowego bazy danych. Jeśli wartość kluczowa będzie występowała więcej niż jeden raz, zwrócone zostaną wszystkie możliwe kombinacje. Na przykład, jeśli rekordy zawierające tą samą wartość zmiennej kluczowej *A* zawierają różniące się wartości *B*, *C* i *D* w innych zmiennych, połączone zmienne utworzą osobny rekord dla każdej kombinacji *A* z wartością *B*, *A* z wartością *C* oraz *A* z wartością *D*.

  Uwaga: W przypadku metody łączenia według klucza wartości null nie są uznawane za identyczne i nie zostaną złączone.
- **Warunek** Ta opcja umożliwia określenie warunku łączenia. Aby uzyskać więcej informacji, zobacz “Określanie warunków dla łączenia” na stronie 88.
- **Warunek z rangowaniem** Ta opcja pozwala określić, czy każdy wiersz tworzący parę w podstawowym i dodatkowym zbiorze danych ma zostać połączony; do posortowania wielu dopasowań w kolejności rosnącej należy użyć wyrażenia rangującego. Aby uzyskać więcej informacji, zobacz “Określanie warunków z rangowaniem dla łączenia” na stronie 89.

Dostępne klucze Wyświetlane są tylko te zmienne, których nazwy są dokładnie dopasowane we wszystkich wejściowych źródłach danych. Należy wybrać zmienną z tej listy i za pomocą przycisku strzałki dodać ją jako zmienną kluczową używaną do łączenia rekordów. Możliwe jest użycie więcej niż jednej zmiennej kluczowej. Nazwy niedopasowanych zmiennych wejściowych można zmienić za pomocą węzła filtrowania lub karty Filtrowanie w węźle źródłowym.

Użyte klucze Wyświetlane są wszystkie zmienne użyte do połączenia rekordów ze wszystkich wejściowych źródeł danych na podstawie wartości zmiennych kluczowych. Aby usunąć klucz z listy, należy go wybrać i za pomocą przycisku strzałki przenieść z powrotem na listę Dostępne klucze. Jeśli wybranych zostanie kilka zmiennych kluczowych, aktywowana jest poniższa opcja.

Połącz duplikaty zmiennych kluczowych Jeśli na powyższym etapie wybrano kilka zmiennych kluczowych, ta opcja sprawia, że dostępna jest tylko jedna zmienna wyjściowa o danej nazwie. Ta opcja jest włączona domyślnie, z wyjątkiem przypadków, w których strumienie zostały zaimportowane z wcześniejszych wersji programu IBM SPSS Modeler. Jeśli ta opcja jest wyłączona, zduplikowane zmienne kluczowe wymagają zmiany nazwy lub muszą zostać wykluczone za pośrednictwem karty Filtrowanie w oknie dialogowym węzła łączenia.

Połącz tylko pasujące rekordy (złączenie wewnętrzne) Tę opcję należy zaznaczyć, aby połączyć tylko kompletne rekordy.

Połącz pasujące i niepasujące rekordy (pełne złączenie zewnętrzne) Po wybraniu tej opcji wykonywane jest „pełne złączenie zewnętrzne". Oznacza to, że jeśli wartości zmiennej kluczowej nie znajdują się we wszystkich tabelach wejściowych, niekompletne rekordy będą nadal zachowywane. Wartość niezdefiniowana ($null$) jest dodawana do zmiennej kluczowej i zostaje uwzględniona w rekordzie wynikowym.

Połącz pasujące i wybrane niepasujące rekordy (częściowe złączenie zewnętrzne) Tę opcję należy wybrać, aby wykonać „częściowe złączenie zewnętrzne" tabel wybranych w podoknie dialogowym. Należy kliknąć przycisk **Wybierz**, aby określić tabele, dla których niekompletne rekordy będą zachowywane po połączeniu.

Uwzględnij rekordy z pierwszego źródła niepasujące do innych (anty złączenie) Ta opcja pozwala wykonać „anty złączenie", w którym tylko niedopasowane rekordy z pierwszego zbioru danych są przekazywane do dalszej części strumienia. Można określić kolejność wejściowych zbiorów danych, używając strzałek na karcie danych wejściowych. Ten typ złączenia nie uwzględnia kompletnych rekordów w wynikowym zbiorze danych. Aby uzyskać więcej informacji, zobacz "Typy złączeń" na stronie 85.

Wybór danych dla złączeń częściowych

W przypadku częściowych złączeń zewnętrznych należy wybrać tabele, dla których niekompletne rekordy zostaną zachowane. Przykładowo, użytkownik chce zachować wszystkie rekordy z tabeli Klient, zachowując tylko dopasowane rekordy z tabeli Kredyt hipoteczny.

Kolumna złączenia zewnętrznego. W kolumnie *Outer Join* (Złączenie zewnętrzne) należy wybrać zbiory danych, aby uwzględnić ich całą zawartość. W przypadku złączenia częściowego dla wybranych tutaj zbiorów danych zachowane zostaną nakładające się rekordy, jak również rekordy niekompletne. Więcej informacji można znaleźć w temacie "Typy złączeń" na stronie 85.

Określanie warunków dla łączenia

Ustawienie metody łączenia na **Warunek** pozwala określić co najmniej jeden warunek, jaki musi zostać spełniony, aby nastąpiło połączenie.

Warunki można wprowadzić bezpośrednio w zmiennej warunku lub można je utworzyć za pośrednictwem konstruktora wyrażeń, klikając symbol kalkulatora po prawej stronie zmiennej.

Dodaj znaczniki do zduplikowanych nazw zmiennych, aby uniknąć konfliktów przy łączeniu Jeśli co najmniej dwa zbiory danych do połączenia zawierają te same nazwy zmiennych, należy zaznaczyć to pole wyboru w celu dodania różnych znaczników przedrostka na początku nagłówków kolumn zmiennych. Przykładowo, jeśli występują dwie zmienne o nazwie *Name* (Nazwa), w wynikach łączenia znajdą się zmienne *1_Name* (1_Nazwa) i *2_Name* (2_Nazwa). Jeśli nazwa znacznika została zmieniona w źródle danych, zamiast liczbowego znacznika prefiksu użyta zostanie nowa nazwa. Jeśli to pole wyboru nie zostanie zaznaczone, a w danych będą występować zduplikowane nazwy, po prawej stronie zmiennej wyświetlone zostanie ostrzeżenie.

Określanie warunków z rangowaniem dla łączenia

Połączenie warunkowe z rangowaniem można traktować jako lewostronne złączenie zewnętrzne według warunku; po lewej stronie łączenia znajduje się główny zbiór danych, w którym każdy rekord jest zdarzeniem. Na przykład, w modelu używanym do znajdowania wzorców w danych dotyczących przestępstwa każdy rekord w głównym zbiorze danych będzie zawierał informacje o przestępstwie oraz informacje z nim powiązane (lokalizacja, rodzaj itd.). W tym przykładzie po prawej stronie mogą znajdować się odpowiednie zbiory danych geoprzestrzennych.

Podczas łączenia zastosowane zostaną warunek łączenia oraz wyrażenie rangujące. Warunek łączenia może korzystać z funkcji geoprzestrzennej, takiej jak *within* (zawiera) lub *close_to* (zbliżone do). Podczas łączenia wszystkie zmienne w zbiorach danych po prawej stronie są dodawane do zbioru danych po lewej stronie; w zmiennej listy znajdą się wielokrotne dopasowania. Na przykład:

- Lewa strona: dane dotyczące przestępstwa
- Prawa strona: zbiór danych dot. regionów i zbiór danych dot. dróg
- Warunki łączenia: dane dotyczące przestępstwa *within* (zawierające) regiony i *close_to* (zbliżone do) dróg wraz z definicją określającą, co oznacza *close_to*.

W tym przykładzie, jeśli przestępstwo miało miejsce w wymaganej odległości, określonej jako *close_to*, od trzech dróg (a liczba dopasowań, jakie mają zostać zwrócone, została ustawiona na minimum trzy), wszystkie trzy drogi zostaną zwrócone jako elementy listy.

Ustawienie metody łączenia na **Warunek z rangowaniem** pozwala określić co najmniej jeden warunek, jaki musi zostać spełniony, aby nastąpiło połączenie.

Główny zbiór danych Należy wybrać główny zbiór danych dla łączenia; zmienne z pozostałych zbiorów danych zostaną dodane do wybranego zbioru danych. Można to uznać za lewą stronę złączenia zewnętrznego.

Po wybraniu głównego zbioru danych wszystkie pozostałe wejściowe zbiory danych połączone z węzłem łączenia są automatycznie wyświetlane w tabeli **Połączenia**.

Dodaj znaczniki do zduplikowanych nazw zmiennych, aby uniknąć konfliktów przy łączeniu Jeśli co najmniej dwa zbiory danych do połączenia zawierają te same nazwy zmiennych, należy zaznaczyć to pole wyboru w celu dodania różnych znaczników przedrostka na początku nagłówków kolumn zmiennych. Przykładowo, jeśli występują dwie zmienne o nazwie *Name* (Nazwa), w wynikach łączenia znajdą się zmienne *1_Name* (1_Nazwa) i *2_Name* (2_Nazwa). Jeśli nazwa znacznika została zmieniona w źródle danych, zamiast liczbowego znacznika prefiksu użyta zostanie nowa nazwa. Jeśli to pole wyboru nie zostanie zaznaczone, a w danych będą występować zduplikowane nazwy, po prawej stronie zmiennej wyświetlone zostanie ostrzeżenie.

Połączenia

Zbiór danych

Przedstawia nazwę dodatkowych zbiorów danych połączonych z węzłem łączenia jako źródło danych wejściowych. Domyślnie istnieje więcej niż jeden zbiór danych, wyświetlane są one w kolejności, w jakiej zostały połączone z węzłem Łączenie.

Warunek łączenia

Należy wprowadzić unikalne warunki łączenia poszczególnych zbiorów danych w tabeli z głównym zbiorem danych. Warunki można wpisać bezpośrednio w komórce lub można je utworzyć za pośrednictwem konstruktora wyrażeń, klikając symbol kalkulatora po prawej stronie komórki. Przykładowo można użyć predykatów geoprzestrzennych do utworzenia warunku łączenia, który będzie umieszczał dane dotyczące przestępstwa z jednego zbioru danych w danych dotyczących regionu z innego zbioru danych. Domyślny warunek łączenia zależy od geoprzestrzennego poziomu pomiaru, co przedstawiono na liście poniżej.

- Punkt, Łańcuch, Multipunkt, Multiłańcuch — warunek domyślny dla *close_to*.
- Wielokąt, Multiwielokąt — warunek domyślny dla *within*.

Więcej informacji na temat tych poziomów zawiera temat “Geoprzestrzenne podpoziomy pomiarów” na stronie 141.

Jeśli zbiór danych zawiera wiele zmiennych geoprzestrzennych różnego typu, domyślny warunek, jaki zostanie zastosowany, zależy od pierwszego poziomu pomiaru, jaki zostanie znaleziony w danych, przy ustawieniu w następującym porządku malejącym.

- Punkt
- Łańcuch
- Wielokąt

Uwaga: Ustawienia domyślne są dostępne wyłącznie wówczas, kiedy w dodatkowej bazie danych znajdują się zmienne danych geoprzestrzennych.

Wyrażenie rangujące

Należy określić wyrażenie rangowania łączonych zbiorów danych; to wyrażenie służy do sortowania wielu dopasowań w kolejności ustalonej na podstawie kryterium rangowania. Warunki można wpisać bezpośrednio w komórce lub można je utworzyć za pośrednictwem konstruktora wyrażeń, klikając symbol kalkulatora po prawej stronie komórki.

Domyślnie za pośrednictwem konstruktora wyrażeń wprowadzane są wyrażenia rangowania odległości i powierzchni; w obu przypadkach rangowanie jest przeprowadzane rosnąco, co oznacza, przykładowo, że najwyższym dopasowaniem dla odległości będzie najmniejsza wartość. Przykładowe rangowanie według odległości: główny zbiór danych zawiera dane dotyczące przestępstwa i powiązaną z nim lokalizację, a drugi zbiór danych zawiera obiekty z lokalizacjami; w takim przypadku odległość pomiędzy miejscami przestępstwa a obiektami mogą stanowić kryterium rangowania. Domyślne wyrażenie rangujące zależy od geoprzestrzennego poziomu pomiaru, co przedstawiono na liście poniżej.

- Punkt, Łańcuch, Multipunkt, Multiłańcuch — wyrażenie domyślne to *distance* (odległość).
- Wielokąt, Multiwielokąt — wyrażenie domyślne to *area* (obszar).

Uwaga: Ustawienia domyślne są dostępne wyłącznie wówczas, kiedy w dodatkowej bazie danych znajdują się zmienne danych geoprzestrzennych.

Liczba dopasowań

Należy określić liczbę dopasowań, jakie są zwracane, na podstawie warunku i wyrażeń rangowania. Domyślna liczba dopasowań zależy od geoprzestrzennego poziomu pomiaru w dodatkowym zbiorze danych, co przedstawiono na liście poniżej; można jednak dwukrotnie kliknąć komórkę, aby wprowadzić własną wartość, wynoszącą maksymalnie 100.

- Punkt, Łańcuch, Multipunkt, Multiłańcuch — wartość domyślna to 3.
- Wielokąt, Multiwielokąt — wartość domyślna to 1.
- Zbiór danych nie zawiera zmiennych geoprzestrzennych — wartość domyślna to 1.

Przykładowo, jeśli skonfigurowane zostanie łączenie na podstawie **warunku łączenia** *close_to* oraz **wyrażenia rangującego** *distance*, trzy pierwsze (najbliższe) dopasowania z dodatkowych zbiorów danych dla każdego rekordu w głównym zbiorze danych zostaną zwrócone jako wartości w wynikowej zmiennej listy.

Filtrowanie zmiennych za pośrednictwem węzła Łączenie

Węzły Łączenie zapewniają wygodny sposób filtrowania zduplikowanych zmiennych w wyniku połączenia kilku źródeł danych. Aby wybrać opcje filtrowania, należy kliknąć zakładkę **Filtrowanie** w oknie dialogowym.

Dostępne tutaj opcje są niemal identyczne jak w węźle Filtrowanie. W menu filtrowania są jednak dostępnie dodatkowe opcje, które nie zostały tutaj omówione. Więcej informacji można znaleźć w temacie “Filtrowanie lub zmiana nazw zmiennych” na stronie 151.

Zmienna. Wyświetla zmienne wejściowe z aktualnie połączonych źródeł danych.

Znacznik. Wyświetla listę nazw znaczników (lub numerów) powiązanych z łączem źródła danych. Należy kliknąć zakładkę **Wejścia**, aby zmienić aktywne łącza dla tego węzła łączenia.

Węzeł źródłowy. Wyświetla węzeł źródłowy, którego dane będą łączone.

Węzeł podłączony. Wyświetla nazwę węzła, który jest połączony z węzłem łączenia. Często eksploracja danych złożonych wymaga kilku operacji łączenia lub dołączania, które mogą obejmować ten sam węzeł źródłowy. Nazwa połączonego węzła umożliwia ich rozróżnianie.

Filtrowanie. Wyświetla bieżące połączenia pomiędzy zmienną wejściową i wyjściową. Dla połączeń aktywnych wyświetlana jest niezłamana strzała. Połączenia oznaczone czerwonym znakiem X oznaczają zmienne z zastosowanym filtrem.

Zmienna. Wyświetla listę zmiennych wyjściowych po połączeniu lub dołączeniu. Zmienne zduplikowane są wyświetlane w kolorze czerwonym. Należy kliknąć zmienną filtrowania powyżej, aby wyłączyć duplikowanie zmiennych.

Widok aktualnych zmiennych. Tę opcję należy wybrać, aby wyświetlić informacje o zmiennych wybranych jako zmienne kluczowe.

Widok ustawień niewykorzystanych zmiennych. Tę opcję należy wybrać, aby wyświetlać informacje o zmiennych, które obecnie nie są używane.

Ustawianie porządku danych wejściowych i dodawanie znaczników

Korzystając z karty danych wejściowych w oknach dialogowych węzłów Łączenie i Dołączanie, można określić porządek wejściowych źródeł danych i wprowadzić zmiany w nazwach znaczników dla każdego źródła.

Znaczniki i porządek wprowadzania zbiorów danych. Tę opcję należy wybrać, aby połączyć lub dołączyć tylko kompletne rekordy.

- **Znacznik.** Wyświetla listę bieżących nazw znaczników dla każdego wejściowego źródła danych. Nazwy znaczników lub **znaczniki** umożliwiają jednoznaczną identyfikację łączy danych wybranych dla operacji łączenia lub dołączania. Przykładowo, wyobraźmy sobie wodę z różnych rurociągów, które w jednym punkcie łączą się, w wyniku czego woda płynie w jednym rurociągu. Przepływ danych w programie IBM SPSS Modeler jest podobny, a punktem łączącym jest często złożona interakcja pomiędzy różnymi źródłami danych. Znaczniki umożliwiają zarządzanie danymi wejściowymi („rurociągami”) w węźle Łączenie lub Dołączanie, dzięki czemu w przypadku zapisania lub odłączenia węzła łącza pozostają i można je łatwo zidentyfikować.

Po połączeniu dodatkowych źródeł danych do węzła Łączenie lub Dołączanie automatycznie tworzone są znaczniki domyślne, w których cyfry reprezentują kolejność, w jakiej węzły zostały połączone. Ten porządek nie ma związku z porządkiem zmiennych w wejściowych lub wyjściowych zbiorach danych. Znacznik domyślny można zmienić, wprowadzając nową nazwę w kolumnie *Znacznik*.

- **Węzeł źródłowy.** Wyświetla węzeł źródłowy, którego dane będą połączone.
- **Węzeł podłączony.** Wyświetla nazwę węzła, który jest połączony z węzłem łączenia lub dołączania. Często eksploracja danych złożonych wymaga kilku operacji łączenia, które mogą obejmować ten sam węzeł źródłowy. Nazwa połączonego węzła umożliwia ich rozróżnianie.
- **Zmienne.** Wyświetla liczbę zmiennych w każdym źródle danych.

Widok aktualnych znaczników. Po wybraniu tej opcji wyświetlane są aktywne znaczniki, które są używane w węźle Łączenie lub Dołączanie. Innymi słowy, bieżące znaczniki identyfikują łącza do węzła, przez który przepływają dane. Posługując się metaforą rurociągu, bieżące znaczniki są analogią dla rurociągów, w których przepływa woda.

Widok ustawień niewykorzystanych znaczników. Tę opcję należy wybrać, aby wyświetlić znaczniki lub łącza, które wcześniej były użyte do połączenia z węzłem Łączenie lub Dołączanie, ale nie są obecnie połączone ze źródłem

danych. Analogią są puste rurociągi, które nadal pozostają w instalacji wodociągowej. „Rurociągi” te można połączyć z nowym źródłem lub je usunąć. Aby usunąć nieużywane znaczniki z węzła, należy kliknąć opcję **Wyczyść**. Spowoduje to usunięcie nieużywanych znaczników.

Ustawienia optymalizacji łączenia

System udostępnia dwie opcje, które mogą pomóc zwiększyć wydajność łączenia danych w niektórych sytuacjach. Opcje te umożliwiają zoptymalizowanie łączenia, kiedy jeden wejściowy zestaw danych jest znacznie większy od pozostałych zbiorów danych lub kiedy dane zostały już posortowane według wszystkich lub niektórych zmiennych kluczowych, jakie są używane do łączenia.

Uwaga: Optymalizacje wprowadzone za pomocą tej karty mają zastosowanie tylko do wykonywania węzła rodzimego IBM SPSS Modeler; czyli w sytuacji, kiedy węzeł łączenia nie zostaje przekazany do bazy danych SQL. Ustawienia optymalizacji nie wpływają na generowanie kodu SQL.

Jeden z wejściowych zbiorów danych jest relatywnie duży. Tę opcję należy wybrać, aby wskazać, że jeden z wejściowych zbiorów danych jest znacznie większy od pozostałych. System zapisze mniejsze zbiory danych w pamięci podręcznej, a następnie dokona połączenia poprzez przetworzenie dużego zbioru danych bez zapisywania go w pamięci podręcznej lub sortowania. Tego typu złączenie będzie najczęściej używane w przypadku danych korzystających ze schematu gwiazdy lub podobnej konstrukcji, w których znajduje się duża tabela centralna zawierająca współużytkowane dane (np. dane transakcyjne). Jeśli ta opcja zostanie zaznaczona, należy kliknąć przycisk **Wybierz**, aby określić duży zbiór danych. Należy pamiętać, że można wybrać tylko *jeden* duży zbiór danych. Poniższa tabela zawiera podsumowanie dotyczące złączeń, które można zoptymalizować za pomocą tej metody.

Tabela 16. Podsumowanie optymalizacji złączeń

Typ złączenia	Czy możliwa jest optymalizacja dla dużego zbioru danych?
Wewnętrzne	Tak
Częściowe	Tak, jeśli w dużym zbiorze danych nie ma żadnych niekompletnych rekordów.
Pełne	Nie
Anty złączenie	Tak, jeśli duży zbiór danych stanowi pierwsze źródło danych wejściowych.

Wprowadzane dane zostały posortowane według zmiennych kluczowych. Tę opcję należy wybrać, aby wskazać, że dane wejściowe są już posortowane według co najmniej jednej zmiennej kluczowej, jaka jest używana do łączenia. Należy upewnić się, czy *wszystkie* wejściowe zbiory danych zostały posortowane.

Określ porządek istniejącego sortowania. Należy określić zmienne, które zostały już posortowane. Korzystając z okna dialogowego Wybierz zmienne, można dodać zmienne do listy. Wybrać można tylko te zmienne kluczowe, które są używane do łączenia (określone na karcie Łączenie). W kolumnie *Porządek* należy określić, czy dana zmienna jest posortowana w porządku rosnącym czy malejącym. Jeśli określanych jest wiele zmiennych, należy upewnić się, czy wyświetlane są we właściwym porządku. Strzałki po prawej stronie listy pozwalają ustawić zmienne w poprawnym porządku. Jeśli istniejący porządek sortowania zostanie błędnie określony, po uruchomieniu strumienia wyświetlony zostanie błąd, wskazujący numer rekordu, w którym sortowanie jest niezgodne z dokonanym wyborem.

W zależności od tego, czy w metodzie scalania używanej przez bazę danych rozróżniana jest wielkość liter, optymalizacja może nie działać poprawnie, jeśli co najmniej jedna zmienna jest sortowana przez bazę danych. Na przykład, jeśli dostępne są dwie zmienne i tylko w jednej z nich wielkość liter jest rozróżniana, wyniki sortowania mogą się różnić. Optymalizacja łączenia sprawia, że rekordy są przetwarzane według ich porządku sortowania. W wyniku tego, jeśli dane wejściowe są posortowane przy użyciu różnych metod scalania, węzeł Łączenie zgłosi błąd i wyświetli numer rekordu, w którym sortowanie jest niezgodne. Jeśli wszystkie dane wejściowe pochodzą z jednego źródła lub jeśli są posortowane z zastosowaniem metod scalania, które wzajemnie się przenikają, rekordy mogą zostać pomyślnie połączone.

Uwaga: Szybkość łączenia może zostać zwiększona poprzez aktywowanie przetwarzania równoległego.

Węzeł Dołączanie

Węzły Dołączanie umożliwiają łączenie zestawów rekordów. W odróżnieniu od węzłów Łączenie, które umożliwiają łączenie rekordów z różnych źródeł, węzły Dołączanie odczytują i przekazują do dalszej części strumienia wszystkie rekordy z jednego źródła, aż do ich wyczerpania. Następnie odczytywane są rekordy z kolejnego źródła, z zastosowaniem tej samej struktury danych (liczba rekordów, liczba zmiennych itd.), stanowiące pierwsze lub główne dane wejściowe. Jeśli w główne źródło zawiera więcej zmiennych niż kolejne źródło danych wejściowych, wówczas dla wszystkich niekompletnych wartości zastosowany zostanie systemowy łańcuch null ($null$).

Węzły Dołączanie są przydatne do łączenia zbiorów danych z podobnymi strukturami zawierającymi inne dane. Na przykład dostępne są dane transakcyjne zapisane w różnych plikach dla różnych okresów, np. w pliku danych sprzedaży z marca i w osobnym dla kwietnia. Jeśli mają taką samą strukturę (te same zmienne ustawione w takim samym porządku), węzeł Dołączanie połączy je w jeden duży plik, który można następnie poddać analizie.

Uwaga: Aby dołączyć pliki, poziomy pomiaru zmiennych muszą być podobne. Na przykład zmiennej *nominalnej* nie można dołączyć do zmiennej, której poziom pomiaru jest *ilościowy*.

Ustawianie opcji dołączania

Dopasuj zmienne według. Należy wybrać metodę, jaka będzie stosowana do dopasowania zmiennych, jakie mają zostać dołączone.

- **Położenie.** Tę opcję należy wybrać, aby dołączać zbiory danych na podstawie położenia zmiennych w głównym zbiorze danych. W przypadku zastosowania tej metody dane powinny być posortowane, zapewniając prawidłowe dołączenie.
- **Nazwa.** Tę opcję należy wybrać, aby dołączać zbiory danych na podstawie nazwy zmiennych w wejściowych zbiorach danych. Należy również wybrać opcję **Uwzględnij wielkość liter**, aby podczas dopasowywania nazw zmiennych uwzględniana była wielkość liter.

Zmienna wyjściowa. Wyświetla listę węzłów źródłowych, jakie są połączone z węzłem Dołączanie. Pierwszy węzeł na liście jest głównym źródłem wejściowym. Wyświetlane zmienne można posortować, klikając nagłówek kolumny. Takie sortowanie w rzeczywistości nie zmienia porządku zmiennych w zbiorze danych.

Dołącz zmienne z. Wybranie opcji **Tylko główny zbiór danych** spowoduje utworzenie zmiennych wyjściowych na podstawie zmiennych z głównego zbioru danych. Główny zbiór danych stanowi pierwsze źródło danych, określone na karcie danych wejściowych. Po wybraniu opcji **Wszystkie zbiory danych** tworzone są zmienne wyjściowe dla wszystkich zmiennych we wszystkich zbiorach danych, niezależnie od tego, czy we wszystkich wejściowych zbiorach danych znajduje się pasująca zmienna.

Zapisz informację według źródłowego zbioru danych w zmiennej. Tę opcję należy wybrać, aby dodać dodatkową zmienną do pliku wynikowego, którego wartości wskazują źródłowy zbiór danych dla każdego rekordu. Nazwę należy określić w polu tekstowym. Domyślna nazwa zmiennej to *Input* (Dane wejściowe).

Węzeł Powtórzenia

Przed rozpoczęciem eksploracji danych konieczne jest usunięcie duplikowanych rekordów ze zbioru danych. Na przykład w bazie danych marketingowych osoby mogą występować klika razy dla różnych adresów lub danych firmy. Węzeł Powtórzenia umożliwia znalezienie lub usunięcie zduplikowanych rekordów z danych lub utworzenie pojedynczego, złożonego rekordu na podstawie grupy zduplikowanych rekordów.

Aby użyć węzła Powtórzenia, najpierw należy zdefiniować zestaw zmiennych kluczowych, który określi, kiedy dwa rekordy zostaną uznane za zduplikowane.

Jeśli nie wszystkie zmienne zostaną wybrane jako zmienne kluczowe, wówczas dwa „zduplikowane” rekordy mogą nie być rzeczywiście identyczne, ponieważ wartości pozostałych zmiennych nadal mogą się różnić. W takim przypadku można również zdefiniować porządek sortowania, jaki zostanie zastosowany w każdej grupie zduplikowanych rekordów. Porządek sortowania zapewnia dokładną kontrolę nad tym, który rekord będzie traktowany jako pierwszy w

grupie. W przeciwnym razie wszystkie duplikaty będą traktowane jako zamienne i wybrany może zostać dowolny rekord. Porządek wejściowy rekordów nie jest brany pod uwagę, dlatego nie pomaga w użyciu wcześniejszego węzła Sortowanie (patrz temat „Sortowanie rekordów w węźle Powtórzenia" poniżej).

Dominanta. Należy określić, czy ma zostać utworzony rekord złożony, czy też konieczne jest uwzględnienie lub wykluczenie (odrzucenie) pierwszego rekordu.

- **Utwórz rekord złożony dla każdej grupy.** Zapewnia sposób zagregowania zmiennych nienumerycznych. Po zaznaczeniu tej opcji udostępniana jest karta Złożone, na której można określić sposób, w jaki tworzone będą rekordy złożone. Więcej informacji można znaleźć w "Ustawienia złożonego rekordu w węźle Powtórzenia" na stronie 96.
- **W każdej z grup dołącz tylko pierwszy rekord.** Powoduje wybranie pierwszego rekordu z każdej grupy zduplikowanych rekordów i odrzucenie pozostałych. *Pierwszy* rekord jest ustalany na podstawie porządku sortowania zdefiniowanego poniżej, a nie na podstawie wejściowego porządku rekordów.
- **W każdej z grup odrzuć tylko pierwszy rekord.** Powoduje odrzucenie pierwszego rekordu z każdej grupy zduplikowanych rekordów i wybranie pozostałych. *Pierwszy* rekord jest ustalany na podstawie porządku sortowania zdefiniowanego poniżej, a nie na podstawie wejściowego porządku rekordów. Ta opcja jest przydatna do *wyszukiwania* duplikatów w danych, tak aby możliwe było ich zbadanie w dalszej części strumienia.

Pola kluczowe dla grupowania. Wyświetla listę zmiennych użytych do określenia, czy rekordy są identyczne. Można:

- Dodawać zmienne za pomocą przycisku wybierania zmiennych, który znajduje się po prawej stronie.
- Usuwać zmienne z listy za pomocą czerwonego przycisku X (Usuń).

W obrębie grup uporządkuj rekordy ze względu na. Wyświetla listę zmiennych używanych do określenia sposobu sortowania rekordów w każdej grupie duplikatów oraz porządku sortowania (malejąco lub rosnąco). Można:

- Dodawać zmienne za pomocą przycisku wybierania zmiennych, który znajduje się po prawej stronie.
- Usuwać zmienne z listy za pomocą czerwonego przycisku X (Usuń).
- Przenosić zmienne za pomocą przycisków strzałki w górę lub w dół, o ile sortowanie jest przeprowadzane na podstawie więcej niż jednej zmiennej.

Jeśli wybrano opcję uwzględniania pierwszego rekordu w grupie lub wykluczania go z grupy i jeśli ma znaczenie to, który rekord jest traktowany jak pierwszy, należy określić porządek sortowania.

Porządek sortowania można również określić, jeśli wybrano opcję utworzenia złożonego rekordu (niektóre opcje na karcie Złożony). Więcej informacji można znaleźć w "Ustawienia złożonego rekordu w węźle Powtórzenia" na stronie 96.

Domyślny porządek sortowania. Należy określić, czy rekordy (domyślnie) będą sortowane w porządku **rosnącym** czy **malejącym** ich wartości kluczowych sortowania.

Sortowanie rekordów w węźle Powtórzenia

Jeśli porządek rekordów w grupie duplikatów jest istotny, należy go określić, używając opcji **W obrębie grup uporządkuj rekordy ze względu na** w węźle Powtórzenia. Nie należy polegać na ustawieniach we wcześniejszym węźle Sortowanie. Należy pamiętać, że wejściowy porządek rekordów nie jest brany pod uwagę — ważny jest tylko porządek określony w danym węźle.

Jeśli nie zostaną określone żadne zmienne sortowania (lub jeśli określona zostanie niedostateczna liczba zmiennych sortowania), wówczas rekordy w każdej grupie duplikatów będą nieuporządkowane (lub będą uporządkowane w sposób niekompletny), a wyniki mogą być nieprzewidywalne.

Załóżmy na przykład, że dostępny jest bardzo duży zbiór rekordów dziennika odnoszący się do wielu komputerów. Dziennik zawiera dane, takie jak:

Tabela 17. Dane dziennika wg komputera

Znacznik czasu	Komputer	Temperatura
17:00:22	Machine A	31
13:11:30	Machine B	26
16:49:59	Machine A	30
18:06:30	Machine X	32
16:17:33	Machine A	29
19:59:04	Machine C	35
19:20:55	Machine Y	34
15:36:14	Machine X	28
12:30:41	Machine Y	25
14:45:49	Machine C	27
19:42:00	Machine B	34
20:51:09	Machine Y	36
19:07:23	Machine X	33

Aby zmniejszyć liczbę rekordów do najnowszego rekordu dla każdego komputera, należy użyć zmiennej Machine(Komputer) jako zmiennej kluczowej oraz zmiennej Timestamp (Znacznik czasu) jako zmiennej sortowania (w porządku malejącym). Porządek wprowadzania danych wejściowych nie wpływa na wynik, ponieważ wybór sortowania określa, które z wierszy dla danego komputera mają zostać zwrócone; ostatecznie dane wyjściowe będą wyglądały następująco.

Tabela 18. Posortowane dane dziennika wg komputera

Znacznik czasu	Komputer	Temperatura
17:00:22	Machine A	31
19:42:00	Machine B	34
19:59:04	Machine C	35
19:07:23	Machine X	33
20:51:09	Machine Y	36

Ustawienia optymalizacji w węźle Powtórzenia

Jeśli dane zawierają tylko niewielką liczbę rekordów lub zostały już posortowane, można zoptymalizować sposób ich obsługi, umożliwiając programowi IBM SPSS Modeler bardziej efektywne przetwarzanie danych.

Uwaga: Jeśli wybrana zostanie opcja **Wejściowy zbiór danych ma niewielką liczbę niezależnych kluczy** lub użyte zostanie generowanie kodu SQL dla węzła, mogą zostać zwrócone wszystkie wiersze tworzące wartość klucza niezależnego; aby kontrolować, który wiersz zostanie zwrócony dla klucza niezależnego, należy określić porządek sortowania, używając zmiennych **W obrębie grup uporządkuj rekordy ze względu na** na karcie Ustawienia. Opcje optymalizacji nie wpływają na wyniki dla węzła Powtórzenia, o ile określono porządek sortowania na karcie Ustawienia.

Wejściowy zbiór danych ma niewielką liczbę niezależnych kluczy. Tę opcję należy wybrać, jeśli dostępna jest niewielka liczba rekordów i/lub nie wielka liczba unikalnych wartości zmiennych kluczowych. Dzięki temu można zwiększyć wydajność.

Wejściowy zbiór danych jest już uporządkowany ze względu na pola grupowania i pola porządkowania zdefiniowane w zakładce Ustawienia. Tę opcję należy wybrać tylko wówczas, gdy dane zostały już posortowane

według wszystkich zmiennych wymienionych w obszarze **W obrębie grup uporządkuj rekordy ze względu na** na karcie Ustawienia, oraz jeśli porządek sortowania danych rosnąco lub malejąco jest taki sam. Dzięki temu można zwiększyć wydajność.

Wyłącz generowanie kodu SQL. Tę opcję należy wybrać, aby wyłączyć generowanie kodu SQL dla węzła.

Ustawienia złożonego rekordu w węźle Powtórzenia

Jeśli dane zawierają wiele rekordów, na przykład dotyczących tej samej osoby, można zoptymalizować sposób obsługi danych poprzez utworzenie pojedynczego, złożonego lub zagregowanego rekordu do przetworzenia.

Uwaga: Ta karta jest dostępna tylko w przypadku wybrania opcji **Utwórz rekord złożony dla każdej grupy** na karcie Ustawienia.

Ustawianie opcji dla karty Złożone

Zmienna. W tej kolumnie wyświetlane są wszystkie zmienne, z wyjątkiem zmiennych kluczowych z modelu danych, w rzeczywistym porządku sortowania. jeśli węzeł nie jest połączony, żadna zmienna nie jest wyświetlana. Aby posortować wiersze alfabetycznie według nazwy zmiennej, należy kliknąć nagłówek kolumny. Korzystając z klawiszy Shift+kliknięcie lub Ctrl+kliknięcie można wybrać więcej niż jeden wiersz naraz. Ponadto po kliknięciu zmiennej prawym przyciskiem myszy wyświetlane jest menu, w którym można zaznaczyć wszystkie wiersze, sortować wiersze rosnąco lub malejąco według nazwy zmiennej lub wartości, wybierać zmienne według miary lub typu składowania lub wybrać wartość pozwalającą na automatyczne dodawanie tego samego wpisu **Wypełnij wartościami w oparciu o** do każdego wybranego wiersza.

Wypełnij wartościami w oparciu o. Należy wybrać typ wartości, jaki będzie stosowany w rekordzie złożonym dla opcji **Zmienna**. Dostępne opcje zależą od typu zmiennej.

- W przypadku zmiennych zakresów liczbowych dostępne są następujące opcje:
  - Pierwszy rekord w grupie
  - Ostatni rekord w grupie
  - Suma
  - Średnia
  - Minimum
  - Maksimum
  - Użytkownika
- Dla zmiennych czasu lub daty dostępne są następujące opcje:
  - Pierwszy rekord w grupie
  - Ostatni rekord w grupie
  - Najwcześniejszy
  - Najnowszy
  - Użytkownika
- Dla zmiennych łańcuchowych lub bez określonego typu dostępne są następujące opcje:
  - Pierwszy rekord w grupie
  - Ostatni rekord w grupie
  - Pierwsze alfanumeryczne
  - Ostatnie alfanumeryczne
  - Użytkownika

W każdym przypadku można użyć opcji **Użytkownika**, aby dokładniej kontrolować, która wartość zostanie użyta do wypełnienia złożonego rekordu. Więcej informacji można znaleźć w “Rekord złożony w węźle Powtórzenia — karta Użytkownika” na stronie 97.

Dołącz liczebność rekordów w zmiennej. Należy wybrać tę opcję, aby domyślnie w każdym rekordzie wynikowym o nazwie Record_Count (Liczebność rekordu) uwzględnić dodatkową zmienną. To pole wskazuje, ile rekordów wejściowych zostało zagregowanych w celu utworzenia poszczególnych rekordów agregacji. Aby utworzyć niestandardową nazwę dla tej zmiennej, należy wpisać ją w polu edycji.

Rekord złożony w węźle Powtórzenia — karta Użytkownika

Okno dialogowe Wypełnienie niestandardowe zapewnia większą kontrolę nad tym, która wartość zostanie użyta w nowym rekordzie złożonym. Jeśli dostosowany ma zostać tylko jeden wiersz zmiennej na karcie Złożone, należy pamiętać, aby przed użyciem tej opcji określić dane.

Uwaga: To okno dialogowe jest dostępne tylko po wybraniu wartości Użytkownika w kolumnie **Wypełnij wartościami w oparciu o** na karcie Złożone.

W zależności od typu zmiennej można wybrać jedną z następujących opcji.

- **Wybierz wg częstotliwości.** Należy wybrać wartość na podstawie częstości jej występowania w rekordzie danych.

  Uwaga: Opcja jest niedostępna dla zmiennych typu ilościowa, bez określonego typu lub data/czas.
  - **Wykorzystanie.** Można wybrać jedną z dwóch opcji: Najczęstszy lub Najrzadszy.
  - **Wiązania.** Jeśli dostępne są co najmniej dwa rekordy o takiej samej częstości występowania, należy określić, w jaki sposób wymagany rekord ma zostać wybrany. Można wybrać jedną z czterech opcji: Użyj pierwszego, Użyj ostatniego, Użyj najniższego lub Użyj najwyższego.
- **Zawiera wartość (T/F).** Tę opcję należy wybrać, aby przekształcić zmienną na flagę, która będzie sprawdzała, czy w grupie znajduje się rekord zawierający określoną wartość. Następnie można wybrać **wartość** z listy dostępnych dla wybranej zmiennej.

  Uwaga: Opcja nie jest dostępna, jeśli na karcie Złożone wybrany zostanie więcej niż jeden wiersz Zmienna.
- **Pierwsze dopasowanie na liście.** Należy zaznaczyć tę opcję, aby określić priorytet dla wartości, która ma zostać dodana do złożonego rekordu. Następnie można wybrać jeden z **elementów** z listy dostępnych dla wybranej zmiennej.

  Uwaga: Opcja nie jest dostępna, jeśli na karcie Złożone wybrany zostanie więcej niż jeden wiersz Zmienna.
- **Wartości połączenia.** Tę opcję należy wybrać, aby zachować wszystkie wartości w grupie poprzez połączenie ich w łańcuch. Należy wybrać separator, jaki będzie umieszczany pomiędzy wartościami.

  Uwaga: Ta opcja jest dostępna, jeśli wybrany zostanie co najmniej jeden wiersz Zmienna typu ilościowego, bez określonego typu lub typu data/czas.
- **Stosuj separator.** Jako wartość separatora w połączonym łańcuchu można wybrać opcję **Spacja** lub **Przecinek**. Alternatywnie, w polu **Inne** można wprowadzić własny znak separatora.

  Uwaga: Opcja jest dostępna, jeśli zaznaczono opcję **Wartości połączenia**.

Węzeł Szeregi czasowe

Węzeł Szeregi czasowe umożliwia budowanie i ocenianie modeli szeregów czasowych w jednym kroku. Dla każdej zmiennej przewidywanej tworzony jest osobny model szeregów czasowych, jednak do wygenerowanej palety modeli nie są dodawane modele użytkowe i nie można przeglądać informacji o modelu.

Metody modelowania danych szeregów czasowych wymagają równomiernych przedziałów między poszczególnymi pomiarami oraz reprezentacji braków danych w postaci pustych wierszy. Jeśli dane nie spełniają jeszcze tych wymogów, konieczne będzie odpowiednie przekształcenie wartości.

Pozostałe kwestie związane ze stosowaniem węzłów Szereg czasowy:

- Zmienne muszą być liczbowe.
- Zmienne typu data nie mogą być danymi wejściowymi.

- Podzbiory są ignorowane.

Węzeł Szeregi czasowe umożliwia estymację modelu wygładzania wykładniczego, modelu autoregresyjnej zintegrowanej średniej ruchomej (ARIMA) jednej zmiennej oraz modelu ARIMA wielu zmiennych (lub funkcji przenoszenia) dla danych szeregów czasowych i generuje prognozy w oparciu o dane szeregu czasowego. Funkcja jest również dostępna za pośrednictwem automatycznego doboru modelu, który podejmuje próby automatycznego zidentyfikowania i oszacowania modelu ARIMA lub modelu wygładzania wykładniczego o najlepszym dopasowaniu dla co najmniej jednej zmiennej przewidywanej.

Więcej informacji na temat modelowania szeregów czasowych zawiera sekcja Modele szeregów czasowych w publikacji SPSS Modeler — węzły modelowania.

Węzeł Szeregi czasowe jest używany w środowisku wdrożenia strumienia, za pośrednictwem programu IBM SPSS Modeler Solution Publisher, z wykorzystaniem aplikacji IBM SPSS Collaboration and Deployment Services Scoring Service..

Węzeł Szeregi czasowe — opcje zmiennych

Użyj wstępnie zdefiniowanych ról Ta opcja korzysta z ustawień roli (zmienne przewidywane, predyktory itd.) z poprzedzającego węzła Typy (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Tę opcję należy wybrać, aby ręcznie przypisać zmienne przewidywane, predyktory i inne role.

Uwaga: Jeśli dane zostały podzielone na podzbiory, podzbiory te są uwzględniane po wybraniu opcji **Użyj wstępnie zdefiniowanych ról**, jednak nie są uwzględniane po wybraniu opcji **Użyj niestandardowych przypisań**.

Pola. Aby ręcznie przypisać pozycje z tej listy do różnych zmiennych ról po prawej stronie ekranu, należy użyć klawiszy strzałek. Ikony wskazują prawidłowe poziomy pomiaru dla każdego pola roli.

Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Przewidywane Należy wybrać co najmniej jedną zmienną jako zmienną przewidywaną dla predykcji.

Potencjalne zmienne wejściowe Należy wybrać do najmniej jedną zmienną jako zmienna wejściowa dla predykcji.

Zdarzenia i interwencje Tego obszaru należy użyć, aby wyznaczyć konkretne zmienne jako zmienne zdarzenia lub interwencji. Dzięki temu zmienna jest identyfikowana jako zawierająca dane szeregu czasowego, na które wpływ mogą mieć zdarzenia (przewidywalne sytuacje powtarzalne; na przykład promocje) lub interwencje (jednorazowe incydenty; na przykład przerwa w zasilaniu lub strajk pracowników).

Węzeł Szeregi czasowe — opcje specyfikacji danych

Na karcie Specyfikacja danych można ustawić wszystkie opcje danych, jakie będą uwzględnione w modelu. Jeśli określone zostaną wartości **Zmienna typu data/czas** oraz **Przedział czasowy**, można kliknąć przycisk **Uruchom** w celu zbudowania modelu z wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik zwykle potrzebuje dostosować proces budowy do własnych potrzeb.

Karta zawiera kilka różnych paneli, w których można dostosować ustawienia odpowiednio do specyfiki modelu.

Węzeł Szeregi czasowe — obserwacje

Należy użyć ustawień w tym panelu, aby określić zmienne definiujące obserwacje.

Obserwacje określone przez zmienną typu data/czas

Można określić, czy obserwacje będą definiowane na podstawie zmiennej daty, czasu lub znacznika czasu. Oprócz zmiennej, która definiuje obserwacje, należy wybrać odpowiedni przedział czasowy, w którym będą opisywane obserwacje. W zależności od określonego przedziału czasowego można również dokonać innych ustawień, takich jak przedział między obserwacjami (przyrost) lub liczba dni w tygodniu. Poniższe stwierdzenia mają zastosowanie do przedziałów czasowych:

- Wartości **Nieregularny** należy użyć, jeśli obserwacje są nierównomiernie rozmieszczone w czasie, na przykład są wykonywane w chwili, gdy następuje przetwarzanie zamówienia sprzedaży. Jeśli wybrana zostanie opcja **Nieregularny**, należy określić przedział czasowy, jaki będzie używany do analizy, wybierając ustawienia **Przedział czasowy** na karcie Specyfikacja danych.
- Jeśli obserwacje reprezentują datę i czas, a przedział czasowy to godziny, minuty lub sekundy, wówczas należy wybrać opcję **Godziny dziennie**, **Minuty dziennie** lub **Sekundy dziennie**. Jeśli obserwacje reprezentują czas (trwanie) bez odniesienia do daty, a przedział czasowy to godziny, minuty lub sekundy, wówczas należy użyć opcji **Godziny (nieokresowo)**, **Minuty (nieokresowo)** lub **Sekundy (nieokresowo)**.
- Na podstawie wybranego przedziału czasowego procedura może wykryć brakujące obserwacje. Wykrywanie brakujących obserwacji jest konieczne, ponieważ procedura zakłada, że wszystkie obserwacje są równomiernie rozłożone w czasie i że nie ma brakujących obserwacji. Na przykład, jeśli przedziałem czasu są dni, a po dacie 2015-10-27 występuje 2015-10-29, istnieje brakująca obserwacja dla daty 2015-10-28. Dla wszystkich brakujących obserwacji wprowadzane są wartości; w obszarze **Traktowanie braków danych** na karcie Specyfikacja danych można określić ustawienia obsługi braków danych.
- Określone przedziały czasowe umożliwiają procedurze wykrycie wielu obserwacji w jednym przedziale czasowym, które muszą być zagregowane i które są dopasowane do obserwacji dla granicy przedziału (np. pierwszy dzień miesiąca), dzięki czemu obserwacje będą równomiernie rozmieszczone. Na przykład, jeśli przedziałem czasu są miesiące, to zagregowanych zostanie wiele dat z tego samego miesiąca. Ten typ agregacji jest nazywany *grupowaniem*. Domyślnie obserwacje są sumowane podczas grupowania. Można określić inną metodę grupowania, np. średnia z obserwacji; ustawienia **Agregacja i rozkład** można wprowadzić na karcie Specyfikacja danych.
- Przy niektórych przedziałach czasowych istnieją dodatkowe ustawienia umożliwiające zdefiniowanie odstępów w zwykle równomiernie rozmieszczonych przedziałach. Na przykład, jeśli przedziałem czasu są dni, ale istotne są tylko dni robocze, można określić pięciodniowy tydzień rozpoczynający się od poniedziałku.

Obserwacje zdefiniowane jako okresy lub okresy cykliczne

Obserwacje mogą być zdefiniowane przez co najmniej jedną zmienną całkowitą, która reprezentuje okresy lub powtarzające się cyklicznie okresy, aż do dowolnej liczby poziomów cyklicznych. Taka struktura pozwala opisać serie obserwacji, które nie pasują do jednego ze standardowych przedziałów czasowych. Przykładowo, rok fiskalny trwający tylko 10 miesięcy może być opisany przez zmienną cyklu, która reprezentuje lata, i zmienną okresu, która reprezentuje miesiące, przy czym długość jednego cyklu wynosi 10.

Zmienne, które określają okresy cykliczne, definiują hierarchię poziomów okresowości, w której najniższy poziom jest definiowany przez zmienną **Okres**. Następny wyższy poziom jest określany przez zmienną cyklu z poziomem 1, po której następuje zmienna cyklu z poziomem 2 itd. Wartości zmiennych dla każdego poziomu, z wyjątkiem najwyższego, muszą być okresowe w odniesieniu do kolejnego najwyższego poziomu. Wartości dla najwyższego poziomu nie mogą być okresowe. Na przykład, dla 10-miesięcznego roku fiskalnego miesiące występują okresowo w latach, ale lata nie są okresowe.

- Długość cyklu na poszczególnych poziomach stanowi okresowość dla kolejnego najniższego poziomu. W przykładzie dot. roku fiskalnego istnieje tylko jeden poziom cyklu, a długość cyklu wynosi 10, ponieważ kolejny najniższy poziom reprezentuje miesiące, a w określonym roku fiskalnym jest 10 miesięcy.
- Należy określić wartość początkową dla każdej zmiennej okresowej, która nie rozpoczyna się od 1. To ustawienie jest niezbędne dla wykrywania braków danych. Na przykład, jeśli zmienna okresowa rozpoczyna się od 2, ale wartość początkowa jest określona jako 1, wówczas procedura zakłada, że istnieje brak danych dla pierwszego okresu w każdym cyklu dla tej zmiennej.

Węzeł Szeregi czasowe — przedział czasowy do analizy

Przedział czasowy, który jest używany do analizy, może różnić się od przedziału czasowego dla obserwacji. Na przykład, jeśli przedział czasowy obserwacji to dni, jako przedział czasowy dla analizy można wybrać miesiące. Wówczas przed zbudowaniem modelu dane agregowane są z dziennych na miesięczne. Można również rozłożyć dane z dłuższego przedziału czasu na krótszy. Przykładowo, jeśli obserwacje są przeprowadzane kwartalnie, wówczas można rozłożyć dane z kwartalnych na miesięczne.

Należy użyć ustawień na tym panelu, aby określić przedział czasowy dla analizy. Metoda, którą dane są agregowane lub rozkładane, jest określana w ustawieniach **Agregacja i rozkład** na karcie Specyfikacja danych.

Opcje możliwe do wyboru dla przedziału czasowego, w jakim wykonywana jest analiza, zależą od sposobu zdefiniowania obserwacji oraz wyznaczonego dla nich przedziału czasowego. W szczególności, jeśli obserwacje są zdefiniowane przez okresy cykliczne, wówczas obsługiwana jest tylko agregacja. W takim przypadku przedział czasowy dla analizy musi być większy od przedziału czasowego dla obserwacji lub mu równy.

Węzeł Szeregi czasowe — opcje rozkładu i agregacji

Należy użyć ustawień w tym panelu, aby określić ustawienia agregacji lub rozkładu danych wejściowych odpowiednio do przedziałów czasowych dla obserwacji.

Funkcje agregacji

Jeśli przedział czasowy użyty dla analizy jest dłuższy niż przedział czasowy dla obserwacji, dane wejściowe zostają zagregowane. Przykładowo agregacja jest przeprowadzana, kiedy przedział czasowy dla obserwacji to dni, a przedział czasowy dla analizy to miesiące. Dostępne są następujące funkcje agregacji: średnia, suma, dominanta, wartość minimalna lub maksymalna.

Funkcje rozkładu

Jeśli przedział czasowy użyty dla analizy jest krótszy niż przedział czasowy dla obserwacji, dane wejściowe zostają rozłożone. Przykładowo rozkład jest przeprowadzany, kiedy przedział czasowy dla obserwacji to kwartały, a przedział czasowy dla analizy to miesiące. Dostępne są następujące funkcje rozkładu: średnia lub suma.

Funkcje grupujące

Grupowanie jest stosowane, kiedy obserwacje są definiowane przez datę/czas i w tym samym przedziale czasowym występuje wiele obserwacji. Na przykład, jeśli przedział czasowy dla obserwacji to miesiące, wówczas wiele dat z tego samego miesiąca jest grupowanych i tworzone jest ich powiązanie z miesiącem, w którym występują. Dostępne są następujące funkcje grupowania: średnia, suma, dominanta, wartość minimalna lub maksymalna. Grupowanie jest zawsze przeprowadzane, kiedy obserwacje są zdefiniowane przez datę/czas, a przedział czasowy dla obserwacji jest określony jako Nieregularny.

Uwaga: Chociaż grupowanie jest formą agregacji, jest przeprowadzane przed rozpoczęciem obsługi braków danych, podczas gdy formalna agregacja jest wykonywana po zakończeniu obsługi braków danych. Jeśli przedział czasowy dla obserwacji jest określony jako Nieregularny, agregacja jest wykonywana tylko za pomocą funkcji grupowania.

Agreguj obserwacje przekraczające granice dnia do dnia poprzedniego

Określa, czy obserwacje, których czas przekracza granicę dnia, są agregowane na wartości dla dnia poprzedniego. Przykładowo, dla obserwacji godzinowych trwających osiem godzin dziennie i rozpoczynających się o godzinie 20:00, to ustawienie określi, czy obserwacje od godziny 00:00 do 04:00 będą uwzględniane w zagregowanych wynikach dla poprzedniego dnia. To ustawienie ma zastosowanie tylko w przypadku, kiedy przedział czasowy dla obserwacji to Godziny dziennie, Minuty dziennie lub Sekundy dziennie, a przedział czasowy dla analizy to Dni.

Ustawienia niestandardowe dla określonych zmiennych

Funkcje agregacji, rozkładu i grupowania dla zmiennej można określić na podstawie zmiennej. Ustawienia te zastępują domyślne ustawienia funkcji agregacji, rozkładu i grupowania.

Węzeł Szeregi czasowe — opcje braków danych

Ustawienia w tym panelu umożliwiają określenie sposobu zastępowania braków danych w danych wejściowych przez wartość podstawianą. Dostępne są następujące metody zastępowania:

Interpolacja liniowa
Powoduje zastąpienie braków danych przy wykorzystaniu interpolacji liniowej. W interpolacji używana jest ostatnia ważna wartość przed brakiem danych oraz pierwsza ważna za brakiem. Jeśli pierwsza lub ostatnia obserwacja w szeregu zawiera brakujące wartości, wówczas używane są dwie najbliższe niebrakujące wartości na początku i na końcu serii.

Średnia szeregu
Zastępuje braki danych średnią obliczoną ze wszystkich obserwacji.

Średnia z sąsiednich punktów
Powoduje zastąpienie braków danych średnią z ważnych wartości sąsiednich. Rozpiętość sąsiednich punktów, to liczba poprawnych wartości występujących przed brakującą wartością i po niej, jakie są wykorzystywane do obliczenia średniej.

Mediana z sąsiednich punktów
Powoduje zastąpienie braków danych medianą ważnych wartości sąsiednich. Rozpiętość sąsiednich punktów, to liczba poprawnych wartości występujących przed brakującą wartością i po niej, jakie są wykorzystywane do obliczenia mediany.

Trend liniowy
Ta opcja wykorzystuje niebrakujące obserwacje w szeregu do dopasowania prostego modelu regresji liniowej, który jest następnie używany w celu przypisania braków danych.

Inne ustawienia:

Najniższy wynik jakości danych (%)
Przelicza miary jakości danych dla zmiennej czasu i dla danych wejściowych odpowiadających poszczególnym szeregom czasowym. Jeśli wynik jakości danych jest niższy niż wyznaczony przez próg, odpowiednie szeregi czasowe zostaną odrzucone.

Węzeł Szeregi czasowe — okres szacowania

W panelu Okres oszacowania można określić zakres rekordów, jakie będą użyte do oszacowania modelu. Domyślnie okres estymacji zaczyna się od czasu z najwcześniejszą obserwacją, a kończy w czasie z najpóźniejszą obserwacją we wszystkich szeregach.

Wyznaczony przez czas początkowy i końcowy
Można określić datę rozpoczęcia i zakończenia okresu estymacji lub można określić tylko datę rozpoczęcia lub tylko datę zakończenia. Jeśli rozpoczęcie lub zakończenie okresu estymacji zostanie pominięte, użyta będzie wartość domyślna.

- Jeśli obserwacje są zdefiniowane przez określenie zmiennej daty/czasu, wówczas wartości rozpoczęcia i zakończenia okresu należy wprowadzić w takim samym formacie, jaki został użyty dla zmiennej daty/czasu.
- W przypadku obserwacji definiowanych na podstawie okresów cyklicznych należy określić wartość dla każdej zmiennej okresu cyklicznego. Każda zmienna jest wyświetlana w osobnej kolumnie.

Wyznaczony przez najwcześniejszy lub najpóźniejszy przedział czasowy

Definiuje okres estymacji jako określoną liczbę przedziałów czasowych, która rozpoczyna się od najwcześniejszego przedziału czasowego lub kończy na najpóźniejszym przedziale czasowym określonym w danych, z opcjonalnym przesunięciem. W tym kontekście przedział czasowy odnosi się do przedziału czasowego dla analizy. Przykładowo, załóżmy, że obserwacje są przeprowadzane miesięcznie, ale przedział czasowy dla analizy to kwartały. Określenie wartości **Najpóźniejszy** i wartości 24 dla opcji **Liczba przedziałów czasowych** będzie oznaczało ostatnie 24 kwartały.

Opcjonalnie można wykluczyć określoną liczbę przedziałów czasowych. Przykładowo, określenie ostatnich 24 przedziałów czasowych i 1 do wykluczenia oznacza, że okres oszacowania składa się z 24 przedziałów, które poprzedzają ostatni.

Węzeł Szeregi czasowe — opcje budowania

Na karcie Opcje budowania można ustawić wszystkie opcje budowania modelu. Można oczywiście po prostu kliknąć przycisk **Uruchom** w celu zbudowania modelu z wszystkimi opcjami domyślnymi, lecz w normalnej sytuacji użytkownik potrzebuje zwykle dostosować proces budowy do swoich celów.

Karta zawiera dwa różne panele, w których można dostosować ustawienia odpowiednio do specyfiki modelu.

Węzeł Szeregi czasowe — ogólne opcje budowania

Opcje dostępne w tym panelu zależą tego, które z trzech ustawień zostanie wybrane z listy **Metoda**:

- **Automatyczny dobór modelu.** Wybór tej opcji powoduje użycie funkcji Automatyczny dobór modelu, która automatycznie znajduje model o najlepszym dopasowaniu dla każdego szeregu zależnego.
- **Wygładzanie wykładnicze.** Ta opcja umożliwia określenie niestandardowego modelu wygładzania wykładniczego.
- **ARIMA.** Ta opcja umożliwia określenie niestandardowego modelu ARIMA.

Automatyczny dobór modelu

Opcja **Typ modelu** umożliwia wybranie typów modeli, jakie mają zostać zbudowane:

- **Wszystkie modele.** Funkcja automatycznego doboru modeli uwzględnia zarówno modele ARIMA, jak i modele wygładzania wykładniczego.
- **Tylko modele wygładzania wykładniczego.** Automatyczny dobór modeli uwzględnia tylko modele wygładzania wykładniczego.
- **Tylko modele ARIMA.** Automatyczny dobór modeli uwzględnia tylko modele ARIMA.

Automatyczny dobór modelu uwzględnia modele sezonowe. Ta opcja jest aktywna tylko wówczas, jeśli dla aktywnego zbioru danych zdefiniowano okresowość. Po zaznaczeniu tej opcji automatyczny dobór modeli uwzględnia zarówno modele sezonowe, jak i niesezonowe. Jeśli ta opcja nie zostanie zaznaczona, automatyczny dobór modeli uwzględnia tylko modele niesezonowe.

Automatyczny dobór modelu uwzględnia wyrafinowane metody wygładzania wykładniczego. Po wybraniu tej opcji Automatyczny dobór modelu przeszukuje łącznie 13 modeli wygładzania wykładniczego (7 z nich znajduje w pierwotnym węźle szeregów czasowych, a 6 zostało dodanych w wersji 18.1). Jeśli ta opcja nie zostanie zaznaczona, Automatyczny dobór modelu będzie przeszukiwał tylko 7 pierwotnych modeli wygładzania wykładniczego.

Funkcja **Wartości odstające** umożliwia wybranie następujących opcji:

Automatycznie wykryj wartości odstające. Domyślnie automatyczne wykrywanie wartości odstających nie jest przeprowadzane. Zaznacz tę opcję, aby automatycznie wykrywać wartości odstające, a następnie wybierz żądany typ wartości odstających.

Aby zmienne wejściowe zostały uwzględnione na tej liście, muszą mieć poziom pomiaru *Flaga*, *Nominalny* lub *Porządkowy* i muszą być numeryczne (na przykład 1/0 zamiast Prawda/Fałsz w przypadku zmiennej flagi).

Automatyczny dobór modeli uwzględnia tylko regresję prostą, pomijając dowolne funkcje przenoszenia dla danych wejściowych zidentyfikowanych jako zmienne typu zdarzenie czy interwencja na karcie **Zmienne**.

Wygładzanie wykładnicze

Typ modelu. Modele wygładzania wykładniczego są klasyfikowane jako sezonowe lub niesezonowe.[1] Modele sezonowe są dostępne tylko wówczas, jeśli okresowość zdefiniowana za pomocą panelu Przedziały czasowe na karcie Specyfikacja danych jest sezonowa. Okresowości sezonowe to: okresy cykliczne, lata, kwartały, miesiące, dni tygodnia, godziny dnia, minuty dnia oraz sekundy dnia. Dostępne są następujące typy modeli:

- **Prosty.** Ten model jest odpowiedni w przypadku szeregu, w którym brak trendu lub sezonowości. Jedynym odpowiednim parametrem wygładzania jest poziom. Proste wygładzanie wykładnicze jest najbardziej podobne do modelu ARIMA bez rzędów autoregresji, z jednym rzędem różnicowania, jednym rzędem ruchomej średniej i brakiem stałych.
- **Trend liniowy Holta.** Model ten jest odpowiedni dla szeregów, w których istnieje trend liniowy i nie ma sezonowości. Właściwe dla niego parametry wygładzania to poziom i trend; w przypadku tego modelu nie są one ograniczone wzajemnie swoimi wartościami. Model Holta jest bardziej ogólny niż model Browna, ale w przypadku dłuższych szeregów przeliczanie zajmuje więcej czasu. Wygładzanie wykładnicze Holta jest najbardziej podobne do modelu ARIMA bez rzędu autoregresji, z dwoma rzędami różnicowania i dwoma rzędami średniej ruchomej.
- **Trend wygasający.** Model ten jest odpowiedni dla szeregów, w których istnieje trend liniowy, który wygasa i nie ma sezonowości. Właściwe dla niego parametry wygładzania to: poziom, trend i trend osłabiający. Wygasające wygładzanie wykładnicze jest najbardziej podobne do modelu ARIMA z jednym rzędem autoregresji, jednym rzędem różnicowania i dwoma rzędami średniej ruchomej.
- **Trend multiplikatywny.** Ten model jest odpowiedni dla szeregów, w których występuje trend zmieniający się wraz z modułem szeregu bez efektu sezonowości. Właściwe dla niego parametry wygładzania to: poziom i trend. Wygładzanie wykładnicze trendu multiplikatywnego nie jest podobne do żadnego modelu ARIMA.
- **Trend liniowy Browna.** Model ten jest odpowiedni dla szeregów, w których istnieje trend liniowy i nie ma sezonowości. Właściwe dla niego parametry wygładzania to poziom i trend; w tym modelu zakłada się jednak, że są one sobie równe. Zatem model Browna jest specjalnym przypadkiem modelu Holta. Wygładzanie wykładnicze Browna jest podobne do modelu ARIMA z zerowym rzędem autoregresji, dwoma rzędami różnic oraz dwoma rzędami ruchomych średnich ze współczynnikiem drugiego rzędu ruchomej średniej równym kwadratowi połowy współczynnika pierwszego rzędu.
- **Prosty sezonowy.** Model ten jest odpowiedni dla szeregów, w których nie ma trendu, a efekt sezonowości jest stały w czasie. Właściwe dla niego parametry wygładzania to: poziom i sezon. Proste sezonowe wygładzanie wykładnicze jest podobne do modelu ARIMA bez rzędów autoregresji, z jednym rzędem różnicowania oraz jednym rzędem różnicowania sezonowego i rzędami 1, p i p+1 średniej ruchomej, gdzie p jest liczbą okresów w przedziale sezonowym. W przypadku danych miesięcznych $p = 12$.
- **Addytywny model Wintersa.** Model ten jest odpowiedni dla szeregów, w których nie ma trendu liniowego, a efekt sezonowości jest stały w czasie. Właściwe dla niego parametry wygładzania to: poziom, trend i sezon. Wygładzanie wykładnicze addytywnego modelu Wintersa jest podobne do modelu ARIMA bez rzędu autoregresji, z jednym rzędem różnicowania oraz jednym rzędem różnicowania sezonowego i p+1 rzędami średniej ruchomej, gdzie p jest liczbą okresów w przedziale sezonowym. W przypadku danych miesięcznych $p = 12$.
- **Trend wygasający ze składnikiem sezonowym addytywnym.** Model ten jest odpowiedni dla szeregów, w których istnieje wygasający trend liniowy, a efekt sezonowości jest stały w czasie. Właściwe dla niego parametry wygładzania to: poziom, trend, trend gasnący i sezon. Wygładzanie wykładnicze trendu wygasającego i sezonowości addytywnej nie jest podobne do żadnego modelu ARIMA.
- **Trend multiplikatywny ze składnikiem sezonowym addytywnym.** Ten model jest odpowiedni dla szeregów, w których występuje trend zmieniający się wraz z modułem szeregu, a efekt sezonowości jest stały w czasie. Właściwe dla niego parametry wygładzania to: poziom, trend i sezon. Wygładzanie wykładnicze trendu multiplikatywnego i sezonowości addytywnej nie jest podobne do żadnego modelu ARIMA.
- **Multiplikatywny sezonowo.** Model ten jest odpowiedni dla szeregów, w których nie ma trendu, a efekt sezonowości jest zmienny wraz z modułem szeregu. Właściwe dla niego parametry wygładzania to: poziom i sezon. Wygładzanie wykładnicze sezonowości multiplikatywnej nie jest podobne do żadnego modelu ARIMA.

1. Gardner, E. S. 1985. Exponential smoothing: The state of the art. *Journal of Forecasting,* 4, 1-28.

- **Multiplikatywny model Wintersa.** Model ten jest odpowiedni dla szeregów, w których występuje trend liniowy oraz efekt sezonowości zmienny wraz z modułem szeregu. Właściwe dla niego parametry wygładzania to: poziom, trend i sezon. Wygładzanie wykładnicze multiplikatywnego modelu Wintersa nie jest podobne do modelu ARIMA.
- **Trend wygasający ze składnikiem sezonowym multiplikatywnym.** Model ten jest odpowiedni dla szeregów, w których występuje gasnący trend liniowy oraz efekt sezonowości zmienny wraz z modułem szeregu. Właściwe dla niego parametry wygładzania to: poziom, trend, trend gasnący i sezon. Wygładzanie wykładnicze trendu wygasającego i sezonowości multiplikatywnej nie jest podobne do żadnego modelu ARIMA.
- **Trend multiplikatywny ze składnikiem sezonowym multiplikatywnym.** Model ten jest odpowiedni dla szeregów, w których występuje trend oraz efekt sezonowości zmienne wraz z modułem szeregu. Właściwe dla niego parametry wygładzania to: poziom, trend i sezon. Wygładzanie wykładnicze trendu multiplikatywnego i sezonowości multiplikatywnej nie jest podobne do żadnego modelu ARIMA.

Transformacja zmiennej przewidywanej. Istnieje możliwość określenia transformacji do wykonania dla każdej zmiennej zależnej przed jej zamodelowaniem.

- **Brak.** Nie jest wykonywana żadna transformacja.
- **Pierwiastek kwadratowy.** Wykonywana jest transformacja pierwiastkiem kwadratowym.
- **Logarytm naturalny.** Wykonywana jest transformacja logarytmem naturalnym.

ARIMA

Należy określić strukturę niestandardowego modelu ARIMA.

Rzędy ARIMA. Należy wprowadzić wartości dla różnych składników ARIMA modelu do odpowiednich komórek siatki. Wszystkie wartości muszą być nieujemnymi liczbami całkowitymi. W przypadku składników autoregresja i średnia ruchoma wartość ta reprezentuje rząd maksymalny. W modelu zostaną uwzględnione wszystkie niższe rzędy dodatnie. Na przykład w przypadku podania wartości 2 model będzie obejmował rzędy 2 i 1. Komórki w kolumnie Sezonowa są aktywne tylko, jeśli dla aktywnego zbioru danych zdefiniowano okresowość.

- **Autoregresja (p).** Liczba rzędów autoregresji w modelu. Rzędy autoregresji określają, które z poprzednich wartości są używane do przewidywania bieżących wartości. Na przykład rząd autoregresji 2 oznacza, że do przewidywania bieżącej wartości zostanie użyta wartość dwu okresów czasu szeregu w przeszłości.
- **Różnica (d).** Określa rząd różnicowania stosowany względem szeregu przed oszacowaniem modeli. Różnicowanie jest niezbędne w przypadku obecności trendów (szeregi z trendami są zwykle niestacjonarne, zaś modelowanie ARIMA zakłada stacjonarność) i służy do usuwania ich wpływu. Rząd różnicowania odpowiada stopniowi trendu szeregu — różnicowanie pierwszego rzędu jest uwzględniane w przypadku trendów liniowych, różnicowanie drugiego rzędu w przypadku trendów kwadratowych itd.
- **Średnia ruchoma (q).** Liczba rzędów średniej ruchomej w modelu. Rzędy średniej ruchomej określają, w jaki sposób odchylenia od średniej szeregu dla poprzednich wartości są używane do przewidywania bieżących wartości. Na przykład rzędy średniej ruchomej 1 i 2 oznaczają, że odchylenia od wartości średniej szeregu z każdego z dwu ostatnich okresów czasu będą uwzględniane podczas przewidywania bieżących wartości szeregu.

Sezonowe. Autoregresja sezonowa, średnia ruchoma i różnicowanie odgrywają takie same role, jak ich niesezonowe odpowiedniki. W przypadku rzędów sezonowości na bieżące wartości szeregu wpływają jednak poprzednie wartości szeregu, rozdzielone jednym lub większą liczbą okresów sezonowości. Na przykład w przypadku danych miesięcznych (okres sezonowości 12) rząd sezonowości 1 oznacza, że na bieżącą wartość szeregu wpływa 12 okresów wartości szeregu poprzedzających okres bieżący. Rząd sezonowości 1, w przypadku danych miesięcznych, jest wówczas taki sam, jak w przypadku rzędu niesezonowości wynoszącego 12.

Automatycznie wykryj wartości odstające. Wybór tej opcji pozwala automatycznie wykrywać wartości odstające i wybrać jeden lub więcej spośród dostępnych typów wartości odstających.

Typy wykrywanych wartości odstających. Wybierz typ(y) wartości odstających, które chcesz wykrywać. Obsługiwane typy to:

- Addytywne (domyślny)

- Przesunięcie poziomu (domyślny)
- Innowacyjne
- Przemijające
- Sezonowo addytywne
- Trend lokalny
- Wiązka addytywna

Rzędy i transformacje funkcji przenoszenia. Aby określić przekształcenia i zdefiniować funkcje przenoszenia dla dowolnych lub wszystkich zmiennych wejściowych w modelu ARIMA, należy kliknąć przycisk **Ustaw**; zostanie wyświetlone osobne okno dialogowe, w którym można wprowadzić szczegóły dotyczące przenoszenia i przekształcenia.

Dołącz stałe do modelu. Uwzględnienie stałej to praktyka standardowa, o ile użytkownik ma pewność, że ogólna wartość średniej szeregu wynosi 0. Wykluczenie stałej zaleca się w przypadku stosowania różnicowania.

Funkcje przenoszenia i transformacji: Okno dialogowe Rzędy i transformacje funkcji przenoszenia umożliwia określenie przekształceń i zdefiniowanie funkcji przenoszenia dla dowolnych lub wszystkich zmiennych wejściowych w modelu ARIMA.

Transformacja zmiennej przewidywanej. Istnieje możliwość określenia transformacji do wykonania dla każdej zmiennej przewidywanej przed jej zamodelowaniem.

- **Brak.** Nie jest wykonywana żadna transformacja.
- **Pierwiastek kwadratowy.** Wykonywana jest transformacja pierwiastkiem kwadratowym.
- **Logarytm naturalny.** Wykonywana jest transformacja logarytmem naturalnym.

Funkcje przenoszenia i transformacje potencjalnych zmiennych wejściowych. Funkcje przenoszenia umożliwiają określenie sposobu, w jaki wartości z przeszłości zmiennych wejściowych są używane do prognozowania przyszłych wartości przewidywanego szeregu. Lista po lewej stronie panelu przedstawia wszystkie zmienne wejściowe. Pozostałe informacje w tym panelu są specyficzne dla wybranej zmiennej wejściowej.

Polecenia wykonania funkcji transferu. Należy wprowadzić wartości dla różnych składników funkcji przenoszenia do odpowiednich komórek siatki **Struktura** . Wszystkie wartości muszą być nieujemnymi liczbami całkowitymi. W przypadku składników takich jak licznik i mianownik wartość ta reprezentuje rząd maksymalny. W modelu zostaną uwzględnione wszystkie niższe rzędy dodatnie. Ponadto w przypadku składników typu licznik zawsze uwzględniany jest rząd 0. Na przykład w przypadku wskazania dla licznika wartości 2 model będzie obejmował rzędy 2, 1 i 0. W przypadku wskazania wartości 3 dla mianownika model będzie obejmował rzędy 3, 2 i 1. Komórki w kolumnie Sezonowy są aktywne tylko wówczas, jeśli okresowość jest zdefiniowana dla aktywnego zbioru danych.

Licznik. Rząd licznika funkcji przenoszenia określa, które poprzednie wartości z wybranego szeregu niezależnego (predyktora) są używane do przewidywania bieżących wartości szeregu zależnego. Na przykład rząd licznika równy 1 oznacza, że wartość szeregu niezależnego jeden okres wstecz — a także bieżąca wartość szeregu niezależnego — służą do przewidywania bieżącej wartości każdego szeregu zależnego.

Mianownik. Rząd mianownika funkcji przenoszenia określa, jak odchylenia od średniej szeregu, dla poprzednich wartości wybranych szeregów niezależnych (predyktora) są używane do przewidywania bieżących wartości szeregu zależnego. Na przykład rząd mianownika 1 oznacza, że odchylenia od wartości średniej szeregu niezależnego jeden okres wstecz zostaną uwzględnione podczas przewidywania bieżącej wartości każdego szeregu zależnego.

Różnica. Określa rząd różnicowania stosowany względem wybranego szeregu niezależnego (predykcyjnego) przed oszacowaniem modeli. Różnicowanie jest niezbędne w przypadku obecności trendów i służy do usuwania ich wpływu.

Sezonowe. Składniki takie jak licznik, mianownik i różnicowanie odgrywają takie same role, jak ich niesezonowe odpowiedniki. W przypadku rzędów sezonowości na bieżące wartości szeregu wpływają jednak poprzednie wartości szeregu, rozdzielone jednym lub większą liczbą okresów sezonowości. Na przykład w przypadku danych miesięcznych

(okres sezonowości 12) rząd sezonowości 1 oznacza, że na bieżącą wartość szeregu wpływa 12 okresów wartości szeregu poprzedzających okres bieżący. Rząd sezonowości 1, w przypadku danych miesięcznych, jest wówczas taki sam, jak w przypadku rzędu niesezonowości wynoszącego 12.

Opóźnienie. Ustawienie opóźnienia powoduje opóźnienie wpływu zmiennej wejściowej o podaną liczbę przedziałów. Na przykład ustawienie opóźnienia na wartość 5 oznacza, że wartość zmiennej wejściowej w chwili *t* nie wpływa na prognozy aż do chwili upłynięcia pięciu okresów (*t* + 5).

Transformacja. Specyfikacja funkcji przenoszenia dla zestawu zmiennych niezależnych obejmuje także opcjonalną transformację do wykonania na tych zmiennych.

- **Brak.** Nie jest wykonywana żadna transformacja.
- **Pierwiastek kwadratowy.** Wykonywana jest transformacja pierwiastkiem kwadratowym.
- **Logarytm naturalny.** Wykonywana jest transformacja logarytmem naturalnym.

Węzeł Szeregi czasowe — opcje modelu

Szerokość przedziału ufności (%). Przedziały ufności są obliczane dla predykcji modelu i autokorelacji reszt. Można określić dowolną wartość dodatnią mniejszą od 100. Domyślnie ustawiony jest 95% przedział ufności.

Opcja **Rozszerz rekordy na przedziały z przyszłości** pozwala ustawić liczbę przedziałów do prognozowania poza koniec okresu estymacji. Przedział czasowy jest w tym przypadku przedziałem czasowym dla analizy, określonym na karcie Specyfikacja danych. Z chwilą wywołania prognoz następuje automatyczna budowa modeli autoregresji dla szeregów wejściowych niebędących jednocześnie wartościami przewidywanymi. Modele te są następnie używane do generowania wartości dla tych szeregów wejściowych w okresie prognozy. Dla tego ustawienia nie ma maksymalnego limitu.

Wartości przyszłe używane w prognozowaniu

- **Oblicz przyszłe wartości zmiennych wejściowych** Jeśli ta opcja zostanie wybrana, automatycznie obliczane są wartości prognozy dla predykcji, predykcje szumów, oszacowania wariancji i przyszłe wartości czasu. Z chwilą wywołania prognoz następuje automatyczna budowa modeli autoregresji dla szeregów wejściowych niebędących jednocześnie wartościami przewidywanymi. Modele te są następnie używane do generowania wartości dla tych szeregów wejściowych w okresie prognozy.
- **Wybierz zmienne, których wartości mają być dodane do danych.** W przypadku każdego rekordu, jaki ma zostać objęty prognozą (z wyjątkiem wstrzymań), korzystanie ze zmiennych predykcyjnych (z rolą ustawioną na Dane wejściowe) pozwala określić oszacowane wartości dla okresu prognozy każdego predyktora. Wartości można określić ręcznie lub wybrać je z listy.
  - **Zmienna.** Należy kliknąć przycisk Selektor zmiennych i wybrać zmienne, które mogą być używane jako predyktory. Należy pamiętać, że wybrane tutaj zmienne mogą ale nie muszą być używane w modelowaniu; aby rzeczywiście użyć zmiennej jako predyktora, musi być ona wybrana w dalszym węźle modelowania. To okno dialogowe pozwala na określanie przyszłych wartości w wygodny sposób, dzięki czemu mogą one być współużytkowane przez wiele kolejnych węzłów modelowania bez konieczności określania ich w każdym węźle osobno. Należy również pamiętać, że lista dostępnych zmiennych może być ograniczona na skutek opcji wybranych na karcie Opcje budowania.

    Należy pamiętać, że jeśli przyszłe wartości są określone dla zmiennej, które nie jest już dostępna w strumieniu (ponieważ została usunięta lub na karcie Budowanie dokonano aktualizacji dokonanych wyborów), wówczas na karcie Prognoza zmienna jest wyświetlana w kolorze czerwonym.
  - **Wartości.** Dla każdej zmiennej można wybrać funkcje z listy lub kliknąć **Określ**, aby wprowadzić wartości ręcznie lub wybrać je z listy predefiniowanych wartości. Jeśli zmienne predykcyjne odnoszą się do elementów objętych kontrolą użytkownika lub elementów, które są wcześniej znane z innych powodów, wartości należy wprowadzić ręcznie. Przykładowo, jeśli tworzona jest prognoza przychodów hotelu dla kolejnego miesiąca na podstawie liczby zarezerwowanych pokoi, można określić liczbę rezerwacji w rzeczywistości dokonanych dla tego okresu. I odwrotnie, jeśli zmienna predykcyjna odnosi się do elementów poza kontrolą użytkownika, na przykład cena akcji, należy użyć funkcji, takich jak ostatnia wartość lub średnia z ostatnich punktów.

    Dostępne funkcje zależą pod poziomu pomiaru zmiennej.

Tabela 19. Funkcje dostępne dla poziomów pomiaru

Poziom pomiaru	Funkcje
Zmienna ilościowa lub nominalna	Pusta Średnia z ostatnich punktów Ostatnia wartość Określ
Zmienna flagi	Pusta Ostatnia wartość Prawda Fałsz Określ

Średnia z ostatnich punktów umożliwia obliczenie przyszłej wartości na podstawie średniej z trzech ostatnich punktów danych.

Ostatnia wartość ustawia przyszłą wartość na tę z ostatniego punktu danych.

Prawda/Fałsz ustawia przyszłą wartość zmiennej flagi na Prawda lub Fałsz.

Określ otwiera okno dialogowe umożliwiające ręczne określenie przyszłych wartości lub wybranie ich z predefiniowanej listy.

Udostępnij do oceniania

W tym miejscu można ustawić wartości domyślne dla opcji oceniania, które będą widoczne w oknie dialogowym dla modelu użytkowego.

- **Oblicz górny i dolny przedział ufności.** Po zaznaczeniu ta opcja powoduje utworzenie nowych zmiennych (z domyślnymi prefiksami $TSLCI- i $TSUCI-) dla dolnego i górnego przedziału ufności (dla każdej zmiennej przewidywanej).
- **Oblicz reszty szumów.** Zaznaczenie tej opcji powoduje utworzenie nowej zmiennej (z prefiksem domyślnym $TSResidual-) dla reszt modelu dla każdej zmiennej przewidywanej, wraz z sumą tych wartości.

Ustawienia modelu

Maksymalna liczba modeli wyświetlanych w wynikach. Określa maksymalną liczbę modelów, które mają zostać zawarte w danych wynikowych. Należy pamiętać, że jeśli liczba zbudowanych modelów przekracza ten próg, modele nie zostaną pokazane w danych wynikowych, ale nadal będą dostępne do oceny. Wartość standardowa to 10. Wyświetlanie dużej liczby modelów może skutkować słabszą wydajnością lub brakiem stabilności.

Węzeł SMOTE

Węzeł SMOTE (Synthetic Minority Over-sampling Technique — generowanie próbek syntetycznych z klasy mniejszościowej) realizuje algorytm nadpróbkowania przydatny w pracy z niezrównoważonymi zbiorami danych. Udostępnia on zaawansowaną metodę równoważenia danych. Węzeł procesowy SMOTE jest zaimplementowany w języku Python i wymaga biblioteki Python imbalanced-learn©. Szczegółowe informacje o bibliotece imbalanced-learn można znaleźć na stronie http://contrib.scikit-learn.org/imbalanced-learn/about.html[1].

Karta Python na palecie węzłów zawiera węzeł SMOTE i inne węzły Python.

[1]Lemaître, Nogueira, Aridas. "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research*, vol. 18, no. 17, 2017, pp. 1-5. (http://jmlr.org/papers/v18/16-365.html)

Węzeł SMOTE — Ustawienia

Na karcie **Ustawienia** węzła SMOTE zdefiniuj następujące ustawienia.

Ustawienia zmiennej przewidywanej

Zmienna przewidywana. Wybierz zmienną przewidywaną. Obsługiwane są wszystkie typy pomiaru: flaga, nominalne, porządkowe i dyskretne. Jeżeli w sekcji Podział zaznaczona jest opcja **Użyj danych podzielonych na podzbiory**, to nadpróbkowane będą tylko dane uczące.

Współczynnik próbek syntetycznych

Wybierz opcję **Automatycznie**, aby automatycznie wybrać współczynnik próbek syntetycznych, albo wybierz opcję **Ustaw współczynnik (mniejszość do większości)**, aby określić niestandardowy współczynnik. Współczynnik to liczba próbek w klasie mniejszościowej podzielona przez liczbę próbek w klasie większościowej. Wartość współczynnika musi być większa od **0** i mniejsza lub równa **1**.

Wartość startowa generatora liczb losowych

Ustaw wartość początkową generatora liczb losowych. Wybierz tę opcję i kliknij przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Metody

Rodzaj algorytmu. Wybierz typ algorytmu SMOTE, który ma być używany.

Reguły próbek

K-sąsiedzi. Określ liczbę najbliższych sąsiadów, która ma być używana do tworzenia próbek syntetycznych.

M-sąsiedzi. Określ liczbę najbliższych sąsiadów, która ma być używana do określenia, czy próbka mniejszościowa jest zagrożona. Ten parametr jest używany tylko w przypadku wybrania typu algorytmu SMOTE **Borderline1** lub **Borderline2**.

Podział

Użyj danych podzielonych na podzbiory. Wybierz tę opcję, jeśli tylko dane uczące mają być nadpróbkowane.

Węzeł SMOTE wymaga biblioteki Python imbalanced-learn©. W poniższej tabeli przedstawiono relacje między ustawieniami w oknie dialogowym węzła SMOTE w programie SPSS Modeler a parametrami algorytmu w języku Python.

Tabela 20. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Nazwa parametru w interfejsie API środowiska Python
Współczynnik próbek syntetycznych (pole do wprowadzania liczby)	sample_ratio_value	ratio
Wartość początkowa	random_seed	random_state
K-sąsiedzi	k_neighbours	k
M-sąsiedzi	m_neighbours	m
Rodzaj algorytmu	algorithm_kind	kind

Węzeł importowania przez rozszerzenie

Dzięki węzłowi transformacji przez rozszerzenie można pobrać dane ze strumienia programu IBM SPSS Modeler, a następnie zastosować do przekształcenia danych za pomocą skryptów R lub Python for Spark. Zmodyfikowane dane są zwracane do strumienia celem dalszego przetwarzania, utworzenia modelu lub oceny modelu. Węzeł Rozszerzenie Przekształcenia umożliwia przekształcanie danych za pomocą algorytmów napisanych w języku R lub Python for Spark, a tym samym opracowywanie metod przekształcania danych dopasowanych do specyfiki problemu.

Aby używać tego węzła z językiem R, należy zainstalować IBM SPSS Modeler - Essentials for R. Informacje o instalacji i kompatybilności zawiera publikacja *IBM SPSS Modeler - Essentials for R: instrukcja instalacji*. Na komputerze musi być także zainstalowana kompatybilna wersja środowiska R.

Węzeł Rozszerzenie Przekształcenia — karta Polecenia

Wybierz język poleceń: **R** albo **Python for Spark**. Więcej informacji można znaleźć w następujących sekcjach. Gdy polecenia będą gotowe, można kliknąć przycisk **Wykonaj**, aby wykonać węzeł Rozszerzenie Przekształcenia.

Polecenia R

Polecenia R. Do tego pola można wpisać lub wkleić własny skrypt R służący do analizy danych.

Konwertuj zmienne typu flaga. Określa sposób traktowania zmiennych typu flaga. Dostępne są dwie opcje: **Łańcuchy na czynnik, liczby całkowite i rzeczywiste na liczby typu double** oraz **Wartości logiczne (Prawda, Fałsz)**. W przypadku wybrania opcji **Wartości logiczne (Prawda, Fałsz)** pierwotne wartości zmiennych typu flaga zostaną utracone. Na przykład, jeśli zmienna ma wartości Mężczyzna i Kobieta, to zostaną zamienione na Prawda i Fałsz.

Konwertuj brakujące wartości na wartość niedostępności danych (NA) pakietu R. Gdy ta opcja jest wybrana, wszelkie brakujące wartości są przekształcane w wartość NA w języku R. W języku R wartość NA oznacza brakujące wartości. Niektóre funkcje R przyjmują argument sterujący zachowaniem funkcji w przypadku, gdy dane zawierają wartość NA. Na przykład funkcja może oferować opcję automatycznego wykluczania rekordów zawierających wartość NA. Jeśli ta opcja nie będzie wybrana, wszelkie brakujące wartości będą przekazywane do skryptu R bez zmian, co może powodować błędy podczas jego wykonywania.

Konwertuj zmienne daty/czasu na klasy pakietu R ze specjalną kontrolą stref czasowych. Gdy ta opcja jest wybrana, zmienna typu data lub data/czas są przekształcane w obiekty date/time języka R. Należy wybrać jedną z następujących opcji:

- **R POSIXct.** Zmienne typu data lub data/czas są przekształcane w obiekty POSIXct języka R.
- **R POSIXlt (lista).** Zmienne typu data lub data/czas są przekształcane w obiekty POSIXlt języka R.

Uwaga: Formaty POSIX są opcjami zaawansowanymi. Opcji tych należy używać tylko wtedy, gdy w skrypcie R nakazano traktowanie zmiennych daty/czasu w sposób wymagający zastosowania tych formatów. Formaty POSIX nie mają zastosowania względem zmiennych z formatami czasu.

Polecenia Python

Polecenia Python. Do tego pola można wpisać lub wkleić własny skrypt Python służący do analizy danych. Aby uzyskać więcej informacji na temat języka Python for Spark, patrz Python for Spark i Pisanie skryptów w języku Python for Spark.

Węzeł Rozszerzenie Przekształcenia — karta Wynik z konsoli

Karta **Wynik z konsoli** zawiera wszelkie wyniki odbierane podczas wykonywania skryptu w języku R lub Python for Spark na karcie Polecenia (na przykład, jeśli używany jest skrypt R, to na tej karcie wyświetlane są wyniki odbierane z konsoli R podczas wykonywania skryptu z pola **Polecenia R** na karcie **Polecenie**). Wyniki te mogą zawierać

komunikaty o błędach lub ostrzeżenia generowane podczas wykonywania skryptu w języku R lub Python. Wyniki można wykorzystać przede wszystkim do debugowania skryptu. Karta **Wynik z konsoli** zawiera także skrypt z pola **Polecenia R** lub **Polecenia Python**.

Po każdym wykonaniu skryptu Rozszerzenie Przekształcenia zawartość karty **Wynik z konsoli** jest nadpisywana wynikami z konsoli R lub środowiska Python for Spark. Wyników nie można edytować.

Węzeł Siatka czasoprzestrzeni

Siatka czasoprzestrzeni stanowi rozszerzenie geokodowanych lokalizacji przestrzennych. Mówiąc dokładniej, siatka czasoprzestrzeni to łańcuch alfanumeryczny reprezentujący obszar czasu i przestrzeni o regularnych kształtach.

Przykładowo, siatka czasoprzestrzeni **dr5ru7|2013-01-01 00:00:00|2013-01-01 00:15:00** składa się z następujących trzech części:

- Geokod **dr5ru7**
- Początkowy znacznik czasu **2013-01-01 00:00:00**
- Końcowy znacznik czasu **2013-01-01 00:15:00**

Przykładowo można użyć informacji dotyczących przestrzeni i czasu w celu zwiększenia ufności, że dwie jednostki są takie same, ponieważ wirtualnie znajdują się w tym samym miejscu w tym samym czasie. Alternatywnie można zwiększyć dokładność identyfikacji relacji, pokazując, że dwie jednostki są powiązane ze względu na ich bliskość w przestrzeni i czasie.

Można wybrać tryb **Indywidualne rekordy** lub **Przebywania**, odpowiednio do wymagań. Oba tryby wymagają tych samych podstawowych danych szczegółowych, takich jak:

Szerokość geograficzna. Należy wybrać zmienną, która identyfikuje szerokość geograficzną (w układzie współrzędnych WGS84).

Długość geograficzna. Należy wybrać zmienną, która identyfikuje długość geograficzną (w układzie współrzędnych WGS84).

Znacznik czasu. Należy wybrać zmienną, która identyfikuje godzinę lub datę.

Opcje indywidualnych rekordów

Tego trybu należy użyć, aby dodać dodatkową zmienną do rekordu w celu określenia lokalizacji w danym czasie.

Wyliczanie. Należy wybrać do najmniej jedną gęstość dla przestrzeni i czasu, na podstawie której wyliczona zostanie nowa zmienna. Więcej informacji można znaleźć w “Definiowanie gęstości siatki czasoprzestrzeni” na stronie 112.

Rozszerzenie nazwy zmiennej. Należy wpisać rozszerzenie, jakie ma zostać dodane do nazw nowych zmiennych. Można wybrać dodanie rozszerzenia jako **Przyrostek** lub **Przedrostek**.

Opcje przebywania

Przebywanie można określić jako lokalizację i/lub czas, w jakich jednostka stale lub często przebywa. Przykładowo, funkcja ta może zostać użyta do określenia pojazdu, który systematycznie przewozi towar, oraz wszelkich odchyleń od normy.

Detektor przebywania monitoruje ruch jednostek i oznacza warunki, w jakich jednostka „zatrzymuje się” na danym obszarze. Detektor przebywania automatycznie przypisuje oznakowane „przebywanie” w siatce czasoprzestrzeni i wykorzystuje funkcję śledzenia jednostki i zdarzenia w pamięci w celu wykrycia przebywania o optymalnej efektywności.

Gęstość siatki czasoprzestrzeni. Należy wybrać gęstość dla przestrzeni i czasu, na podstawie której wyliczona zostanie nowa zmienna. Przykładowo wartość **STB_GH4_10MINS** będzie odpowiadała czteroznakowemu polu geokodu o wielkości około 20 km na 20 km i 10-minutowemu oknu czasowemu. Więcej informacji można znaleźć w “Definiowanie gęstości siatki czasoprzestrzeni” na stronie 112.

Identyfikacja jednostki. Należy wybrać jednostkę, jaka będzie używana jako identyfikator przebywania. Ta zmienna identyfikacyjna określa zdarzenie.

Minimalna liczba zdarzeń. Zdarzenie jest wierszem danych. Należy wybrać minimalną liczbę wystąpień zdarzenia dla jednostki, aby została uznana jako przebywająca w danym punkcie. Przebywanie musi być również zakwalifikowane na podstawie zmiennej **Czas pozostawania wynosi co najmniej**.

Czas pozostawania wynosi co najmniej. Należy określić minimalny czas trwania, w jakim jednostka musi pozostawać w tej samej lokalizacji. Pozwoli to na przykład wykluczyć uznanie postoju samochodu na światłach za przebywanie. Przebywanie musi być również zakwalifikowane na podstawie wcześniejszej zmiennej **Minimalna liczba zdarzeń**.

Poniżej bardziej szczegółowo omówiono, na jakiej podstawie zdarzenie jest kwalifikowane jako przebywanie:

Niech e_1, ..., e_n oznacza wszystkie uporządkowane według czasu zdarzenia dla danego id. jednostki w określonym czasie trwania (t_1, t_n). Zdarzenia te zostaną zakwalifikowane jako przebywanie, jeśli:

- $n >=$ *minimalna liczba zdarzeń*
- $t_n - t_1 >=$ *minimalny czas pozostawania*
- Wszystkie zdarzenia e_1, ..., e_n występują w tej samej siatce geoprzestrzeni

Rozszerz zasięg przebywania poza sztywne granice siatki czasoprzestrzeni. Jeśli ta opcja zostanie wybrana, definicja przebywania jest mniej ścisła i może na przykład obejmować jednostkę, które przebywa w więcej niż jednej siatce czasoprzestrzeni. Przykładowo, jeśli siatki czasoprzestrzeni są zdefiniowane jako pełne godziny, po zaznaczeniu tej opcji jednostka zostanie rozpoznana jako przebywająca przez godzinę, nawet jeśli godzina obejmowała 30 minut przed północą i 30 minut po północy. Jeśli ta opcja nie zostanie wybrana, 100% czasu przebywania musi przypadać w pojedynczej siatce czasoprzestrzeni.

Min. proporcja zdarzeń w kwalifikującym oknie czasowym (%). Ta opcja jest dostępna tylko po wybraniu opcji **Rozszerz zasięg przebywania poza sztywne granice siatki czasoprzestrzeni**. Należy jej użyć, aby kontrolować stopień, w jakim przebywanie wskazane w siatce czasoprzestrzeni może w rzeczywistości nakładać się na inne. Należy wybrać minimalną proporcję zdarzeń, jaka jest wymagana w pojedynczej siatce czasoprzestrzeni, aby zidentyfikować przebywanie. Jeśli wartość zostanie ustawiona jako 25%, a proporcja zdarzeń wynosi 26%, spowoduje to zakwalifikowanie jako przebywanie.

Załóżmy na przykład, że detektor przebywania został skonfigurowany, tak aby wymagane były co najmniej dwa zdarzenia (minimalna liczba zdarzeń = 2) oraz czas ciągłego zatrzymania wynoszący co najmniej 2 minuty dla pola przestrzeni z 4-bajtowym geokodem i 10-minutowego okna czasu (STB_NAME = STB_GH4_10MINS). Po wykryciu przebywania jednostka pozostaje w tym samym polu przestrzeni z 4-bajtowym geokodem, podczas gdy w ciągu 10 minut, od godziny 16:57 do 17:07, występują trzy zdarzenia kwalifikujące (o 16:58, 17:01 i 17:03). Wartość procentowa kwalifikującego okna czasu określa siatkę czasoprzestrzeni, która może zostać uznana jako przebywanie, w następujący sposób:

- **100%.** Przebywanie jest zgłaszane w oknie czasowym od godziny 17:00 do 17:10, a nie od godziny 16:50 do 17:00 (zdarzenia z godziny 17:01 i 17:03 spełniają wszystkie warunki wymagane dla zakwalifikowania przebywania i 100% tych zdarzeń wystąpiło w oknie czasowym od godziny 17:00 do 17:10).
- **50%.** Zgłaszane są przebywania w obu oknach czasowych (zdarzenia z godziny 17:01 i 17:03 spełniają wszystkie warunki wymagane do zakwalifikowania przebywania oraz co najmniej 50% z tych zdarzeń wystąpiło w oknie czasowym od godziny 16:50 do 17:00 i co najmniej 50% z nich w oknie czasowym od godziny 17:00 do 17:10).
- **0%.** Zgłaszane są przebywania w obu oknach czasowych.

Jeśli określona jest wartość 0%, raporty dot. przebywania obejmują siatki czasoprzestrzeni reprezentujące każde okno czasowe, na które nachodzi czas trwania kwalifikowania. Czas trwania kwalifikowania musi krótszy lub taki sam jak odpowiedni czas trwania okna czasowego w siatce czasoprzestrzeni. Innymi słowy, nigdy nie należy tworzyć konfiguracji, w których 10-minutowa siatka czasoprzestrzeni jest konfigurowana w zestawieniu z 20-minutowym czasem trwania kwalifikowania.

Przebywanie jest zgłaszane niezwłocznie po spełnieniu warunków i nie jest zgłaszane więcej niż jeden raz dla danej siatki czasoprzestrzeni. Załóżmy, że trzy zdarzenia powodują zakwalifikowanie jako przebywanie, a dla czasu trwania kwalifikowania w sumie wystąpiło 10 zdarzeń, wszystkie w tej samej siatce czasoprzestrzeni. W takim przypadku przebywanie zostaje zgłoszone po wystąpieniu trzeciego zdarzenia kwalifikującego. Żadne z pozostałych siedmiu zdarzeń nie uruchomi raportu przebywania.

Uwaga:

- Dane o zdarzeniach z pamięci detektora przebywania nie są udostępniane w innych procesach. Dlatego poszczególne jednostki są spowinowacone z konkretnym węzłem detektora przebywania (afiniczność jednostki). Oznacza to, że wejściowe dane dotyczące ruchu jednostki zawsze muszą być w sposób spójny przekazywane do węzła detektora przebywania śledzącego jednostkę, który jest zwykle tym samym węzłem w całym przebiegu.
- Dane dotyczące zdarzenia z pamięci detektora przebywania są nietrwałe. Po każdym zamknięciu i ponownym uruchomieniu detektora przebywania następuje utrata niezakończonych danych dot. przebywania. Oznacza to, że zatrzymanie i ponownie uruchomienie procesu może spowodować w systemie utworzenie błędnych raportów dotyczących rzeczywistego przebywania. Możliwym środkiem zaradczym będzie odtworzenie niektórych historycznych danych związanych z ruchem (na przykład cofnięcie się o 48 godzin i odtworzenie rekordów ruchu mających zastosowanie do dowolnego węzła, który został ponownie uruchomiony).
- Detektor przebywania musi otrzymywać dane w porządku zachowującym kolejność w czasie.

Definiowanie gęstości siatki czasoprzestrzeni

Należy wybrać wielkość (gęstość) siatek czasoprzestrzeni poprzez określenie fizycznego obszaru i czasu, jaki upłynął, w celu uwzględnienia w każdej z nich.

Gęstość obszarowa. Należy wybrać wielkość obszaru, jaki ma zostać uwzględniony w każdej siatce czasoprzestrzeni.

Przedział czasowy. Należy wybrać liczbę godzin, jaka ma zostać uwzględniona w każdej siatce czasoprzestrzeni.

Nazwa zmiennej. To pole jest automatycznie wypełniane na podstawie wyboru dokonanego w poprzednich dwóch polach, z dodaniem przedrostka STB.

Węzeł Strumień TCM

Węzeł Strumień TCM umożliwia budowanie i ocenianie modeli przyczynowych szeregów czasowych w jednym kroku.

Więcej informacji na temat modelowania przyczynowego szeregów czasowych zawiera temat Modele przyczynowe szeregów czasowych w sekcji Modele szeregów czasowych podręcznika SPSS Modeler — węzły modelowania.

Węzeł Strumień TCM — opcje szeregów czasowych

Na karcie Zmienne należy użyć ustawień **Szeregi czasowe**, aby określić szeregi, jakie będą uwzględniane w systemie modelu.

Należy wybrać opcję dla struktury danych, która jest odpowiednia dla przetwarzanych danych. W przypadku danych wielowymiarowych należy kliknąć opcję **Wybierz wymiary**, aby określić zmienne wymiarów. Kolejność określona w zmiennych wymiaru definiuje kolejność, w jakiej będą wyświetlane wszystkie następne okna dialogowe i wyniki. Kolejność zmiennych wymiarów można zmieniać za pomocą przycisków strzałek w górę i w dół w podrzędnym oknie dialogowym Wybierz wymiary.

W przypadku danych kolumnowych pojęcie *szeregi* ma takie samo znaczenie jak pojęcie *zmienna*. W przypadku danych wielowymiarowych zmienne, które zawierają szeregi czasowe, są nazywane zmiennymi *metryk*. Szeregi czasowe, dla danych wielowymiarowych, są definiowane przez zmienną metryk i wartość dla każdej zmiennej wymiaru. Przedstawione poniżej rozważania mają zastosowanie do danych kolumnowych oraz do danych wielowymiarowych.

- Szeregi określane jako potencjalne zmienne wejściowe oraz jako przewidywane i wejściowe są uwzględniane w celu dołączenia do modelu dla każdej zmiennej przewidywanej. Model dla każdej zmiennej przewidywanej zawsze uwzględnia wartości opóźnione samej zmiennej przewidywanej.
- Szeregi określane jako wymuszone zmienne wejściowe zawsze są uwzględniane w modelu dla każdej zmiennej przewidywanej.
- Co najmniej jeden szereg musi być określony jako przewidywany lub jako przewidywany i wejściowy.
- Jeśli zaznaczono opcję **Użyj wstępnie zdefiniowanych ról**, zmienne z rolą Zmienna wejściowa są ustawiane jako potencjalne zmienne wejściowe. Żadna wstępnie zdefiniowana rola nie jest odwzorowywana na potencjalną zmienną wejściową.

Dane wielowymiarowe

W przypadku danych wielowymiarowych zmienne metryk i powiązane role są określane w siatce, w której każdy wiersz określa pojedynczą metrykę lub rolę. Domyślnie system modelu obejmuje szeregi dla wszystkich kombinacji zmiennych wymiarów dla każdego wiersza w siatce. Jeśli na przykład dostępne są wymiary dla *region* i *brand*, domyślnie określenie metryki *sales* jako przewidywanej oznacza, że istnieje osobny szereg przewidywany sprzedaży dla każdej kombinacji wartości *region* i *brand*.

Dla każdego wiersza w siatce można dostosować zestaw wartości dowolnej zmiennej wymiaru, klikając przycisk wielokropka danego wymiaru. Ta czynność otworzy podrzędne okno dialogowe Wybierz wartości wymiarów. Wiersze siatki można również dodawać, usuwać lub kopiować.

W kolumnie **Liczebność szeregu** wyświetlana jest liczba zestawów wartości wymiarów, jakie są aktualnie określone dla powiązanej metryki. Wyświetlana wartość może być większa niż rzeczywista liczba szeregów (jeden szereg na zestaw). Taka sytuacja ma miejsce, jeśli niektóre z określonych kombinacji wartości wymiarów nie odpowiadają szeregom objętym przez powiązaną metrykę.

Węzeł Strumień TCM — wybór wartości wymiarów

W przypadku danych wielowymiarowych można dostosować analizę poprzez określenie, które wartości wymiarów mają zastosowanie do konkretnej zmiennej metryki dla konkretnej roli. Na przykład, jeśli *sales* jest zmienną metryki a *channel* jest wymiarem z wartościami 'retail' i 'web', można określić, że 'web' sales jest zmienną wejściową, a 'retail' sales jest zmienną przewidywaną. Można również określić podzbiory wymiarów, jakie będą miały zastosowanie do wszystkich zmiennych metryk użytych w analizie. Na przykład, jeśli *region* jest zmienną wymiaru, która określa region geograficzny, wówczas można ograniczyć analizę do konkretnych regionów.

Wszystkie wartości
Ta opcja określa, że uwzględniane są wszystkie wartości bieżącej zmiennej wymiaru. Jest to opcja domyślna.

Wybierz wartości do uwzględnienia lub wykluczenia
Tej opcji należy użyć, aby określić zbiory wartości dla bieżącej zmiennej wymiaru. Jeśli dla opcji **Tryb** wybrana jest wartość **Uwzględnij**, uwzględniane będą tylko wartości określone na liście **Wybrane wartości**. Jeśli dla opcji **Tryb** wybrana jest wartość **Wyklucz**, wówczas uwzględniane są wszystkie wartości inne niż te określone na liście **Wybrane wartości**.

Wartości, z których można dokonywać wyboru, można filtrować. Wartości, które spełniają warunki filtrowania, są wyświetlane na karcie **Dopasowane**, a wartości, które nie spełniają warunków filtrowania, są wyświetlane na karcie **Niedopasowane** w postaci listy **Niewybrane wartości**. Karta **Wszystkie** zawiera listę wszystkich niewybranych wartości, niezależnie od warunków filtrowania.

- Podczas określania filtru można użyć gwiazdki (*) jako symbolu wieloznacznego.
- Aby wyczyścić bieżący filtr, dla poszukiwanego terminu w oknie dialogowym Filtruj wyświetlane wartości należy wprowadzić pustą wartość.

Węzeł Strumień TCM — opcje obserwacji

Na karcie Zmienne należy użyć ustawień **Obserwacje**, aby określić zmienne definiujące obserwacje.

Obserwacje definiowane jako data/czas

Można określić, czy obserwacje będą definiowane na podstawie zmiennej daty, czasu lub daty/czasu. Oprócz zmiennej, która definiuje obserwacje, należy wybrać odpowiedni przedział czasowy, w którym będą opisywane obserwacje. W zależności od określonego przedziału czasowego można również dokonać innych ustawień, takich jak przedział między obserwacjami (przyrost) lub liczba dni w tygodniu. Poniższe stwierdzenia mają zastosowanie do przedziałów czasowych:

- Wartości **Nieregularny** należy użyć, jeśli obserwacje są nierównomiernie rozmieszczone w czasie, na przykład są wykonywane w chwili, gdy następuje przetwarzanie zamówienia sprzedaży. Jeśli wybrana zostanie opcja **Nieregularny**, należy określić przedział czasowy, jaki będzie używany do analizy, wybierając ustawienia **Przedział czasowy** na karcie Specyfikacja danych.
- Jeśli obserwacje reprezentują datę i czas, a przedział czasowy to godziny, minuty lub sekundy, wówczas należy wybrać opcję **Godziny dziennie**, **Minuty dziennie** lub **Sekundy dziennie**. Jeśli obserwacje reprezentują czas (trwanie) bez odniesienia do daty, a przedział czasowy to godziny, minuty lub sekundy, wówczas należy użyć opcji **Godziny (nieokresowo)**, **Minuty (nieokresowo)** lub **Sekundy (nieokresowo)**.
- Na podstawie wybranego przedziału czasowego procedura może wykryć brakujące obserwacje. Wykrywanie brakujących obserwacji jest konieczne, ponieważ procedura zakłada, że wszystkie obserwacje są równomiernie rozłożone w czasie i że nie ma brakujących obserwacji. Na przykład, jeśli przedziałem czasu są dni, a po dacie 2014-10-27 występuje 2014-10-29, istnieje brakująca obserwacja dla 2014-10-28. Dla wszystkich brakujących obserwacji wprowadzane są wartości. Ustawienia obsługi braków danych można wprowadzić na karcie Specyfikacja danych.
- Określone przedziały czasowe umożliwiają procedurze wykrycie wielu obserwacji w jednym przedziale czasowym, które muszą być zagregowane i które są dopasowane do obserwacji dla granicy przedziału (np. pierwszy dzień miesiąca), dzięki czemu obserwacje będą równomiernie rozmieszczone. Na przykład, jeśli przedziałem czasu są miesiące, to zagregowanych zostanie wiele dat z tego samego miesiąca. Ten typ agregacji jest nazywany *grupowaniem*. Domyślnie obserwacje są sumowane podczas grupowania. Można określić inną metodę grupowania, np. średnia z obserwacji; ustawienia **Agregacja i rozkład** można wprowadzić na karcie Specyfikacja danych.
- Przy niektórych przedziałach czasowych istnieją dodatkowe ustawienia umożliwiające zdefiniowanie odstępów w zwykle równomiernie rozmieszczonych przedziałach. Na przykład, jeśli przedziałem czasu są dni, ale istotne są tylko dni robocze, można określić pięciodniowy tydzień rozpoczynający się od poniedziałku.

Obserwacje zdefiniowane jako okresy lub okresy cykliczne

Obserwacje mogą być zdefiniowane przez co najmniej jedną zmienną całkowitą, która reprezentuje okresy lub powtarzające się cyklicznie okresy, aż do dowolnej liczby poziomów cyklicznych. Taka struktura pozwala opisać serie obserwacji, które nie pasują do jednego ze standardowych przedziałów czasowych. Przykładowo, rok fiskalny trwający tylko 10 miesięcy może być opisany przez zmienną cyklu, która reprezentuje lata, i zmienną okresu, która reprezentuje miesiące, przy czym długość jednego cyklu wynosi 10.

Zmienne, które określają okresy cykliczne, definiują hierarchię poziomów okresowości, w której najniższy poziom jest definiowany przez zmienną **Okres**. Następny wyższy poziom jest określany przez zmienną cyklu z poziomem 1, po której następuje zmienna cyklu z poziomem 2 itd. Wartości zmiennych dla każdego poziomu, z wyjątkiem najwyższego, muszą być okresowe w odniesieniu do kolejnego najwyższego poziomu. Wartości dla najwyższego poziomu nie mogą być okresowe. Na przykład, dla 10-miesięcznego roku fiskalnego miesiące występują okresowo w latach, ale lata nie są okresowe.

- Długość cyklu na poszczególnych poziomach stanowi okresowość dla kolejnego najniższego poziomu. W przykładzie dot. roku fiskalnego istnieje tylko jeden poziom cyklu, a długość cyklu wynosi 10, ponieważ kolejny najniższy poziom reprezentuje miesiące, a w określonym roku fiskalnym jest 10 miesięcy.
- Należy określić wartość początkową dla każdej zmiennej okresowej, która nie rozpoczyna się od 1. To ustawienie jest niezbędne dla wykrywania braków danych. Na przykład, jeśli zmienna okresowa

rozpoczyna się od 2, ale wartość początkowa jest określona jako 1, wówczas procedura zakłada, że istnieje brak danych dla pierwszego okresu w każdym cyklu dla tej zmiennej.

Węzeł Strumień TCM — opcje przedziałów czasowych

Przedział czasowy, który jest używany do analizy, może różnić się od przedziału czasowego dla obserwacji. Na przykład, jeśli przedział czasowy obserwacji to dni, jako przedział czasowy dla analizy można wybrać miesiące. Wówczas przed zbudowaniem modelu dane agregowane są z dziennych na miesięczne. Można również rozłożyć dane z dłuższego przedziału czasu na krótszy. Przykładowo, jeśli obserwacje są przeprowadzane kwartalnie, wówczas można rozłożyć dane z kwartalnych na miesięczne.

Opcje możliwe do wyboru dla przedziału czasowego, w jakim wykonywana jest analiza, zależą od sposobu zdefiniowania obserwacji oraz wyznaczonego dla nich przedziału czasowego. W szczególności, jeśli obserwacje są zdefiniowane przez okresy cykliczne, wówczas obsługiwana jest tylko agregacja. W takim przypadku przedział czasowy dla analizy musi być większy od przedziału czasowego dla obserwacji lub mu równy.

Przedział czasowy dla analizy można określić w ustawieniach **Przedział czasowy** na karcie Specyfikacja danych. Metoda, którą dane są agregowane lub rozkładane, jest określana w ustawieniach **Agregacja i rozkład** na karcie Specyfikacja danych.

Węzeł Strumień TCM — opcje agregacji i rozkładu

Funkcje agregacji

Jeśli przedział czasowy użyty dla analizy jest dłuższy niż przedział czasowy dla obserwacji, dane wejściowe zostają zagregowane. Przykładowo agregacja jest przeprowadzana, kiedy przedział czasowy dla obserwacji to dni, a przedział czasowy dla analizy to miesiące. Dostępne są następujące funkcje agregacji: średnia, suma, dominanta, wartość minimalna lub maksymalna.

Funkcje rozkładu

Jeśli przedział czasowy użyty dla analizy jest krótszy niż przedział czasowy dla obserwacji, dane wejściowe zostają rozłożone. Przykładowo rozkład jest przeprowadzany, kiedy przedział czasowy dla obserwacji to kwartały, a przedział czasowy dla analizy to miesiące. Dostępne są następujące funkcje rozkładu: średnia lub suma.

Funkcje grupujące

Grupowanie jest stosowane, kiedy obserwacje są definiowane przez datę/czas i w tym samym przedziale czasowym występuje wiele obserwacji. Na przykład, jeśli przedział czasowy dla obserwacji to miesiące, wówczas wiele dat z tego samego miesiąca jest grupowanych i tworzone jest ich powiązanie z miesiącem, w którym występują. Dostępne są następujące funkcje grupowania: średnia, suma, dominanta, wartość minimalna lub maksymalna. Grupowanie jest zawsze przeprowadzane, kiedy obserwacje są zdefiniowane przez datę/czas, a przedział czasowy dla obserwacji jest określony jako Nieregularny.

Uwaga: Chociaż grupowanie jest formą agregacji, jest przeprowadzane przed rozpoczęciem obsługi braków danych, podczas gdy formalna agregacja jest wykonywana po zakończeniu obsługi braków danych. Jeśli przedział czasowy dla obserwacji jest określony jako Nieregularny, agregacja jest wykonywana tylko za pomocą funkcji grupowania.

Agreguj obserwacje przekraczające granice dnia do dnia poprzedniego

Określa, czy obserwacje, których czas przekracza granicę dnia, są agregowane na wartości dla dnia poprzedniego. Przykładowo, dla obserwacji godzinowych trwających osiem godzin dziennie i rozpoczynających się o godzinie 20:00, to ustawienie określi, czy obserwacje od godziny 00:00 do 04:00 będą uwzględniane w zagregowanych wynikach dla poprzedniego dnia. To ustawienie ma zastosowanie tylko w przypadku, kiedy przedział czasowy dla obserwacji to Godziny dziennie, Minuty dziennie lub Sekundy dziennie, a przedział czasowy dla analizy to Dni.

Ustawienia niestandardowe dla określonych zmiennych

Funkcje agregacji, rozkładu i grupowania dla zmiennej można określić na podstawie zmiennej. Ustawienia te zastępują domyślne ustawienia funkcji agregacji, rozkładu i grupowania.

Węzeł Strumień TCM — opcje braków danych

Brakujące wartości w danych wejściowych są zastępowane przez wartość podstawianą. Dostępne są następujące metody zastępowania:

Interpolacja liniowa
Powoduje zastąpienie braków danych przy wykorzystaniu interpolacji liniowej. W interpolacji używana jest ostatnia ważna wartość przed brakiem danych oraz pierwsza ważna za brakiem. Jeśli pierwsza lub ostatnia obserwacja w szeregu zawiera brakujące wartości, wówczas używane są dwie najbliższe niebrakujące wartości na początku i na końcu serii.

Średnia szeregu
Zastępuje braki danych średnią obliczoną ze wszystkich obserwacji.

Średnia z sąsiednich punktów
Powoduje zastąpienie braków danych średnią z ważnych wartości sąsiednich. Rozpiętość sąsiednich punktów, to liczba poprawnych wartości występujących przed brakującą wartością i po niej, jakie są wykorzystywane do obliczenia średniej.

Mediana z sąsiednich punktów
Powoduje zastąpienie braków danych medianą ważnych wartości sąsiednich. Rozpiętość sąsiednich punktów, to liczba poprawnych wartości występujących przed brakującą wartością i po niej, jakie są wykorzystywane do obliczenia mediany.

Trend liniowy
Ta opcja wykorzystuje niebrakujące obserwacje w szeregu do dopasowania prostego modelu regresji liniowej, który jest następnie używany w celu przypisania brakujących wartości.

Inne ustawienia:

Maksymalny procent braków danych (%)
Określa maksymalną wartość procentową braków danych, jaka jest dozwolona w szeregu. Szeregi z większą liczbą braków danych od określonego maksimum są wykluczane z analizy.

Węzeł Strumień TCM — ogólne opcje danych

Maksymalna liczba odrębnych wartości na zmienną wymiaru
To ustawienie dotyczy danych wielowymiarowych i określa maksymalną liczbę odrębnych wartości, jaka jest dozwolona dla dowolnej zmiennej wymiaru. Domyślnie ograniczenie jest ustawione na 10000, ale wartość tę można zmienić na dowolnie dużą liczbę.

Węzeł Strumień TCM — ogólne opcje budowania

Szerokość przedziału ufności (%)
To ustawienie decyduje o przedziałach ufności dla prognoz i parametrów modelu. Można określić dowolną wartość dodatnią mniejszą od 100. Domyślnie ustawiony jest 95-procentowy przedział ufności.

Maksymalna liczba zmiennych wejściowych na zmienną przewidywaną
To ustawienie określa maksymalną liczbę zmiennych wejściowych, jaka jest dozwolona w modelu dla każdej zmiennej przewidywanej. Można określić liczbę całkowitą z zakresu od 1 do 20. Model dla każdej zmiennej przewidywanej zawsze uwzględnia wartości opóźnione samej zmiennej przewidywanej; ustawienie tej wartości na 1 spowoduje, że tylko zmienna wejściowa jest samą zmienną przewidywaną.

Tolerancja modelu
To ustawienie kontroluje proces iteracyjny, jaki jest stosowany do określenia najlepszego zestawu zmiennych wejściowych dla każdej zmiennej przewidywanej. Można określić dowolną wartość większą od zera. Domyślną wartością jest 0,001. Tolerancja modelu jest kryterium zatrzymania wyboru predyktorów. Może wpłynąć na liczbę predyktorów uwzględnionych w ostatecznym modelu. Jeśli jednak zmienna przewidywana bardzo dobrze przewiduje sama siebie, to pozostałe predyktory mogą nie zostać uwzględnione w ostatecznym modelu. Konieczne może być postępowanie metodą prób i błędów (np. jeśli wartość tego ustawienia jest wysoka, można ją zmniejszyć, aby sprawdzić, czy inne predyktory mogą zostać uwzględnione, czy nie).

Próg (%) dla wartości odstającej
Obserwacja jest oznaczana jako odstająca, jeśli obliczone wg modelu prawdopodobieństwo, że jest odstająca, przekracza wyznaczony próg. Można określić wartość z zakresu od 50 do 100.

Liczba opóźnień dla każdej zmiennej wejściowej
To ustawienie określa liczbę opóźnień dla każdej zmiennej wejściowej w modelu dla każdej zmiennej przewidywanej. Domyślnie liczba opóźnień jest określana automatycznie na podstawie przedziału czasowego używanego do analizy. Przykładowo, jeśli przedział czasowy to miesiące (z przyrostem o jeden miesiąc), wówczas liczba opóźnień wynosi 12. Opcjonalnie można jawnie określić liczbę opóźnień. Podana wartość musi być liczbą całkowitą z zakresu od 1 do 20.

Kontynuuj oszacowanie, używając istniejących modeli
Jeśli wygenerowano już model przyczynowy szeregów czasowych, tę opcję należy wybrać, aby zamiast budowania nowego modelu, ponownie użyć ustawień kryteriów, jakie są określone dla tego modelu. Można w ten sposób zaoszczędzić czas, ponownie oszacowując i tworząc nową prognozę w oparciu o te same ustawienia modelu co wcześniej, lecz na podstawie bardziej aktualnych danych.

Węzeł Strumień TCM — opcje okresu estymacji

Domyślnie okres estymacji zaczyna się od czasu z najwcześniejszą obserwacją, a kończy w czasie z najpóźniejszą obserwacją we wszystkich szeregach.

Wyznaczony przez czas początkowy i końcowy
Można określić datę rozpoczęcia i zakończenia okresu estymacji lub można określić tylko datę rozpoczęcia lub tylko datę zakończenia. Jeśli rozpoczęcie lub zakończenie okresu estymacji zostanie pominięte, użyta będzie wartość domyślna.

- Jeśli obserwacje są zdefiniowane przez określenie zmiennej daty/czasu, wówczas wartości rozpoczęcia i zakończenia okresu należy wprowadzić w takim samym formacie, jaki został użyty dla zmiennej daty/czasu.
- W przypadku obserwacji definiowanych na podstawie okresów cyklicznych należy określić wartość dla każdej zmiennej okresu cyklicznego. Każda zmienna jest wyświetlana w osobnej kolumnie.

Wyznaczony przez najwcześniejszy lub najpóźniejszy przedział czasowy

Definiuje okres estymacji jako określoną liczbę przedziałów czasowych, która rozpoczyna się od najwcześniejszego przedziału czasowego lub kończy na najpóźniejszym przedziale czasowym określonym w danych, z opcjonalnym przesunięciem. W tym kontekście przedział czasowy odnosi się do przedziału czasowego dla analizy. Przykładowo, załóżmy, że obserwacje są przeprowadzane miesięcznie, ale przedział czasowy dla analizy to kwartały. Określenie wartości **Najpóźniejszy** i wartości 24 dla opcji **Liczba przedziałów czasowych** będzie oznaczało ostatnie 24 kwartały.

Opcjonalnie można wykluczyć określoną liczbę przedziałów czasowych. Przykładowo, określenie ostatnich 24 przedziałów czasowych i 1 do wykluczenia oznacza, że okres oszacowania składa się z 24 przedziałów, które poprzedzają ostatni.

Węzeł Strumień TCM — opcje modelu

Nazwa modelu
Można określić niestandardową nazwę modelu lub zaakceptować nazwę wygenerowaną automatycznie, czyli *TCM*.

Prognoza
Opcja **Rozszerz rekordy na przedziały z przyszłości** pozwala ustawić liczbę przedziałów do prognozowania poza koniec okresu estymacji. Przedział czasowy jest w tym przypadku przedziałem czasowym dla analizy, określonym na karcie Specyfikacja danych. Z chwilą wywołania prognoz następuje automatyczna budowa modeli autoregresji dla szeregów wejściowych niebędących jednocześnie wartościami przewidywanymi. Modele te są następnie używane do generowania wartości dla tych szeregów wejściowych w okresie prognozy. Dla tego ustawienia nie ma maksymalnego limitu.

węzeł optymalizacji CPLEX

Węzeł optymalizacji CPLEX zapewnia możliwość korzystania z zaawansowanej optymalizacji matematycznej (CPLEX) za pośrednictwem pliku modelu OPL (Optimization Programming Language). Ta funkcja jest dostępna w produkcie IBM Analytical Decision Management, jednak teraz węzła CPLEX można również używać w programie SPSS Modeler bez IBM Analytical Decision Management.

Więcej informacji na temat optymalizacji CPLEX i OPL zawiera dokumentacja programu IBM ILOG CPLEX Optimization Studio.

Węzeł Optymalizacja CPLEX obsługuje wiele źródeł danych lub wielowymiarowe dane przychodzące. Do węzła Optymalizacja CPLEX można podłączyć kilka węzłów, a każdy poprzedni węzeł może dostarczać dane do obliczeń modelu OPL — jako odrębny zbiór krotek z odrębnym odwzorowaniem zmiennych.

W wynikach generowanych przez węzeł Optymalizacja CPLEX pierwotne dane ze źródeł danych mogą być ujęte łącznie jako pojedyncze indeksy lub jako wiele indeksów wymiarów wyników.

Uwaga: Gdy w programie **IBM SPSS Modeler Server** wykonywany jest strumień zawierający węzeł Optymalizacja CPLEX, domyślnie używana jest wbudowana edycja Community biblioteki CPLEX. W edycji tej obowiązuje ograniczenie do 1000 zmiennych i 1000 ograniczeń. Jeśli zainstalowana jest pełna edycja programu IBM ILOG CPLEX, w której takie ograniczenie nie obowiązuje, można skorzystać właśnie z niej, wykonując następujące czynności (w zależności od używanej platformy).

- W systemie Windows zmodyfikuj plik options.cfg, dodając ścieżkę do biblioteki OPL. Na przykład:
  ```
  cplex_opl_lib_path="<CPLEX_path>\opl\bin\<Platform_dir>"
  ```
  Gdzie <ścieżka_CPLEX_> jest katalogiem instalacyjnym produktu CPLEX, na przykład C:\Program Files\IBM\ILOG\CPLEX_Studio127, a <katalog_Platformy> jest katalogiem właściwym dla platformy, na przykład x64_win64.
- W systemie Linux zmodyfikuj plik modelersrv.sh, dodając ścieżkę do biblioteki OPL. Na przykład:
  ```
  CPLEX_OPL_LIB_PATH=<ścieżka_CPLEX>/opl/bin/<katalog_Platformy>
  ```
  Gdzie <ścieżka_CPLEX> jest katalogiem instalacyjnym oprogramowania CPLEX, np. /root/Libs_127_FullEdition/Linux_x86_64, a <katalog_Platformy> jest katalogiem właściwym dla platformy, np. x86-64_linux.

Uwaga: Gdy w programie **SPSS Modeler Solution Publisher** wykonywany jest strumień zawierający węzeł optymalizacji CPLEX, domyślnie używana jest wbudowana edycja Community biblioteki CPLEX. W edycji tej obowiązuje ograniczenie do 1000 zmiennych i 1000 ograniczeń. Jeśli zainstalowana jest pełna edycja programu IBM ILOG CPLEX, w której takie ograniczenie nie obowiązuje, można skorzystać właśnie z niej, wykonując następujące czynności (w zależności od używanej platformy).

- W systemie Windows dodaj ścieżkę biblioteki OPL jako argument wywołania programu modelerrun.exe. Na przykład:
  ```
  -o cplex_opl_lib_path="<CPLEX_path>\opl\bin\<Platform_dir>"
  ```
  Gdzie <ścieżka_CPLEX_> jest katalogiem instalacyjnym produktu CPLEX, na przykład C:\Program Files\IBM\ILOG\CPLEX_Studio127, a <katalog_Platformy> jest katalogiem właściwym dla platformy, na przykład x64_win64.
- W systemie Linux zmodyfikuj plik modelerrun, dodając ścieżkę do biblioteki OPL. Na przykład:
  ```
  CPLEX_OPL_LIB_PATH=<ścieżka_CPLEX>/opl/bin/<katalog_Platformy>
  ```
  Gdzie <ścieżka_CPLEX> jest katalogiem instalacyjnym oprogramowania CPLEX, np. /root/Libs_127_FullEdition/Linux_x86_64, a <katalog_Platformy> jest katalogiem właściwym dla platformy, np. x86-64_linux.

Opcje ustawień dla węzła optymalizacji CPLEX

Karta Opcje węzła optymalizacji CPLEX zawiera następujące pola.

Plik modelu OPL. Umożliwia wybór pliku modelu OPL (Optimization Programming Language).

Model OPL. Po wybraniu modelu OPL tutaj wyświetlana jest jego zawartość.

Dane wejściowe

Na karcie Dane wejściowe lista rozwijana **Źródło danych** zawiera wszystkie źródła danych (poprzednie węzły) podłączone do bieżącego węzła Optymalizacja CPLEX. Wybranie źródła danych z listy rozwijanej powoduje odświeżenie poniższej sekcji **Odwzorowanie danych wejściowych**. Kliknij opcję **Zastosuj wszystkie zmienne**, aby automatycznie wygenerować odwzorowania wszystkich zmiennych dla wybranego źródła danych. Tabela **Odwzorowanie danych wejściowych** zostanie wypełniona automatycznie.

Wprowadź nazwę zbioru krotek w modelu OPL odpowiadającego danym wejściowym. Następnie, w razie potrzeby, upewnij się, że wszystkie zmienne krotek są odwzorowane na zmienne danych wejściowych zgodnie z ich kolejnością w definicji krotki.

Po skonfigurowaniu odwzorowania danych wejściowych dla źródła danych można wybrać inne źródło danych z listy rozwijanej i powtórzyć proces. Poprzednie odwzorowania źródeł danych zostaną zapisane automatycznie. Po zakończeniu kliknij przycisk **Zastosuj** lub **OK**.

Inne dane

Na karcie Inne dane w sekcji **Dane OPL** można określić pozostałe dane dotyczące optymalizacji.

Wynik

Gdy wynik jest zmienną decyzyjną, musi korzystać z poprzednich źródeł danych (danych przychodzących) jako indeksów, a indeksy te muszą być predefiniowane w sekcji **Odwzorowanie danych wejściowych** na karcie Dane wejściowe. Obecnie nie są obsługiwane żadne inne typy zmiennych decyzyjnych. Zmienna decyzyjna może mieć jeden indeks lub wiele indeksów. SPSS Modeler wygeneruje wyniki CPLEX z całością lub częścią pierwotnych danych przychodzących, podobnie jak inne węzły SPSS Modeler. Przywoływane odpowiednie indeksy muszą być określone w polu **Krotka wynikowa** opisanym poniżej.

Na karcie Wynik można wybrać tryb wynikowy (**Nieprzetworzone wyniki** lub **Zmienna decyzyjna**) i określić inne wymagane opcje. Opcja Nieprzetworzone wyniki pozwala na uzyskanie bezpośrednio wartości funkcji celu, niezależnie od nazwy.

Nazwa zmiennej wartości funkcji celu w OPL. To pole jest aktywowane, jeśli wybrano tryb wynikowy **Zmienna decyzyjna**. Należy wprowadzić nazwę zmiennej wartości funkcji celu z modelu OPL.

Nazwa wyjściowej zmiennej wartości funkcji celu. Należy wprowadzić nazwę zmiennej, jaka będzie używana w wyniku. Wartością domyślną jest _OBJECTIVE.

Krotka wynikowa. Wprowadź nazwę predefiniowanej krotki z danych przychodzących. Pełni rolę indeksów zmiennej decyzyjnej i oczekuje się, że będzie generowana w wynikach ze zmiennymi wynikowymi. Krotka wynikowa powinna zgadzać się z definicją zmiennej decyzyjnej w języku OPL. Jeśli istnieje więcej niż jeden indeks, nazwy krotek muszą być połączone przecinkiem (,).

Zmienne wynikowe. Należy dodać co najmniej jedną zmienną, jaka będzie dołączana do wyniku.

Rozdział 4. Węzły operacji na zmiennych

Przegląd operacji na zmiennych

Po przeprowadzeniu początkowej eksploracji danych prawdopodobnie konieczne będzie wybranie, wyczyszczenie lub utworzenie danych w celu przygotowania ich do analizy. Paleta Operacje na zmiennych zawiera wiele węzłów przydatnych do transformacji i przygotowania.

Przykładowo, korzystając z węzła wyliczeń, można utworzyć atrybut, który obecnie nie jest reprezentowany w danych. Można również użyć węzła kategoryzacji, który umożliwia automatyczne ponowne kodowanie wartości zmiennych na potrzeby docelowej analizy. Prawdopodobnie często używany będzie węzeł typu — umożliwia on przypisywanie poziomu pomiaru, wartości i roli modelowania dla każdej zmiennej w zbiorze danych. Jego operacje są przydatne do obsługi braków danych i modelowania kolejnych węzłów strumienia.

Paleta Operacje na zmiennych zawiera następujące węzły:

Węzeł Auto Przygotowanie (ADP — Automated Data Preparation) umożliwia analizę danych oraz identyfikowanie poprawek, odsiewanie problematycznych zmiennych lub zmiennych, które prawdopodobnie nie będą przydatne, wyliczanie nowych atrybutów, o ile to konieczne, oraz zwiększanie wydajności dzięki zastosowaniu inteligentnych technik badań przesiewowych i doboru próby. Węzła można użyć w sposób w pełni zautomatyzowany, pozwalając mu na wybór i zastosowanie poprawek lub można przejrzeć zmiany przed ich wprowadzeniem, aby je zaakceptować, odrzucić lub zmienić, jeśli będzie to koniczne.

Węzeł Typy określa metadane i właściwości zmiennej. Przykładowo można określić nominalny poziom pomiaru (ilościowy, nominalny, porządkowy lub flaga) dla każdej zmiennej, ustawić opcje obsługi braków danych oraz systemowych wartości null, ustawić rolę zmiennej na potrzeby modelowania, określić etykiety zmiennej i wartości oraz określić wartości dla zmiennej.

Węzeł Filtrowanie filtruje (odrzuca) zmienne, zmienia nazwy zmiennych i mapuje zmienne z jednego węzła źródłowego do drugiego.

Węzeł Wyliczanie modyfikuje wartości danych lub tworzy nowe zmienne z co najmniej jednej istniejącej zmiennej. Tworzy pola typu formuła, flaga, nominalne, stan, liczebność i warunkowe.

Węzeł Zespół łączy co najmniej dwa modele użytkowe w celu uzyskania bardziej dokładnych predykcji, jakie można uzyskać z dowolnego modelu.

Węzeł wypełniania zastępuje wartości zmiennych i zmienia typ składowania. Wartości mogą być zastępowane na podstawie warunku CLEM, np. @BLANK(@FIELD). Alternatywnie można wybrać, aby wszystkie wartości puste lub null zastępowane były konkretną wartością. Węzeł wypełniania często jest używany z węzłem Typy do zastępowania braków danych.

Węzeł Anonimizacja przekształca sposób, w jaki nazwy i wartości zmiennych są reprezentowane w dalszej części strumienia, maskując oryginalne dane. Może to być przydatne, jeśli inni użytkownicy mają mieć możliwość budowania modeli z wykorzystaniem danych poufnych, takich jak nazwiska klientów lub inne szczegóły.

Węzeł Rekodowanie przekształca jeden zestaw wartości jakościowych w inny. Rekodowanie jest przydatne do zwijania kategorii lub ponownego pogrupowania danych do analizy.

Węzeł Kategoryzacja automatycznie tworzy nowe zmienne nominalne (zbioru) na podstawie wartości z jednej lub większej liczby istniejących zmiennych ilościowych (zakres liczbowy). Można na przykład przekształcić ilościową zmienną przychodu na nową zmienną jakościową zawierającą grupy przychodu stanowiące odchylenia od średniej. Po utworzeniu kategorii dla nowej zmiennej na podstawie punktu podziału można wygenerować węzeł Wyliczanie.

Węzeł analizy RFM (Recency — Aktualność, Frequency — Częstość, Monetary — Kwota) umożliwia określenie ilościowo, którzy klienci najprawdopodobniej będą najlepszymi, poprzez dokonanie oceny, kiedy ostatnio dokonali zakupu (aktualność), jak często dokonują zakupu (częstość) i jak dużo wydali na wszystkie transakcje (kwota).

Węzeł Partycja generuje zmienną dzielącą na podzbiory, która dzieli dane na osobne podzbiory dla etapów do uczenia, testowania i walidacji podczas budowania modelu.

Węzeł Flagowanie służy do wyliczania zmiennych flag na podstawie zmiennych wartości jakościowych zdefiniowanych dla co najmniej jednej zmiennej nominalnej.

Węzeł Restrukturyzacja przekształca zmienną nominalną lub typu flaga na grupę zmiennych, które mogą być wypełnione wartościami jeszcze innej zmiennej. Na przykład, dana jest zmienna o nazwie *payment type* (rodzaj płatności), której wartości to *credit* (kredyt), *cash* (gotówka) i *debit* (debet) i utworzone zostaną trzy nowe zmienne (*credit*, *cash*, *debit*), a każda z nich może zawierać wartość dla rzeczywiście dokonanej płatności.

Węzeł Transpozycja zamienia miejscami dane w wierszach i kolumnach, w wyniku czego rekordy stają się zmiennymi, a zmienne rekordami.

Węzeł Przedziały czasowe może być używany do określenia przedziałów i wyliczenia nowej zmiennej czasu na potrzeby oszacowania lub prognozowania. Obsługiwany jest cały zakres przedziałów czasowych, od sekund po lata.

Węzeł Historia tworzy nowe zmienne zawierające dane ze zmiennych z wcześniejszych rekordów. Węzły historii są najczęściej używane w przypadku danych sekwencyjnych, takich jak dane szeregu czasowego. Przed użyciem węzła historii można posortować dane za pomocą węzła Sortowanie.

Węzeł Reorganizacja definiuje rzeczywistą kolejność, w jakiej wyświetlane są zmienne w dalszej części strumienia. Ta kolejność wpływa na wyświetlanie zmiennych w różnych obszarach, takich jak tabele, listy i selektor zmiennych. Ta operacja jest przydatna podczas pracy z obszernymi bazami danych w celu zapewnienia lepszej widoczności zmiennych, które interesują użytkownika.

W programie SPSS Modeler składniki, takie jak funkcje przestrzenne w Konstruktorze wyrażeń, węzeł STP i węzeł Wizualizacja na mapie, używają rzutowanego układu współrzędnych. Węzeł Odwzorowanie umożliwia zmianę układu współrzędnych każdego rodzaju importowanych danych, które korzystają z układu współrzędnych geograficznych.

Niektóre z tych węzłów można wygenerować bezpośrednio z audytu danych utworzonego przez węzeł Audyt danych. Więcej informacji można znaleźć w temacie “Generowanie innych węzłów w celu przygotowania danych” na stronie 320.

Automatyczne przygotowywanie danych

Przygotowywanie danych do analizy jest jednym z najbardziej istotnych kroków w każdym projekcie — i również jednym z najbardziej czasochłonnych. Automatyczne przygotowanie danych (Automated Data Preparation — ADP) ma za zadanie analizę danych i identyfikację stałych, klasyfikację pól (zmiennych), które są problematyczne lub mają małe prawdopodobieństwo bycia użytecznymi, w razie potrzeby obliczanie nowych atrybutów i zwiększanie wydajności poprzez wykorzystywanie inteligentnych technik klasyfikowania. Można używać tego algorytmu w sposób **w pełni automatyczny**, pozwalając mu na wybór i zastosowanie stałych, lub korzystać z niego w sposób **interaktywny**, przeglądając zmiany przed ich dokonaniem i zaakceptować je lub odrzucać.

Użycie funkcji automatycznego przygotowania danych (ADP) umożliwia przygotowanie danych do szybkiego i łatwego budowania modelu, bez konieczności uzyskiwania wiedzy na temat użytych koncepcji statystycznych. Budowa i ocena modeli będzie odbywać się szybciej; ponadto, korzystanie z automatycznego przygotowywania danych (ADP) zwiększa elastyczność procesów automatycznego modelowania, takich jak odświeżenie modelu i Champion Challenger.

Uwaga: Kiedy proces automatycznego przygotowywania danych przygotowuje zmienną do analizy, tworzona jest nowa zmienna zawierająca korekty lub transformacje, zamiast zastępowania istniejących wartości i właściwości starej zmiennej. Stara zmienna nie jest używana w dalszej analizie; jej rola jest ustawiona na Brak.

Przykład. Firma ubezpieczeniowa o ograniczonych środkach na sprawdzenie roszczeń chce zbudować model do flagowania podejrzanych, potencjalnie oszukańczych roszczeń. Przed utworzeniem modelu dane będą przygotowane do modelowania przy użyciu automatycznego przygotowywania danych. Ponieważ firma chce mieć możliwość przejrzenia zaproponowanych transformacji przed ich zastosowaniem, użyte zostanie automatyczne przygotowywanie danych w trybie interaktywnym.

Grupa z branży motoryzacyjnej śledzi sprzedaż różnych pojazdów osobowych. Podejmując próbę zidentyfikowania najbardziej i najmniej rentownych modeli, chcą ustalić relacje pomiędzy sprzedażą pojazdów a charakterystykami pojazdów. Automatyczne przygotowywanie danych umożliwia przygotowanie danych do analizy, a utworzenie modeli z użyciem danych „przed" i „po" przygotowaniu pozwoli zobaczyć różnice w wynikach.

Jaki jest cel? Automatyczne przygotowywanie danych rekomenduje kroki przygotowania danych, które będą wpływały na szybkość, z jaką inne algorytmy mogą budować modele i które ulepszą jakość predykcji tych modeli. Możliwe działania to przykładowo przekształcanie, tworzenie i wybór predyktorów. Możliwe jest również przekształcenie zmiennej przewidywanej. Można określić priorytety budowania modelu, na jakich proces przygotowywania danych powinien się skoncentrować.

- **Zrównoważenie szybkości i dokładności.** Ta opcja umożliwia przygotowanie danych, tak aby nadać jednakowy priorytet szybkości przetwarzania danych przez algorytmy budowania modelu oraz dokładności predykcji.
- **Optymalizacja dla szybkości.** Ta opcja umożliwia przygotowanie danych, tak aby nadać priorytet szybkości przetwarzania danych przez algorytmy budowania modelu. Opcję tę należy wybrać w przypadku pracy z dużymi zbiorami danych lub poszukiwania szybkiej odpowiedzi.
- **Optymalizacja dla dokładności.** Ta opcja umożliwia przygotowywanie danych, taka aby nadać priorytet dokładności predykcji tworzonych przez algorytmy budowania modelu.
- **Analiza użytkownika.** Opcję tę należy wybrać, aby ręcznie zmienić algorytm na karcie Ustawienia. Jeśli w późniejszym czasie na karcie Ustawienia dokonane zostaną zmiany opcji, które są niekompatybilne z jednym z pozostałych celów, należy pamiętać, że to ustawienie jest zaznaczane automatycznie.

Uczenie węzła

Węzeł automatycznego przygotowywania danych (ADP) jest implementowany jako węzeł procesowy i działa podobnie, jak węzeł typu; **uczenie** węzła ADP odnosi się do określania węzła typu. Po przeprowadzeniu analizy określone przekształcenia są stosowane do danych bez przeprowadzania kolejnej analizy, dopóki model danych we wcześniejszej części strumienia nie ulegnie zmianie. Podobnie jak w przypadku węzłów typu i filtrowania, jeśli węzeł automatycznego przygotowywania danych zostanie odłączony, będzie pamiętał model danych ii transformacje, dzięki czemu po ponownym podłączeniu nie ma konieczności ponownego uczenia go; pozwala to na uczenie węzła na podzbiorze typowych danych, a następnie skopiowanie lub wdrożenie do użycia z danymi rzeczywistymi tak często, jak to konieczne.

Korzystanie z paska narzędzi

Pasek narzędzi umożliwia uruchamianie i aktualizowanie sposobu wyświetlania analizy danych oraz generowanie węzłów, jakie mogą zostać użyte w połączeniu z oryginalnymi danymi.

- **Utwórz** Za pomocą tego menu można wygenerować węzeł filtrowania lub wyliczeń. Należy pamiętać, że to menu jest dostępne tylko w przypadku, gdy na karcie Analiza wyświetlana jest analiza.

  Węzeł filtrowania usuwa przekształcone zmienne wejściowe. Jeśli węzeł automatycznego przygotowywania danych zostanie skonfigurowany, tak aby pozostawiał oryginalne zmienne wejściowe w zbiorze danych, spowoduje to odtworzenie oryginalnego zbiory danych wejściowych, umożliwiając interpretację zmiennej oceny w odniesieniu do danych wejściowych. Przykładowo, może to być przydatne do utworzenia wykresu dla zmiennej oceny w odniesieniu do różnych danych wejściowych.

  Węzeł wyliczeń może przywrócić oryginalny zbiór danych i jednostki zmiennych przewidywanych. Węzeł wyliczeń można wygenerować tylko wówczas, gdy węzeł automatycznego przygotowywania danych (ADP) obejmuje analizę, która umożliwia zmianę skali przewidywanej ilościowej (czyli na panelu Zmienne wejściowe i przewidywana wybrano metodę przeskalowania Boxa-Coxa). Węzła wyliczeń nie można wygenerować, jeśli zmienna przewidywana nie jest ilościowa lub jeśli nie wybrano metody przeskalowania Boxa-Coxa. Więcej informacji można znaleźć w temacie "Generowanie węzła Wyliczanie" na stronie 136.
- **Widok** Zawiera opcje, które umożliwiają kontrolowanie elementów wyświetlanych na karcie Analiza. Są to na przykład elementy sterujące umożliwiające edytowanie wykresu i wyświetlanie opcji wybranych dla panelu głównego oraz widoków powiązanych.
- **Podgląd** Wyświetla przykładowe transformacje, jakie zostaną wykonane dla danych wejściowych.

- **Analizuj dane** Inicjuje analizy z zastosowaniem bieżących ustawień i wyświetla wyniki na karcie Analiza.
- **Wyczyść analizę** Usuwa istniejącą analizę (opcja jest dostępna, tylko w przypadku, gdy istnieje bieżąca analiza).

Status węzła

Status węzła automatycznego przygotowywania danych w obszarze roboczym programu IBM SPSS Modeler jest wskazywany przez strzałkę lub zaznaczenie na ikonie, które wskazuje, czy analiza miała miejsce.

Więcej informacji na temat obliczeń wykonywanych z użyciem Automatyczne przygotowywanie danych zawiera sekcja *Algorytmy automatycznego przygotowania danych* w publikacji *IBM SPSS Modeler — podręcznik algorytmów*. Publikacja jest dostępna w formacie PDF w katalogu \Documentation na dysku instalacyjnym i stanowi część materiałach do pobrania dla produktu lub na stronie WWW.

Karta Zmienne

Aby możliwe było zbudowanie modelu, konieczne jest określenie, które zmienne mają być używane jako zmienne przewidywane, a które jako dane wejściowe. We wszystkich węzłach modelowania (z kilkoma wyjątkami) stosowane są informacje na temat zmiennych z wcześniejszego węzła Typy. Korzystając z węzła Typy do wyboru zmiennych wejściowych i przewidywanych, nie trzeba zmieniać żadnych ustawień na tej karcie.

Użyj ustawień węzła Typy. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej z wcześniejszego węzła Typy. Jest to ustawienie domyślne.

Użyj ustawień niestandardowych. Ta opcja stanowi dla węzła instrukcję o konieczności użycia informacji o zmiennej określonych w tym miejscu, a nie w żadnym wcześniejszym węźle Typy. Po wybraniu tej opcji określ poniższe zmienne odpowiednio do potrzeb.

Zmienna przewidywana. W przypadku modeli, które wymagają jednej lub większej liczby zmiennych przewidywanych, należy wybrać zmienną lub zmienne przewidywane. Działanie jest podobne, jak w przypadku ustawienia roli zmiennej na wartość *Zmienna przewidywana* w węźle Typy.

Zmienne wejściowe. Umożliwia wybór zmiennej wejściowej lub kilku zmiennych. Działanie jest podobne, jak w przypadku ustawienia roli zmiennej na *Zmienna wejściowa* w węźle Typy.

Karta Ustawienia

Karta Ustawienia zawiera wiele różnych grup ustawień, które można zmieniać w celu precyzyjnego określenia sposobu przetwarzania danych użytkownika przez algorytm. W przypadku wprowadzenia zmian w ustawieniach domyślnych, które są niezgodne z innymi celami, zakładka Cel zostanie automatycznie zaktualizowana do zaznaczenia opcji **Analiza niestandardowa**.

Ustawienia zmiennych

Użyj zmiennej częstości. Ta opcja umożliwia wybranie zmiennej jako wagi częstości. Tej opcji należy użyć, jeśli każdy rekord w danych uczących reprezentuje więcej niż jedną jednostkę — na przykład, jeśli stosowane są dane zagregowane. Wartości zmiennych powinny odpowiadać liczbom jednostek reprezentowanych przez poszczególne rekordy.

Użyj zmiennej ważącej. Ta opcja umożliwia wybranie zmiennej jako wagi obserwacji. Wagi obserwacji są stosowane w celu uwzględniania różnic w wariancji między poziomami zmiennej wyjściowej.

Sposób obsługi zmiennych, które są wyłączone z modelowania. Należy określić, co stanie się z wykluczonymi zmiennymi; można odfiltrować je z danych lub ustawić ich *rolę* na wartość **Brak**.

Uwaga: Ta czynność zostanie również zastosowana do zmiennej przewidywanej, jeśli zostanie ona poddana transformacji. Przykładowo, jeśli nowo wyliczona wersja zmiennej przewidywanej zostanie użyta jako zmienna **Przewidywana**, oryginalna zmienna przewidywana zostanie odfiltrowana lub ustawiona jako **Brak**.

Jeśli zmienne wejściowe nie pokrywają się z poprzednią analizą. Należy określić, co stanie się, jeśli co najmniej jedna wymagana zmienna wejściowa będzie niedostępna w wejściowym zbiorze danych w czasie wykonywania węzła wyuczonego ADP.

- **Zatrzymaj wykonywanie i zachowaj poprzednią analizę.** Ta opcja zatrzymuje proces wykonywania, zachowuje informacje o bieżącej analizie i wyświetla błędy.
- **Usuń poprzednią analizę i dokonaj analizy nowych danych.** Ta opcja usuwa istniejącą analizę, analizuje dane wejściowe i stosuje zalecane przekształcenia do tych danych.

Przygotowanie daty i czasu

Wiele algorytmów modelowania nie jest w stanie bezpośrednio obsługiwać danych daty i czasu; te ustawienia umożliwiają wyliczenie nowych danych czasu trwania, które mogą być użyte jako dane wejściowe modelu, na podstawie informacji o dacie i czasie dostępnych w istniejących danych. Zmienne zawierające datę i czas muszą zostać wstępnie zdefiniowane z użyciem typów składowania data lub czas. Oryginalne zmienne daty i czasu nie będą zalecane przez proces automatycznego przygotowywania danych jako dane wejściowe modelu.

Przygotuj datę i czas do modelowania. Usunięcie zaznaczenia tej opcji wyłączy pozostałe elementy sterujące w obszarze Przygotowanie daty i czasu podczas dokonywania wyboru ustawień.

Wylicz czas jaki upłynął od daty odniesienia. Ta opcja wyznacza liczbę lat/miesięcy/dni od daty odniesienia dla każdej zmiennej zawierającej daty.

- **Data odniesienia.** Umożliwia określenie daty, od której obliczany będzie czas trwania, z odniesieniem do informacji na temat daty dostępnych w danych wejściowych. Wybranie opcji **Dzisiejsza data** oznacza, że podczas wykonywania automatycznego przygotowywania danych zawsze używana będzie bieżąca data systemowa. Aby użyć konkretnej daty, należy zaznaczyć opcję **Ustalona data** i wprowadzić wymaganą datę. Podczas tworzenia węzła po raz pierwszy w polu **Ustalona data** automatycznie wprowadzana jest bieżąca data.
- **Jednostki czasu trwania.** Należy określić, czy proces automatycznego przygotowywania danych będzie automatycznie dobierał jednostkę czasu trwania, lub wybrać opcję **Ustalone jednostki** i zaznaczyć Lata, Miesiące lub Dni.

Wylicz czas jaki upłynął od czasu odniesienia. Ta opcja powoduje wyznaczenie liczby godzin/minut/sekund od czasu odniesienia dla każdej zmiennej zawierającej dane o czasie.

- **Czas odniesienia.** Umożliwia określenie czasu, od którego obliczany będzie czas trwania, z odniesieniem do informacji na temat czasu dostępnych w danych wejściowych. Wybranie opcji **Bieżący czas** oznacza, że podczas wykonywania automatycznego przygotowywania danych zawsze używany będzie bieżący czas systemowy. Aby użyć konkretnego czasu, należy wybrać opcję **Ustalony czas** i wprowadzić odpowiednie szczegóły. Podczas tworzenia węzła po raz pierwszy w polu **Ustalony czas** automatycznie wprowadzany jest bieżący czas.
- **Jednostki czasu trwania.** Należy określić, czy proces automatycznego przygotowywania danych będzie automatycznie dobierał jednostkę czasu trwania, lub wybrać opcję **Ustalone jednostki** i zaznaczyć Godziny, Minuty lub Sekundy.

Wyodrębnij cykliczne elementy czasu. Te ustawienia umożliwiają podzielenie pojedynczej zmiennej daty lub czasu na jedną lub więcej zmiennych. Przykładowo, jeśli zaznaczone zostaną wszystkie trzy pola wyboru dla daty, wejściowa zmienna daty "1954-05-23" zostanie podzielona na trzy zmienne: 1954, 5 i23, a dla każdej z nich zastosowany zostanie przedrostek zdefiniowany w panelu **Nazwy zmiennych**, a oryginalna zmienna daty zostanie zignorowana.

- **Pobierz z daty.** Dla dowolnych danych wejściowych daty należy określić lata, miesiące lub dni do wyodrębnienia lub dowolną ich kombinację.
- **Pobierz z czasu.** Dla dowolnych danych wejściowych czasu należy określić godziny, minuty lub sekundy do wyodrębnienia lub dowolną ich kombinację.

Wykluczanie zmiennych

Słaba jakość danych może wpływać na dokładność predykcji; dlatego można określić akceptowalny poziom jakości dla wejściowych predykcji. Wszystkie zmienne, które są stałe lub mają 100% braków danych, zostają automatycznie wykluczone.

Wyklucz zmienne wejściowe niskiej jakości. Usunięcie zaznaczenia tej opcji wyłączy pozostałe elementy sterujące w obszarze Statystyki wykluczonych zmiennych podczas dokonywania wyboru ustawień.

Wyklucz zmienne wejściowe o zbyt dużej liczbie braków. Zmienne, w których liczba braków danych przekracza określoną wartość procentową, zostają usunięte z dalszej analizy. Określenie wartości większej niż lub równej 0 jest równoznaczne z usunięciem zaznaczenia tej opcji, a wartości mniejszej niż lub równej 100 spowoduje, że zmienne ze wszystkimi brakami danych będą automatycznie wykluczane. Domyślną wartością jest 50.

Wyklucz zmienne nominalne o zbyt dużej liczbie unikatowych kategorii. Zmienne nominalne z liczbą kategorii większą od podanej będą wykluczone z dalszej analizy. Określ dodatnią liczbę całkowitą. Domyślną wartością jest 100. Ta opcja jest przydatna do automatycznego usuwania z modelowania zmiennych zawierających informacje unikalne dla rekordu, takie jak identyfikator, adres lub nazwa.

Wyklucz zmienne kategorialne o zbyt dużej liczbie wartości w jednej kategorii. Zmienne porządkowe i nominalne z kategorią, która zawiera więcej rekordów od określonej wartości procentowej, są usuwane z dalszej analizy. Określenie wartości większej niż lub równej 0 jest równoznaczne z usunięciem zaznaczenia tej opcji, a wartości mniejszej niż lub równej 100 spowoduje, że zmienne o stałych wartościach będą automatycznie wykluczane. Domyślną wartością jest 95.

Przygotowywanie zmiennych wejściowych i przewidywanych

Ponieważ nie istnieją dane, które są w stanie idealnym do przetwarzania, przed uruchomieniem analizy można dokonać pewnych ustawień. Przykładowo może to obejmować usuwanie wartości odstających, określanie sposobu postępowania z brakami danych lub skorygowanie typu.

Uwaga: Jeśli wartości w tym panelu zostaną zmienione, karta **Cele** zostanie automatycznie zaktualizowana, aby została wybrana opcja **Analiza niestandardowa**.

Przygotowanie predyktorów i zmiennej przewidywanej do modelowania. Powoduje włączenie lub wyłączenie wszystkich zmiennych na panelu.

Dostosowanie typu i poprawa jakości danych. W przypadku zmiennych wejściowych i przewidywanych można określić kilka transformacji danych do wykonania osobno; dzięki temu wartości zmiennej przewidywanej mogą pozostać niezmienione. Na przykład predykcja przychodu w dolarach jest bardziej znacząca niż predykcja mierzona za pomocą logarytmu przychodu (w dolarach). Ponadto, jeśli zmienna przewidywana zawiera braki danych, nie ma dostępnych przewidywanych korzyści do wypełnienia braków danych, podczas gdy wypełnienie braków danych w zmiennych wejściowych może aktywować pewne algorytmy umożliwiające przetworzenie informacji, w które w przeciwnym razie zostaną utracone.

Dodatkowe ustawienia dla tych transformacji, takie jak wartość odcięcia dla wartości odstających, są wspólne dla zmiennych przewidywanych i wejściowych.

Można wybrać następujące ustawienia dla zmiennych wejściowych i/lub przewidywanych:

- **Skoryguj typ zmiennych numerycznych.** Tę opcję należy wybrać, aby określić, czy zmienne numeryczne z poziomem pomiaru *Porządkowy* mogą być przekształcane na zmienne typu *Ilościowy* lub odwrotnie. Istnieje możliwość określenia minimalnych i maksymalnych wartości granicznych do sterowania przekształceniem.
- **Zmień kolejność zmiennych nominalnych.** Tę opcję należy wybrać, aby sortować zmienne nominalne (zbiór) w porządku od najmniejszej do największej kategorii.
- **Zastępowanie wartości odstających w ilościowych.** Tę opcję należy wybrać, aby zastąpić wartości odstające; należy jej użyć razem z opcją **Metoda zastępowania wartości odstających** poniżej.

- **Zmienne ilościowe: zastępowanie wartości brakujących średnimi.** Tę opcję należy wybrać, aby zastąpić braki danych funkcjami ilościowymi (zakres).
- **Zmienne nominalne: zastępowanie wartości brakujących dominantą.** Tę opcję należy wybrać, aby zastąpić braki danych funkcjami nominalnymi (zbiór).
- **Zmienne porządkowe: zastępowanie wartości brakujących medianą.** Tę opcję należy wybrać, aby zastąpić braki danych funkcjami porządkowymi (zbiór uporządkowany).

Maksymalna liczba wartości dla zmiennych porządkowych. Należy określić wartość graniczną dla ponownego zdefiniowania zmiennych porządkowych (zbiór uporządkowany) jako ilościowe (zakres). Domyślne ustawienie to 10; dlatego jeśli zmienna porządkowa ma więcej niż 10 kategorii, zostanie ponownie zdefiniowana jako ilościowa (zakres).

Minimalna liczba wartości dla zmiennych ilościowych. Należy określić wartość graniczną dla ponownego zdefiniowania zmiennych skali lub ilościowych (zakres) jako porządkowe (zbiór uporządkowany). Wartością domyślną jest 5; dlatego jeśli zmienna ilościowa zawiera więcej niż 5 wartości, zostaje ponownie zdefiniowana jako porządkowa (zbiór uporządkowany).

Wartość odcięcia dla wartości odstających. Należy określić kryterium odcięcia dla wartości odstających, mierzone jako odchylenie standardowe; wartość domyślna to 3.

Metoda zastępowania wartości odstających. Należy wybrać, czy wartości odstające mają być zastępowane przez przycięcie (wymuszanie) z użyciem wartości odstających, czy też mają być usuwane i ustawiane jako braki danych. Dla wszystkich wartości odstających ustawionych jako braki danych obowiązują ustawienia postępowania wybrane powyżej.

Sprowadź wszystkie wejściowe zmienne ilościowe do wspólnej skali. Aby znormalizować wejściowe zmienne ilościowe, należy zaznaczyć to pole wyboru i wybrać metodę normalizacji. Domyślnie ustawiona jest opcja **Standaryzacja**, w której można określić wartość **Końcowa średnia**, domyślnie ustawioną jako 0, oraz wartość **Końcowe odchylenie standardowe**, domyślnie ustawioną jako 1. Alternatywnie można wybrać opcję **Transformacja Min/Maks** i określić wartości minimalne i maksymalne, domyślnie ustawione odpowiednio jako 0 i 100.

To pole jest szczególnie przydatne po wybraniu opcji **Konstruuj predyktory** na panelu Tworzenie i wybór predyktorów.

Przeskaluj docelową wartość ilościową za pomocą transformacji Boxa-Coxa. Aby znormalizować docelową wartość ilościową (skala lub zakres), należy zaznaczyć to pole wyboru. Transformacja Boxa-Coxa ma wartości domyślne wynoszące 0 dla opcji **Końcowa średnia** i 1 dla opcji **Końcowe odchylenie standardowe**.

Uwaga: Jeśli wybrana zostanie opcja normalizacji zmiennej przewidywanej, wymiar zmiennej przewidywanej zostanie przekształcony. W takim przypadku konieczne może być wygenerowanie węzła wyliczeń, aby zastosować przekształcenie odwrotne w celu przywrócenia przekształconych jednostek do formatu możliwego do rozpoznania w celu przeprowadzenia dalszego przetwarzania. Więcej informacji można znaleźć w temacie “Generowanie węzła Wyliczanie” na stronie 136.

Wybór i tworzenie predyktorów

Aby zwiększyć jakość predykcji danych, można przekształcić zmienne wejściowe lub utworzyć nowe w oparciu o zmienne istniejące.

Uwaga: Jeśli wartości w tym panelu zostaną zmienione, karta **Cele** zostanie automatycznie zaktualizowana, aby została wybrana opcja **Analiza niestandardowa**.

Transformacja, tworzenie i wybór zmiennych wejściowych w celu poprawy jakości predykcji. Powoduje włączenie lub wyłączenie wszystkich zmiennych na panelu.

Połącz małoliczne kategorie w celu zwiększenia związku ze zmienną przewidywaną. Tę opcję należy wybrać, aby model stał się skromniejszy poprzez zmniejszenie liczby zmiennych do przetworzenia w powiązaniu z docelową. W razie konieczności należy zmienić domyślną wartość prawdopodobieństwa wynoszącą 0,05.

Należy pamiętać, że jeśli wszystkie kategorie są scalone w jedną, oryginalne i wyliczone wersje zmiennej są wyłączone z modelu, ponieważ nie mają żadnej wartości jako predyktora.

Przy braku zmiennej przewidywanej połącz małoliczne kategorie na podstawie liczebności. W przypadku danych, które nie mają zmiennej przewidywanej, można połączyć małoliczne kategorie właściwości porządkowych (zbiór uporządkowany) i/lub nominalnych (zbiór). Należy określić minimalny procent obserwacji lub rekordów w danych, który identyfikuje kategorie, jakie mają zostać połączone; ustawienie domyślne to 10.

Kategorie są łączone z zastosowaniem następujących reguł:

- Łączenie nie jest wykonywane dla zmiennych binarnych.
- Jeśli w czasie łączenia dostępne są tylko dwie kategorie, łączenie zostaje zatrzymane.
- Jeśli nie jest dostępna żadna kategoria porządkowa ani żadna kategoria utworzona podczas łączenia, dla których minimalny procent obserwacji jest niższy niż określony, łączenie zostaje zatrzymane.

Kategoryzuj zmienne ilościowe przy zachowaniu jakości predykcji. Jeśli dostępne są dane, które obejmują przewidywaną zmienną jakościową, można skategoryzować ilościowe dane wejściowe z silnymi powiązaniami w celu zwiększenia wydajności przetwarzania. W razie konieczności należy zmienić wartość prawdopodobieństwa dla jednorodnych podzbiorów, która domyślnie jest ustawiona na 0,05.

Jeśli w wyniku kategoryzacji tworzona jest pojedyncza kategoria dla konkretnej zmiennej, oryginalne i skategoryzowane wersje zmiennej są wykluczane, ponieważ nie mają wartości jako predyktora.

Uwaga: Kategoryzacja w ADP różni się od kategoryzacji optymalnej użytej w innych częściach programu IBM SPSS Modeler. Kategoryzacja optymalna korzysta z informacji o entropii w celu przekształcenia zmiennych ilościowych na zmienne jakościowe; wymaga to posortowania danych i zapisania ich wszystkich w pamięci. ADP korzysta z jednorodnych podzbiorów w celu skategoryzowania zmiennej ilościowej, co oznacza, że kategoryzacja ADP nie wymaga sortowania danych i nie zapisuje ich wszystkich w pamięci. Użycie metody jednorodnych podzbiorów do skategoryzowania zmiennych ilościowych oznacza, że liczba kategorii po kategoryzacji zawsze będzie mniejsza od lub równa liczbie kategorii zmiennej przewidywanej.

Dokonaj wyboru predyktora. Tę opcję należy wybrać, aby usunąć funkcje z niskim współczynnikiem korelacji. W razie konieczności należy zmienić domyślną wartość prawdopodobieństwa wynoszącą 0,05.

Ta opcja ma zastosowanie tylko do funkcji ilościowych zmiennych wejściowych, w których zmienna przewidywana jest ilościowa, oraz do wejściowych zmiennych jakościowych.

Konstruuj predyktory. Tę opcję należy wybrać, aby wyliczyć nowe właściwości poprzez połączenie kilku istniejących właściwości (które są następnie odrzucane z modelowania).

Ta opcja ma zastosowanie tylko do ilościowych zmiennych wejściowych, w których zmienna przewidywana jest ilościowa lub w których nie ma zmiennych przewidywanych.

Nazwy zmiennych

Aby w prosty sposób zidentyfikować nowe i przekształcone predyktory, proces automatycznego przygotowywania danych tworzy i stosuje podstawowe nowe nazwy, przedrostki lub przyrostki. Nazwy te można zmienić, tak aby były bardziej odpowiednie do potrzeb użytkownika i danych, którymi dysponuje. Aby określić inne etykiety, należy to zrobić w węźle typu w dalszej części strumienia.

Zmienne transformowane i konstruowane. Należy określić rozszerzenie nazw, jakie będą stosowane do przekształcanych zmiennych przewidywanych i zmiennych wejściowych.

Należy pamiętać, że w węźle automatycznego przygotowywania danych ustawienie zmiennych łańcuchowych, które nie będą nic zawierały, może spowodować błąd, w zależności od opcji wybranej dla sposobu postępowania z nieużywanymi zmiennymi. Jeśli opcja **Sposób obsługi zmiennych, które są wyłączone z modelowania** jest ustawiona na wartość **Odfiltruj nieużywane zmienne** w panelu Ustawienia zmiennych na karcie Ustawienia, wówczas

rozszerzenia nazw dla zmiennych wejściowych i przewidywanych mogą być ustawione, tak aby nic nie zawierały. Oryginalne zmienne zostają odfiltrowane, a zmienne przekształcone będą zapisane nad nimi; w takim przypadku nowe zmienne przekształcone będą miały taką samą nazwę jak oryginalne.

Jeśli jednak dla nieużywanych zmiennych wybrana zostanie opcja **Ustaw kierunek nieużywanych zmiennych na 'Pomiń'**, wówczas rozszerzenia nazw puste lub null dla zmiennych przewidywanych lub wejściowych będą powodowały błąd, ponieważ podejmowana będzie próba utworzenia zduplikowanych nazw zmiennych.

Ponadto, należy określić nazwę przedrostka, jaki będzie stosowany do dowolnych predyktorów tworzonych za pośrednictwem ustawień tworzenia i wyboru. Nowa nazwa jest tworzona poprzez dołączenie numerycznego przyrostka do trzonu nazwy z przyrostkiem. Format liczbowy zależy od liczby wyznaczanych nowych predyktorów, na przykład:

- Utworzone predyktory od 1 do 9 będą miały następujące nazwy: od feature1 (predyktor1) do feature9 (predyktor9).
- Utworzone predyktory od 10 do 99 będą miały następujące nazwy: od feature01 (predyktor01) do feature99 (predyktor99).
- Utworzone predyktory od 100 do 999 będą miały nazwy: od feature001 (predyktor001) do feature999 (predyktor999) itd.

Dzięki temu utworzone predyktory będą posortowane w sensowny sposób, niezależnie od ich liczby.

Czasy trwania obliczone na podstawie daty i czasu. Należy określić nazwy rozszerzeń, jakie będą stosowane do czasów trwania obliczonych na podstawie daty i czasu.

Elementy cykliczne wyodrębnione z daty i czasu. Należy określić nazwy rozszerzeń, jakie będą stosowane do elementów cyklicznych wyodrębnionych na podstawie daty i czasu.

Karta Analiza

1. Jeśli ustawienia automatycznego przygotowywania danych są zadowalające, w tym zmiany dokonane na kartach Cele, Zmienne i Ustawienia, należy kliknąć przycisk **Analizuj dane**; algorytm zastosuje ustawienia do danych wejściowych i wyświetli wyniki na karcie Analiza.

Karta Analiza zwiera wyniki w postaci tabeli i wykresu, podsumowujące przetwarzanie danych oraz wyświetla zalecenia co do możliwych sposobów modyfikacji lub udoskonalenia danych do przeprowadzenia oceny. Następnie można wyświetlić podgląd i zaakceptować lub odrzucić zalecenia.

Karta Analiza składa się z dwóch paneli, widoku głównego z lewej strony i powiązanego lub dodatkowego widoku z prawej strony. Istnieją trzy główne widoki:

- Podsumowanie przetwarzania zmiennych (ustawienie domyślne). Więcej informacji można znaleźć w temacie "Podsumowanie przetwarzania zmiennej" na stronie 131.
- Zmienne. Więcej informacji można znaleźć w temacie "Zmienne" na stronie 131.
- Podsumowanie kroku. Więcej informacji można znaleźć w temacie "Podsumowanie kroku" na stronie 132.

Istnieją cztery połączone/dodatkowe widoki:

- Jakość predykcji (ustawienie domyślne). Więcej informacji można znaleźć w temacie "Jakość predykcji" na stronie 133.
- Tabela zmiennych. Więcej informacji można znaleźć w temacie "Tabela zmiennych" na stronie 133.
- Szczegóły zmiennej. Więcej informacji można znaleźć w temacie "Szczegóły zmiennej" na stronie 133.
- Szczegóły działania. Więcej informacji można znaleźć w temacie "Szczegóły działania" na stronie 134.

Łącza pomiędzy widokami

W widoku głównym podkreślony tekst w tabelach umożliwia sterowanie wyświetlaniem w powiązanym widoku. Kliknięcie tekstu umożliwia uzyskanie szczegółowych informacji o konkretnej zmiennej, zestawie zmiennych lub

kroku przetwarzania. Ostatnio wybrane łącze jest wyświetlane w ciemniejszym kolorze; ułatwi to identyfikację połączenia pomiędzy zawartością dwóch paneli widoku.

Resetowanie widoków

Aby ponownie wyświetlić oryginalne rekomendacje z karty Analiza i zrezygnować ze zmian wprowadzonych w widokach analizy, należy kliknąć przycisk **Resetuj** w dolnej części panelu głównego widoku.

Podsumowanie przetwarzania zmiennej

Tabela podsumowania przetwarzania zmiennej stanowi obraz stanu planowanego ogólnego wpływu przetwarzania, z uwzględnieniem zmian stanu predyktorów i liczby tworzonych predyktorów.

Należy pamiętać, że w rzeczywistości nie jest tworzony żaden model, dlatego nie jest dostępna żadna miara ani wykres zmiany ogólnej jakości predykcji przed przygotowaniem danych i po ich przygotowaniu; można jednak wyświetlić wykresy jakości predykcji dla pojedynczych zalecanych predyktorów.

Tabela zawiera następujące informacje:

- Liczba zmiennych przewidywanych.
- Liczba oryginalnych (wejściowych) predyktorów.
- Predyktory zalecane do użycia w analizie i modelowaniu. W tym łączna liczba zalecanych zmiennych; liczba zalecanych oryginalnych, nieprzekształconych zmiennych; liczba zalecanych przekształconych zmiennych (z wykluczeniem pośrednich wersji zmiennych, zmiennych pochodnych wyznaczonych na podstawie predyktorów typu data/czas oraz utworzonych predyktorów); liczba zalecanych zmiennych pochodnych, które zostały wyznaczone na podstawie zmiennej typu data/czas; oraz liczba zalecanych utworzonych predyktorów.
- Liczba predyktorów wejściowych niezalecanych do użycia w żadnej postaci, niezależnie od tego, czy są w oryginalnej postaci, takiej jak zmienna pochodna, czy stanowią dane wejściowe dla utworzonego predyktora.

Jeśli informacje w tabeli **Zmienne** są podkreślone, można je kliknąć, aby w powiązanym widoku wyświetlić dodatkowe szczegóły. W powiązanym widoku tabeli Zmienne wyświetlane są szczegóły na temat **zmiennej przewidywanej**, **predyktorów wejściowych** i **nieużywanych predyktorów wejściowych**. Więcej informacji można znaleźć w temacie “Tabela zmiennych” na stronie 133. **Predyktory zalecane do użycia w analizie** są wyświetlane w powiązanym widoku jakości predykcji. Więcej informacji można znaleźć w temacie “Jakość predykcji” na stronie 133.

Zmienne

W widoku głównym Zmienne wyświetlane są zmienne przetworzone oraz informacja, czy proces automatycznego przygotowania danych zaleca ich użycie w dalszej części modeli. Zalecenia dla każdej zmiennej można zastąpić; na przykład, aby wykluczyć utworzone predyktory lub uwzględnić predyktory, które proces automatycznego przygotowania danych zaleca wykluczyć. Jeśli zmienna została przekształcona, można zdecydować, czy zaakceptować zalecane przekształcenie, czy też użyć oryginalnej wersji.

Widok Zmienne składa się z dwóch tabel, jednej dla zmiennych przewidywanych, drugiej dla predyktorów, które zostały przetworzone lub utworzone

Tabela zmiennych przewidywanych

Tabela **Zmienna przewidywana** jest wyświetlana tylko po zdefiniowaniu zmiennej w danych.

Tabela składa się z dwóch kolumn:

- **Nazwa.** Jest to nazwa lub etykieta zmiennej przewidywanej; zawsze używana jest oryginalna nazwa, nawet jeśli zmienna została przekształcona.
- **Poziom pomiaru.** Wyświetla ikonę reprezentującą poziom pomiaru; należy ustawić wskaźnik myszy nad ikoną, aby wyświetlić etykietę (ilościowy, porządkowy, nominalny itd.), która opisuje dane.

  Jeśli zmienna przewidywana została przekształcona, ostatnia przekształcona wersja ma odzwierciedlenie w kolumnie **Poziom pomiaru**. *Uwaga*: nie można wyłączyć przekształceń dla zmiennych przewidywanych.

Tabela predyktorów

Tabela **Predyktory** jest zawsze wyświetlana. Każdy wiersz w tabeli odpowiada jednej zmiennej. Domyślnie wiersze są posortowane w porządku malejącym jakości predykcji.

W przypadku predyktorów porządkowych nazwa oryginalna zawsze używana jest jako nazwa wiersza. W tabeli wyświetlane są wersje oryginalne i pochodne zmiennej typu data/czas (w osobnych wierszach); tabela zawiera również predyktory utworzone.

Należy pamiętać, że przekształcone wersje zmiennych wyświetlane w tabeli zawsze reprezentują wersje końcowe.

Domyślnie w tabeli predyktorów wyświetlane są tylko zmienne zalecane. Aby wyświetlić pozostałe zmienne, należy zaznaczyć pole **Umieść nierekomendowane zmienne w tabeli** nad tabelą; zmienne te zostaną wówczas wyświetlone u dołu tabeli.

Tabela zawiera następujące kolumny:

- **Wersja do użycia.** Wyświetla listę rozwijaną, która pozwala kontrolować, czy zmienna będzie używana w dalszej części strumienia i czy użyć sugerowanych przekształceń. Domyślnie lista rozwijana odzwierciedla zalecenia.

  W przypadku predyktorów porządkowych, które zostały przekształcone, lista rozwijana zawiera trzy opcje do wyboru: **Przekształcone**, **Oryginalne** i **Nie używaj**.

  W przypadku predyktorów porządkowych, które nie zostały przekształcone, dostępne opcje to: **Oryginalne** i **Nie używaj**.

  Dla zmiennych pochodnych typu data/czas i predyktorów tworzonych dostępne są następujące opcje: **Przekształcone** i **Nie używaj**.

  Dla oryginalnych zmiennych daty lista rozwijana jest wyłączona i ustawiona jest opcja **Nie używaj**.

  Uwaga: W przypadku predyktorów oryginalnych i przekształconych zmiana wersji pomiędzy **Oryginalne** i **Przekształcone** powoduje automatyczną aktualizację ustawień **Poziom pomiaru** i **Jakość predykcji** dla tych zmiennych.
- **Nazwa.** Każda nazwa zmiennej stanowi odsyłacz. Kliknięcie nazwy pozwala wyświetlić dodatkowe informacje na temat zmiennej w powiązanym widoku. Więcej informacji można znaleźć w temacie “Szczegóły zmiennej” na stronie 133.
- **Poziom pomiaru.** Wyświetla ikonę reprezentującą typ danych; należy ustawić wskaźnik myszy nad ikoną, aby wyświetlić etykietę (ilościowy, porządkowy, nominalny itd.), która opisuje dane.
- **Jakość predykcji.** Jakość predykcji jest wyświetlana tylko dla zmiennych zalecanych przez proces automatycznego przygotowania danych. Ta kolumna nie jest wyświetlana, jeśli nie zdefiniowano żadnej zmiennej przewidywanej. Jakość predykcji może należeć do zakresu od 0 do 1, przy czym wyższe wartości oznaczają „lepsze” predyktory. Ogólnie jakość predykcji jest przydatna do porównywania predyktorów z analizą procesu automatycznego przygotowania danych, ale wartości jakości predykcji nie powinny być porównywane w ramach analizy.

Podsumowanie kroku

W każdym kroku wykonywanym w ramach automatycznego przygotowania danych predyktory wejściowe są przekształcane i/lub filtrowane; zmienne, które pozostaną po wykonaniu jednego kroku, są wykorzystywane w kolejnym. Zmienne, które pozostaną do ostatniego kroku, są wówczas zalecane do użycia w modelowaniu, natomiast wartości wejściowe dla przekształconych i tworzonych predyktorów zostają odfiltrowane.

Podsumowanie kroku to prosta tabela, która zawiera listę czynności przetwarzania wykonywanych w procesie automatycznego przygotowania danych. Jeśli jakiś **krok** jest podkreślony, można go kliknąć, aby w powiązanym widoku wyświetlić dodatkowe szczegóły na temat wykonywanej czynności. Więcej informacji można znaleźć w temacie “Szczegóły działania” na stronie 134.

Uwaga: Wyświetlane są tylko oryginalne i końcowe przekształcone wersje poszczególnych zmiennych; wersje pośrednie, używane w czasie analizy, nie są wyświetlane.

Jakość predykcji

Wykres jest wyświetlany domyślnie po pierwszym uruchomieniu analizy lub po wybraniu opcji **Predyktory zalecane do wykorzystania w analizie** w widoku głównym Podsumowanie przetwarzania zmiennych; przedstawia jakość predykcji zalecanych predyktorów. Zmienne są uporządkowane według jakości predykcji, przy czym zmienna o najwyższej wartości jest wyświetlana na samej górze.

W przypadku przekształconych wersji predyktorów porządkowych nazwa zmiennej zawiera przyrostek wybrany w panelu Nazwy zmiennych na karcie Ustawienia; przykładowo: *_transformed*.

Ikony poziomów pomiaru są wyświetlane za nazwami poszczególnych zmiennych.

Jakość predykcji każdego zalecanego predyktora jest obliczana na podstawie regresji liniowej lub modelu Naïve Bayes, w zależności od tego, czy zmienna przewidywana jest ilościowa, czy jakościowa.

Tabela zmiennych

Widok Tabela zmiennych jest wyświetlany po kliknięciu opcji **Zmienna przewidywana**, **Predyktory** lub **Niewykorzystane predyktory** w widoku głównym Podsumowanie przetwarzania zmiennych; zawiera prostą tabelę z listą odpowiednich funkcji.

Tabela składa się z dwóch kolumn:

- **Nazwa.** Nazwa predyktora.

  W przypadku zmiennych przewidywanych używana jest oryginalna nazwa lub etykieta zmiennej, nawet jeśli zmienna została przekształcona.

  W przypadku przekształconych wersji predyktorów porządkowych nazwa zawiera przedrostek wybrany w panelu Nazwy zmiennych na karcie Ustawienia; przykładowo: *_transformed*.

  W przypadku zmiennych pochodnych wyznaczonych na podstawie daty i czasu używana jest nazwa końcowej przekształconej wersji; na przykład: *bdate_years*.

  W przypadku tworzonych predyktorów używana jest nazwa utworzonego predyktora; na przykład: *Predictor1*.
- **Poziom pomiaru.** Wyświetla ikonę reprezentującą typ danych.

  Dla zmiennych przewidywanych opcja **Poziom pomiaru** zawsze odzwierciedla wersję przekształconą (o ile zmienna przewidywana została przekształcona); na przykład, zmiana typu z porządkowego (zbiór uporządkowany) na ilościowy (zakres, skala) lub odwrotnie.

Szczegóły zmiennej

Widok Szczegóły zmiennej jest wyświetlany po kliknięciu dowolnej **nazwy** w widoku głównym Zmienne; zawiera wykres rozkładu, braków danych i jakości predykcji (o ile ma zastosowanie) dla wybranej zmiennej. Ponadto wyświetlana jest również historia przetwarzania zmiennej oraz nazwa zmiennej przekształconej (o ile ma zastosowanie).

Dla każdego zestawu wykresów wyświetlane są obok siebie dwie wersje do porównania zmiennej: z zastosowanym przekształceniem i bez przekształcenia; jeśli wersja zmiennej po przekształceniu nie istnieje, wyświetlany jest tylko wykres dla oryginalnej zmiennej. Dla zmiennych pochodnych typu data lub czas oraz utworzonych predyktorów wyświetlane są tylko wykresy dla nowego predyktora.

Uwaga: Jeśli zmienna została wykluczona z powodu zbyt dużej liczby kategorii, wyświetlana jest tylko historia przetwarzania.

Wykres rozkładu

Rozkład zmiennych ilościowych jest wyświetlany w postaci histogramu, z nałożoną krzywą normalną i pionową linią odniesienia dla wartości średniej; zmienne jakościowe są wyświetlane w postaci wykresu słupkowego.

Histogramy są opatrzone etykietami, które wskazują odchylenie standardowe i skośność; skośność nie jest jednak wyświetlana, jeśli liczba wartości wynosi 2 lub mniej lub wariancja oryginalnej zmiennej jest mniejsza niż 10-20.

Ustawienie wskaźnika myszy nad wykresem pozwala wyświetlić średnią dla histogramów lub liczebność i wartość procentową łącznej liczby rekordów dla kategorii na wykresach słupkowych.

Wykres braków danych

Na wykresach kołowych porównywana jest wartość procentowa braków danych z zastosowanym przekształceniem lub bez przekształcenia; etykiety na wykresie wskazują wartość procentową.

Jeśli proces automatycznego przygotowania danych obejmował traktowanie braków danych, na wykresie kołowym po przekształceniu w postaci etykiety przedstawiana jest również wartość zastępcza — czyli wartość użyta zamiast braków danych.

Ustawienie wskaźnika myszy nad wykresem spowoduje wyświetlenie liczebności oraz wartości procentowej braków danych dla łącznej liczby rekordów.

Wykres jakości predykcji

Dla zmiennych zalecanych wykresy słupkowe przedstawiają jakość predykcji przed przekształceniem i po przekształceniu. Jeśli zmienna przewidywana została przekształcona, obliczona jakość predykcji odnosi się do przekształconej zmiennej.

Uwaga: Wykresy jakości predykcji nie są wyświetlane, jeśli nie zdefiniowano żadnej zmiennej przewidywanej lub jeśli zmienna przewidywana została kliknięta w panelu widoku głównego.

Ustawienie wskaźnika myszy nad wykresem pozwala wyświetlić wartość jakości predykcji.

Tabela historii przetwarzania

Tabela przedstawia, w jaki sposób wyliczona została przekształcona wersja zmiennej. Kroki wykonane w procesie automatycznego przygotowania danych są wyświetlane w kolejności, w jakiej zostały wykonane; jednak w przypadku niektórych kroków dla danej zmiennej wykonanych mogło być kilka czynności.

Uwaga: Ta tabela nie jest wyświetlana dla zmiennych, które nie zostały przekształcone.

Informacje w tabeli są podzielone na dwie lub trzy kolumny:

- **Podjęte działania.** Nazwa podjętego działania. Na przykład Predyktory ilościowe. Więcej informacji można znaleźć w temacie “Szczegóły działania”.
- **Szczegóły.** Lista przeprowadzonych procesów. Na przykład Transformuj do jednostek standardowych.
- **Funkcja.** Opcja wyświetlana tylko dla utworzonych predyktorów; wyświetla kombinację liniową zmiennych wejściowych, na przykład: 0,06*age + 1,21*height, gdzie age oznacza wiek, a height wzrost.

Szczegóły działania

Widok powiązany Szczegóły działania jest wyświetlany po wybraniu dowolnego podkreślonego **działania** w widoku głównym Podsumowanie kroku; wyświetla informacje specyficzne dla działania oraz informacje wspólne dla wszystkich wykonanych kroków przetwarzania; szczegóły specyficzne dla działania są wyświetlane jako pierwsze.

Dla każdego kroku w górnej części powiązanego widoku zamieszczany jest opis, który stanowi jego tytuł. Szczegóły specyficzne dla działania są wyświetlane pod tytułem i mogą zawierać informacje, takie jak liczba wyliczonych predyktorów, zmienne rekategoryzowane, przekształcenia zmiennych przewidywanych, kategorie połączone lub uporządkowane oraz predyktory utworzone lub wykluczone.

W trakcie przetwarzania poszczególnych działań liczba użytych predyktorów może ulec zmianie, na przykład po wykluczeniu lub połączeniu predyktorów.

Uwaga: Jeśli działanie zostało wyłączone lub nie określono żadnej zmiennej przewidywanej, po kliknięciu działania w widoku głównym Podsumowanie kroku zamiast szczegółów działania wyświetlany jest komunikat o błędzie.

Dostępnych jest dziewięć działań; jednak nie wszystkie muszą być aktywowane dla każdej analizy.

Tabela zmiennych tekstowych

W tabeli wyświetlane są:

- Wartości spacji końcowych są przycinane.
- Predyktory wykluczone z analizy.

Tabela predyktorów daty i czasu

W tabeli wyświetlane są:

- Czasy trwania wyznaczone na podstawie predyktorów daty i czasu.
- Elementy daty i czasu.
- Wyliczone predyktory daty i czasu, łącznie.

Data lub czas odniesienia są wyświetlane jako przypis, o ile czasy trwania zostały obliczone.

Tabela monitorowania predyktorów

W tabeli wyświetlana jest liczba następujących predyktorów wykluczonych z przetwarzania:

- Stałe.
- Predyktory ze zbyt dużą liczbą braków danych.
- Predyktory ze zbyt dużą obserwacji w jednej kategorii.
- Zmienne nominalne (zbiory) ze zbyt dużą liczbą kategorii.
- Predyktory monitorowane, łącznie.

Tabela sprawdzania poziomu pomiaru

W tabeli wyświetlana jest liczba rekategoryzacji zmiennych, podzielona w następujący sposób:

- Zmienne porządkowe (zbiór uporządkowany) uznane za ilościowe.
- Zmienne ilościowe uznane za porządkowe.
- Łączna liczba rekategoryzacji.

Jeśli żadna zmienna wejściowa (zmienne przewidywane lub predyktory) nie była ilościowa lub porządkowa, informacja ta jest wyświetlana jako przypis.

Tabela wartości odstających

W tej tabeli przedstawiana jest liczebność wartości odstających, jakie były obsługiwane.

- Liczba zmiennych ilościowych, dla których wartości odstające zostały wykryte i odcięte, lub liczba zmiennych ilościowych, dla których wartości odstające zostały wykryte i ustawione jako braki danych, w zależności od ustawień w panelu Zmienne wejściowe i przewidywana na karcie Ustawienia.
- Liczba zmiennych ilościowych wykluczonych, ponieważ były stałe, po zakończeniu działań związanych z obsługą wartości odstających.

Jeden przypis przedstawia wartość odcięcia dla wartości odstających; drugi przypis jest wyświetlany, jeśli żadna zmienna wejściowa (przewidywana lub predyktor) nie była ilościowa.

Tabela braków danych

W tabeli wyświetlana jest liczba zmiennych z zastąpionymi brakami danych, podzielona na następujące części:

- Zmienna przewidywana. Ten wiersz nie jest wyświetlany, jeśli nie określono żadnej zmiennej przewidywanej.
- Predyktory. Ten obszar jest następnie podzielony na liczbę predyktorów nominalnych (zbiór), porządkowych (zbiór uporządkowany) i ilościowych.
- Łączna liczba zastąpionych braków danych.

Tabela zmiennych przewidywanych

Ta tabela przedstawia, czy zmienna przewidywana została przekształcona, w następujący sposób:

- Transformacja Boxa-Coxa na normalność. Ten obszar jest podzielony na kolumny, w których przedstawiane są określone kryteria (średnia i odchylenie standardowe) oraz parametr Lambda.
- Kategorie zmiennej przewidywanej ze zmienioną kolejnością w celu zwiększenia stabilności.

Tabela predyktorów jakościowych

W tej tabeli wyświetlana jest liczba predyktorów jakościowych:

- Których kolejność kategorii została zmieniona z najniższej na najwyższą w celu zwiększenia stabilności.
- Których kategorie zostały połączone w celu zmaksymalizowania powiązań ze zmienną przewidywaną.
- Których kategorie zostały połączone w celu umożliwienia obsługi kategorii małolicznych.
- Wykluczonych z powodu słabego powiązania ze zmienną przewidywaną.
- Wykluczonych, ponieważ były stałe po połączeniu.

Przypis jest wyświetlany, jeśli nie wystąpiły żadne predyktory jakościowe.

Tabela predyktorów ilościowych

Dostępne są dwie tabele. Pierwsza wyświetla jedno z następujących przekształceń:

- Wartości predyktora przekształcone na jednostki standardowe. Ponadto przedstawia liczbę przekształconych predyktorów, określoną średnią oraz odchylenie standardowe.
- Wartości predyktora zmapowane do wspólnego przedziału. Ponadto przedstawia liczbę predyktorów przekształconych z zastosowaniem transformacji min.-maks. oraz określone wartości minimalne i maksymalne.
- Wartości predykcyjne skategoryzowane oraz liczba skategoryzowanych predyktorów.

W drugiej tabeli przedstawiane są szczegóły dotyczące tworzenia przestrzeni predyktorów, wyświetlane jako liczba predyktorów:

- Utworzonych.
- Wykluczonych z powodu słabego powiązania ze zmienną przewidywaną.
- Wykluczonych, ponieważ były stałe po kategoryzacji.
- Wykluczonych, ponieważ były stałe po utworzeniu.

Przypis jest wyświetlany, jeśli żaden z predyktorów wejściowych nie był ilościowy.

Generowanie węzła Wyliczanie

Po wygenerowaniu węzła wyliczeń stosowane jest odwrotne przekształcenie zmiennej przewidywanej na zmienną oceny. Domyślnie węzeł wprowadza nazwę zmiennej wyniku, jaka powinna zostać utworzona przez węzeł automatycznego modelowania (np. Auto Klasyfikacja lub Auto Predykcja) lub węzeł zespolenia. Jeśli zmienna przewidywana skali (ilościowa) została przekształcona, zmienna oceny jest wyświetlana w przekształconych jednostkach; przykładowo, log($) zamiast $. Aby możliwe było interpretowanie i korzystanie z wyników, należy przekonwertować wartość predykcyjną z powrotem do oryginalnej skali.

Uwaga: Węzeł wyliczeń można wygenerować tylko wówczas, gdy węzeł automatycznego przygotowywania danych (ADP) obejmuje analizę, która umożliwia zmianę skali przewidywanej ilościowej (czyli na panelu Zmienne wejściowe i przewidywana wybrano metodę przeskalowania Boxa-Coxa). Węzła wyliczeń nie można wygenerować, jeśli zmienna przewidywana nie jest ilościowa lub jeśli nie wybrano metody przeskalowania Boxa-Coxa.

Węzeł wyliczeń jest tworzony w trybie wielomodalnym i używa w wyrażeniu obiektu @FIELD, dzięki czemu można w razie potrzeby dodać przekształconą zmienną przewidywaną. Przykładowo możliwe jest użycie następujących danych:

- Nazwa zmiennej przewidywanej: response (odpowiedź)
- Nazwa przekształconej zmiennej przewidywanej: response_transformed (odpowiedź_przekształcona)
- Nazwa zmiennej oceny: $XR-response_transformed ($XR-odpowiedź_przekształcona)

Węzeł wyliczeń utworzy nową zmienną: $XR-response_transformed_inverse ($XR-odwrócona_odpowiedź_przekształcona).

Uwaga: Jeśli nie jest używany węzeł automatycznego modelowania lub węzeł Zespół, konieczne będzie przeprowadzenie edycji węzła wyliczeń w celu przekształcenia go na poprawną zmienną oceny dla modelu.

Znormalizowane docelowe wartości ilościowe

Domyślnie, po zaznaczeniu pola wyboru **Przeskaluj docelową wartość ilościową za pomocą transformacji Boxa-Coxa** w panelu Zmienne wejściowe i przewidywana przeprowadzane jest przekształcenie zmiennej przewidywanej i użytkownik tworzy nową zmienną, która będzie zmienną przewidywaną w czasie budowania modelu. Przykładowo, jeśli oryginalna zmienna przewidywana to *response* (odpowiedź), nową zmienną przewidywaną będzie *response_transformed* (odpowiedź_przekształcona); modele poniżej węzła ADP będą automatycznie wybierać nową zmienną przewidywaną.

Może to jednak powodować problemy, w zależności od oryginalnej zmiennej przewidywanej. Przykładowo, jeśli zmienną przewidywaną była zmienna *Age* (Wiek), wartościami nowej zmiennej przewidywanej nie będą *Years* (Lata), ale przekształcona wersja zmiennej *Years*. Oznacza to, że nie można patrzeć na wyniki i interpretować ich, ponieważ ich jednostki są nierozpoznawalne. W takim przypadku można zastosować przekształcenie odwrotne, które przywróci przekształcone jednostki do poprzedniej postaci. W tym celu:

1. Po kliknięciu przycisku **Analizuj dane** w celu uruchomienia analizy automatycznego przygotowywania danych należy wybrać opcję *Węzeł wyliczeń* z menu *Utwórz*.
2. Węzeł wyliczeń należy umieścić po modelu użytkowym w obszarze roboczym modelu.

Węzeł wyliczeń przywróci dla zmiennej oceny oryginalne wymiary, dzięki czemu predykcja będzie przeprowadzana dla oryginalnych wartości *Years*.

Domyślnie węzeł wyliczeń przekształca zmienną oceny wygenerowaną przez automatyczne modelowanie lub model zespolony. Jeśli tworzony jest pojedynczy model, konieczne jest przeprowadzenie edycji węzła wyliczeń, aby obliczenia były wykonywane na podstawie rzeczywistej zmiennej oceny. Jeśli model ma zostać oceniony, należy dodać przekształconą zmienną przewidywaną w węźle wyliczeń w polu **Wylicz z**. Spowoduje to takie same przekształcenie odwrotne dla zmiennej przewidywanej, a każdy węzeł oceny lub analizy w dalszej części strumienia będzie używał poprawnie przekształconych danych, dopóki węzły nie zostaną zmienione w celu użycia nazw zmiennych zamiast metadanych.

Aby przywrócić również oryginalną nazwę, można użyć węzła filtrowania w celu usunięcia oryginalnej zmiennej przewidywanej, o ile jest jeszcze dostępna, i zmienić nazwy zmiennej przewidywanej i zmiennej oceny.

Węzeł Typy

Właściwości zmiennej można określić w węźle źródłowym lub w osobnym węźle typu. Działanie jest podobne w przypadku obu węzłów. Dostępne są następujące właściwości:

- **Zmienna** Należy dwukrotnie kliknąć dowolną nazwę zmiennej, aby określić etykiety wartości i zmiennej dla danych w programie IBM SPSS Modeler. Przykładowo można tutaj wyświetlić lub zmodyfikować metadane zmiennej zaimportowane z programu IBM SPSS Statistics. Podobnie, można utworzyć nowe etykiety dla zmiennych i ich wartości. Określone tutaj etykiety są wyświetlane w programie IBM SPSS Modeler w zależności od opcji wybranych w oknie dialogowym właściwości strumienia.
- **Poziom pomiaru** Jest to poziom pomiaru używany do opisania charakterystyk danych w określonej zmiennej. Jeśli wszystkie szczegóły zmiennej są znane, jest ona nazywana **w pełni określoną**. Aby uzyskać więcej informacji, zobacz “Poziomy pomiaru” na stronie 139.

  Uwaga: Poziom pomiaru zmiennej różni się od typu składowania, który wskazuje, że dane są zapisywane jako łańcuchy, liczby całkowite, liczby rzeczywiste, daty, godziny, znaczniki czasu lub listy.
- **Wartości** Ta kolumna umożliwia określenie opcji do odczytywania wartości danych ze zbiory danych lub użycie opcji **Określ** w celu określenia poziomów pomiaru i wartości w osobnym oknie dialogowym. Można również wybrać opcję przepuszczenia zmiennych bez odczytywania ich wartości. Aby uzyskać więcej informacji, zobacz “Wartości danych” na stronie 143.

  Uwaga: Nie można poprawiać wartości komórek w tej kolumnie, jeśli odpowiadająca jej pozycja **Zmienna** zawiera listę.
- **Braki** Służy do określania sposoby traktowania braków danych dla zmiennej. Aby uzyskać więcej informacji, zobacz “Definiowanie braków danych” na stronie 148.

  Uwaga: Nie można poprawiać wartości komórek w tej kolumnie, jeśli odpowiadająca jej pozycja **Zmienna** zawiera listę.
- **Sprawdź** W tej kolumnie można ustawić opcje, dzięki którym wartości zmiennej będą zgodne z określonymi wartościami lub zakresami. Aby uzyskać więcej informacji, zobacz “Sprawdzanie wartości typu” na stronie 148.

  Uwaga: Nie można poprawiać wartości komórek w tej kolumnie, jeśli odpowiadająca jej pozycja **Zmienna** zawiera listę.
- **Rola** Służy do przekazania węzłom modelowania informacji, czy zmienne będą **wejściowe** (zmienne predykcyjne) czy **przewidywane** (zmienne przewidywane), na potrzeby procesu uczenia maszynowego. Dostępne są również role **Łącznie** i **Brak** oraz **Partycja**, która wskazuje zmienną użytą do podziału rekordów na osobne próby na potrzeby uczenia, testowania i walidacji. Wartość **Podział** określa, że dla każdej możliwej wartości zmiennej tworzone będą osobne modele. Aby uzyskać więcej informacji, zobacz “Ustawianie roli zmiennej” na stronie 149.

Korzystając z okna węzła typu, można określić kilka innych opcji:

- Korzystając z przycisku menu narzędzi, po określeniu węzła typu (za pośrednictwem specyfikacji, poprzez odczytanie wartości lub uruchomienie strumienia) można wybrać opcję **Ignoruj zmienne z jedną wartością**. Ignorowanie unikalnych zmiennych spowoduje automatyczne ignorowanie zmiennych z jedną wartością.
- Korzystając z przycisku menu narzędzi, po określeniu węzła typu można wybrać opcję **Ignoruj zmienne z licznymi kategoriami**. Ignorowanie zmiennych z licznymi kategoriami spowoduje automatyczne ignorowanie zbiorów z dużą liczbą elementów.
- Po określeniu węzła typu za pośrednictwem przycisku menu narzędzi można wybrać opcję **Zamień ilościowe zmienne całkowite na zmienne porządkowe**. Więcej informacji można znaleźć w temacie “Przekształcanie danych ilościowych” na stronie 142.
- Za pomocą przycisku menu narzędzi można wygenerować węzeł filtrowania w celu odrzucenia wybranych zmiennych.
- Korzystając z przycisków przełączania sunglasses (okulary), można ustawić wartości domyślne dla wszystkich zmiennych na Odczyt lub Przepuść. Karta typów w węźle źródłowym przepuszcza zmienne domyślnie, o ile węzeł typu domyślnie sam odczytuje wartości.
- Użycie przycisku **Wyczyść wartości** umożliwia wyczyszczenie zmian wartości zmiennych, jakie zostały wprowadzone w tym węźle (wartości nieodziedziczone), i odczytanie wartości z operacji wykonanych we wcześniejszej części strumienia. Ta opcja jest przydatna w celu zresetowania zmian wprowadzonych dla konkretnych zmiennych we wcześniejszej części strumienia.

- Użycie przycisku **Wyczyść wszystkie** pozwala ponownie ustawić wartości dla **wszystkich** zmiennych wczytanych do węzła. Ta opcja ustawia kolumnę **Wartości** na **Odczytaj** dla wszystkich zmiennych. Opcja ta jest przydatna w celu ponownego ustawienia wartości dla wszystkich zmiennych i ponownego odczytania wartości i typów na podstawie operacji we wcześniejszej części strumienia.
- Korzystając z menu kontekstowego, można wybrać opcję **kopiowania** atrybutów z jednej zmiennej do drugiej. Więcej informacji można znaleźć w temacie “Kopiowanie atrybutów typu” na stronie 149.
- Opcja **Widok ustawień niewykorzystanych zmiennych** umożliwia wyświetlenie ustawień typu wyświetlania dla zmiennych, których nie ma już w danych lub które zostały połączone z tym węzłem typu. Jest to przydatne w przypadku ponownego użycia węzła typu w zbiorach danych, które uległy zmianie.

Poziomy pomiaru

Poziom pomiaru (dawniej znany jako „typ danych” lub „typ użycia”) opisuje użycie zmiennych danych w programie IBM SPSS Modeler. Poziom pomiaru może być określony na karcie Typy w węźle źródłowym lub w węźle typu. Można na przykład ustawić poziom pomiaru dla zmiennej całkowitej zawierającej wartości 1 i 0 jako *Flaga*. Zwykle oznacza to, że 1 = *Prawda*, a 0 = *Fałsz*.

Składowanie a pomiar. Należy pamiętać, że poziom pomiaru zmiennej różni się od jej typu składowania, który wskazuje, czy dane są składowane jako łańcuch, liczba całkowita, liczba rzeczywista, data, czas lub znacznik czasu. Typy danych można modyfikować w dowolnym punkcie strumienia za pomocą węzła typu, natomiast składowanie należy określić na poziomie źródła podczas wczytywania danych do programu IBM SPSS Modeler (chociaż można je później zmienić za pomocą funkcji przekształcenia). Więcej informacji można znaleźć w temacie “Ustawienia składowania i formatowania zmiennej” na stronie 9.

Niektóre węzły modelowania wskazują dozwolone typy poziomu pomiaru dla zmiennych wejściowych i przewidywanych za pośrednictwem ikon na kacie Zmienne.

Ikony poziomów pomiaru

Tabela 21. Ikony poziomów pomiaru

Ikona	Poziom pomiaru
	Domyślny
	Ilościowy
	Jakościowy
	Flaga
	Nominalny
	Porządkowy
	Nieokreślony

Tabela 21. Ikony poziomów pomiaru (kontynuacja)

Ikona	Poziom pomiaru
[ ]	Przedziałowy
	Geoprzestrzenny

Dostępne są następujące poziomy pomiaru:

- **Domyślne** Data, których typ składowania i wartości są nieznane (ponieważ przykładowo nie zostały jeszcze odczytane) są wyświetlane jako **<Domyślne>**.
- **Ilościowy** Umożliwia opisanie wartości liczbowych, takich jak zakres od 0 do 100 lub od 0,75 do 1,25. Wartość ilościowa może być liczbą całkowitą, liczbą rzeczywistą lub wartością typu data/czas.
- **Jakościowy** Stosowany w przypadku wartości łańcuchowych, kiedy dokładna liczba odmiennych wartości jest nieznana. Jest to **nieokreślony** typ danych, co oznacza, że żądne informacje na temat składowania i użycia danych nie są jeszcze znane. Po odczytaniu danych poziom pomiaru zostanie ustawiony jako *Flaga*, *Nominalny* lub *Nieokreślony*, w zależności od maksymalnej liczby elementów zmiennych nominalnych określonej w oknie dialogowym węzła Właściwości strumienia.
- **Flaga** Ten typ jest używany w przypadku danych z dwoma odmiennymi wartościami, które wskazują na obecność lub nieobecność danej cechy, np. prawda i fałsz, Tak i Nie lub 0 i 1. Wartości mogą różnić się, ale jedna z nich musi być zawsze wartością "true" a druga — "false". Dane mogą być reprezentowane jako tekst, liczba całkowita, liczba rzeczywista, data, czas lub znacznik czasu.
- **Nominalny** Służy do opisu danych z wieloma odmiennymi wartościami, a każda z nich jest traktowana jako element zbioru, np. mała/średnia/duża. Dane nominalne mogą mieć dowolny typ składowania — numeryczny, łańcuch lub data/czas. Należy pamiętać, że ustawienie poziomu pomiaru jako *Nominalny* nie powoduje automatycznej zmiany wartości na składowanie łańcuchowe.
- **Porządkowy** Służy do opisywania danych z wieloma odmiennymi wartościami, które mają dziedziczną kolejność. Przykładowo, typ danych porządkowych można przypisać do kategorii wynagrodzenia lub stopni zadowolenia. Porządek jest definiowany zgodnie z rzeczywistym porządkiem sortowania elementów danych. Przykładowo, 1, 3, 5 to domyślny porządek sortowania dla zbioru liczb całkowitych, a HIGH, LOW, NORMAL (Wysoki, Niski, W normie) (rosnąco w kolejności alfabetycznej) to porządek dla zbioru łańcuchów. Porządkowy poziom pomiaru umożliwia zdefiniowanie zbioru danych jakościowych jako dane porządkowe na potrzeby wizualizacji, budowania modelu i eksportowania do innych aplikacji (takich jak IBM SPSS Statistics), które jako typ rozróżniania przyjmują dane porządkowe. Zmiennej porządkowej można użyć wszędzie tam, gdzie może być użyta zmienna nominalna. Ponadto, jako porządkowe można zdefiniować zmienne z dowolnym typem składowania (liczba rzeczywista, liczba całkowita, łańcuch, data, czas itd.).
- **Nieokreślony** Używany w przypadku danych, które nie odpowiadają żadnemu z powyższych typów, zmiennych z pojedynczą wartością lub danych nominalnych, w których zbiór zawiera więcej elementów niż zdefiniowane maksimum. Jest również przydatny, jeśli w przeciwnym razie poziom pomiaru byłby zbiorem dowolnych elementów (np. numer konta). Po wybraniu typu **Nieokreślony** dla zmiennej jej rola jest automatycznie ustawiana ba **Brak**, z opcją **ID rekordów** jako jedyna alternatywa. Domyślnie maksymalna wielkość zbiorów jest ustawiona na 250 unikalnych wartości. Liczbę tę można skorygować lub wyłączyć on na karcie Opcje w oknie dialogowym Właściwości strumienia, do którego można uzyskać dostęp z menu Narzędzia.
- **Przedziałowy** Umożliwia identyfikowanie danych zapisanych w postaci listy, które nie są danymi geoprzestrzennymi. Przedział to w rzeczywistości zmienna listy o głębokości zero, w której elementy z listy mają przypisany jeden z pozostałych poziomów pomiaru.

  Więcej informacji o listach zawiera sekcja Składowanie listy i powiązane poziomy pomiaru w publikacji SPSS Modeler — węzły źródłowe, procesu i wyników
- **Dane geoprzestrzenne** Służy do identyfikowania danych geoprzestrzennych z typem składowania Lista. Listy mogą być zmiennymi Lista liczb całkowitych lub Lista liczb rzeczywistych, a ich głębokość może być określona w przedziale od zero do dwóch włącznie.

Więcej informacji zawiera temat Geoprzestrzenne podpoziomy pomiarów w sekcji Węzeł Typy w publikacji SPSS Modeler — węzły źródłowe, procesowe i wyników.

Istnieje możliwość ręcznego określenia poziomów pomiaru lub można zezwolić, aby oprogramowanie odczytało dane i ustaliło poziom pomiaru na podstawie odczytanych wartości.

Alternatywnie, jeśli dostępnych jest kilka zmiennych danych ilościowych, które powinny być traktowane jako zmienne jakościowe, można wybrać opcję pozwalającą na ich przekształcenie. Więcej informacji można znaleźć w temacie “Przekształcanie danych ilościowych” na stronie 142.

Aby użyć automatycznego wpisywania

1. W węźle typu lub na karcie Typy węzła źródłowego należy ustawić kolumnę **Wartości** na **<Odczyt>** dla wybranych zmiennych. Dzięki temu metadane będą dostępne dla wszystkich węzłów poniżej bieżącego. Można szybko ustawić wszystkie zmienne na **<Odczyt>** lub **<Przepuść>**, używając przycisków okularów w oknie dialogowym.
2. Kliknięcie przycisku **Odczytaj wartości** umożliwia odczytanie wartości bezpośrednio ze źródła danych.

Aby ręcznie ustawić poziom pomiaru dla zmiennej

1. Wybierz zmienną z tabeli.
2. Z listy rozwijanej w kolumnie **Poziom pomiaru** wybierz poziom pomiaru dla zmiennej.
3. Alternatywnie można użyć metody Ctrl+A lub Ctrl+kliknięcie, aby wybrać wiele zmiennych przed użyciem listy rozwijanej w celu wybrania poziomu pomiaru.

Geoprzestrzenne podpoziomy pomiarów

Geoprzestrzenny poziom pomiaru, który jest używany z typem składowania Lista, obejmuje sześć podpoziomów, które są używane do określania różnych typów danych geoprzestrzennych.

- **Punkt** — Określa konkretną lokalizację; na przykład centrum miasta.
- **Wielokąt** — Seria punktów, które identyfikują pojedynczą granicę regionu i jego lokalizację; na przykład kraj.
- **Łańcuch** — Znany również jako łamana lub linia; łańcuch jest serią punktów, które określają przebieg linii. Na przykład, Łańcuch może być elementem, takim jak droga lub rzeka; może również śledzić obiekt, który się porusza, np. może to być trasa lotu samolotu lub rejsu statku.
- **Multipunkt** — Jest używany, kiedy każdy wiersz danych zawiera wiele punktów w regionie. Na przykład, jeśli każdy wiersz reprezentuje ulicę w miejscowości, wiele punktów dla każdej ulicy może określić każdą latarnię uliczną.
- **Multiwielokąt** — Jest używany, kiedy każdy wiersz w danych zawiera kilka wielokątów. Na przykład, jeśli każdy wiersz reprezentuje obrys kraju, Stany Zjednoczone mogą być zapisane jako kilka wielokątów identyfikujących różne obszary, takie jak stały ląd, Alaska i Hawaje.
- **Multiłańcuch** — Jest używany, kiedy każdy wiersz w danych zawiera kilka linii. Ponieważ linie nie mogą się rozgałęziać, należy użyć multiłańcucha, aby określić grupę linii. Na przykład dane, takie jak szlaki żeglowne lub sieć kolejowa w danym kraju.

Te podpoziomy pomiaru są używane z typem składowania Lista. Aby uzyskać więcej informacji, zobacz “Składowanie listy i powiązane poziomy pomiaru” na stronie 12.

Ograniczenia

W przypadku korzystania z danych geoprzestrzennych należy pamiętać możliwych ograniczeniach.

- Układ współrzędnych może wpływać na format danych. Przykładowo, rzutowany układ współrzędnych korzysta z wartości współrzędnych x, y i (o ile jest to konieczne) z, podczas gdy w układzie współrzędnych geograficznych używane są wartości współrzędnych dla długości i szerokości geograficznej oraz (o ile jest to wymagane) wartości dla wysokości lub głębokości.

  Więcej informacji na temat układów współrzędnych zawiera temat Konfigurowanie opcji geoprzestrzennych strumieni w sekcji Praca ze strumieniami w dokumentacji SPSS Modeler — podręcznik użytkownika.

- Łańcuch nie może przecinać samego siebie.
- Wielokąt nie jest obiektem samozamykającym się; dla każdego wielokąta należy upewnić się, że pierwszy i ostatni punkt zostały zdefiniowane jako ten sam punkt.
- W multiwielokącie kierunek danych jest istotny; zgodnie z ruchem wskazówek zegara oznacza formę pełną, a kierunek przeciwny do ruchu wskazówek zegara oznacza otwór. Przykładowo, jeśli zapisywana jest powierzchnia kraju, która obejmuje jeziora, granica obszaru lądu stałego może zostać zapisana w kierunku zgodnym z ruchem wskazówek zegara, a kształt każdego jeziora w kierunku przeciwnym do ruchu wskazówek zegara.
- Wielokąt nie może przecinać samego siebie. Przykładem takiego przecięcia jest próba wykreślenia granicy wielokąta jako linia ciągła w takiej postaci jak na rysunku 8.
- Multiwielokąty nie mogą na siebie nachodzić.
- W przypadku zmiennych geoprzestrzennych możliwe typy składowania to **Liczba rzeczywista** i **Liczba całkowita** (ustawienie domyślne to **Liczba rzeczywista**).

Ikony geoprzestrzennego podpoziomu poziom pomiaru

Tabela 22. Ikony geoprzestrzennego podpoziomu poziom pomiaru

Ikona	Poziom pomiaru
	Punkt
	Wielokąt
	Łańcuch
	Multipunkt
	Multiwielokąt
	Multiłańcuch

Przekształcanie danych ilościowych

Traktowanie danych jakościowych jako ilościowe ma poważny wpływ na jakość modelu, szczególnie w przypadku zmiennych przewidywanych; przykładowo, utworzenie modelu regresji zamiast modelu binarnego. Aby tego uniknąć, można przekształcić zakresy liczb całkowitych na typ jakościowy, taki jak *porządkowy* lub *flaga*.

1. Za pomocą przycisku menu operacji i generowania (oznaczonym symbolem narzędzia) należy wybrać opcję **Zamień ilościowe na porządkowe**. Zostanie wyświetlone okno dialogowe przekształcania wartości.
2. Należy określić wielkość zakresu, jaki zostanie automatycznie przekształcony; ma to zastosowanie w przypadku każdego zakresu maksymalnie do i z uwzględnieniem wprowadzonej wielkości.
3. Kliknij przycisk **OK**. Odpowiednie zakresy są przekształcane na typ *Flaga* lub *Porządkowy* i są wyświetlane na karcie Typy węzła Typy.

Wyniki przekształcenia

- Podczas gdy zmienna *ilościowa* z typem składowania Liczba całkowita jest zmieniana na zmienną *porządkową*, dolne i górne wartości są rozwijane, tak aby obejmowały wszystkie wartości całkowite od dolnej do górnej granicy. Przykładowo, jeśli zakres to 1, 5, zbiór będzie obejmował wartości 1, 2, 3, 4, 5.

- Jeśli zmienna *ilościowa* jest zamieniana na *flagę*, the dolne i górne wartości stają się wartościami fałsz i prawda zmiennej typu Flaga.

Co to jest określanie?

Określanie to proces odczytu lub podania informacji, takich jak typ składowania i wartości zmiennych danych. W celu zoptymalizowania zasobów systemowych określanie stanowi proces ukierunkowany na użytkownika — użytkownik może wskazać w oprogramowaniu, aby wartości były odczytywane poprzez określenie opcji na karcie Typy w węźle źródłowym lub poprzez uruchomienie danych za pośrednictwem węzła Typy.

- Dane o nieznanych typach są również nazywane danymi *nieokreślonymi*. Dane, których typ składowania i wartości są nieznane, są wyświetlane w kolumnie *Poziom pomiaru* na karcie Typy jako **<Domyślne>**.
- Jeśli dostępne są jakieś informacje na temat typu składowania zmiennych, np. łańcuch lub numeryczny, dane są nazywane *częściowo określonymi*. **Jakościowy** lub **ilościowy** to częściowo określone poziomy pomiaru. Przykładowo **jakościowy** określa, że zmienna jest symboliczna, ale nie wiadomo, czy jest ona nominalna, porządkowa, czy jest to flaga.
- Jeśli wszystkie szczegóły na temat typu są znane, w tym wartości, w tej kolumnie wyświetlany jest *w pełni określony* poziom pomiaru — nominalny, porządkowy, flaga lub ilościowy. Należy zwrócić uwagę, że typ *ilościowy*jest używany zarówno w przypadku zmiennych częściowo określonych, jak i w pełni określonych. Dane ilościowe mogą być liczbami całkowitymi lub liczbami rzeczywistymi.

W trakcie wykonywania strumienia danych z użyciem węzła Typy typy nieokreślone stają się od razu częściowo określone, w oparciu o początkowe wartości danych. Kiedy wszystkie dane przejdą przez węzeł, stają się w pełni określone, o ile dla wartości wybrano ustawienie **<Przepuść>**. Jeśli wykonywanie zostanie przerwane, dane pozostaną częściowo określone. Po określeniu karty Typy wartości zmiennych są statyczne w tym punkcie strumienia. Oznacza to, że wszelkie zmiany we wcześniejszej części strumienia nie będą wpływały na wartości konkretnej zmiennej, nawet jeśli strumień zostanie ponownie uruchomiony. Aby zmienić lub zaktualizować wartości w oparciu o nowe dane lub dodane operacje, należy dokonać edycji w samym węźle typu lub ustawić wartość dla zmiennej jako **<Odczytaj>** lub **<Odczytaj +>**.

Kiedy przeprowadzić określanie

Ogólnie, jeśli zbiór danych nie jest bardzo duży i jeśli w dalszej części strumienia zmienne nie będą dodawane, najwygodniejszą metodą jest określenie w węźle źródłowym. Określenie w osobnym węźle typu jest jednak przydatne, jeśli:

- Zbiór danych jest duży i strumień filtruje podzbiór przed węzłem typu.
- Przeprowadzono filtrowanie danych w strumieniu.
- Dane w strumieniu zostały połączone lub dodane.
- W czasie przetwarzania wyliczane są nowe zmienne.

Uwaga: W przypadku eksportu danych w węźle eksportu do bazy danych wymagane jest pełne określenie danych.

Wartości danych

Używając kolumny **Wartości** węzła Typy, można automatycznie odczytać wartości z danych lub określić poziomy pomiaru oraz wartości w osobnym oknie dialogowym.

Opcje z listy rozwijanej Wartości udostępniają instrukcje dotyczące automatycznego wpisywania przedstawione w poniższej tabeli.

Tabela 23. Instrukcje automatycznego wpisywania

Opcja	Funkcja
<Odczytaj>	Dane są odczytywane w czasie wykonywania węzła.
<Odczytaj+>	Dane są odczytywane i dodawane do danych bieżących (o ile istnieją).
<Przepuść>	Żadne dane nie są odczytywane.

Tabela 23. Instrukcje automatycznego wpisywania (kontynuacja)

Opcja	Funkcja
<Bieżący>	Zachowywane są bieżące wartości danych.
Określ...	Otwierane jest osobne okno dialogowe, umożliwiające określenie opcji wartości i poziomu pomiaru.

Wykonanie węzła Typy lub kliknięcie przycisku **Odczytaj wartości** powoduje automatyczne wpisanie i odczytanie wartości ze źródła danych na podstawie wybranych opcji. Wartości te można również określić ręcznie, używając opcji Określ lub klikając dwukrotnie komórkę w kolumnie **Zmienna**.

Po wprowadzeniu zmian zmiennych w węźle Typy można ponownie ustawić informacje dotyczące wartości, używając następujących przycisków dostępnych na pasku narzędzi okna dialogowego:

- Użycie przycisku **Wyczyść wartości** umożliwia wyczyszczenie zmian wartości zmiennych, jakie zostały wprowadzone w tym węźle (wartości nieodziedziczone), i odczytanie wartości z operacji wykonanych we wcześniejszej części strumienia. Ta opcja jest przydatna w celu zresetowania zmian wprowadzonych dla konkretnych zmiennych we wcześniejszej części strumienia.
- Użycie przycisku **Wyczyść wszystkie** pozwala ponownie ustawić wartości dla **wszystkich** zmiennych wczytanych do węzła. Ta opcja ustawia kolumnę *Wartości* na **Odczytaj** dla wszystkich zmiennych. Opcja ta jest przydatna w celu ponownego ustawienia wartości dla wszystkich zmiennych i ponownego odczytania wartości i poziomów pomiaru na podstawie operacji w wcześniejszej części strumienia.

Wyszarzony tekst w kolumnie Wartości

W węźle typu lub w węźle źródłowym, jeśli dane w kolumnie **Wartości** są wyświetlane jako czarny tekst, oznacza to, że wartości danej zmiennej zostały odczytane i są zapisane w danym węźle. Jeśli w tym polu nie ma danych zapisanych na czarno, oznacza to, że dana zmienna nie została odczytana i jest określona we wcześniejszej części strumienia.

Czasami dane są wyświetlane jako szary tekst. Taka sytuacja ma miejsce, kiedy program SPSS Modeler może określić lub wywnioskować poprawne wartości zmiennych bez rzeczywistego odczytywania i zapisywania danych. Najczęściej dzieje się tak w przypadku użycia jednego z następujących węzłów:

- Węzeł Dane niestandardowe. Ponieważ dane są definiowane w węźle, zakres wartości dla zmiennej jest zawsze znany, nawet jeśli wartości nie zostały zapisane w węźle.
- Węzeł źródłowy Plik Statistics. Jeśli dla typów danych dostępne są metadane, program SPSS Modeler może wywnioskować możliwy zakres wartości bez odczytywana lub zapisywania danych.

W dowolnym węźle wartości są wyświetlane jako szary tekst, dopóki użytkownik nie kliknie przycisku **Odczytaj wartości**.

Uwaga: Jeśli dane nie zostaną określone w strumieniu i wartości danych są wyświetlane na szaro, nie zostanie wykonane żadne sprawdzenie typów wartości ustawione w kolumnie **Sprawdź**.

Użycie okna dialogowego Wartości

Kliknięcie kolumny **Wartości** lub **Braki** na karcie Typy powoduje wyświetlenie listy rozwijanej z predefiniowanymi wartościami. Po wybraniu opcji **Określ...** na liście otwarte zostaje osobne okno dialogowe, w którym można ustawić opcje dotyczące odczytywania, określania, etykietowania i obsługi wartości dla wybranej zmiennej.

Wiele elementów sterujących jest wspólnych dla wszystkich typów danych. Poniżej omówiono te wspólne elementy sterujące.

Poziom pomiaru Wyświetla aktualnie wybrany poziom pomiaru. Można zmienić to ustawienie, tak aby było odpowiednie dla sposobu użycia danych. Na przykład, jeśli zmienna o nazwie day_of_week (dzień tygodnia) zawiera liczby reprezentujące poszczególne dni, można zmienić je na dane nominalne, aby utworzyć węzeł rozkładu sprawdzający pojedynczo każdą kategorię.

Składowanie Wyświetla typ składowania, o ile jest on znany. Wybrany poziom pomiaru nie wpływa na typy składowania. Aby zmienić typ składowania, można użyć karty Dane w węźle źródłowym Plik kolumnowy i Plik zmiennych lub funkcji przekształcania w węźle wypełniania.

Zmienna modelu W przypadku zmiennych wygenerowanych po przeprowadzeniu oceny modelu użytkowego, można również wyświetlić szczegóły zmiennej modelu. Dotyczy to nazwy zmiennej przewidywanej, jak również roli zmiennej w modelowaniu (czy jest to wartość przewidywana, prawdopodobieństwo, skłonność itd.).

Wartości Należy wybrać metodę określania wartości dla wybranej zmiennej. Opcje tutaj wybrane zastępują ustawienia wykonane wcześniej w kolumnie **Wartości** w oknie dialogowym węzła typu. Możliwe do wybrania opcje dla odczytu wartości:

- **Odczytaj z danych** Tę opcję należy wybrać, aby wartości były odczytywane podczas wykonywania węzła. Ta opcja ma takie samo zastosowanie jak opcja **<Odczytaj>**.
- **Przepuść** Tę opcję należy wybrać, aby nie odczytywać danych dla bieżącej zmiennej. Ta opcja ma takie samo zastosowanie jak opcja **<Przepuść>**.
- **Określ wartości i etykiety** Dostępne tutaj opcje są używane do określania wartości i etykiet dla wybranej zmiennej. Użycie tej opcji z opcją sprawdzania wartości umożliwia określenie wartości w oparciu o wiedzę użytkownika na temat bieżącej zmiennej. Ta opcja powoduje aktywowanie unikalnych elementów sterujących dla zmiennej każdego typu. Opcje dla wartości i etykiet zostały omówione osobno w kolejnych tematach.

  Uwaga: Nie można określić wartości lub etykiet dla zmiennej, której poziom pomiaru został ustawiony jako Nieokreślony lub <Domyślny>.
- **Rozszerz przedział wartości o tutaj zdefiniowane** Należy wybrać tę opcję, aby połączyć bieżące dane z wartościami tutaj wprowadzonymi. Przykładowo, jeśli field_1 (zmienna_1) mieści się w zakresie (0,10) i wprowadzone zostaną wartości z zakresu (8,16), wówczas zakres zostanie rozszerzony poprzez dodanie wartości 16, bez usuwania oryginalnej wartości minimalnej. Nowy zakres zatem to: (0,16). Wybranie tej opcji powoduje automatyczne ustawienie opcji automatycznego wpisywania w polu **<Odczytaj+>**.

Maksymalna długość listy Opcja jest dostępna tylko dla danych z poziomem pomiaru Geoprzestrzenny lub Przedziałowy. Należy ustawić maksymalną długość listy, ustawiając dozwoloną dla listy liczbę elementów.

Maksymalna długość łańcucha Zmienna dostępna tylko dla danych bez określonego typu; należy jej użyć podczas generowania kodu SQL w celu utworzenia tabeli. Należy wprowadzić wartość największego łańcucha w danych; spowoduje to wygenerowanie kolumny w tabeli, która ma odpowiednią wielkość dla łańcucha. Jeśli wartość długości łańcucha jest niedostępna, zastosowana zostanie domyślna wielkość łańcucha, która może być nieodpowiednia dla danych (przykładowo, jeśli wartość jest zbyt mała, wystąpią błędy podczas zapisywania danych w tabeli; wartość zbyt duża może negatywnie wpłynąć na wydajność).

Sprawdzanie wartości Należy wybrać metodę koercji wartości do zastosowania dla określonych wartości ilościowych, typu flaga lub nominalnych. Ta opcja odnosi się do kolumny **Sprawdź** w oknie dialogowym węzła typu i ustawienia tutaj dokonane zastępują te z okna dialogowego. Użycie opcji sprawdzania wartości z opcją **Określ wartości i etykiety** umożliwia zastosowanie w danych oczekiwanych wartości. Przykładowo, jeśli wybrane zostaną wartości 1, 0, a następnie użyta zostanie opcja **Odrzuć**, wówczas można odrzucić wszystkie rekordy z wartościami różniącymi się od 1 lub 0.

Zdefiniuj puste Tę opcję należy wybrać, aby aktywować następujące elementy sterujące, które umożliwiają deklarowanie braków danych lub pustych wartości w danych.

- **Brakujące wartości** Ta tabela umożliwia zdefiniowanie konkretnych wartości (takich jak 99 lub 0) jako oznaczających brak danych. Wartość powinna być odpowiednia dla typu składowania zmiennej.
- **Przedział** Umożliwia określenie przedziału braków danych, na przykład wiek od 1 do 17 lub powyżej 65. Jeśli wartość graniczna pozostanie pusta, wówczas przedział jest nieograniczony; na przykład, jeśli dolna granica zostanie określona jako 100, a górna nie będzie określona, wówczas wszystkie wartości większe niż lub równe 100 zostaną zdefiniowane jako braki. Wartości graniczne należą do przedziału; na przykład, przedział z dolną granicą wynoszącą 5 i górną granicą wynoszącą 10 będzie w definicji przedziału zawierał wartości 5 i 10. Przedział brakujących

wartości można zdefiniować dla każdego typu składowania, z uwzględnieniem daty/czasu i łańcucha (w tym przypadku określenie, czy wartość należy do przedziału, odbywa się w oparciu o alfabetyczny porządek).

- **Wartość null/Białe znaki** Jako wartości puste można również określić systemowe wartości null (wyświetlane w danych jako $null$) oraz białe znaki (łańcuchy wartość z niewidocznymi znakami).

  Uwaga: Węzeł Typy również traktuje puste łańcuchy jako białe znaki w przypadku analizy, chociaż wewnętrzne są one zapisywane w inny sposób i mogą być różnie obsługiwane w określonych przypadkach.

Uwaga: Aby zakodować wartości puste jako niezdefiniowane lub wartości $null$, należy użyć węzła wypełniania.

Opis To pole tekstowe umożliwia określenie etykiety zmiennej. Etykiety te są wyświetlane w różnych miejscach, takich wykresy, tabele, wyniki i przeglądarki modeli, w zależności od opcji wybranych w oknie dialogowym Właściwości strumienia.

Określanie wartości i etykiet dla danych ilościowych

Poziom pomiaru *Ilościowe* jest używany dla zmiennych numerycznych. Dla danych ilościowych dostępne są trzy typy składowania:

- Liczba rzeczywista
- Liczba całkowita
- Data/czas

To samo okno dialogowe jest używane do edytowania wszystkich zmiennych ilościowych; typ składowania jest wyświetlany tylko w celach referencyjnych.

Określanie wartości

Przedstawione poniżej elementy sterujące są unikalne dla zmiennych ilościowych i są używane do określania zakresu wartości:

Dolne. Należy określić dolny limit zakresu wartości.

Górne. Należy określić górny limit zakresu wartości.

Określanie etykiet

Użytkownik może określić etykiety dowolnej wartości zmiennej przedziału. Kliknij przycisk **Etykiety**, aby otworzyć oddzielne okno dialogowe umożliwiające określenie etykiet wartości.

Podokno dialogowe Wartości i etykiety: Kliknięcie opcji **Etykiety** w oknie dialogowym Wartości dla zmiennej przedziału spowoduje otwarcie nowego okna dialogowego, w którym można określić etykiety dla dowolnej wartości z przedziału.

Kolumny *Wartości* i *Etykiety* w tej tabeli umożliwiają zdefiniowanie par wartość-etykieta. Obecnie zdefiniowane pary są tutaj wyświetlane. Można dodać nowe pary etykiet, klikając pustą komórkę i wprowadzając wartość oraz odpowiadającą jej etykietę. *Uwaga*: Dodawanie par wartość/wartość-etykieta do tej tabeli nie spowoduje dodania nowych wartości do zmiennej. Spowoduje jednak utworzenie metadanych dla wartości zmiennej.

Etykiety określone w węźle typu są wyświetlane w wielu miejscach (jako podpowiedzi, etykiety wyników itd.), w zależności od opcji wybranych w oknie dialogowym właściwości strumienia.

Określanie wartości i etykiet dla zmiennych nominalnych i porządkowych

Poziomy pomiaru nominalny (zbiór) i porządkowy (uporządkowany zbiór) wskazują, że wartości danych są używane w sposób dyskretny jako elementy zbioru. Możliwe typy składowania dla zbioru to: łańcuch, liczba całkowita, liczba rzeczywista lub data/czas.

Przedstawione poniżej elementy sterujące są unikalne dla zmiennych nominalnych i porządkowych i służą do określana wartości i etykiet:

Wartości. W kolumnie *Wartości* w tabeli można określić podstawowe wartości w oparciu o własną wiedzę na temat bieżącej zmiennej. Korzystając z tej tabeli, można wprowadzić oczekiwane wartości dla zmiennej i sprawdzić zgodność zbiorów danych z tymi wartościami, używając listy rozwijanej Sprawdzanie wartości. Używając przycisków strzałek i usuwania, można modyfikować istniejące wartości oraz zmieniać ich kolejność i usuwać je.

Etykiety. Kolumna *Etykiety* umożliwia określenie etykiet dla każdej wartości w zbiorze. Etykiety te są wyświetlane w różnych miejscach, takich jak wykresy, tabele, wyniki i przeglądarki modeli, w zależności od opcji wybranych w oknie dialogowym właściwości strumienia.

Określanie wartości dla zmiennych typu flaga

Zmienne typu flaga są używane do wyświetlania danych, które mają dwie odmienne wartości. Możliwe typy składowania dla flag to: łańcuch, liczba całkowita, liczba rzeczywista lub data/czas.

Prawda. Należy określić wartość flagi dla zmiennej po spełnieniu warunku.

Fałsz. Należy określić wartość flagi dla zmiennej, jeśli warunek nie jest spełniony.

Etykiety. Określa etykiety dla każdej wartości w zmiennej typu flaga. Etykiety te są wyświetlane w różnych miejscach, takich jak wykresy, tabele, wyniki i przeglądarki modeli, w zależności od opcji wybranych w oknie dialogowym właściwości strumienia.

Określanie wartości dla danych przedziałowych

Zmienne przedziałowe służą do wyświetlania danych z listy, które nie są geoprzestrzenne.

Dla przedziałowego **poziomu pomiaru** można wybrać jedynie opcję **Miara listy**. Domyślnie ten pomiar jest ustawiony jako Nieokreślony, ale można wybrać inną wartość, ustawiając w ten sposób poziom pomiaru dla elementów z listy. Można wybrać jedną z następujących opcji:

- Nieokreślony
- Ilościowy
- Nominalny
- Porządkowy
- Flaga

Określanie wartości dla danych geoprzestrzennych

Zmienne geoprzestrzenne są używane do wyświetlania danych geoprzestrzennych z listy.

Dla geoprzestrzennego **poziomu pomiaru** można wybrać jedną z następujących opcji, aby ustawić poziom pomiaru elementów z listy:

Typ Należy wybrać podpoziom pomiaru dla zmiennej geoprzestrzennej. Dostępne podpoziomy są określane przez głębokość zmiennej listy; ustawienia domyślne to: Punkt (głębokość zerowa), Łańcuch (głębokość jeden) i Wielokąt (głębokość jeden).

Więcej informacji o podpoziomach zawiera temat “Geoprzestrzenne podpoziomy pomiarów” na stronie 141.

Więcej informacji o tworzeniu głębokości listy zawiera temat “Składowanie listy i powiązane poziomy pomiaru” na stronie 12.

Układ współrzędnych Ta opcja jest dostępna tylko w przypadku zmiany poziomu pomiaru na geoprzestrzenny (z innego niż geoprzestrzenny). To pole należy zaznaczyć, aby zastosować układ współrzędnych do danych geoprzestrzennych. Domyślnie wyświetlany jest układ współrzędnych ustawiony w panelu **Narzędzia > Właściwości**

strumienia > Opcje > Geoprzestrzenne. Aby użyć innego układu współrzędnych, należy kliknąć przycisk **Zmień**; spowoduje to wyświetlenie okna dialogowego Wybierz układ współrzędnych, w którym można wybrać wymagany system.

Więcej informacji na temat układów współrzędnych zawiera temat Konfigurowanie opcji geoprzestrzennych strumieni w sekcji Praca ze strumieniami w dokumentacji SPSS Modeler — podręcznik użytkownika.

Definiowanie braków danych

W kolumnie **Braki** na karcie Typy znajduje się informacja, czy dla danej zmiennej zdefiniowano obsługę braków danych. Możliwe ustawienia to:

Wł. (*). Wskazuje, że dla danej zmiennej zdefiniowana została obsługa braków danych. Jest to możliwe za pośrednictwem węzła wypełniania wstawionego poniżej bieżącego węzła lub poprzez jawną specyfikację za pomocą opcji Określ (patrz poniżej).

Wył. Dla zmiennej nie zdefiniowano obsługi braków danych.

Określ. Tę opcję należy wybrać, aby wyświetlić okno dialogowe, w którym można zadeklarować jawne wartości, jakie mają być uznawane za braki danych dla tej zmiennej.

Sprawdzanie wartości typu

Włączenie opcji sprawdzania dla każdej zmiennej powoduje zbadanie wszystkich wartości danej zmiennej w celu ustalenia, czy są one zgodne z bieżącym typem ustawień lub wartości, jakie określono w oknie dialogowym Określ wartości. Jest to przydatne podczas czyszczenia zbiorów danych i redukowania wielkości zbioru danych w ramach jednej operacji.

Wybranie kolumny *Sprawdź* w oknie dialogowym węzła typu określa zadania, jakie zostaną wykonane w przypadku wykrycia wartości spoza określonych limitów dla danego typu. Aby zmienić ustawienia sprawdzania dla zmiennej, należy użyć listy rozwijanej dla tej zmiennej w kolumnie *Sprawdź*. Aby wprowadzić ustawienia sprawdzania dla wszystkich zmiennych, należy kliknąć kolumnę *Zmienna* i nacisnąć kombinację klawiszy Ctrl+A. Następnie należy użyć listy rozwijanej dowolnej zmiennej w kolumnie *Sprawdź*.

Poniżej przedstawiono dostępne ustawienia sprawdzania:

Brak. Wartości będą przekazywane bez sprawdzenia. Jest to ustawienie domyślne.

Wyzeruj. Zmienia wartości znajdujące się poza limitami na systemową wartość null ($null$).

Wymuś. Zmienne, których poziomu pomiaru są w pełni określone, zostaną sprawdzone w celu wykrycia wartości znajdujących się poza określonymi zakresami. Wartości nieokreślone będą przekształcone na poprawną wartość dla danego poziomu pomiaru z zastosowaniem następujących reguł:

- W przypadku flag wszystkie wartości inne niż prawda lub fałsz są przekształcane na wartość fałsz.
- W przypadku zbiorów (nominalne lub porządkowe) wszystkie nieznane wartości są przekształcane na pierwszy element z wartości zbioru.
- Liczby większe od wartości górnego limitu zakresu są zastępowane wartością górnego limitu.
- Liczby mniejsze niż wartość dolnego limitu zakresu są zastępowane wartością dolnego limitu.
- Wartości null dla danego zakresu uzyskują wartość punktu środkowego tego zakresu.

Odrzuć. Jeśli wykryte zostaną wartości niedozwolone, cały rekord zostaje odrzucony.

Ostrzegaj. Liczba niedozwolonych wartości zostaje zliczona i zgłoszona w oknie dialogowym właściwości strumienia po odczytaniu wszystkich danych.

Przerwij. Pierwsza wykryta wartość niedozwolona powoduje przerwanie działania strumienia. W oknie dialogowym właściwości strumienia zostaje wyświetlony błąd.

Ustawianie roli zmiennej

Role zmiennych określają sposób, w jaki będą one używane podczas tworzenia modelu — na przykład, czy zmienna będzie wejściowa, czy przewidywana (element przewidywany).

Uwaga: Role Podział, Częstość i ID rekordów mogą być stosowane tylko do pojedynczych zmiennych.

Dostępne są następujące role:

Zmienna wejściowa. Zmienna będzie używana jako wartość wejściowa dla uczenia maszynowego (zmienna predykcyjna).

Zmienna przewidywana. Zmienna będzie używana jako wartość wynikowa lub przewidywana dla uczenia maszynowego (jedna ze zmiennych w modelu będzie podejmowała próbę przewidywania).

Łącznie. Zmienna zostanie użyta jako dane wejściowe i wyjściowe przez węzeł Apriori. We wszystkich pozostałych węzłach modelowania zmienna zostanie zignorowana.

Brak. Zmienna zostanie zignorowana przez uczenie maszynowe. Dla zmiennych, dla których poziom pomiaru został ustawiony jako **Nieokreślony**, w kolumnie **Rola** automatycznie ustawiana jest wartość **Brak**.

Podział. Wskazuje zmienną użytą do podziału danych na osobne próby przeznaczone do uczenia, testowania i (opcjonalnie) walidacji. Zmienna musi być określonym typem zbioru z dwoma lub trzema możliwymi wartościami (jak zdefiniowano w oknie dialogowym Wartości zmiennych). Pierwsza wartość reprezentuje próbę uczącą, druga próbę testującą, a trzecia (o ile jest obecna) reprezentuje próbę walidacyjną. Wszystkie dodatkowe wartości są ignorowane; zmiennych typu flaga nie można używać. Należy pamiętać, że aby korzystać z dzielenia na podzbiory w analizie, należy aktywować dzielenie na podzbiory na karcie Opcje modelu w odpowiednim węźle budowania modelu lub analizy. Jeśli podział na podzbiory jest włączony, rekordy z wartościami null dla zmiennej dzielącej na podzbiory są wykluczane z analizy. Jeśli w strumieniu zdefiniowano kilka zmiennych dzielących na podzbiory, na karcie Zmienne należy określić pojedynczą zmienną dzielącą na podzbiory w każdym zastosowanym węźle modelowania. Jeśli w danych nie istnieje jeszcze odpowiednia zmienna, można ją utworzyć za pośrednictwem węzła podziału na podzbiory lub wyliczania. Więcej informacji można znaleźć w temacie “Węzeł Partycja” na stronie 176.

Podział. (Tylko zmienne nominalne, porządkowe i faga) Określa, że model będzie tworzony dla każdej możliwej wartości zmiennej.

Częstość. (Tylko zmienne numeryczne) Ustawienie tej roli umożliwia użycie wartości zmiennej jako współczynnika ważenia częstości dla rekordu. Ta funkcja jest obsługiwana tylko przez modele C&R Tree, CHAID, QUEST i modele liniowe; wszystkie pozostałe węzły ignorują tę rolę. Ważenie częstości jest aktywowane za pośrednictwem opcji **Użyj wagi częstości** na karcie Zmienne tych węzłów modelowania, które te funkcję obsługują.

Identyfikator rekordu. Zmienna będzie używana jako unikatowy identyfikator rekordu. Ta funkcja jest ignorowana przez większość węzłów; jest jednak obsługiwana przez modele liniowe i jest wymagana dla węzłów eksploracji w bazie danych IBM Netezza.

Kopiowanie atrybutów typu

Atrybuty typu, takie jak wartości, opcje sprawdzania i braki danych, można w prosty sposób kopiować pomiędzy zmiennymi:

1. Kliknij prawym przyciskiem myszy zmienną, której atrybuty zamierzasz skopiować.
2. Z menu kontekstowego wybierz opcję **Kopiuj**.
3. Kliknij prawym przyciskiem myszy zmienne, których atrybuty zamierzasz zmienić.

4. W menu kontekstowym wybierz polecenie **Wklej specjalne**. *Uwaga*: Można wybrać kilka zmiennych, korzystając z metody Ctrl+kliknięcie lub opcji **Wybierz zmienne** dostępnej w menu kontekstowym.

Zostanie otwarte nowe okno dialogowe, w którym można wybrać atrybuty, jakie mają zostać wklejone. Jeśli wklejanie obejmuje wiele zmiennych, wybrane tutaj opcje będą miały zastosowanie do wszystkich zmiennych przewidywanych.

Wklej następujące atrybuty. Należy wybrać atrybuty do wklejenia z jednej zmiennej do drugiej.

- **Typ.** Należy wybrać tę opcję, aby wkleić poziom pomiaru.
- **Wartości.** Tę opcję należy wybrać, aby wkleić wartości zmiennej.
- **Braki danych.** Ta opcja pozwala wkleić ustawienia braków danych.
- **Sprawdź.** Tę opcję należy wybrać, aby wkleić opcje sprawdzania wartości.
- **Rola.** Ta opcja umożliwia wklejenie roli zmiennej.

Karta ustawień formatu zmiennej

Karta Format w węzłach Tabela i Typy wyświetla listę bieżących lub nieużywanych zmiennych oraz opcje formatowania dla każdej zmiennej. Poniżej zamieszczono opis każdej kolumny w tabeli formatowania zmiennej:

Zmienna. Wyświetla nazwę wybranej zmiennej.

Format. Dwukrotne kliknięcie komórki w tej kolumnie pozwala określić formatowanie dla pojedynczych zmiennych; umożliwia to okno dialogowe, które zostaje otwarte. Więcej informacji można znaleźć w temacie “Ustawianie opcji formatu zmiennej” na stronie 151. Określone tutaj formatowanie zastępuje formatowanie ustawione we wszystkich właściwościach strumienia.

Uwaga: Węzły Eksport Statistics i Wynik Statistics eksportują pliki *.sav*, które obejmują formatowanie na podstawie zmiennej w metadanych. Jeśli określony format na podstawie zmiennej nie jest obsługiwany przez format pliku IBM SPSS Statistics *.sav*, wówczas węzeł będzie używał domyślnego formatowania IBM SPSS Statistics.

Wyrównanie. Ta kolumna umożliwia określenie sposobu wyrównania wartości w kolumnie tabeli. Ustawienie domyślne to **Automatycznie**, które powoduje wyrównanie wartości symbolicznych do lewej a wartości numerycznych do prawej strony. Ustawienie domyślne można zastąpić, wybierając opcję **Do lewej**, **Do prawej** lub **Do środka**.

Szerokość kolumny. Domyślnie szerokości kolumn są wyznaczane automatycznie na podstawie wartości zmiennej. Aby zastąpić automatyczne obliczanie szerokości, należy kliknąć komórkę tabeli i korzystając z listy rozwijanej, wybrać nową szerokość. Aby wprowadzić szerokość niestandardową, która nie jest tutaj wyświetlana, należy otworzyć podokno dialogowe formatów zmiennych, klikając dwukrotnie komórkę tabeli w kolumnie Zmienna lub Format. Alternatywnie można kliknąć prawym przyciskiem myszy komórkę i wybrać opcję **Ustaw format**.

Widok aktualnych zmiennych. Domyślnie w tum oknie dialogowym wyświetlana jest lista aktualnie aktywnych zmiennych. Aby wyświetlić listę nieużywanych zmiennych, należy wybrać opcję **Widok ustawień niewykorzystanych zmiennych**.

Menu kontekstowe. Menu kontekstowe tej karty udostępnia różne opcje wyboru i aktualizacji ustawień. Aby wyświetlić to menu, należy kliknąć kolumnę prawym przyciskiem myszy.

- **Wybierz wszystkie.** Powoduje zaznaczenie wszystkich zmiennych.
- **Anuluj wybór wszystkich.** Czyści zaznaczone opcje.
- **Wybierz zmienne.** Umożliwia wybranie zmiennych w oparciu o typ lub charakterystykę składowania. Możliwe opcje to: **Wybierz jakościowe**, **Wybierz zmienne ilościowe** (numeryczne), **O typie nieokreślonym**, **Zmienne tekstowe**, **Zmienne numeryczne** lub **W formacie daty i czasu**. Więcej informacji można znaleźć w temacie “Poziomy pomiaru” na stronie 139.
- **Ustaw format.** Otwiera podokno dialogowe, umożliwiające określenie opcji daty, godziny i wartości dziesiętnych na podstawie zmiennej.

- **Ustaw wyrównanie.** Ustawia wyrównanie dla wybranych zmiennych. Możliwe opcje to: **Automatycznie**, **Do środka**, **Do lewej** lub **Do prawej**.
- **Ustaw szerokość kolumny.** Ustawia szerokość kolumny dla wybranych zmiennych. Wybranie opcji **Automatycznie** umożliwia odczyt szerokości z danych. Można również ustawić szerokość zmiennej na 5, 10, 20, 30, 50, 100 lub 200.

Ustawianie opcji formatu zmiennej

Formatowanie zmiennej jest określane w podoknie dialogowym dostępnym z karty Format w węzłach Typy i Tabela. Jeśli przed otwarciem tego okna dialogowego wybrano więcej niż jedną zmienną, wówczas ustawienia z pierwszej zmiennej będą zastosowane dla wszystkich. Kliknięcie przycisku **OK** po dokonaniu tutaj specyfikacji spowoduje zastosowanie tych ustawień do wszystkich zmiennych wybranych na karcie Format.

Poniżej przedstawiono opcje dostępne na podstawie zmiennej. Wiele z tych ustawień można również określić w oknie dialogowym właściwości strumienia. Wszelkie ustawienia wprowadzone na poziomie zmiennej zastępują wartości domyślne określone dla strumienia.

Format daty. Wybierz format daty, który będzie używany w przypadku zmiennych składowania daty lub gdy łańcuchy są interpretowane jako daty przez funkcje dat CLEM.

Format czasu. Wybierz format czasu, który będzie używany w przypadku zmiennych składowania czasu lub gdy łańcuchy są interpretowane jako czas przez funkcje czasu CLEM.

Format wyświetlania liczb. Użytkownik może wybrać format wyświetlania: standardowy (**####.###**), naukowy (**#.###E+##**) lub walutowy (**$###.##**).

Separator dziesiętny. Należy wybrać przecinek (,) lub kropkę (.) jako separator miejsc dziesiętnych.

Symbol grupowania. Aby określić formaty wyświetlania liczb, należy wybrać symbol użyty do pogrupowania wartości (na przykład, przecinek w zapisie *3,000.00*). Dostępne opcje to: brak, kropka, przecinek, spacja, symbol zdefiniowany w ustawieniach regionalnych (wtedy wybierane jest ustawienie domyślne opcji regionalnych).

Miejsca dziesiętne (standardowe, naukowe, walutowe oraz dla eksportu). W przypadku formatów wyświetlania liczb określana jest liczba miejsc dziesiętnych, jaka zostanie użyta podczas wyświetlania liczb rzeczywistych. Tę opcję należy ustawić oddzielnie dla każdego formatu wyświetlania. Należy pamiętać, że ustawienie **Eksportuj pozycje dziesiętne** dotyczy tylko eksportu do pliku płaskiego i zastępuje właściwości strumienia. Strumień domyślny dla eksportu do pliku płaskiego jest określany dla ustawienia **Standardowe miejsca dziesiętne** we właściwościach strumienia. Liczba miejsc dziesiętnych wyeksportowanych przez węzeł eksportu XML to zawsze 6.

Wyrównanie. Określa sposób wyrównania wartości w kolumnie. Ustawienie domyślne to **Automatycznie**, które powoduje wyrównanie wartości symbolicznych do lewej a wartości numerycznych do prawej strony. Można zastąpić domyślne ustawienie, wybierając wyrównanie do lewej, do prawej lub do środka.

Szerokość kolumny. Domyślnie szerokości kolumn są wyznaczane automatycznie na podstawie wartości zmiennej. Używając strzałek po prawej stronie pola listy, można określić szerokość niestandardową w przedziałach co pięć.

Filtrowanie lub zmiana nazw zmiennych

Istnieje możliwość zmiany nazwy lub wykluczenia zmiennych w dowolnym punkcie strumienia. Przykładowo, pracownik naukowo-badawczy może nie być zainteresowany poziomem potasu (dane na poziomie zmiennej) u pacjentów (dane na poziomie rekordu); dlatego może odfiltrować zmienną K (potas). Można to zrobić, używając osobnego węzła filtrowania lub karty Filtrowanie w węźle źródłowym lub wynikowym. Działanie jest takie samo, niezależnie od węzła wybranego do uzyskania dostępu.

- Za pośrednictwem węzłów źródłowych, takich jak Plik zmiennych, Plik kolumnowy, Plik Statistics, XML lub Importowanie przez rozszerzenie, można zmienić nazwę lub odfiltrować zmienne podczas odczytywania danych w programie IBM SPSS Modeler.

- Korzystając z węzła filtrowania można zmienić nazwy lub odfiltrować zmienne w dowolnym punkcie strumienia.
- Węzły Eksport Statistics, Transformacja Statistics, Model Statistics i Wynik Statistics umożliwiają zmianę nazwy lub filtrowanie zmiennych, tak aby były zgodne ze standardami nadawania nazw w programie IBM SPSS Statistics. Więcej informacji można znaleźć w temacie “Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics” na stronie 364.
- Karta Filtrowanie w dowolnym z powyższych węzłów umożliwia zdefiniowanie lub edytowanie zestawów wielokrotnych odpowiedzi. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi” na stronie 153.
- Węzła Filtrowanie można użyć do mapowania zmiennych z jednego źródła na inne.

Ustawianie opcji filtrowania

Tabela na karcie Filtrowanie wyświetla nazwę każdej zmiennej, która jest wprowadzana do węzła, a także nazwy zmiennych wychodzących z węzła. Opcje w tej tabeli umożliwiają zmianę nazwy lub filtrowanie zmiennych zduplikowanych lub zmiennych, które nie są potrzebne w czasie wykonywania operacji w dalszej części strumienia.

- **Zmienna.** Wyświetla zmienne wejściowe z aktualnie połączonych źródeł danych.
- **Filtrowanie.** Wyświetla status filtrowania wszystkich zmiennych wejściowych. W tej kolumnie znajduje się czerwony znak X dla odfiltrowanych zmiennych, który oznacza, że zmienna nie została przepuszczona do dalszej części strumienia. Kliknięcie w kolumnie *Filtrowanie* dla wybranej zmiennej umożliwia włączanie i wyłączanie filtrowania. Opcje można również wybierać dla kilku zmiennych jednocześnie, używając do zaznaczania metody Shift+kliknięcie.
- **Zmienna.** Wyświetla zmienne, które opuściły węzeł filtrowania. Nazwy zduplikowane są wyświetlane w kolorze czerwonym. Nazwy zmiennych można edytować, klikając w tej kolumnie i wprowadzając nową nazwę. Lub można usunąć zmienne, klikając w kolumnie *Filtrowanie* w celu wyłączenia zmiennych zduplikowanych.

Wszystkie kolumny w tabeli można posortować, klikając nagłówek kolumny.

Widok aktualnych zmiennych. Tę opcję należy wybrać, aby wyświetlić zmienne dla zbiorów danych aktywnie połączonych z węzłem Filtrowanie. Ta opcja jest wybrana domyślnie i jest to najczęściej stosowana metoda użycia węzłów filtrowania.

Widok ustawień niewykorzystanych zmiennych. Tę opcję należy wybrać, aby wyświetlić zmienne dla zbiorów danych, które kiedyś były, ale już nie są połączone z węzłem filtrowania. Ta opcja jest przydatna podczas kopiowania węzłów filtrowania z jednego strumienia do innego lub podczas zapisywania i ponownego wczytywania węzłów filtrowania.

Menu przycisku filtrowania

Należy kliknąć przycisk filtrowania w lewym górnym rogu okna dialogowego, aby uzyskać dostęp do menu, które udostępnia kilka skrótów i innych opcji.

Można:

- Usunąć wszystkie zmienne.
- Uwzględnić wszystkie zmienne.
- Przełączyć wszystkie zmienne.
- Usunąć duplikaty. Należy pamiętać, że zaznaczenie tej opcji powoduje usunięcie wszystkich wystąpień zduplikowanej nazwy, z uwzględnieniem pierwszej.
- Zmienić nazwy zmiennych i zestawów wielokrotnych odpowiedzi, tak aby były zgodne z innymi aplikacjami. Więcej informacji można znaleźć w temacie “Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics” na stronie 364.
- Przyciąć nazwy zmiennych.
- Przeprowadzić anonimizację zmiennych i zestawów wielokrotnych odpowiedzi.
- Użyć nazw zmiennych wejściowych.

- Edytować zestawy wielokrotnych odpowiedzi. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi”.
- Ustawić domyślny stan filtru.

Można również użyć przycisków przełączania w postaci strzałek w górnej części okna dialogowego, aby określić, czy zmienne mają być domyślnie uwzględniane czy odrzucane. Jest to przydatne w przypadku dużych zbiorów danych, w których tylko kilka zmiennych ma zostać uwzględnionych w dalszej części strumienia. Przykładowo można wybrać tylko te zmienne, które mają być zachowane i określić, że pozostałe zmienne powinny zostać odrzucone (zamiast wybierania pojedynczo wszystkich zmiennych do odrzucenia).

Przycinanie nazw zmiennych

Korzystając z menu przycisku filtrowania (górny lewy róg karty Filtrowanie), można wybrać opcję przycięcia nazw zmiennych.

Maksymalna długość. Należy określić liczbę znaków, aby ograniczyć długość nazw zmiennych.

Liczba cyfr. Jeśli po przycięciu nazwy zmiennych nadal nie są unikalne, zostaną ponownie przycięte i rozróżnione poprzez dodanie cyfr do nazwy. Można określić liczbę cyfr, jaka będzie użyta. Liczbę można ustawić, korzystając z przycisków strzałek.

Przykładowo, w poniższej tabeli przedstawiono, w jaki sposób nazwy zmiennych w medycznym zbiorze danych są przycinane przy użyciu ustawień domyślnych (długość maksymalna = 8 a liczba cyfr = 2).

Tabela 24. Przycinanie nazwy zmiennej

Nazwy zmiennych	Przycięte nazwy zmiennych
Patient Input 1	Patien01
Patient Input 2	Patien02
Heart Rate	HeartRat
CIŚNIENIE KRWI	CIŚNIENIE KRWI

Anonimizacja nazw zmiennych

Istnieje możliwość anonimizacji nazw zmiennych z dowolnego węzła zawierającego kartę Filtrowanie; należy w tym celu kliknąć menu przycisku filtrowania w lewym górnym rogu i wybrać opcję **Anonimizuj nazwy zmiennych**. Anonimizowane nazwy zmiennych składają się z łańcucha przedrostka oraz unikalnej wartości liczbowej.

Anonimizuj nazwy. Należy wybrać opcję **Tylko wybrane zmienne**, aby przeprowadzić anonimizację tylko nazw zmiennych, które zostały już wybrane na karcie Filtrowanie. Domyślnie zaznaczona jest opcja **Wszystkie zmienne**, co powoduje anonimizację wszystkich nazw zmiennych.

Przedrostek nazw zmiennych. Domyślnym przedrostkiem dla nazw zmiennych poddanych anonimizacji jest **anon_**; aby zastosować inny przedrostek, należy wybrać opcję **Użytkownika** i wpisać go.

Zakoduj zestawy wielokrotnych odpowiedzi. Umożliwia anonimizację zestawów wielokrotnych odpowiedzi w taki sam sposób, jak w przypadku zmiennych. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi”.

Aby przywrócić oryginalne nazwy zmiennych, należy wybrać opcję **Użyj nazw zmiennych wejściowych** z menu przycisku filtrowania.

Edytowanie zestawów wielokrotnych odpowiedzi

Możliwe jest dodawanie lub edytowanie zestawów wielokrotnych odpowiedzi z dowolnego węzła, który zawiera kartę Filtrowanie; w tym celu należy kliknąć menu przycisku filtrowania w lewym górnym rogu i wybrać opcję **Edytuj zestawy wielokrotnych odpowiedzi**.

Zestawy wielokrotnych odpowiedzi służą do rejestrowania danych, które mogą mieć więcej niż jedną wartość dla każdej obserwacji — na przykład po zadaniu respondentom ankiety pytania, w których byli muzeach lub które czasopisma czytają. Zaimportowanie zestawów wielokrotnych odpowiedzi do programu IBM SPSS Modeler umożliwia węzeł źródłowy Data Collection lub węzeł źródłowy Plik Statistics; definiowanie tych zestawów w programie IBM SPSS Modeler odbywa się za pośrednictwem węzła Filtrowanie.

Należy kliknąć przycisk **Nowy**, aby utworzyć nowy zestaw wielokrotnych odpowiedzi lub przycisk **Edytuj**, aby zmodyfikować istniejący.

Nazwa i etykieta. Określa nazwę i opis zestawu.

Typ. Z pytaniami z wielokrotnymi odpowiedziami można postępować na dwa sposoby:

- **Zestaw wielokrotnych dychotomii.** Dla każdej możliwej odpowiedzi tworzona jest osobna zmienna flagi; dlatego, jeśli występuje 10 czasopism, zmiennych flagi jest 10, a każda z nich może mieć wartości, takie jak 0 lub 1, oznaczające *prawdę* lub *fałsz*. Obliczona wartość umożliwia określenie, która wartość jest prawdziwa. Ta metoda jest przydatna, jeśli respondenci mają mieć możliwość wybrania wszystkich opcji, jakie mają zastosowanie.
- **Zestaw wielokrotnych kategorii.** Dla każdej odpowiedzi tworzona jest zmienna nominalna; zakres obejmuje maksymalną liczbę odpowiedzi udzielonych przez konkretnego respondenta. Każda zmienna nominalna ma wartości reprezentujące możliwe odpowiedzi, takie jak *1* dla magazynu *Time*, *2* dla magazynu *Newsweek* i *3* dla magazynu *PC Week*. Ta metoda jest najbardziej przydatna, jeśli konieczne jest ograniczenie liczby odpowiedzi — na przykład, zadanie respondentom pytania, aby wybrali trzy czasopisma, które czytają najczęściej.

Zmienne w zbiorze. Ikony po prawej stronie umożliwiają dodawanie lub usuwanie zmiennych.

Komentarze

- Wszystkie zmienne uwzględnione w zestawie wielokrotnych odpowiedzi muszą mieć taki sam typ składowania.
- Zestawy są niezależne od zmiennych, które zawierają. Przykładowo, usunięcie zestawu nie spowoduje usunięcia zmiennych, które zawiera — a jedynie złączy pomiędzy tymi zmiennymi. Zestaw nadal będzie widoczny w strumieniu przed punktem usunięcia, ale nie będzie widoczny za nim.
- Jeśli nazwy zmiennych zostaną zmienione za pomocą węzła Filtrowanie (bezpośrednio na karcie lub po wybraniu opcji zmiany nazwy dla programu IBM SPSS Statistics, czyli **Przytnij** lub **Anonimizacja** w menu filtrowania), wszelkie odniesienia do tych zmiennych użyte w zestawach wielokrotnych odpowiedzi również zostaną zaktualizowane. Jednak zmienne w zestawach wielokrotnych odpowiedzi, które zostaną usunięte za pomocą węzła Filtrowanie, nie zostaną usunięte z zestawu wielokrotnych odpowiedzi. Zmienne te, chociaż nie będą już widoczne w strumieniu, nadal będą stanowiły odniesienie do zestawu wielokrotnych odpowiedzi; należy o tym pamiętać na przykład podczas eksportowania.

węzeł wyliczeń

Jedną z najbardziej zaawansowanych właściwości programu IBM SPSS Modeler jest możliwość modyfikowania wartości danych i wyliczania nowych zmiennych na podstawie istniejących danych. W czasie realizacji długotrwałych projektów eksploracji danych często wykonywanych jest kilka wyliczeń, takich jak wyodrębnianie identyfikatora klienta z łańcucha danych dziennika sieciowego lub tworzenie wartości czasu życia klienta na podstawie danych dotyczących transakcji i danych demograficznych. Wszystkie te przekształcenia można wykonać, korzystając z różnych węzłów operacji na zmiennych.

Kilka węzłów umożliwia wyliczenie nowych zmiennych:

Węzeł Wyliczanie modyfikuje wartości danych lub tworzy nowe zmienne z co najmniej jednej istniejącej zmiennej. Tworzy pola typu formuła, flaga, nominalne, stan, liczebność i warunkowe.

Węzeł Rekodowanie przekształca jeden zestaw wartości jakościowych w inny. Rekodowanie jest przydatne do zwijania kategorii lub ponownego pogrupowania danych do analizy.

Węzeł Kategoryzacja automatycznie tworzy nowe zmienne nominalne (zbioru) na podstawie wartości z jednej lub większej liczby istniejących zmiennych ilościowych (zakres liczbowy). Można na przykład przekształcić ilościową zmienną przychodu na nową zmienną jakościową zawierającą grupy przychodu stanowiące odchylenia od średniej. Po utworzeniu kategorii dla nowej zmiennej na podstawie punktu podziału można wygenerować węzeł Wyliczanie.

Węzeł Flagowanie służy do wyliczania zmiennych flag na podstawie zmiennych wartości jakościowych zdefiniowanych dla co najmniej jednej zmiennej nominalnej.

Węzeł Restrukturyzacja przekształca zmienną nominalną lub typu flaga na grupę zmiennych, które mogą być wypełnione wartościami jeszcze innej zmiennej. Na przykład, dana jest zmienna o nazwie *payment type* (rodzaj płatności), której wartości to *credit* (kredyt), *cash* (gotówka) i *debit* (debet) i utworzone zostaną trzy nowe zmienne (*credit*, *cash*, *debit*), a każda z nich może zawierać wartość dla rzeczywiście dokonanej płatności.

Węzeł Historia tworzy nowe zmienne zawierające dane ze zmiennych z wcześniejszych rekordów. Węzły historii są najczęściej używane w przypadku danych sekwencyjnych, takich jak dane szeregu czasowego. Przed użyciem węzła historii można posortować dane za pomocą węzła Sortowanie.

Użycie węzła Wyliczanie

Za pomocą węzła wyliczeń na podstawie jednej lub kilku istniejących zmiennych można utworzyć sześć typów nowych zmiennych:

- **Formuła.** Nowa zmienna jest wynikiem dowolnego wyrażenia CLEM.
- **Flaga.** Nowa zmienna jest flagą reprezentującą określony warunek.
- **Nominalne.** Nowa zmienna jest nominalna, co znaczy, że jej elementy stanowią grupę określonych wartości.
- **Stan.** Nowa zmienna jest jednym z dwóch stanów. Przełączanie pomiędzy tymi stanami jest wyzwalane przez określony warunek.
- **Liczebności.** Nowa zmienna tworzona jest w oparciu o liczbę określającą, ile razy dany warunek został spełniony.
- **Warunkowe.** Nowa zmienna jest wartością jednego z dwóch wyrażeń, w zależności od wartości warunku.

Każdy z tych węzłów zawiera zestaw specjalnych opcji wyświetlanych w oknie dialogowym węzła wyliczeń. Opcje te zostały omówione w kolejnych tematach.

Należy pamiętać, że zastosowanie poniższych rozwiązań może spowodować zmianę kolejności wierszy:

- Wykonywanie w bazie danych za pośrednictwem analizy wstępnej SQL
- Wykonywanie za pośrednictwem zdalnego serwera IBM SPSS Analytic Server
- Korzystanie z funkcji uruchamianych w osadzonego serwera IBM SPSS Analytic Server
- Wyliczanie listy (przykładowo patrz “Wyliczanie zmiennej listy lub geoprzestrzennej” na stronie 158)
- Wywoływanie funkcji opisanych w sekcji Funkcje przestrzenne

Ustawianie podstawowych opcji dla węzła Wyliczanie

W górnej części okna dialogowego węzłów wyliczeń znajdują się opcje umożliwiające wybór odpowiedniego typu węzła Wyliczanie.

Dominanta. Należy wybrać opcję **Pojedyncze** lub **Wielokrotne**, w zależności od tego, czy mają być wyliczane zmienne wielokrotne. Po zaznaczeniu opcji **Wielokrotne** okno dialogowe ulega zmianie, wyświetlając opcje dla wielokrotnych zmiennych wyliczanych.

Zmienna wyliczana. W przypadku prostych zmiennych wyliczanych należy określić nazwę zmiennej, jaka ma zostać wyliczona dla i dodana do każdego rekordu. Domyślna nazwa to Wyliczanie*N*, gdzie *N* oznacza liczbę węzłów wyliczeń, jakie zostały utworzone do tej pory w czasie bieżącej sesji.

Wylicz jako. Z listy rozwijanej należy wybrać węzeł wyliczeń, taki jak Formuła lub Nominalny. Dla każdego typu w oparciu o warunki określone w oknie dialogowym specyficznym dla typu tworzona jest nowa zmienna.

Wybór opcji z listy rozwijanej spowoduje dodanie zestawu nowych elementów sterujących do głównego okna dialogowego, odpowiednio do właściwości każdego typu węzła wyliczeń.

Typ zmiennej. Należy wybrać poziom pomiaru, taki jak ilościowy, jakościowy lub flaga, dla nowo wyliczonego węzła. Ta opcja jest wspólna dla wszystkich formularzy węzła Wyliczanie.

Uwaga: Wyliczanie nowych zmiennych często wymagana użycia funkcji specjalnych lub wyrażeń matematycznych. Aby łatwiej było tworzyć tego typu wyrażenia, w oknie dialogowym dla wszystkich węzłów Wyliczanie dostępny jest konstruktor wyrażeń, który udostępnia funkcje sprawdzania reguł oraz pełną listę wyrażeń CLEM.

Wyliczanie wielu zmiennych

Ustawienie trybu **Wielokrotne** w węźle wyliczeń zapewnia możliwość wyliczenia wielu zmiennych na podstawie tego samego warunku w tym samym węźle. Ta funkcja pozwala zaoszczędzić czas w przypadku wykonywania identycznych przekształceń dla kilku zmiennych w zbiorze danych. Przykładowo, aby zbudować model regresji przewidujący aktualne wynagrodzenie na podstawie wynagrodzenia początkowego i wcześniejszego doświadczenia, korzystne może być zastosowanie transformacji logarytmicznej do wszystkich trzech zmiennych skośnych. Zamiast dodawać węzeł wyliczeń do każdej transformacji, można zastosować tę samą funkcję jednocześnie do wszystkich zmiennych. Wystarczy wybrać wszystkie zmienne, z których ma zostać wyliczona nowa zmienna, a następnie wpisać wyrażenie wyliczania, korzystając z funkcji @FIELD, przy czym zmienną należy ująć w nawiasy.

Uwaga: Funkcja @FIELD jest ważnym narzędziem do wyliczania wielu zmiennych w tym samym czasie. Umożliwia odniesienie treści bieżącej zmiennej lub zmiennych bez określania dokładnej nazwy zmiennej. Na przykład, wyrażenie CLEM użyte do zastosowania transformacji logarytmicznej do wielu zmiennych to: log(@FIELD).

Po wybraniu trybu **Wielokrotne** do okna dialogowego dodawane są następujące opcje:

Wylicz z. Selektor zmiennych umożliwia wybranie zmiennych, z których mają zostać wyliczone nowe zmienne. Dla każdej zmiennej wygenerowana zostanie jedna zmienna wyjściowa. *Uwaga*: Wybrane zmienne nie muszą mieć tego samego typu składowania; jednak operacja wyliczania nie powiedzie się, jeśli warunek nie będzie poprawny dla *wszystkich* zmiennych.

Rozszerzenie nazwy zmiennej. Należy wpisać rozszerzenie, jakie ma zostać dodane do nazw nowych zmiennych. Przykładowo, dla nowej zmiennej zawierającej dziennik *Current Salary* (Bieżące wynagrodzenie) można dodać rozszerzenie *log_* do nazwy zmiennej, tworząc nazwę *log_Current Salary*. Korzystając z przycisków opcji, należy wybrać, czy rozszerzenie ma zostać dodane jako przedrostek (na początku), czy jako przyrostek (na końcu) nazwy zmiennej. Domyślna nazwa to Wyliczanie*N*, gdzie *N* oznacza liczbę węzłów wyliczeń, jakie zostały utworzone do tej pory w czasie bieżącej sesji.

Podobnie jak w węźle wyliczeń w trybie pojedynczym teraz konieczne jest utworzenie wyrażenia służącego do wyliczenia nowej zmiennej. W zależności od typu wybranej operacji wyliczania dostępne są różne opcje umożliwiające

utworzenie warunku. Opcje te zostały omówione w kolejnych tematach. Aby utworzyć wyrażenie, można po prostu wpisać zmienne formuły lub użyć kreatora wyrażeń, klikając przycisk kalkulatora. Należy pamiętać, aby użyć funkcji @FIELD w celu odniesienia manipulacji do wielu zmiennych.

Wybór wielu zmiennych

W przypadku wszystkich węzłów, które wykonują operacje na wielu zmiennych wejściowych, takich jak Wyliczanie (tryb wielokrotny), Agregacja, Sortowanie, Wykres wielokrotny i Wykres sekwencyjny, można w prosty sposób wybrać wiele zmiennych, korzystając z okna dialogowego Wybierz zmienne.

Sortuj według. Dostępne zmienne można sortować podczas wyświetlania, wybierając jedną z następujących opcji:

- **Naturalnie.** Zmienne wyświetlane są w kolejności, w jakiej zostały wprowadzone w dół strumienia do bieżącego węzła.
- **Nazwa.** Zmienne wyświetlane są w porządku alfabetycznym.
- **Typ.** Wyświetlane zmienne są posortowane według poziomu pomiaru. Ta opcja jest przydatna w przypadku wyboru zmiennych z konkretnym poziomem pomiaru.

Zmienne można wybierać pojedynczo lub można użyć metody Shift+kliknięcie i Ctrl+kliknięcie, aby wybrać wiele zmiennych. Można również użyć przycisków poniżej listy, aby wybrać grupy zmiennych na podstawie ich poziomu pomiaru, zaznaczyć wszystkie zmienne w tabeli lub usunąć zaznaczenie wszystkich zmiennych w tabeli.

Ustawianie opcji węzła wyliczeń — Formuła

Węzły wyliczania typu Formuła tworzą nową zmienną dla każdego rekordu w zbiorze danych na podstawie wyników wyrażenia CLEM. To wyrażenie nie może być wyrażeniem warunkowym. Aby wyliczyć wartości na podstawie wyrażenia warunkowego, należy użyć węzła wyliczeń typu flaga lub warunkowego.

Formuła Należy określić formułę, używając do wyliczenia wartości nowej zmiennej języka CLEM.

Uwaga: Ponieważ program SPSS Modeler nie wie, jaki poziom podpomiaru ma zostać użyty dla wyliczonej zmiennej listy, w przypadku przedziałowego i geoprzestrzennego poziomu pomiaru można kliknąć przycisk **Określ...**, aby otworzyć okno dialogowe Wartość i ustawić wymagany poziom podpomiaru. Aby uzyskać więcej informacji, zobacz “Ustawianie wyliczonych wartości listy”.

W przypadku zmiennych geoprzestrzennych możliwe typy składowania to **Liczba rzeczywista** i **Liczba całkowita** (ustawienie domyślne to **Liczba rzeczywista**).

Ustawianie wyliczonych wartości listy

Po wybraniu opcji **Określ...** z listy rozwijanej **Typ zmiennej** dla formuły w węźle wyliczeń wyświetlane jest okno dialogowe Wartość. W tym oknie dialogowym można ustawić wartości poziomu podpomiaru, jakie będą używane dla poziomów pomiaru **Typ zmiennej** Przedziałowy lub Geoprzestrzenne (Formuła).

Poziom pomiaru Należy wybrać **Przedziałowy** lub **Geoprzestrzenne**. Jeśli wybrany zostanie inny poziom pomiaru, w oknie dialogowym zostanie wyświetlony komunikat, że nie ma wartości do edycji.

Przedziałowy

Dla przedziałowego **poziomu pomiaru** można wybrać jedynie opcję **Miara listy**. Domyślnie ten pomiar jest ustawiony jako Nieokreślony, ale można wybrać inną wartość, ustawiając w ten sposób poziom pomiaru dla elementów z listy. Można wybrać jedną z następujących opcji:

- Nieokreślony
- Jakościowy
- Ilościowy
- Nominalny
- Porządkowy

- Flaga

Geoprzestrzenny

Dla geoprzestrzennego **poziomu pomiaru** można wybrać jedną z następujących opcji, aby ustawić poziom pomiaru elementów z listy:

Typ Należy wybrać podpoziom pomiaru dla zmiennej geoprzestrzennej. Dostępne podpoziomy są określane przez głębokość zmiennej listy; ustawienia domyślne to:

- Punkt (głębokości zerowa)
- Łańcuch (głębokość jeden)
- Wielokąt (głębokość jeden)
- Multipunkt (głębokość jeden)
- Multiłańcuch (głębokość dwa)
- Multiwielokąt (głębokość dwa)

Więcej informacji o podpoziomach zawiera temat Geoprzestrzenne podpoziomy pomiarów w sekcji Węzeł Typy w publikacji SPSS Modeler — węzły źródłowe, procesowe i wyników.

Więcej informacji o głębokościach list zawiera temat Składowanie listy i powiązane poziomy pomiaru w sekcji Węzeł źródłowy w publikacji SPSS Modeler — węzły źródłowe, procesowe i wyników.

Układ współrzędnych Ta opcja jest dostępna tylko w przypadku zmiany poziomu pomiaru na geoprzestrzenny (z innego niż geoprzestrzenny). To pole należy zaznaczyć, aby zastosować układ współrzędnych do danych geoprzestrzennych. Domyślnie wyświetlany jest układ współrzędnych ustawiony w panelu **Narzędzia > Właściwości strumienia > Opcje > Geoprzestrzenne**. Aby użyć innego układu współrzędnych należy kliknąć przycisk **Zmień**; spowoduje to wyświetlenie okna dialogowego Wybierz układ współrzędnych, w którym można wybrać system odpowiedni do danych.

Więcej informacji na temat układów współrzędnych zawiera temat Konfigurowanie opcji geoprzestrzennych strumieni w sekcji Praca ze strumieniami w dokumentacji SPSS Modeler — podręcznik użytkownika.

Wyliczanie zmiennej listy lub geoprzestrzennej

Zdarza się, że dane, które powinny być zapisane jako element listy, zostają zaimportowane do programu SPSS Modeler z niewłaściwymi atrybutami. Przykładowo osobne zmienne geoprzestrzenne, takie jak współrzędna x i współrzędna y lub długość i szerokość geograficzna, jako osobne wiersze w pliku .csv. W takiej sytuacji należy połączyć osobne zmienne, tworząc jedną zmienną listy; jedynym sposobem, w jaki można to zrobić, jest użycie węzła Wyliczanie.

Uwaga: Podczas łączenia danych geoprzestrzennych użytkownik musi wiedzieć, która zmienna to x (lub długość geograficzna), a która to y (lub szerokość geograficzna). Dane należy tak połączyć, aby w wynikowej zmiennej listy kolejność elementów była następująca: [x, y] lub [Długość geograficzna, Szerokość geograficzna] — są to standardowe formaty współrzędnych geoprzestrzennych.

Poniżej przedstawiono prosty przykład wyliczania zmiennej listy.

1. W strumieniu do węzła źródłowego dołącz węzeł wyliczenia.
2. Na karcie Ustawienia węzła wyliczeń wybierz opcję **Formuła** z listy **Wylicz jako**.
3. W obszarze **Typ zmiennej** wybierz opcję **Przedziałowy** dla listy niegeoprzestrzennej lub **Geoprzestrzenne**. Domyślnie program SPSS Modeler do ustawienia szczegółów listy stosuje rozwiązanie „najlepszego przypuszczenia”; wybranie opcji **Określ...** umożliwia otwarcie okna dialogowego Wartość. To okno dialogowe może być używane w przypadku zbiorów do wprowadzania dodatkowych informacji o danych na liście, w przypadku danych geoprzestrzennych umożliwia ustawienie typu danych i określenie układu współrzędnych danych.

Uwaga: W przypadku danych geoprzestrzennych określany układ współrzędnych musi być zgodny z układem współrzędnych danych. W przeciwnym razie funkcje geoprzestrzenne będą generowały nieprawidłowe wyniki.

4. W panelu **Formuła** należy wprowadzić formułę umożliwiającą połączenie danych, tak aby utworzyły poprawny format listy. Alternatywnie można kliknąć przycisk kalkulatora, aby otworzyć konstruktora wyrażeń.

   Prostym przykładem formuły do wyliczenia listy jest [x, y], gdzie x i y to osobne zmienne w źródle danych. Utworzona nowa zmienna wyliczona jest listą, w której wartość każdego rekordu stanowi połączone wartości x i y dla tego rekordu.

   Uwaga: Zmienne łączone jako lista w ten sposób muszą mieć taki sam typ składowania.

Więcej informacji o listach i głębokościach listy zawiera temat “Składowanie listy i powiązane poziomy pomiaru” na stronie 12.

Ustawianie opcji węzła wyliczeń — Flaga

Węzły wyliczeń typu Flaga służą do wskazywania konkretnego warunku, np. wysokie ciśnienie krwi lub brak aktywności na koncie klienta. Zmienna typu flaga jest tworzona dla każdego rekordu, a po spełnieniu warunku typu „prawda”, wartość flagi dla prawdy jest dodawana do zmiennej.

Wartość prawdziwa. Należy określić wartość, jaka ma być uwzględniona w zmiennej typu flaga dla rekordów, które spełniają określony niżej warunek. Ustawienie domyślne to T (prawda).

Wartość fałszywa. Należy określić wartość, jaka ma być uwzględniona w zmiennej typu flaga dla rekordów, które *nie* spełniają określonego niżej warunku. Ustawienie domyślne to F (fałsz).

Jest prawdziwa pod warunkiem. Należy określić warunek CLEM, aby wyznaczyć określone wartości poszczególnych rekordów i przypisać do rekordu wartość prawdziwą lub fałszywą (definicja powyżej). Należy pamiętać, że wartość prawdziwa zostanie przypisana do rekordów w przypadku niefałszywych wartości liczbowych.

Uwaga: Aby zwrócony został pusty łańcuch, należy wpisać cudzysłów otwierający i zamykający, pomiędzy którymi nie będzie żadnej wartości, czyli "". Pusty łańcuch jest często używany na przykład jako wartość fałszywa, aby wartości prawdziwe były lepiej czytelne w tabeli. Podobnie cudzysłowów należy użyć, jeśli wartość łańcucha ma być traktowana jako liczba

Przykład

W wersjach programu IBM SPSS Modeler wcześniejszych niż 12.0 wiele odpowiedzi było importowanych do pojedynczej zmiennej, a wartości były rozdzielane przecinkami. Na przykład:

```
museum_of_design,institute_of_textiles_and_fashion
museum_of_design
archeological_museum
$null$
national_art_gallery,national_museum_of_science,other
```

Aby przygotować takie dane do analizy, można użyć funkcji hassubstring, która pozwoli wygenerować osobną zmienną flagi dla każdej odpowiedzi, używając wyrażenia:

```
hassubstring(museums,"museum_of_design")
```

Ustawianie opcji węzła wyliczeń — Nominalny

Węzły wyliczeń typu Nominalny służą do wykonywania zestawu warunków CLEM w celu ustalenia, które rekordy spełniają określone warunki. Jeśli warunek jest spełniony dla każdego rekordu, wartość (wskazująca zestaw spełnionych warunków) zostanie dodana do nowej wyliczonej zmiennej.

Wartość standardowa. Należy określić wartość, jaka będzie używana dla nowej zmiennej, jeśli żaden z warunków nie zostanie spełniony.

Wartość zmiennej. Należy określić wartość, jaka będzie wstawiona w nowej zmiennej, jeśli spełniony zostanie określony warunek. Każda wartość z listy ma powiązany warunek, jaki jest określony w sąsiedniej kolumnie.

Warunek do spełnienia. Należy określić warunek dla każdego elementu na liście Wartość zmiennej. Konstruktor wyrażeń ułatwi wybór z dostępnych funkcji i zmiennych. Aby zmieniać kolejność lub usuwać warunki, można użyć przycisków strzałek i usuwania.

Warunek polega na testowaniu wartości poszczególnych zmiennych w zbiorze danych. Po sprawdzeniu każdego warunku określone powyżej wartości zostaną przypisane do nowej zmiennej, aby wskazać, który warunek został spełniony (o ile jakiś został spełniony). Jeśli żaden z warunków nie jest spełniony, używana jest wartość domyślna.

Ustawianie opcji węzła wyliczeń — Stan

Węzły wyliczeń typu Stan są dość podobne do węzłów wyliczeń typu Flaga. Węzeł Flaga ustawia wartości w zależności od tego, czy spełniają *pojedynczy* warunek dla bieżącego rekordu, a węzeł wyliczeń typu Stan może zmienić wartości zmiennej w zależności od tego, w jaki sposób spełnia ona *dwa niezależne* warunki. Oznacza to, że wartość zostanie zmieniona (włączona lub wyłączona), jeśli każdy warunek zostanie spełniony.

Stan początkowy. Należy określić, czy każdy rekord z nową zmienną ma początkowo mieć wartość **Wł.** czy **Wył.** Należy zwrócić uwagę, że ta wartość może ulec zmianie po spełnieniu poszczególnych warunków.

Wartość „Wł.”. Należy określić wartość dla nowej zmiennej po spełnieniu warunku Wł.

Przełącz na „Wł.”, kiedy. Należy określić warunek CLEM, który spowoduje zmianę stanu na Wł., jeśli warunek jest prawdziwy. Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

Wartość „Wył.”. Należy określić wartość dla nowej zmiennej po spełnieniu warunku Wył.

Przełącz na Wył.”, kiedy. Należy określić warunek CLEM, który spowoduje zmianę stanu na Wył., jeśli warunek jest fałszywy. Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

Uwaga: Aby określić pusty łańcuch, należy wpisać cudzysłów otwierający i zamykający, pomiędzy którymi nie będzie żadnej wartości, czyli "". Podobnie cudzysłowów należy użyć, jeśli wartość łańcucha ma być traktowana jako liczba.

Ustawianie opcji węzła wyliczeń — Liczebność

Węzeł wyliczeń typu Liczebność jest używany do zastosowania serii warunków do wartości zmiennej numerycznej w zbiorze danych. Jeśli poszczególny warunek zostanie spełniony, wartość wyliczonej zmiennej liczebności jest zwiększana o ustawioną wartość przyrostu. Tego typu węzeł wyliczeń jest przydatny w przypadku danych szeregów czasowych.

Wartość początkowa. Ustawia wartość używaną podczas wykonywania działań dla nowej zmiennej. Wartość początkowa musi być stałą wartością liczbową. Aby zwiększyć lub zmniejszyć wartość, można użyć przycisków strzałek.

Zwiększ zliczaną wartość, gdy. Należy określić warunek CLEM, którego spełnienie spowoduje zmianę wyliczonej wartości na podstawie liczby określonej w polu Przyrost. Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

Przyrost. Należy ustawić wartość używaną do zwiększania liczebności. Może to być stała wartość liczbowa lub wynik wyrażenia CLEM.

Zresetuj do początkowej wartości, gdy. Należy określić warunek, którego spełnienie spowoduje zresetowanie wyliczonej wartości i ustawienie wartości początkowej. Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

Ustawianie opcji węzła wyliczeń — Warunkowe

Węzły wyliczeń typu Warunkowe do wyliczenia nowej zmiennej używają serii instrukcji Jeżeli-To-Inaczej.

Jeżeli. Należy określić warunek CLEM, jaki podczas wykonywania będzie oceniany dla każdego rekordu. Jeśli warunek będzie prawdziwy (lub niefałszywy w przypadku liczb), wówczas nowej zmiennej zostanie przypisana wartość określona poniżej przez wyrażenie To. Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

To. Należy określić wartość lub wyrażenie CLEM dla nowej zmiennej, jeśli instrukcja Jeżeli opisana powyżej będzie prawdziwa (lub niefałszywa). Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

Inaczej. Należy określić wartość lub wyrażenie CLEM dla nowej zmiennej, jeśli instrukcja Jeżeli opisana powyżej będzie fałszywa. Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

Rekodowanie wartości za pomocą węzła wyliczeń

Węzłów wyliczeń można także użyć do rekodowania wartości, na przykład przez przekształcanie zmiennej łańcuchowej z wartościami jakościowymi na liczbową zmienną nominalną (zbiór).

1. Dla opcji Wylicz jako wybierz odpowiednio typ zmiennej (Nominalna, Flaga itd.) .
2. Określ warunki rekodowania wartości. Przykładowo można ustawić wartość na 1, jeśli Drug='drugA', na 2, jeśli Drug='drugB' itd.

---

węzeł Wypełnianie

Węzły Wypełnianie służą do zastępowania wartości zmiennych i zmiany typu składowania. Wartości mogą być zastępowane na podstawie określonego warunku CLEM, np. @BLANK(@FIELD). Alternatywnie można wybrać, aby wszystkie wartości puste lub null zastępowane były konkretną wartością. W celu zastąpienia braków danych węzły wypełniania często są używane w połączeniu z węzłem typu. Na przykład można wypełnić puste wartości średnią wartością zmiennej, określając wyrażenie, takie jak @GLOBAL_MEAN. To wyrażenie spowoduje wypełnienie wszystkich wartości pustych wartością średnią obliczoną przez węzeł ustawień globalnych.

Wypełnij zmienne. Korzystając z selektora zmiennych (przycisk po prawej stronie pola tekstowego), należy wybrać zmienne ze zbioru danych, którego wartości zostaną sprawdzone i zastąpione. Działanie domyślne polega na zastąpieniu wartości w zależności od warunku i zastąpieniu ich przez wyrażenia określone poniżej. Można także wybrać alternatywną metodę zastępowania przy użyciu opcji Zastąp.

Uwaga: W przypadku wybierania wielu zmiennych do zastąpienia wartością zdefiniowaną przez użytkownika istotne jest, aby typy zmiennych były podobne (wszystkie numeryczne lub wszystkie symboliczne).

Zamień. Tę opcję należy wybrać, aby zamienić wartości wybranych zmiennych przy użyciu jednej z następujących metod:

- **W oparciu o warunek.** Ta opcja aktywuje zmienną Warunek oraz konstruktora wyrażeń w celu utworzenia wyrażenia stanowiącego warunek zastąpienia określoną wartością.
- **Zawsze.** Zastępuje wszystkie wartości wybranej zmiennej. Przykładowo można użyć tej opcji, aby przekształcić typ składowania przychodu na łańcuch, używając następującego wyrażenia CLEM: (to_string(income)).
- **Puste wartości.** Zastępuje wszystkie określone przez użytkownika puste wartości w wybranej zmiennej. Do wybrania pustych wartości używany jest standardowy warunek @BLANK(@FIELD). *Uwaga*: Puste wartości można zdefiniować za pomocą karty Typy węzła źródłowego lub za pomocą węzła Typy.
- **Wartości null.** Zastępuje wszystkie systemowe wartości null w wybranej zmiennej. Do wybrania wartości null używany jest standardowy warunek @NULL(@FIELD).
- **Puste i wartości null.** Zastępuje zarówno puste wartości, jak i systemowe wartości null w wybranej zmiennej. Ta opcja jest przydatna, jeśli użytkownik nie ma pewności, czy wartość null zostały zdefiniowane jako braki danych.

Warunek. Ta opcja jest dostępna po wybraniu opcji **W oparciu o warunek**. To pole tekstowe umożliwia określenie wyrażenia CLEM do oceny wybranych zmiennych. Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

Zamień na. Należy określić wyrażenie CLEM, aby nadać wybranym zmiennym nową wartość. Wartość można również zastąpić wartością null, wpisując w polu tekstowym wyrażenie undef. Aby otworzyć konstruktora wyrażeń, należy kliknąć przycisk kalkulatora.

Uwaga: Jeśli wybranymi zmiennymi są zmienne łańcuchowe, należy je zastąpić wartością łańcuchową. Użycie wartości domyślnej wynoszącej 0 lub innej wartości liczbowej jako wartości zastępującej zmienne łańcuchowe spowoduje wystąpienie błędu.

Należy pamiętać, że zastosowanie poniższych rozwiązań może spowodować zmianę kolejności wierszy:

- Wykonywanie w bazie danych za pośrednictwem analizy wstępnej SQL
- Wykonywanie za pośrednictwem zdalnego serwera IBM SPSS Analytic Server
- Korzystanie z funkcji uruchamianych w osadzonego serwera IBM SPSS Analytic Server
- Wyliczanie listy (przykładowo patrz “Wyliczanie zmiennej listy lub geoprzestrzennej” na stronie 158)
- Wywoływanie funkcji opisanych w sekcji Funkcje przestrzenne

Przekształcanie sposobu składowania za pomocą węzła Wypełnianie

Użycie warunku Zamień węzła wypełniania pozwala w prosty sposób przekształcić sposób składowania zmiennej lub wielu zmiennych. Przykładowo, używając funkcji przekształcania to_integer, można przekształcić zmienną *income* (przychód) z łańcucha na liczbę całkowitą za pomocą następującego wyrażenia CLEM: to_integer(income).

Do wyświetlania dostępnych funkcji przekształcania oraz automatycznego tworzenia wyrażeń CLEM służy konstruktor wyrażeń. Z listy rozwijanej Funkcje należy wybrać opcję **Konwersja**, aby wyświetlić listę funkcji przekształcania sposobu składowania. Dostępne są następujące funkcje przekształcania:

- to_integer(ELEMENT)
- to_real(ELEMENT)
- to_number(ELEMENT)
- to_string(ELEMENT)
- to_time(ELEMENT)
- to_timestamp(ELEMENT)
- to_date(ELEMENT)
- to_datetime(ELEMENT)

Konwertowanie wartości daty i czasu. Funkcje przekształcania (i inne funkcje wymagające określonego typu danych wejściowych, np. wartość daty lub czasu) są uzależnione od bieżących formatów określonych w oknie dialogowym opcji strumienia. Przykładowo: aby wykonać przekształcenie zmiennej łańcuchowej o wartościach *Sty 2003*, *Lut 2003* itd. do postaci składowania daty, jako domyślny format daty strumienia należy wybrać **MIE RRRR**.

Funkcje przekształcania są również dostępne z węzła wyliczeń i umożliwiają tymczasowe przekształcenie podczas wyliczania. Węzła wyliczeń można także użyć do wykonywania innych działań, takich jak rekodowanie zmiennych łańcuchowych przez wartości jakościowe. Więcej informacji można znaleźć w temacie “Rekodowanie wartości za pomocą węzła wyliczeń” na stronie 161.

Węzeł Rekodowanie

Węzeł Rekodowanie umożliwia transformację z jednego zbioru wartości jakościowych na inny. Rekodowanie jest przydatne do zwijania kategorii lub ponownego pogrupowania danych do analizy. Na przykład można rekodować wartości dla zmiennej *Product* (Produkt) na trzy grupy, takie jak *Kitchenware* (Naczynia kuchenne), *Bath and Linens* (Ręczniki i pościele) oraz *Appliances* (Urządzenia). Często ta operacja jest wykonywana bezpośrednio z węzła Rozkład poprzez pogrupowanie wartości i wygenerowanie węzła rekodowania. Więcej informacji można znaleźć w temacie “Użycie węzła rozkładu” na stronie 238.

Rekodowanie można przeprowadzić dla jednej lub kilku zmiennych symbolicznych. Można również zastąpić nowe wartości istniejącej zmiennej lub wygenerować nową zmienną.

Kiedy użyć węzła Rekodowanie

Przed użyciem węzła Rekodowanie należy zastanowić się, czy inny węzeł Operacje na zmiennych nie byłby bardziej odpowiedni do wykonania danego zadania:

- Aby przekształcić przedziały liczbowe na zestawy za pomocą zautomatyzowanej metody (rangi lub percentyle), należy użyć węzła Kategoryzacja. Więcej informacji można znaleźć w temacie “Węzeł Kategoryzacja” na stronie 167.
- W celu ręcznego sklasyfikowania przedziałów liczbowych jako zestawy należy użyć węzła Wyliczanie. Przykładowo, aby zwinąć wartości wynagrodzenia do określonych kategorii przedziału wynagrodzenia, należy użyć węzła Wyliczanie w celu ręcznego zdefiniowania poszczególnych kategorii.
- Aby utworzyć co najmniej jedną zmienną flagi w oparciu o wartości zmiennej jakościowej, takiej jak *Mortgage_type* (Typ hipoteki), należy użyć węzła Flagowanie.
- Aby przekształcić typ składowania zmiennej jakościowej na numeryczną, można użyć węzła Wyliczanie. Przykładowo można przekształcić wartości *No* (Nie) i *Yes* (Tak) odpowiednio na 0 i 1. Więcej informacji można znaleźć w temacie “Rekodowanie wartości za pomocą węzła wyliczeń” na stronie 161.

Ustawianie opcji dla węzła Rekodowanie

Użycie węzła rekodowania odbywa się w trzech etapach:

1. Najpierw należy wybrać, czy rekodowanie ma dotyczyć wielu zmiennych, czy pojedynczej zmiennej.
2. Następnie należy wybrać, czy ma zostać wykonane przekodowanie na istniejącą zmienną, czy ma zostać utworzona nowa zmienna.
3. I wreszcie należy użyć opcji dynamicznych dostępnych w oknie dialogowym węzła rekodowania w celu odwzorowania zestawów w odpowiedni sposób.

   Dominanta. Należy wybrać opcję **Pojedynczy**, aby rekodować kategorie dla jednej zmiennej. Wybranie opcji **Wielokrotny** pozwala aktywować opcje umożliwiające transformację więcej niż jednej zmiennej na raz.

   Rekoduj na. Należy wybrać opcję **Nowa zmienna**, aby zachować oryginalną zmienną nominalną i utworzyć dodatkową zmienną zawierającą rekodowane wartości. Opcję **Istniejąca zmienna** należy wybrać, aby zastąpić wartości oryginalnej zmiennej nowymi klasyfikacjami. Zasadniczo jest to operacja „wypełniania”.

   Po określeniu trybu i opcji zastępowania należy wybrać zmienną transformacji i określić nowe wartości klasyfikacji przy użyciu opcji dynamicznych w dolnej części okna dialogowego. Opcje te różnią się w zależności od trybu wybranego powyżej.

   Rekoduj zmienną/zmienne. Za pomocą przycisku Selektor zmiennych po prawej stronie wybierz jedną (tryb pojedynczy) lub więcej (tryb wielokrotny) zmiennych jakościowych.

   Nazwa nowej zmiennej. Należy określić nazwę nowej zmiennej nominalnej zawierającej przekodowane wartości. Ta opcja jest dostępna tylko w przypadku trybu pojedynczego po wybraniu opcji **Nowa zmienna**. Jeśli wybrana zostanie opcja **Istniejąca zmienna**, oryginalna nazwa zmiennej jest zachowywana. W przypadku korzystania z trybu wielokrotnego ta opcja jest zastępowana elementami sterującymi umożliwiającymi określenie rozszerzenia dodawanego do każdej nowej zmiennej. Więcej informacji można znaleźć w temacie “Rekodowanie wielu zmiennych” na stronie 164.

   Rekodowane wartości. Ta tabela umożliwia usunięcie z mapowania starych zestawów wartości.

- **Wartość oryginalna.** W tej kolumnie wyświetlane są istniejące wartości wybranych zmiennych.
- **Nowa wartość.** Ta kolumna służy do wpisania wartości nowej kategorii lub wybrania jej z listy rozwijanej. W przypadku automatycznego generowania węzła rekodowania na podstawie wartości z wykresu rozkładu wartości te są uwzględniane na liście rozwijanej. Dzięki temu można szybko odwzorować istniejące wartości na zbiór znanych wartości. Przykładowo organizacje zajmujące się opieką medyczną niekiedy grupują diagnozy w różny sposób na podstawie sieci lub ustawień regionalnych. Po scaleniu lub zgromadzeniu danych wszystkie strony będą musiały rekodować nowe, a nawet istniejące dane w spójny sposób. Zamiast ręcznego wpisywania z długiej listy każdej wartości przewidywanej można wczytać nadrzędną listę wartości do programu IBM SPSS Modeler, uruchomić wykres rozkładu dla zmiennej *Diagnosis* (Diagnoza) i wygenerować węzeł rekodowania (wartości) dla tej zmiennej bezpośrednio z wykresu. Ten proces udostępni wszystkie wartości przewidywane diagnozy z listy rozwijanej nowych wartości.

4. Kliknij przycisk **Uzyskaj**, aby odczytać oryginalne wartości dla co najmniej jednej zmiennej wybranej powyżej.
5. Kliknij przycisk **Kopiuj**, aby wkleić oryginalne wartości w kolumnie *Nowa wartość* dla zmiennych, które nie zostały jeszcze zmapowane. Niezmapowane oryginalne wartości zostają dodane do listy rozwijanej.
6. Kliknij przycisk **Wyczyść nowe**, aby wykasować wszystkie specyfikacje w kolumnie *Nowa wartość*. *Uwaga*: Ta opcja nie spowoduje skasowania wartości z listy rozwijanej.
7. Kliknij przycisk **Automatycznie**, aby automatycznie wygenerować kolejne liczby całkowite dla każdej oryginalnej wartości. Wygenerowane mogą zostać tylko wartości całkowite (nie wartości rzeczywiste, takie jak 1,5, 2,5 itd.).

Przykładowo można automatycznie wygenerować kolejne numery identyfikacyjne produktów dla nazw produktów lub numery kursów dla oferowanych zajęć na uniwersytecie. Ta funkcja odpowiada transformacji w wyniku automatycznego rekodowania dla zestawów w programie IBM SPSS Statistics.

Dla wartości nieokreślonych użyj. Ta opcja służy do wypełniania nieokreślonych wartości w nowej zmiennej. Można zachować oryginalną wartość, wybierając opcję **Wartość oryginalna**, lub określić wartość domyślną.

Rekodowanie wielu zmiennych

Aby jednorazowo zmapować wartości kategorii na więcej niż jedną zmienną, należy ustawić tryb na **Wielokrotny**. Spowoduje to włączenie nowych ustawień w oknie dialogowym Rekodowanie, które opisano poniżej.

Rekoduj zmienne. Przycisk Selektor zmiennych po prawej stronie umożliwia wybranie zmiennych, jakie mają zostać poddane transformacji. Korzystając z selektora zmiennych można wybrać wszystkie zmienne naraz lub zmienne podobnego typu, np. nominalne lub flaga.

Rozszerzenie nazwy zmiennej. W przypadku rekodowania wielu zmiennych jednocześnie bardziej efektywne będzie określenie wspólnego rozszerzenia dodawanego do wszystkich nowych zmiennych niż do nazw pojedynczych zmiennych. Należy określić rozszerzenie, takie jak _recode, i wybrać, czy ma ono znajdować się na początku, czy na końcu oryginalnej nazwy zmiennej.

Składowanie i poziom pomiaru dla zmiennych rekodowanych

Węzeł Rekodowanie w wyniku operacji rekodowania zawsze tworzy zmienną nominalną. W niektórych przypadkach, jeśli używany jest tryb rekodowania **Istniejąca zmienna**, może dojść do zmiany poziomu pomiaru zmiennej.

Nowy sposób składowania zmiennej (jak zmienne są *składowane*, a nie w jaki sposób są *używane*) jest wyznaczany na podstawie następujących opcji na karcie Ustawienia:

- Jeśli ustawiono, że dla nieokreślonych wartości ma być używana wartość domyślna, typ składowania jest ustalany poprzez zbadanie nowych wartości oraz wartości domyślnej i określenie odpowiedniego składowania. Przykładowo, jeśli wartości mogą być przeanalizowane jak liczby całkowite, typem składowania dla tej zmiennej będzie składowanie z użyciem liczby całkowitej.
- Jeśli dla wartości nieokreślonych mają być stosowane oryginalne wartości, typ składowania zostanie określony na podstawie typu składowania zmiennej oryginalnej. Jeśli wszystkie wartości mogą zostać przeanalizowane zgodnie z typem składowania oryginalnej zmiennej, wówczas ten typ składowania jest zachowywany; w przeciwnym razie zostaje określony poprzez wyszukanie odpowiedniego typu składowania obejmującego stare i nowe wartości. Na

przykład rekodowanie zbioru liczb całkowitych { 1, 2, 3, 4, 5 }, gdzie 4 => 0, 5 => 0, spowoduje wygenerowanie nowego zbioru liczb całkowitych { 1, 2, 3, 0 }, natomiast przy założeniu, że 4 => “Powyżej 3”, 5 => “Powyżej 3” wygenerowany zostanie zbiór łańcuchowy { “1”, “2”, “3”, “Powyżej 3” }.

Uwaga: Jeśli oryginalny typ nie był określony, nowy typ również będzie nieokreślony.

Węzeł Anonimizacja

Węzeł Anonimizacja umożliwia maskowanie nazw zmiennych i/lub wartości zmiennych podczas pracy z danymi, jakie mają zostać uwzględnione w modelu za węzłem. W ten sposób wygenerowany model może być swobodnie rozdzielany (np. do działu wsparcia technicznego) bez ryzyka, że nieuprawnieni użytkownicy będą mogli zobaczyć poufne dane, takie jak rekordy pracowników czy dokumentacja medyczna pacjentów.

W zależności od umiejscowienia węzła Anonimizacja w strumieniu konieczne może być wprowadzenie zmian w innych węzłach. Przykładowo, jeśli węzeł Anonimizacja zostanie wstawiony przed węzłem Selekcja, kryteria wyboru w węźle selekcji będą wymagały zmiany, jeśli oddziałują na wartości, które teraz zostaną poddane anonimizacji.

Metoda, jakiej należy użyć do anonimizacji, zależy od różnych czynników. W przypadku nazw zmiennych i wszystkich wartości zmiennych z wyjątkiem ciągłych poziomów pomiarów dane są zastępowane przez łańcuch w następującej postaci:

*prefix*_S*n*

gdzie *prefix_* to łańcuch określony przez użytkownika lub domyślny łańcuch anon_, a *n* jest wartością całkowitą rozpoczynającą się od 0 i zwiększaną dla każdej unikalnej wartości (np. anon_S0, anon_S1 itd.).

Wartości zmiennych typu ilościowego muszą zostać poddane transformacji, ponieważ przedziały liczbowe mogą być przetwarzane z wartościami całkowitymi lub rzeczywistymi, a nie z łańcuchami. Dlatego mogą zostać poddane anonimizacji poprzez transformację przedziału na inny przedział, który zamaskuje oryginalne dane. Transformacja wartości *x* w przedziale odbywa się w następujący sposób:

A*(*x* + B)

gdzie:

A jest czynnikiem skalującym, który musi być większy od 0.

B jest przesunięciem translacji, jakie zostanie dodane do wartości.

Przykład

W przypadku zmiennej *AGE* (Wiek), w której czynnik skalujący *A* jest ustawiany na wartość 7, a przesunięcie translacji *B* jest ustawiane na wartość 3, wartości *AGE* zostaną przetransformowane w następujący sposób:

7*(AGE + 3)

Ustawianie opcji dla węzła Anonimizacja

Tutaj można wybrać, wartości których zmiennych zostaną zamaskowane w dalszej części strumienia.

Należy pamiętać, że zmienne danych muszą zostać określone przed węzłem Anonimizacja, aby można było zrealizować operacje anonimizacji. Dane można określić klikając przycisk **Odczytaj wartości** w węźle Typy lub na karcie Typy w węźle źródłowym.

Zmienna. Wyświetla listę zmiennych w bieżącym zbiorze danych. Jeśli jakieś nazwy zmiennych zostały już poddane anonimizacji, nazwy te są tutaj wyświetlane.

Poziom pomiaru. Poziom pomiaru zmiennej.

Anonimizacja wartości. Należy wybrać co najmniej jedną zmienną, kliknąć tę kolumnę i wybrać opcję **Tak**, aby przeprowadzić anonimizację wartości zmiennej z zastosowaniem domyślnego przedrostka **anon_**; wybranie opcji **Określ** pozwala wyświetlić okno dialogowe, w którym można wprowadzić własny przedrostek, lub, jeśli wartości zmiennych są typu *ilościowego*, określić, czy podczas transformacji wartości zmiennych mają być używane wartości losowe, czy zdefiniowane przez użytkownika. Należy pamiętać, że typów zmiennych *ilościowych* i nie*ilościowych* nie można określić w tej samej operacji; należy to zrobić osobno dla każdego typu zmiennej.

Widok aktualnych zmiennych. Tę opcję należy wybrać, aby wyświetlić zmienne dla zbiorów danych aktywnie połączonych z węzłem Anonimizacja. Ta opcja jest wybrana domyślnie.

Widok ustawień niewykorzystanych zmiennych. Opcję tę należy wybrać, aby wyświetlić zmienne dla zbiorów danych, które kiedyś były, ale już nie są połączone z węzłem. Ta opcja jest przydatna podczas kopiowania węzłów z jednego strumienia do innego lub podczas zapisywania i ponownego wczytywania węzłów.

Określanie sposobu anonimizacji wartości zmiennych

W oknie Zastąp wartości można zdecydować, czy dla wartości zmiennych poddanych anonimizacji ma być używany domyślny przedrostek, czy też przedrostek niestandardowy. Kliknięcie przycisku **OK** w tym oknie dialogowym spowoduje zmianę ustawienia Anonimizacja wartości na karcie Ustawienia na **Tak** dla wybranych zmiennych.

Przedrostek wartości zmiennej. Domyślnym przedrostkiem dla wartości zmiennych poddanych anonimizacji jest **anon_**; po wybraniu opcji **Użytkownika** można wprowadzić własny przedrostek.

Okno dialogowe Transformuj wartości jest wyświetlane tylko dla zmiennych ilościowych i umożliwia określenie, czy podczas transformacji wartości zmiennych mają być używane wartości losowe, czy wartości zdefiniowane przez użytkownika.

Losowa. Tę opcję należy wybrać, aby podczas transformacji używane były wartości losowe. Opcja **Ustaw wartość początkową generatora liczb losowych** jest domyślnie zaznaczona; należy określić wartość w polu **Wartość początkowa** lub użyć wartości domyślnej.

Stała. Tę opcję należy wybrać, aby określić własne wartości dla transformacji.

- **Czynnik skalujący.** Liczba, wg której wartości zmiennych będą pomnożone w czasie transformacji. Wartość minimalna wynosi 1; wartość maksymalna to zwykle 10, ale można ją obniżyć, aby uniknąć przepełnienia.
- **Współczynnik translacji.** Liczba, jaka zostanie dodana do wartości zmiennych podczas transformacji. Wartość minimalna wynosi 0; wartość maksymalna to zwykle 1000, ale można ją obniżyć, aby uniknąć przepełnienia.

Anonimizacja wartości zmiennych

W przypadku zmiennych wybranych do anonimizacji na karcie Ustawienia ich wartości zostają poddane anonimizacji:

- Po uruchomieniu strumienia zawierającego węzeł Anonimizacja
- Po wyświetleniu podglądu wartości

Aby wyświetlić podgląd wartości, należy kliknąć przycisk **Anonimizacja wartości** na karcie Anonimizacja. Następnie należy wybrać nazwę zmiennej z listy rozwijanej.

Jeśli poziom pomiaru ustawiono jako ilościowy, wyświetlane są następujące wartości:

- Wartości minimalne i maksymalne z oryginalnego przedziału
- Równanie użyte do transformacji wartości

Jeśli poziom pomiaru jest inny niż ilościowy, na ekranie wyświetlane są oryginalne i anonimizowane wartości zmiennej.

Jeśli na ekranie wyświetlane jest żółte tło, oznacza to, że ustawienie dla wybranej zmiennej zostało zmienione od czasu ostatniej anonimizacji wartości lub że wprowadzone zostały zmiany w danych powyżej węzła Anonimizacja, w wyniku

których wartości anonimizowane mogą już nie być poprawne. Wyświetlany jest bieżący zbiór wartości; należy ponownie kliknąć przycisk **Anonimizacja wartości**, aby wygenerować nowy zbiór wartości odpowiedni dla bieżących ustawień.

Anonimizacja wartości. Tworzy anonimizowane wartości dla wybranej zmiennej i wyświetla je w tabeli. Jeśli dla zmiennej typu ilościowego wartości początkowe są określane w sposób losowy, kliknięcie tego przycisku kilkakrotnie spowoduje utworzenie za każdym razem innego zbioru wartości.

Wyczyść wartości. Usuwa z tabeli oryginalne i anonimizowane wartości.

Węzeł Kategoryzacja

Węzeł Kategoryzacja umożliwia automatyczne utworzenie nowych zmiennych nominalnych na podstawie wartości z jednej lub większej liczby istniejących zmiennych ilościowych (zakres liczbowy). Można na przykład przekształcić ilościową zmienną przychodu na nową zmienną jakościową zawierającą grupy przychodu o równej szerokości lub stanowiące odchylenia od średniej. Alternatywnie można wybrać jakościową zmienną nadzorującą, aby zachować siłę oryginalnego powiązania pomiędzy dwiema zmiennymi.

Kategoryzacja może być użyteczna z wielu powodów, takich jak:

- **Wymagania algorytmu.** Niektóre algorytmy, np. Naive Bayes, regresja logistyczna, wymagają jakościowych danych wejściowych.
- **Wydajność.** Algorytmy, takie jak wielomianowa regresja logistyczna, mogą działać lepiej, jeśli liczba odmiennych wartości zostanie zredukowana. Na przykład w każdym przedziale należy użyć mediany lub wartości średniej zamiast oryginalnych wartości.
- **Ochrona danych.** W celu ochrony prywatności poufne dane osobowe, takie jak wynagrodzenia, mogą być zgłaszane jako przedziały, a nie jako rzeczywiste wartości wynagrodzenia.

Dostępnych jest wiele metod kategoryzacji. Po utworzeniu kategorii dla nowej zmiennej na podstawie punktu podziału można wygenerować węzeł Wyliczanie.

Kiedy użyć węzła Kategoryzacja

Przed użyciem węzła Kategoryzacja należy zastanowić się, czy inna technika nie byłaby bardziej odpowiednia do wykonania zadania:

- Aby ręcznie określić punkty podziału dla kategorii, np. konkretne wstępnie zdefiniowane przedziały wynagrodzenia, należy użyć węzła Wyliczanie. Więcej informacji można znaleźć w temacie “węzeł wyliczeń” na stronie 154.
- Aby utworzyć nowe kategorie dla istniejących zestawów, należy użyć węzła Rekodowanie. Więcej informacji można znaleźć w temacie “Węzeł Rekodowanie” na stronie 163.

Traktowanie braków danych

Węzeł Kategoryzacja traktuje braki danych w następujący sposób:

- **Puste wartości określone przez użytkownika.** Braki danych określone jako puste wartości są uwzględniane podczas transformacji. Na przykład, jeśli za pomocą węzła Typy wyznaczono, że wartość –99 oznacza pustą wartość, wartość ta będzie uwzględniana w procesie kategoryzacji. Aby zignorować puste wartości podczas kategoryzacji, należy użyć węzła wypełniania w celu zastąpienia wartości pustych systemową wartością null.
- **Systemowe braki danych ($null$).** Wartości null są ignorowane podczas przekształcania kategoryzacji i pozostają wartościami null po zakończeniu przekształcania.

Karta Ustawienia udostępnia opcje dla dostępnych technik. Na karcie Widok wyświetlane są punkty podziału ustalone dla danych, które wcześniej były uruchomione w węźle.

Ustawianie opcji dla węzła Kategoryzacja

Korzystając z węzła Kategoryzacja, można automatycznie generować przedziały (kategorie), korzystając z następujących technik:

- Ustalona szerokość przedziałów
- N-tyle (równa liczebność lub suma)
- Średnia i odchylenie standardowe
- Rangi
- Zoptymalizowane wartości względem jakościowej zmiennej nadzorującej

Dolna część okna dialogowego zmienia się w sposób dynamiczny w zależności od wybranej metody kategoryzacji.

Zmienne poddane podziałowi. Tutaj wyświetlane są zmienne ilościowe (zakres liczbowy) oczekujące na transformację. Węzeł Kategoryzacja umożliwia podział wielu zmiennych jednocześnie. Dodawanie lub usuwanie zmiennych umożliwiają przyciski po prawej stronie.

Metoda kategoryzacji. Należy wybrać metodę, jaka została użyta do ustalenia punktów podziału dla nowych przedziałów zmiennych (kategorii). W kolejnych tematach opisano opcje dostępne w poszczególnych przypadkach.

Wartości graniczne kategorii. Należy określić, w jaki sposób obliczane będą wartości graniczne kategorii.

- **Zawsze przeliczaj.** Punkty podziału i alokacje kategorii są zawsze przeliczane po uruchomieniu węzła.
- **Odczytaj z karty Wartości podziałów, jeśli są tam zdefiniowane.** Punkty podziału i alokacje kategorii są obliczane tylko w razie konieczności (na przykład, jeśli dodano nowe dane).

W poniższych tematach omówiono opcje dla dostępnych metod kategoryzacji.

Przedziały o ustalonej szerokości

Jeśli jako metoda kategoryzacji wybrana zostanie opcja **Ustalona szerokość przedziałów**, w oknie dialogowym wyświetlany jest nowy zestaw opcji.

Rozszerzenie nazwy. Należy określić rozszerzenie, jakie będzie używane dla wygenerowanych zmiennych. *_BIN* to rozszerzenie domyślne. Można również określić, czy rozszerzenie ma być dodawane na początku (**Przedrostek**), czy na końcu (**Przyrostek**) nazwy zmiennej. Można na przykład wygenerować nową zmienną o nazwie *income_BIN*.

Szerokość kategorii. Należy określić wartość (liczba całkowita lub rzeczywista), jaka będzie używana do obliczenia „szerokości" przedziału. Można na przykład użyć wartości domyślnej 10, aby dokonać podziału dla zmiennej *Age* (Wiek). Ponieważ zmienna *Age* mieści się w zakresie od 18 do 65, wygenerowane zostaną przedziały przedstawione w poniższej tabeli.

Tabela 25. Przedziały dla zmiennej Age (Wiek) mieszczącej się w zakresie od 18 do 65

Przedział 1	Przedział 2	Przedział 3	Przedział 4	Przedział 5	Przedział 6
>=13 do <23	>=23 do <33	>=33 do <43	>=43 do <53	>=53 do <63	>=63 do <73

Początek przedziałów jest obliczany na podstawie najniższej zeskanowanej wartości minus połowa szerokości przedziału (jak określono). Na przykład w przypadku przedziałów pokazanych powyżej jako początek przedziałów użyta została wartość 13, zgodnie z następującym obliczeniem: 18 *[najniższa wartość danych]* – 5 *[0,5 × (Szerokość przedziału równa 10)]* = 13.

Liczba przedziałów. Ta opcja umożliwia określenie liczby całkowitej używanej do określenia liczby przedziałów o ustalonej szerokości (kategorii) dla nowych zmiennych.

Po wykonaniu węzła kategoryzacji w strumieniu można wyświetlić wygenerowane wartości graniczne kategorii, klikając zakładkę **Podgląd** w oknie dialogowym węzła kategoryzacji. Więcej informacji można znaleźć w temacie "Podgląd wygenerowanych przedziałów" na stronie 172.

N-tyle (równa liczebność lub suma)

Metoda kategoryzacji powoduje utworzenie zmiennych nominalnych, jakie mogą zostać użyte do rozdziału skanowanych rekordów na grupy percentyli (lub kwartyli, decyli itd.), w wyniku czego każda grupa będzie zawierała taką samą liczbę rekordów lub suma wartości w każdej grupie będzie taka sama. Rekordy są rangowane w porządku rosnącym na podstawie wartości określonej w zmiennej poddanej podziałowi, w wyniku czego rekordy o najniższych wartościach dla wybranej zmiennej poddanej podziałowi mają przypisaną rangę 1, kolejny zestaw rekordów ma rangę 2 itd. Wartości graniczne dla każdego przedziału są generowane automatycznie w oparciu o użyte dane i zastosowaną metodę tworzenia N-tyli.

Rozszerzenie nazwy N-tyla. Należy określić rozszerzenie, jakie będzie używane dla zmiennych wygenerowanych z zastosowaniem standardowych p-tyli. Domyślnym rozszerzeniem jest *_TILE* plus *N*, gdzie *N* oznacza liczbę N-tyli. Można również określić, czy rozszerzenie ma być dodawane na początku (**Przedrostek**), czy na końcu (**Przyrostek**) nazwy zmiennej. Można na przykład wygenerować nową zmienną o nazwie *income_BIN4*.

Rozszerzenie N-tyla użytkownika. Należy określić rozszerzenie, jakie będzie używane dla niestandardowego przedziału N-tyli. Wartość domyślna to *_TILEN*. Należy pamiętać, że *N* w tym przypadku *nie* jest zastępowane wartością niestandardową.

Dostępne p-tyle to:

- **Kwartyl.** Generuje 4 przedziały, każdy składający się z 25% obserwacji.
- **Kwintyl.** Generuje 5 przedziałów, każdy składający się z 20% obserwacji.
- **Decyl.** Generuje 10 przedziałów, każdy składający się z 10% obserwacji.
- **Vingtyl.** Generuje 20 przedziałów, każdy składający się z 5% obserwacji.
- **Percentyl.** Generuje 100 przedziałów, każdy składający się z 1% obserwacji.
- **N użytkownika.** Tę opcję należy wybrać, aby określić liczbę przedziałów. Przykładowo wartość 3 spowoduje utworzenie 3 kategorii podziału (2 punkty podziału), każda składająca się z 33,3% obserwacji.

Należy pamiętać, że jeśli w danych jest mniejsza liczba wartości dyskretnych niż liczba określonych N-tyli, nie wszystkie N-tyle zostaną użyte. W takich sytuacjach nowy rozkład prawdopodobnie będzie odzwierciedlał oryginalny rozkład danych.

Metoda tworzenia N-tyli. Określa metodę używaną do przypisywania rekordów do przedziałów.

- **Liczebność rekordów.** Stara się przypisać jednakową liczbę rekordów do każdego przedziału.
- **Suma wartości.** Stara się przypisać rekordy do przedziałów, tak aby suma wartości w każdym przedziale była jednakowa. Na przykład w przypadku konkretnego ukierunkowania działań sprzedażowych ta metoda może być zastosowana w celu przypisania potencjalnych klientów do grup decylowych na podstawie wartości dla rekordu, umieszczając potencjalnych klientów o najwyższej wartości w górnej części przedziału. Przykładowo firma farmaceutyczna może dokonać rangowania lekarzy w postaci grup decylowych w oparciu o liczbę wypisanych recept. Każdy decyl będzie zawierał w przybliżeniu taką samą liczbę recept, jednak liczba osób wystawiających te recepty będzie różna, przy czym osoby, które wypisały najwięcej recept, będą skupione w decylu 10. Należy pamiętać, że przy takim rozwiązaniu zakłada się, że wszystkie wartości są większe od zera; w przeciwnym wypadku może dojść do uzyskania nieoczekiwanych wyników.

Wiązania. Warunek wiązania występuje, kiedy wartości po obu stronach punktu podziału są identyczne. Przykładowo, jeśli przypisywane są decyle i więcej niż 10% rekordów zawiera tą samą wartość dla zmiennej poddanej podziałowi, wówczas nie ma możliwości dopasowania ich do tego samego przedziału bez wymuszenia wartości granicznej. Wiązania można przenieść w górę do następnego przedziału lub mogą pozostać w bieżącym, ale konieczne będzie ich przetworzenie, dlatego wszystkie rekordy z identycznymi wartościami znajdą się w tym samym przedziale, nawet jeśli spowoduje to, że niektóre przedziały będą zawierały więcej rekordów niż oczekiwano. W wyniku tego wartości

graniczne kolejnych przedziałów również mogą być skorygowane, co spowoduje różne przypisanie wartości dla tego samego zbioru liczb z użyciem metody zastosowanej do przetworzenia wiązań.

- **Dodaj do następnej.** Tę opcję należy wybrać, aby przenieść wartości wiązania w górę do następnej kategorii.
- **Pozostaw w bieżącej.** Zachowuje wartości w bieżącej (niższej) kategorii. Zastosowanie tej metody może skutkować utworzeniem mniejszej liczby przedziałów.
- **Przydziel losowo.** Tę opcję należy wybrać, aby przydzielać wartości wiązania do przedziału w sposób losowy. Podejmowana będzie próba zachowania w każdym przedziale takiej samej liczby rekordów.

Przykład: tworzenie N-tyli na podstawie liczebności rekordów

W poniższej tabeli przedstawiono sposób rangowania uproszczonych wartości zmiennych jako kwartyle przy zastosowaniu metody tworzenia N-tyli na podstawie liczebności rekordów. Należy pamiętać, że wyniki będą różniły się w zależności od wybranej opcji N-tyli.

Tabela 26. Przykład tworzenia N-tyli na podstawie liczebności rekordów

Wartości	Dodaj do następnej	Pozostaw w bieżącej
10	1	1
13	2	1
15	3	2
15	3	2
20	4	3

Liczba pozycji w każdej kategorii jest obliczana w następujący sposób:

```
łączna liczba wartości/liczba N-tyli
```

W uproszczonym przykładzie przedstawionym powyżej żądana liczba pozycji w danej kategorii wynosi 1,25 (5 wartości/4 kwartyle). Wartość 13 (wartość numer 2) przekracza 1,25 żądanej wartości granicznej liczebności i dlatego jest traktowana w różny sposób, w zależności od wybranej opcji wiązania. W trybie **Dodaj do następnej** jest dodawana do kategorii 2. W trybie **Zachowaj w bieżącej** pozostaje w kategorii 1, przesuwając zakres wartości dla kategorii 4 poza istniejące wartości danych. W wyniku tego utworzone zostają tylko trzy kategorie, a wartości graniczne dla każdej kategorii zostają odpowiednio skorygowane, co przedstawiono w następującej tabeli.

Tabela 27. Wynik przykładowej kategoryzacji

Kategoria	Dolna	Górna
1	>=10	<15
2	>=15	<20
3	>=20	<=20

Uwaga: Szybkość kategoryzacji wg wiązań może zostać zwiększona poprzez aktywowanie przetwarzania równoległego.

Rangowanie obserwacji

Jeśli jako metoda kategoryzacji wybrana zostanie opcja **Rangi**, w oknie dialogowym wyświetlany jest nowy zestaw opcji.

W wyniku rangowania tworzone są nowe zmienne zawierające rangi, rangi ułamkowe i wartości percentyli dla zmiennych numerycznych, w zależności od opcji wybranych poniżej.

Porządek rang. Można wybrać opcję **Rosnąco** (najniższa wartość jest oznaczana jako 1) lub **Malejąco** (najwyższa wartość jest oznaczana jako 1).

Ranga. Tę opcję należy wybrać, aby rangowanie obserwacji odbywało się w porządku rosnącym lub malejącym w zależności od tego, jak określono powyżej. Przedział wartości w nowej zmiennej będzie wynosił 1–*N*, gdzie *N* oznacza liczbę wartości dyskretnych w oryginalnej zmiennej. Powiązane wartości uzyskują średnie wartości w danej randze.

Ranga ułamkowa. Tę opcję należy wybrać, aby przeprowadzić rangowanie obserwacji, w którym nowa zmienna jest równa randzie podzielonej przez sumę wag obserwacji bez braków danych. Rangi ułamkowe mieszczą się w przedziale od 0 do 1.

Ułamkowa ranga procentowa. Każda ranga jest podzielona przez liczbę rekordów o poprawnych wartościach i pomnożona przez 100. Ułamkowe rangi procentowe mieszczą się w przedziale od 1 do 100.

Rozszerzenie. Dla wszystkich opcji rang można utworzyć rozszerzenia niestandardowe i określić, czy rozszerzenie ma zostać dodane na początku (**Przedrostek**), czy na końcu (**Przyrostek**) nazwy zmiennej. Można na przykład wygenerować nową zmienną o nazwie *income_P_RANK*.

Średnia/Odchylenie standardowe

Jeśli jako metoda kategoryzacji wybrana zostanie opcja **Średnia/Odchylenie standardowe**, w oknie dialogowym wyświetlany jest nowy zestaw opcji.

Ta metoda powoduje wygenerowanie co najmniej jednej zmiennej z kategoriami podziału na podstawie wartości średniej i standardowego odchylenia dla rozkładu określonych zmiennych. Należy wybrać liczbę odchyleń, jaka będzie użyta (poniżej).

Rozszerzenie nazwy. Należy określić rozszerzenie, jakie będzie używane dla wygenerowanych zmiennych. *_SDBIN* to rozszerzenie domyślne. Można również określić, czy rozszerzenie ma być dodawane na początku (**Przedrostek**), czy na końcu (**Przyrostek**) nazwy zmiennej. Można na przykład wygenerować nową zmienną o nazwie *income_SDBIN*.

- **+/– 1 odchylenie standardowe.** Tę opcję należy wybrać, aby wygenerować trzy przedziały.
- **+/– 2 odchylenia standardowe.** Tę opcję należy wybrać, aby wygenerować pięć przedziałów.
- **+/– 3 odchylenia standardowe.** Tę opcję należy wybrać, aby wygenerować siedem przedziałów.

Przykładowo wybór opcji +/–1 odchylenie standardowe spowoduje utworzenie trzech przedziałów, zgodnie z obliczeniami w poniższej tabeli.

Tabela 28. Przykład kategoryzacji wg odchylenia standardowego

Przedział 1	Przedział 2	Przedział 3
x < (Średnia – Odch. std.)	(Średnia – Odch. std.) <= x <= (Średnia + Odch. std.)	x > (Średnia + Odch. std.)

Dla rozkładu normalnego ok. 68% obserwacji znajduje się w przedziale o granicach w punktach oddalonych o jedno odchylenie standardowe od średniej w obie strony, ok. 95% dla dwóch odchyleń standardowych i ok. 99% dla trzech odchyleń standardowych. Należy jednak pamiętać, że tworzenie kategorii podziału na podstawie odchyleń standardowych może spowodować, że niektóre zdefiniowane przedziały znajdą się poza rzeczywistym zakresem danych, a nawet poza zakresem możliwych wartości danych (np. ujemne wartości wynagrodzenia).

Kategoryzacja optymalna

Jeśli zmienna, jaka ma zostać skategoryzowana, jest ściśle powiązana z inną zmienną jakościową, można wybrać zmienną jakościową jako zmienną nadzorującą, aby tworzyć przedziały, zachowując siłę oryginalnego powiązania pomiędzy dwiema zmiennymi.

Załóżmy na przykład, że do pogrupowania regionów na podstawie wskaźników zaległych płatności z tytułu kredytów mieszkaniowych użyto analizy skupień, przy założeniu, że najwyższe wskaźniki przypadają dla pierwszego skupienia.

W takim przypadku można wybrać zmienne *Percent past due* (Procent po terminie) i *Percent of foreclosures* (Procent zajęć) jako zmienne poddane podziałowi oraz zmienną przynależności do skupień wygenerowaną przez model jako zmienną nadzorującą.

Rozszerzenie nazwy Należy określić rozszerzenie, jakie będzie używane dla wygenerowanych zmiennych, oraz ustalić czy ma ono być dodawane na początku (**Przyrostek**), czy na końcu (**Przyrostek**) nazwy zmiennej. Przykładowo można wygenerować nową zmienną o nazwie pastdue_OPTIMAL oraz inną zmienną o nazwie inforeclosure_OPTIMAL.

Zmienna nadzorcy Zmienna jakościowa służąca do budowania przedziałów.

Wstępnie podzielone zmienne w celu zwiększenia wydajności dla dużych zestawów danych Wskazuje, czy należy przeprowadzić wstępne przetwarzanie w celu uproszczenia kategoryzacji optymalnej. Spowoduje to, że wartości skali zostaną pogrupowane na wiele przedziałów z zastosowaniem prostej metody kategoryzacji nienadzorowanej, wartości w każdym każdym przedziale będą reprezentowane według średniej, a waga obserwacji zostanie odpowiednio skorygowana przed przetworzeniem z zastosowaniem kategoryzacji nadzorowanej. W praktyce ta metoda zapewnia kompromis pomiędzy poziomem dokładności a szybkością i jest zalecana w przypadku dużych zbiorów danych. Można również określić liczbę przedziałów, w jakich zmienna powinna się znaleźć po przetworzeniu za pomocą tej opcji.

Łącz kategorie o stosunkowo małej liczbie przypadków z dużym sąsiadem. Jeśli ta opcja zostanie aktywowana, będzie wskazywać, że przedział zostanie połączony, jeśli stosunek jego wielkości (liczba obserwacji) do wielkości sąsiadującego przedziału jest mniejszy niż określona wartość graniczna; należy pamiętać, że większe wartości graniczne mogą powodować więcej połączeń.

Ustawienia punktu odcięcia

Okno dialogowe Ustawienia punktu odcięcia umożliwia określenie zaawansowanych opcji dla algorytmu kategoryzacji optymalnej. Opcje te informują algorytm, w jaki sposób przedziały mają zostać obliczone z zastosowaniem zmiennej przewidywanej.

Punkty końcowe kategorii. Można określić, czy dolne lub górne punkty końcowe powinny być uwzględniane (mniejsze <= *x*), czy wykluczane (mniejsze < *x*).

Pierwsza i ostatnia kategoria. W obu przypadkach dla pierwszej i ostatniej kategorii można określić, czy przedziały powinny być nieograniczone (do plus lub minus nieskończoności), czy też ograniczone przez najniższe lub najwyższe punkty danych.

Podgląd wygenerowanych przedziałów

Karta Wartości podziałów w węźle Kategoryzacja umożliwia wyświetlanie wartości granicznych dla wygenerowanych przedziałów. Korzystając z menu Utwórz, można również wygenerować węzeł Wyliczanie, za pomocą którego można zastosować te wartości graniczne z jednego zbioru danych w innym.

Zmienna kategoryzowana. Korzystając z listy rozwijanej, można wybrać zmienną do wyświetlenia. W celu poprawy czytelności nazwy zmiennych zostały użyte jako nazwy oryginalnych zmiennych.

N-tyl. Korzystając z listy rozwijanej, można wybrać N-tyl, np. 10 lub 100, do wyświetlenia. Ta opcja jest dostępna tylko w przypadku wygenerowania przedziałów przy użyciu metody N-tyle (równa liczebność lub suma).

Wartości graniczne kategorii. Wartości graniczne są tutaj pokazane dla każdego wygenerowanego przedziału, wraz z liczbą rekordów znajdujących się w każdym przedziale. Tylko w przypadku metody Kategoryzacja optymalna: liczba rekordów w każdym przedziale jest wyświetlana jako procent całości. Należy pamiętać, że wartości graniczne nie mają zastosowania w przypadku metody kategoryzacji z rangowaniem.

Odczytaj wartości. Odczytuje podzielone wartości ze zbioru danych. Należy pamiętać, że wartości graniczne będą również zastąpione, jeśli w strumieniu zostaną uruchomione nowe dane.

Generowanie węzła Wyliczanie

Korzystając z menu Utwórz, można utworzyć węzeł Wyliczanie w oparciu o bieżące wartości graniczne. Jest to przydatne do zastosowania ustalonych wartości granicznych z jednego zbioru danych w innym. Ponadto, jeśli te punkty podziału są znane, podczas pracy z dużymi zbiorami danych operacja Wyliczanie jest znacznie bardziej wydajna (czyli szybsza) niż operacja Kategoryzacja.

Węzeł Analiza RFM

Węzeł analizy RFM (Recency — Aktualność, Frequency — Częstość, Monetary — Kwota) umożliwia określenie ilościowo, którzy klienci najprawdopodobniej będą najlepszymi, poprzez dokonanie oceny, kiedy ostatnio dokonali zakupu (aktualność), jak często dokonują zakupu (częstość) i jak dużo wydali na wszystkie transakcje (kwota).

Wnioskowanie, na jakim opiera się analiza RFM, jest następujące: klienci, którzy raz dokonali zakupu produktu lub usługi, z dużo większym prawdopodobieństwem dokonają zakupu ponownie. Podzielone na kategorie dane klientów są dzielone na wiele przedziałów, zgodnie z kryteriami kategoryzacji ustawionymi przez użytkownika. W każdym przedziale do klientów jest przypisywana ocena; oceny te zostają następnie połączone, tworząc ogólną ocenę RFM. Ocena ta stanowi odzwierciedlenie przynależności klienta do przedziałów utworzonych dla poszczególnych parametrów RFM. Tak podzielone dane mogą być wystarczające dla potrzeb użytkownika, np. poprzez zidentyfikowanie najczęściej występujących, najbardziej wartościowych; alternatywnie dane mogą zostać wprowadzone do strumienia w celu przeprowadzenia dalszego modelowania i dokładniejszej analizy.

Należy jednak pamiętać, że chociaż możliwość analizowania i rangowania ocen RFM jest przydatnym narzędziem, podczas korzystania z niego należy mieć świadomość wpływu niektórych czynników. Kuszące może być przypisywanie klientom najwyższych ocen; jednak nadmierne nakłanianie tych klientów może ich urazić i w rzeczywistości spowodować utratę szansy na dokonanie kolejnych zakupów. Warto również pamiętać, że klientów z niższymi ocenami nie należy lekceważyć, ale warto o nich zabiegać, aby stali się lepszymi klientami. I odwrotnie, same wysokie oceny niekonieczne będą oznaczały szansę na dobrą sprzedaż — będzie to zależało od rynku. Przykładowo, klient z przedziału 5 dla aktualności (co oznacza, że dokonał zakupu bardzo niedawno), może w rzeczywistości nie być dobrym klientem docelowym dla kogoś, kto sprzedaje drogie produkty o długim cyklu życia, takie jak samochody czy telewizory.

Uwaga: W zależności od sposobu składowania danych konieczne może być poprzedzenie węzła Analiza RFM węzłem Agregacja RFM w celu przekształcenia danych na format nadający się do użycia. Przykładowo dane wejściowe muszą być przedstawione w formacie klienta (jeden wiersz na klienta); jeśli dane klienta będą zapisane w formacie transakcyjnym, należy wcześniej użyć węzła Agregacja RFM w celu wyznaczenia zmiennych aktualności, częstości i kwoty. Więcej informacji można znaleźć w temacie “Węzeł Agregacja RFM” na stronie 82.

Węzły Agregacja RFM i Analiza RFM w programie IBM SPSS Modeler są skonfigurowane, tak aby korzystały z niezależnej kategoryzacji; oznacza to, że rangowanie i kategoryzacja danych są przeprowadzane dla każdej miary wartości aktualności, częstości i kwoty, bez względu na ich wartości lub pozostałe dwie miary.

Ustawienia węzła Analiza RFM

Aktualność. Korzystając z selektora zmiennych (przycisk po prawej stronie pola tekstowego), należy wybrać zmienną aktualności. Może to być data, znacznik czasu lub zwykła liczba. Należy pamiętać, że jeśli data lub znacznik czasu reprezentują datę najnowszej transakcji, najwyższa wartość jest traktowana jako najnowsza; jeśli określona jest liczba, reprezentuje ona czas, jaki upłynął od najnowszej transakcji, i najniższa wartość będzie oznaczała najnowszą.

Uwaga: Jeśli węzeł Analiza RFM jest poprzedzony w strumieniu węzłem Agregacja RFM, zmienne aktualności, częstości i kwoty generowane przez węzeł Agregacja RFM powinny stanowić dane wejściowe w węźle Analiza RFM.

Częstość. Korzystając z selektora zmiennych, należy wybrać zmienną częstości, jaka będzie używana.

Kwota. Korzystając z selektora zmiennych, należy wybrać zmienną kwoty, jaka będzie używana.

Liczba przedziałów. Dla każdego z trzech typów wyników należy wybrać liczbę przedziałów, jaka ma zostać utworzona. Domyślną wartością jest 5.

Uwaga: Minimalna liczba przedziałów to 2, a maksymalna to 9.

Waga. Domyślnie podczas obliczania ocen największa ważność jest przypisywana danym dotyczącym aktualności, a następnie częstości i kwocie. W razie konieczności można zmodyfikować ważenie wpływające na te dane, aby zmienić im przypisanie najwyższej ważności.

Ocena RFM jest obliczana w następujący sposób: (ocena aktualności x waga aktualności) + (ocena częstości x waga częstości) + (ocena kwoty x waga kwoty).

Wiązania. Należy określić, jak identyczne (jak bardzo powiązane) muszą być oceny, aby zostały podzielone. Opcje są następujące:

- **Dodaj do następnej.** Tę opcję należy wybrać, aby przenieść wartości wiązania w górę do następnej kategorii.
- **Pozostaw w bieżącej.** Zachowuje wartości w bieżącej (niższej) kategorii. Zastosowanie tej metody może skutkować utworzeniem mniejszej liczby przedziałów. (Jest to wartość domyślna).

Wartości graniczne kategorii. Należy określić, czy oceny RFM i alokacje kategorii będą zawsze przeliczane po wykonaniu węzła czy też będą przeliczane tylko w razie konieczności (np. po podaniu nowych danych). Jeśli wybrana zostanie opcja **Odczytaj z zakładki Wartości podziałów, jeśli są tam zdefiniowane**, można edytować górne i dolne punkty podziału dla różnych przedziałów, korzystają z zakładki Wartości podziałów.

Po wykonaniu węzeł Analiza RFM tworzy przedziały surowych zmiennych aktualności, częstości i kwoty i dodaje następujące nowe zmienne do zbioru danych:

- Ocena aktualności. Ranga (wartość podziału) dla aktualności
- Ocena częstości. Ranga (wartość podziału) dla częstości
- Ocena kwoty. Ranga (wartość podziału) dla kwoty
- Ocena RFM. Suma ważona ocen aktualności, częstości i kwoty.

Dodaj wartości odstające do skrajnych kategorii. Po zaznaczeniu tego pola wyboru rekordy znajdujące się poniżej dolnego przedziału będą dodawane do dolnego przedziału, a rekordy znajdujące się powyżej najwyższego przedziału będą dodawane do najwyższego — w przeciwnym razie zostanie im przypisana wartość null. To pole jest dostępne tylko po zaznaczeniu opcji **Odczytaj z zakładki Wartości podziałów, jeśli są tam zdefiniowane**.

Kategoryzacja węzła Analiza RFM

Karta Wartości podziałów umożliwia wyświetlanie i w niektórych przypadkach zmianę wartości granicznych dla wygenerowanych przedziałów.

Uwaga: Wartości na tej karcie można zmienić tylko po zaznaczeniu opcji **Odczytaj z zakładki Wartości podziałów, jeśli są tam zdefiniowane** na karcie Ustawienia.

Zmienna kategoryzowana. Korzystając z listy rozwijanej, można wybrać zmienną do podzielenia na przedziały. Dostępne są wartości, które zostały wybrane na karcie Ustawienia.

Tabela wartości przedziału. Wyświetlane są tutaj wartości graniczne dla każdego wygenerowanego przedziału. Po wybraniu opcji **Odczytaj z zakładki Wartości podziałów, jeśli są tam zdefiniowane** na karcie Ustawienia można zmienić górne i dolne punkty podziału dla każdego przedziału, klikając dwukrotnie odpowiednią komórkę.

Odczytaj wartości. Odczytuje podzielone wartości ze zbioru danych i wypełnia tabelę wartości podziałów. Należy pamiętać, że w przypadku wybrania opcji **Zawsze przeliczaj** na karcie Ustawienia po uruchomieniu nowych danych w strumieniu wartości graniczne kategorii zostaną zastąpione.

Węzeł Zespół

Węzeł Zespół łączy co najmniej dwa modele użytkowe w celu uzyskania bardziej dokładnych predykcji, jakie można uzyskać z dowolnego pojedynczego modelu. Połączenie predykcji z wielu modeli umożliwia obejście ograniczeń w poszczególnych modelach, co często powoduje wyższą ogólną dokładność. Modele połączone w ten sposób zwykle działają tak samo dobrze, jak najlepsze z pojedynczych modeli, a często nawet lepiej.

Połączenie modeli odbywa się automatycznie w zautomatyzowanych węzłach modelowania Auto Klasyfikacja, Auto Predykcja i Auto Grupowanie.

Po użyciu węzła zespolenia można użyć węzła analizy lub ewaluacji w celu porównania dokładności połączonych wyników z każdym z modeli wejściowych. W tym celu należy upewnić się, czy na karcie Ustawienia węzła zespolenia nie została zaznaczona opcja **Odfiltruj zmienne utworzone przez modele zespolone**.

Zmienne wyjściowe

Każdy węzeł zespolenia powoduje wygenerowanie zmiennej zawierającej połączone oceny. Nazwa tworzona jest na podstawie określonej zmiennej przewidywanej, do której dodawany jest przedrostek *$XF_*, *$XS_* lub *$XR_*, w zależności od poziom pomiaru zmiennej — odpowiednio flaga, nominalna (zbiór) lub ilościowa (przedział). Przykładowo, jeśli zmienną przewidywaną jest zmienna typu flaga o nazwie *response* (odpowiedź), zmienna wyjściowa będzie miała nazwę *$XF_response*.

Zmienne ufności lub skłonności. Dla zmiennych flagi i nominalnych na podstawie metody zespolenia tworzone są dodatkowe zmienne ufności lub skłonności; szczegóły przedstawiono w poniższej tabeli.

Tabela 29. Tworzenie zmiennej metodą zespolenia

Metoda zespolenia	Nazwa zmiennej
Głosowanie Głosowanie ważone ufnością Głosowanie ważone surową skłonnością Głosowanie ważone skorygowaną skłonnością Najwyższa ufność zachowuje nadrzędność	*$XFC_<zmienna>*
Średnia surowa skłonność	*$XFRP_<zmienna>*
Średnia skorygowana surowa skłonność	*$XFAP_<zmienna>*

Ustawienia węzła Zespolenie

Zmienna przewidywana dla zespołu. Należy wybrać pojedynczą zmienną, która będzie używana jako przewidywana przez co najmniej dwa modele poprzedzające. Modele poprzedzające mogą używać zmiennych przewidywanych typu flaga, nominalne lub ilościowe, ale co najmniej dwa modele muszą współużytkować tą samą zmienną przewidywaną, aby połączenie ocen było możliwe.

Odfiltruj zmienne utworzone przez modele zespolone Usuwa z danych wynikowych wszystkie dodatkowe zmienne wygenerowane przez poszczególne modele zasilające węzeł zespolenia. To pole wyboru należy zaznaczyć w przypadku zainteresowania tylko oceną zespoloną wszystkich modeli wejściowych. Upewnij się, że ta opcja nie jest zaznaczona, jeśli na przykład chcesz użyć trybu Analiza lub Ewaluacja do porównania dokładności oceny zespolonej z oceną poszczególnych modeli wejściowych.

Dostępne ustawienia zależą od poziomu pomiaru zmiennej wybranej jako przewidywana.

Przewidywane zmienne ilościowe

W przypadku ilościowej zmiennej przewidywanej oceny będą uśrednione. Jest to jedyna dostępna metoda pozwalająca na łączenie ocen.

Podczas uśredniania ocen i oszacowań węzeł zespolenia korzysta z obliczeń błędu standardowego w celu wyznaczenia różnicy między wartościami zmierzonymi lub estymowanymi a rzeczywistymi, a także w celu zaprezentowania stopnia dopasowania tych estymacji. Obliczenia błędu standardowego są generowane domyślnie dla nowych modeli; można jednak usunąć zaznaczenie pola wyboru dla istniejących modeli, na przykład jeśli mają zostać ponownie wygenerowane.

Przewidywane zmienne jakościowe

W przypadku przewidywanych zmiennych jakościowych dostępnych jest kilka metod, w tym **głosowanie**, które polega na obliczeniu, ile razy każda możliwa wartość przewidywana została wybrana, i wybraniu wartości z najwyższym wynikiem. Przykładowo, jeśli trzy z pięciu modeli przewidują odpowiedź *tak*, a pozostałe dwa przewidują *nie*, wówczas odpowiedź *tak* wygrywa w głosowaniu 3 do 2. Alternatywnie głosowania mogą być **ważone** na podstawie wartości ufności lub skłonności dla każdej predykcji. Wagi są następnie sumowane i ponownie wybierana jest wartość z najwyższą sumą. Ufność dla końcowej predykcji jest sumą wag dla wartości wygranej podzieloną przez liczbę modeli uwzględnionych w zespoleniu.

Wszystkie zmienne jakościowe. W przypadku zmiennych typu flaga i nominalnych obsługiwane są następujące metody:

- Głosowanie
- Głosowanie ważone ufnością
- Najwyższa ufność zachowuje nadrzędność

Tylko zmienne flagi. Dla zmiennych flagi dostępnych jest również kilka metod opartych na skłonności:

- Głosowanie ważone surową skłonnością
- Głosowanie ważone skorygowaną skłonnością
- Średnia surowa skłonność
- Średnia skorygowana skłonność

Wiązania głosowania. W przypadku metod głosowania można określić sposób przetwarzania wiązań.

- **Wybór losowy.** Jedna z wartości wiązania jest wybierana w sposób losowy.
- **Najwyższa ufność.** Wartość wiązania, dla której przewidziano najwyższą ufność, wygrywa. Należy zauważyć, że nie musi to być taka sama wartość jak najwyższa ufność dla wszystkich wartości przewidywanych.
- **Surowa lub skorygowana skłonność (tylko zmienne z flagami).** Wartość wiązania, dla której przewidziano najwyższą skłonność bezwzględną, gdzie skłonność bezwzględna jest obliczana w następujący sposób:

```
abs(0,5 - skłonność) *
2
```

Lub w przypadku skłonności skorygowanej:

```
abs(0,5 - skłonność skorygowana) * 2
```

Węzeł Partycja

Węzły podziału na podzbiory służą do generowania zmiennej dzielącej na podzbiory, która dzieli dane na osobne podzbiory lub próby wykorzystywane podczas uczenia, testowania i walidacji w procesie budowania modelu. Korzystając z jednej próby do generowania modelu oraz innej do testowania go, można uzyskać wiarygodne informacje o tym, jak dobrze model pozwala uogólnić większe zbiory danych, podobne do danych bieżących.

Węzeł Partycja generuje zmienną nominalną z rolą ustawioną na **Partycja**. Alternatywnie, jeśli w danych istnieje już odpowiednia zmienna, można ją wyznaczyć jako partycję, korzystając z węzła Typy. W takim przypadku nie jest wymagany osobny węzeł podziału na podzbiory. Określona zmienna nominalna z dwoma lub trzema wartościami może być używana jako podzbiór; zmienne flagi nie mogą być używane. Więcej informacji można znaleźć w temacie “Ustawianie roli zmiennej” na stronie 149.

W strumieniu można zdefiniować wiele zmiennych dzielących na podzbiory, ale w takim przypadku pojedyncza zmienna dzieląca na podzbiory musi być wybrana na karcie Zmienne w każdym węźle modelowania, który korzysta z dzielenia na podzbiory. (Jeśli obecna jest tylko jedna zmienna dzieląca na podzbiory, jest ona automatycznie używana po aktywowaniu dzielenia).

Włączanie dzielenia na podzbiory. Aby korzystać z dzielenia na podzbiory w analizie, należy aktywować dzielenie na podzbiory na karcie Opcje modelu w odpowiednim węźle budowania modelu lub analizy. Usunięcie zaznaczenia tej opcji umożliwia wyłączenie podziału bez usuwania zmiennej.

Aby utworzyć zmienną dzielącą na podzbiory na podstawie innego kryterium, takiego jak zakres dat lub lokalizacja, można także użyć węzła Wyliczanie. Więcej informacji można znaleźć w temacie “węzeł wyliczeń” na stronie 154.

Przykład. Podczas budowania strumienia RFM w celu zidentyfikowania ostatnich klientów, którzy pozytywnie odpowiedzieli na poprzednie kampanie marketingowe, dział marketingu firmy zajmującej się sprzedażą używa węzła podziału na podzbiory do podzielenia danych na podzbiory uczące i testowe.

Opcje węzła podziału na podzbiory

Zmienna dzieląca na podzbiory. Określa nazwę zmiennej utworzonej przez węzeł.

Podzbiory. Można podzielić dane na dwie próby (uczenia i testowania) lub na trzy próby (uczenia, testowania i walidacji).

- **Uczenie i testowanie.** Podzielenie danych na dwie próby umożliwia uczenie modelu za pomocą jednej próby oraz testowanie za pomocą drugiej.
- **Uczenie, testowanie i walidacja.** Podział danych na trzy próby umożliwia uczenie modelu za pomocą jednej próby, testowanie i udoskonalanie za pomocą drugiej próby oraz walidację wyników za pomocą trzeciej. Pozwala to odpowiednio zredukować wielkość każdego podzbioru i może być najlepszym rozwiązaniem podczas pracy z bardzo dużymi zbiorami danych.

Wielkość podzbioru. Określa względną wielkość każdego podzbioru. Jeśli suma wielkości podzbiorów jest mniejsza niż 100%, wówczas rekordy nieuwzględnione w podzbiorze powinny zostać odrzucone. Przykładowo, jeśli użytkownik ma 10 milionów rekordów i określił wielkości podzbiorów jako 5% dla uczenia i 10% dla testowania, po uruchomieniu węzła powinno być w przybliżeniu 500 000 rekordów uczących i milion testujących; pozostałe powinny zostać odrzucone.

Wartości. Określa wartości użyte do reprezentowania każdej próby podziału w danych.

- **Użyj wartości zdefiniowanych przez system ("1", "2" i "3").** Każdy podzbiór jest reprezentowany jako liczba całkowita; na przykład, wszystkie rekordy należące do próby uczącej mają wartość 1 dla zmiennej dzielącej na podzbiory. Dzięki temu dane będzie można przenosić przy różnych ustawieniach regionalnych, a jeśli zmienna dzieląca na podzbiory jest ponownie określona w innym miejscu (na przykład przy ponownym odczytaniu danych z bazy danych), kolejność sortowania zostaje zachowana (czyli wartość 1 nadal będzie reprezentować podzbiór uczący). Wartości wymagają jednak pewnej interpretacji.
- **Dołącz etykiety do wartości zdefiniowanych przez system.** Łączy liczbę całkowitą z etykietą; przykładowo rekordy podzbiorów uczących mają wartość 1_Training (1_Uczenie). Dzięki temu osoba przeglądająca dane może określić, która wartość jest która, a kolejność sortowania zostaje zachowana. Wartości są jednak specyficzne dla określonych ustawień regionalnych.
- **Użyj etykiet jako wartości.** Używa etykiety bez liczby całkowitej; na przykład **Training** (Uczenie). Dzięki temu można określić wartości poprzez edytowanie etykiet. Jednak dane stają się wówczas specyficzne dla ustawień regionalnych i ponowne określenie kolumny podzbioru spowoduje ustawienie wartości w ich pierwotnym porządku sortowania, który może nie odpowiadać ich kolejności „semantycznej”.

Wartość początkowa. Opcja jest dostępna tylko po zaznaczeniu pola **Ustaw wartość początkową generatora liczb losowych**. W przypadku próbkowania lub dzielenia na podzbiory rekordów w oparciu o losową wartość procentową opcja ta pozwala na zduplikowanie tych samych wyników w innej sesji. Określenie wartości początkowej używanej przez generator liczb pseudolosowych zapewni, że podczas każdego wykonywania węzła przypisywane będą te same

rekordy. Wprowadź żądaną wartość początkową generatora lub kliknij przycisk **Utwórz**, aby automatycznie wygenerować wartość losową. Jeśli nie wybrano tej opcji, przy każdej próbie wykonania węzła wygenerowana zostanie inna próba.

Uwaga: Jeśli używana jest opcja **Wartość początkowa** w przypadku rekordów odczytanych z bazy danych, przed przeprowadzeniem próby konieczne może być sortowanie węzła, aby po każdym wykonaniu węzła uzyskany wynik był taki sam. Wynika to z faktu, że wartość początkowa generatora liczb losowych zależy od kolejności rekordów, która w relacyjnej bazie danych nie musi pozostawać jednakowa. Więcej informacji można znaleźć w temacie “Węzeł Sortowanie” na stronie 84.

Użyj zmiennych unikalnych, aby przypisać partycje. Opcja jest dostępna tylko po zaznaczeniu pola **Ustaw wartość początkową generatora liczb losowych**. (Tylko w przypadku baz danych Warstwa 1) Należy zaznaczyć to pole, aby do przypisania rekordów do podzbiorów używać funkcji przekazywania do bazy danych SQL. Z listy rozwijanej należy wybrać zmienną o unikalnych wartościach (np. zmienna identyfikacyjna), aby upewnić się, że rekordy zostaną przypisane w kolejności losowej, ale w sposób powtarzalny.

Warstwy bazy danych zostały objaśnione w opisie węzła źródłowego bazy danych. Więcej informacji można znaleźć w temacie “Węzeł źródłowy bazy danych” na stronie 17.

Generowanie węzłów selekcji

Korzystając z menu Utwórz w węźle podziału na podzbiory można automatycznie wygenerować węzeł selekcji dla każdego podzbioru. Na przykład można wybrać wszystkie rekordy w podzbiorze uczącym, aby przeprowadzić dalszą ocenę lub analizę korzystając wyłącznie z tego podzbioru.

Węzeł Flagowanie

Węzeł Flagowanie jest używany do wyliczania zmiennych flag na podstawie zmiennych wartości jakościowych zdefiniowanych dla co najmniej jednej zmiennej nominalnej. Przykładowo, zbiór danych może zawierać zmienną nominalną, *BP* (ciśnienie krwi), o wartościach *High* (Wysokie), *Normal* (W normie) i *Low* (Niskie). Aby ułatwić manipulowanie danymi, można utworzyć zmienną flagi dla wysokiego ciśnienia krwi, która wskaże, czy pacjent ma wysokie ciśnienie krwi czy nie.

Ustawianie opcji dla węzła Flagowanie

Zmienne nominalne. Wyświetla listę wszystkich zmiennych danych, których poziom pomiaru ustawiono jako *Nominalne* (zbiór). Należy wybrać jedną zmienną z listy, aby wyświetlić wartości w zbiorze. Wartości te można wybrać, aby utworzyć zmienną flagi. Należy pamiętać, że dane muszą być dokładnie określone za pomocą poprzedzającego węzła źródłowego lub węzła Typy, aby dostępne zmienne nominalne mogły zostać wyświetlone (wraz z ich wartościami). Więcej informacji można znaleźć w temacie “Węzeł Typy” na stronie 137.

Rozszerzenie nazwy zmiennej. Należy wybrać tę opcję, aby włączyć elementy sterujące umożliwiające określenie rozszerzenia, jakie zostanie dodane do nowej zmiennej flagi w postaci przedrostka lub przyrostka. Domyślnie nazwy nowych zmiennych są tworzone automatycznie poprzez połączenie nazwy oryginalnej zmiennej z nazwą wartości zmiennej w postaci etykiety, np. *Nazwazmiennej_wartośćzmiennej*.

Wartości zmiennych jakościowych. Tutaj wyświetlane są wartości wybranego wcześniej zbioru. Należy wybrać wartości, dla których mają zostać wygenerowane flagi. Na przykład, jeśli wartości zmiennej o nazwie *blood_pressure* to *High*, *Medium* i *Low*, można wybrać wartość *High* i dodać ją do listy po prawej stronie. Spowoduje to utworzenie zmiennej z flagą dla rekordów zawierających wartości wskazujące na wysokie ciśnienie krwi.

Tworzone zmienne typu flaga. Tutaj wyświetlana jest lista nowo utworzonych zmiennych typu flaga. Można określić opcje tworzenia nazw nowych zmiennych przy użyciu elementów sterujących rozszerzeniami nazwy.

Wartość prawdziwa. Należy określić wartość prawdziwą, jaka będzie używana przez węzeł podczas ustawiania flagi. Domyślnie ta wartość to **T** (Prawda).

Wartość fałszywa. Należy określić wartość fałszywą, jaka będzie używana przez węzeł podczas ustawiania flagi. Domyślnie ta wartość to **F** (Fałsz).

Klucze agregacji. Tę opcję należy wybrać, aby pogrupować rekordy na podstawie zmiennych kluczowych określonych poniżej. Po wybraniu opcji **Klucze agregacji** wszystkie zmienne typu flaga w grupie zostaną „włączone", o ile dla *dowolnego* rekordu ustawiona została wartość „prawda". Selektor zmiennych umożliwia określenie, które zmienne kluczowe zostaną użyte podczas agregacji rekordów.

Węzeł Restrukturyzacja

Węzła Restrukturyzacja można użyć do wygenerowania wielu zmiennych na podstawie wartości zmiennej nominalnej lub typu flaga. Nowo wygenerowane zmienne mogą zawierać wartości z innej flagi lub flag numerycznych (0 i 1). Działanie tego węzła jest podobne do węzła Flagowanie. Zapewnia jednak większą elastyczność. Umożliwia utworzenie zmiennych dowolnego typu (w tym flagi numeryczne), z zastosowaniem wartości z innej zmiennej. Następnie można przeprowadzić agregację lub inne manipulacje, używając innych węzłów poniżej. (Węzeł Flagowanie umożliwia przeprowadzenie agregacji zmiennych w jednym kroku, co może być wygodne w przypadku tworzenia zmiennych typu flaga).

Przykładowo, następujący zbiór danych zawiera zmienną nominalną, *Account* (Konto), której wartości to *Savings* (Oszczędnościowe) i *Draft* (Czekowe). Dla każdego konta rejestrowane jest saldo początkowe i bieżące, a niektórzy klienci mają kilka kont każdego typu. Załóżmy, że użytkownik chce wiedzieć, czy poszczególni klienci mają konto określonego typu, a jeśli tak, ile pieniędzy jest na każdym z kont. Do wygenerowania zmiennej dla każdej wartości *Account* używany jest węzeł Restrukturyzacja, a jako wartość wybierana jest opcja *Current_Balance* (Bieżące saldo). Każda zmienna jest wypełniana przez wartość bieżącego salda dla danego rekordu.

Tabela 30. Przykładowe dane przed restrukturyzacją

CustID	Account	Open_Bal	Current_Bal
12701	Draft	1000	1005,32
12702	Savings	100	144,51
12703	Savings	300	321,20
12703	Savings	150	204,51
12703	Draft	1200	586,32

Tabela 31. Przykładowe dane po restrukturyzacji

CustID	Account	Open_Bal	Current_Bal	Account_Draft_ Current_Bal	Account_Savings _Current_Bal
12701	Draft	1000	1005,32	1005,32	$null$
12702	Savings	100	144,51	$null$	144,51
12703	Savings	300	321,20	$null$	321,20
12703	Savings	150	204,51	$null$	204,51
12703	Draft	1200	586,32	586,32	$null$

Użycie węzła Restrukturyzacja z węzłem Agregacja

W wielu przypadkach uzasadnione może być użycie pary węzłów: Restrukturyzacja i Agregacja. W poprzednim przykładzie jeden klient (z id. 12703) ma trzy konta. Korzystając z węzła agregacji można obliczyć łączne saldo dla każdego typu konta. Zmienna kluczowa to *CustID* (Id. klienta), a zmienne agregacji to nowe zmienne restrukturyzowane, *Account_Draft_Current_Bal* (Bieżące saldo na koncie czekowym) oraz *Account_Savings_Current_Bal* (Bieżące saldo na koncie oszczędnościowym). W poniżej tabeli przedstawiono wyniki.

Tabela 32. Przykładowe dane po restrukturyzacji i agregacji

CustID	Liczba_rekordów	Account_Draft_Current_ Bal_Sum	Account_Savings_Current_ Bal_Sum
12701	1	1005,32	$null$
12702	1	$null$	144,51
12703	3	586,32	525,71

Ustawianie opcji dla węzła Restrukturyzacja

Dostępne zmienne. Wyświetla listę wszystkich zmiennych danych, których poziom pomiaru ustawiono jako *Nominalne* (zbiór) lub *Flaga*. Należy wybrać jedną zmienną z listy, aby wyświetlić wartości nominalne lub typu flaga, a następnie wybrać z nich wartości do utworzenia zmiennych restrukturyzowanych. Należy pamiętać, że dane muszą być dokładnie określone za pomocą poprzedzającego węzła źródłowego lub węzła Typy, aby dostępne zmienne mogły zostać wyświetlone (wraz z ich wartościami). Więcej informacji można znaleźć w temacie “Węzeł Typy” na stronie 137.

Dostępne wartości. Tutaj wyświetlane są wartości wybranego wcześniej zbioru. Należy wybrać wartości, dla których mają zostać wygenerowane zmienne restrukturyzowane. Na przykład, jeśli wartości zmiennej o nazwie *Blood Pressure* to *High*, *Medium* i *Low*, można wybrać wartość *High* i dodać ją do listy po prawej stronie. Spowoduje to utworzenie zmiennej o określonej wartości (patrz niżej) dla rekordów z wartością *High*.

Utwórz restrukturyzowane zmienne. Tutaj wyświetlana jest lista nowo utworzonych zmiennych restrukturyzowanych. Domyślnie nazwy nowych zmiennych są tworzone automatycznie poprzez połączenie nazwy oryginalnej zmiennej z nazwą wartości zmiennej w postaci etykiety, np. *Nazwazmiennej_wartośćzmiennej*.

Uwzględnij nazwy zmiennych. Należy usunąć zaznaczenie tej opcji, aby usunąć oryginalną nazwę zmienne stanowiącą przedrostek z nazw nowych zmiennych.

Użyj wartości z innych zmiennych. Należy określić co najmniej jedną zmienną, których wartości będą wstawiane w zmiennych restrukturyzowanych. Aby wybrać zmienne, należy użyć selektora zmiennych. Dla każdej wybranej zmiennej tworzona jest jedna nowa zmienna. Nazwa zmiennej wartości jest dołączana do nazwy zmiennej restrukturyzowanej — na przykład, *BP_High_Age* (BP_wysoki_wiek) lub *BP_Low_Age* (BP_niski_wiek). Każda nowa zmienna dziedziczy typ od oryginalnej zmiennej wartości.

Utwórz numeryczne flagi wartości. Tę opcję należy wybrać, aby nie używać wartości z innej zmiennej, tylko wypełnić nowe zmienne numerycznymi flagami wartości (0 jako fałsz i 1 jako prawda).

Węzeł Transpozycja

Domyślnie kolumny są zmiennymi, a wiersze są rekordami lub obserwacjami. W razie potrzeby można użyć węzła transpozycji, aby zamienić dane w wierszach i kolumnach, tak aby zmienne stały się rekordami, a rekordy zmiennymi. Przykładowo, jeśli dostępne są dane szeregu czasowego, w których każdy szereg jest wierszem, a nie kolumną, można transponować dane przed rozpoczęciem analizy.

Ustawianie opcji dla węzła Transpozycja

Z menu rozwijanego **Metoda transpozycji** wybierz metodę wykonania dla węzła transpozycji: **Zmienne i rekordy**, **Z rekordów na zmienne** lub **Ze zmiennych na rekordy**. Ustawienia dotyczące wszystkich trzech metod są opisane w poniższych sekcjach.

Ograniczenie: Metody **Z rekordów na zmienne** i **Ze zmiennych na rekordy** są obsługiwane tylko przez systemy Windows x64, Linux x64 oraz komputery Mac.

Zmienne i rekordy

Nowe nazwy zmiennych mogą być generowane automatycznie na podstawie określonego prefiksu lub odczytane z istniejących zmiennych w danych.

Użyj przedrostka. Ta opcja generuje nowe nazwy zmiennych automatycznie na podstawie określonego przedrostka (Field1 (Zmienna1), Field2 (Zmienna2) itd.). Przedrostek można dostosować odpowiednio do potrzeb. Korzystając z tej opcji, należy określić liczbę zmiennych, jaka ma zostać utworzona, niezależnie od liczby wierszy w oryginalnych danych. Przykładowo, jeśli **Liczba nowych zmiennych** jest ustawiona na 100, wszystkie dane poza pierwszymi 100 wierszami zostaną odrzucone. Jeśli w oryginalnych danych jest mniej niż 100 wierszy, niektóre zmienne będą miały wartość null. (W razie potrzeby można zwiększyć liczbę zmiennych, ale zadaniem tego ustawienia jest uniknięcie transponowania milionów rekordów na miliony zmiennych, co mogłoby doprowadzić do powstania nadmiernie skomplikowanego wyniku).

Załóżmy na przykład, że dostępne są dane z szeregami w wierszach i osobną zmienną (kolumna) dla każdego miesiąca. Można je transponować, tak aby każdy szereg znajdował się w osobnej zmiennej, a wiersze stanowiły poszczególne miesiące.

Odczytaj nazwy ze zmiennej. Odczytuje nazwy zmiennych z istniejących zmiennych. W przypadku użycia tej opcji liczba nowych zmiennych jest określana przez dane, do wyznaczonego maksimum. Każda wartość wybranej zmiennej staje się nową zmienną w danych wynikowych. Wybranej zmiennej może być przypisany dowolny typ składowania (liczba całkowita, łańcuch, data itd.), ale w celu uniknięcia duplikowania nazw zmiennych każda wartość wybranej zmiennej musi być unikalna (innymi słowy, liczba wartości powinna być zgodna z liczbą wierszy). W razie wykrycia zduplikowanych nazw wyświetlane jest ostrzeżenie.

- **Odczytaj wartości.** Jeśli wybrana zmienna nie została jeszcze określona, należy wybrać tę opcję, aby wypełnić listę nazw nowych zmiennych. Jeśli zmienna została już określona, wówczas ten krok można pominąć.
- **Maksymalna liczba czytanych wartości.** W przypadku odczytu nazw zmiennych z danych określany jest górny limit, aby uniknąć utworzenia zbyt dużej liczby zmiennych. (Jak wspomniano wcześniej, transponowanie milionów rekordów na miliony zmiennych spowoduje utworzenie nadmiernie skomplikowanego wyniku).

Przykładowo, jeśli pierwsza kolumna danych określa nazwę dla każdego szeregu, można użyć tych wartości jako nazwy zmiennych w danych transponowanych.

Transponuj. Domyślnie transponowane są tylko zmienne ilościowe (zakres liczbowy) (typ składowania: liczba całkowita lub rzeczywista). Opcjonalnie można wybrać podzbiór zmiennych numerycznych lub transponować zmienne łańcuchowe. Wszystkie transponowane zmienne muszą mieć jednak taki sam typ składowania — numeryczny lub łańcuchowy — mieszanie zmiennych wejściowych spowoduje wygenerowanie pomieszanych wartości w każdej kolumnie wyników, co spowoduje naruszenie reguły, że typ składowania wszystkich wartości zmiennych musi być taki sam. Transponowanie innych typów składowania (data, czas, znacznik czasu) jest niemożliwe.

- **Wszystkie numeryczne.** Umożliwia transponowanie wszystkich zmiennych numerycznych (typ składowania: liczba całkowita lub rzeczywista). Liczba wierszy wynikowych jest zgodna z liczbą zmiennych numerycznych w oryginalnych danych.
- **Wszystkie łańcuchowe.** Umożliwia transponowanie wszystkich zmiennych łańcuchowych.
- **Użytkownika.** Umożliwia wybór podzbioru zmiennych numerycznych. Liczba wierszy w wyniku jest zgodna z liczbą wybranych zmiennych. Ta opcja jest dostępna tylko dla zmiennych numerycznych.

Nazwa identyfikatora wiersza. Określa nazwę zmiennej identyfikatora wiersza utworzonej przez węzeł. Wartości tej zmiennej są określane przez nazwy zmiennych w oryginalnych danych.

Wskazówka: Jeśli w przypadku transponowania danych szeregów czasowych do kolumn oryginalne dane obejmują wiersze, takie jak data, miesiąc lub rok, które tworzą etykiety okresu dla każdego pomiaru, należy pamiętać, aby odczytać te etykiety w programie IBM SPSS Modeler jako nazwy zmiennych (w sposób omówiony w powyższych przykładach, w których miesiąc lub data są prezentowane w oryginalnych danych jako nazwy zmiennych), tak aby

etykieta nie była uwzględniana w pierwszym wierszu danych. Pozwoli to uniknąć pomieszania etykiet z wartościami we wszystkich kolumnach (co wymusiłoby odczytanie liczb jako łańcuchy, ponieważ nie można mieszać typów składowania w kolumnie).

Z rekordów na zmienne

Pola. Lista zmiennych zawiera wszystkie zmienne wprowadzone w węźle transpozycji.

Indeks. Sekcja Indeks umożliwia wybranie zmiennych, jakie mają być użyte jako zmienne indeksu.

Pola. Sekcja Zmienne umożliwia wybranie zmiennych, które będą używane jako zmienne.

Wartość. Sekcja Wartość umożliwia wybranie zmiennych, jakie mają być użyte jako zmienne wartości.

Funkcja agregująca. Jeśli dla indeksu istnieje więcej niż jeden rekord, konieczne jest zagregowanie tych rekordów w jeden. Korzystając z listy rozwijanej **Funkcja agregująca** należy określić sposób agregacji rekordów; należy w tym celu użyć jednej z następujących funkcji. Należy pamiętać, że agregacja wpływa na wszystkie zmienne.

- **Średnia.** Zwraca wartości średnie dla każdego połączenia zmiennej kluczowej. Średnia jest miarą tendencji centralnej i jest średnią arytmetyczną (suma podzielona przez liczbę obserwacji).
- **Suma.** Zwraca wartości zsumowane dla każdego połączenia zmiennej kluczowej. Suma to łączne wartości wszystkich obserwacji bez braków danych.
- **Min.** Zwraca wartości minimalne dla każdego połączenia zmiennej kluczowej.
- **Maks.** Zwraca wartości maksymalne dla każdego połączenia zmiennej kluczowej.
- **Mediana.** Zwraca wartości mediany dla każdego połączenia zmiennej kluczowej. W przeciwieństwie do średniej, na którą wpływ może mieć nawet kilka ekstremalnie wysokich lub niskich wartości, mediana jest miarą tendencji centralnej niewrażliwą na wartości odstające. Jest również znana jako 50. percentyl lub 2. kwartyl.
- **Liczebności.** Zwraca liczbę wartości innych niż null dla każdego połączenia zmiennej kluczowej.

Ze zmiennych na rekordy

Pola. Lista zmiennych zawiera wszystkie zmienne wprowadzone w węźle transpozycji.

Indeks. Sekcja Indeks umożliwia wybranie zmiennych, jakie mają być użyte jako zmienne indeksu.

Wartość. Sekcja Wartość umożliwia wybranie zmiennych, jakie mają być użyte jako zmienne wartości. Jeśli nie zostanie wybrana żadna wartość, wówczas wszystkie nieprzypisane zmienne liczbowe zostaną użyte jako wartości. Jeśli jednak nie będzie dostępna żadna zmienna liczbowa, wówczas użyte zostaną wszystkie nieprzypisane zmienne łańcuchowe.

Węzeł Historia

Węzły historii są najczęściej używane w przypadku danych sekwencyjnych, takich jak dane szeregu czasowego. Służą do tworzenia nowych zmiennych zawierających dane ze zmiennych z wcześniejszych rekordów. Korzystając z węzła historii, pomocne może być wstępne posortowanie danych według konkretnej zmiennej. Aby to zrobić, można użyć węzła Sortowanie.

Ustawianie opcji dla węzła Historia

Wybrane zmienne. Korzystając z selektora zmiennych (przycisk po prawej stronie pola tekstowego), należy wybrać zmienne, których dane historyczne są potrzebne. Każda wybrana zmienna spowoduje utworzenie nowych zmiennych dla wszystkich rekordów w zbiorze danych.

Przesunięcie. Należy określić ostatni rekord przed rekordem bieżącym, z którego mają zostać wyodrębnione historyczne wartości zmiennych. Przykładowo, jeśli przesunięcie jest ustawione na wartość 3, przy każdym przejściu

rekordu przez ten węzeł w bieżącym węźle tworzone będą wartości zmiennych dla trzeciego rekordu wstecz. Ustawienia rozpiętości umożliwiają określenie, z ilu rekordów wstecz dane będą wyodrębniane. Wartość przesunięcia można skorygować za pomocą strzałek.

Rozpiętość. Należy określić liczbę wcześniejszych rekordów, z których wartości mają zostać wyodrębnione. Przykładowo, jeśli przesunięcie jest ustawione na wartość 3, a rozpiętość na 5, dla każdego rekordu przechodzącego przez węzeł zostanie dodanych pięć zmiennych dla każdej zmiennej określonej na liście Wybrane zmienne. Oznacza to, że jeśli w węźle przetwarzanych jest 10 rekordów, dodane zostaną zmienne z rekordów od 7 do 3. Wartość rozpiętości można skorygować za pomocą strzałek.

Gdy historia jest niedostępna. Aby określić sposób postępowania z rekordami, które nie mają wartości historycznych, należy wybrać jedną z następujących opcji. Ta opcja zwykle odnosi się do kilku pierwszych rekordów znajdujących się w górnej części zbioru danych, dla których nie istnieją wcześniejsze rekordy mogące stanowić źródło danych historycznych.

- **Odrzuć takie rekordy.** Tę opcję należy wybrać, aby odrzucać rekordy, w których nie ma wartości historycznych dla wybranej zmiennej.
- **Pozostaw niezdefiniowaną historię.** Ta opcja pozwala zachować rekordy, w których nie ma wartości historycznych. Zmienna historyczna zostanie wypełniona niezdefiniowaną wartością i będzie wyświetlana jako $null$.
- **Wypełnij wartościami.** Należy określić wartość lub łańcuch do użycia w przypadku rekordów, w których wartości historyczne są niedostępne. Domyślna wartość zastępcza to *undef*, systemowa null. Wartości null są wyświetlane w postaci łańcucha $null$.

Aby działanie było poprawne, wybierając wartość zastępczą, należy pamiętać o następujących regułach:

- Wybrane zmienne powinny mieć taki sam typ składowania.
- Jeśli typ składowania wszystkich zmiennych jest liczbowy, wartość zastępcza musi być analizowana jako liczba całkowita.
- Jeśli typem składowania wszystkich zmiennych jest liczba rzeczywista, wartość zastępcza musi być analizowana jako liczba rzeczywista.
- Jeśli jedna z wybranych zmiennych ma symboliczny typ składowania, wartość zastępcza musi być analizowana jako łańcuch.
- Jeśli wszystkie wybrane zmienne mają typ składowania data/czas, wartość zastępcza musi być analizowana jako zmienna daty/czasu.

Jeśli któryś z powyższych warunków nie zostanie spełniony, podczas wykonywania węzła Historia wystąpi błąd.

Węzeł Reorganizacja

Węzeł Reorganizacja umożliwia zdefiniowanie naturalnego porządku wyświetlania zmiennych w dalszej części strumienia. Ta kolejność wpływa na wyświetlanie zmiennych w różnych obszarach, takich jak tabele, listy i selektor zmiennych. Operacja ta jest na przykład przydatna podczas pracy z obszernymi bazami danych w celu zapewnienia lepszej widoczności zmiennych, które interesują użytkownika.

Ustawianie opcji węzła Reorganizacja

Dostępne są dwa sposoby reorganizacji zmiennych: porządek użytkownika i automatyczne sortowanie.

Porządek użytkownika

Należy wybrać opcję **Porządek użytkownika**, aby aktywować tabelę z nazwami i typami zmiennych, w której można wyświetlić wszystkie zmienne i za pomocą przycisków strzałek utworzyć porządek użytkownika.

Aby przeprowadzić reorganizację zmiennych:

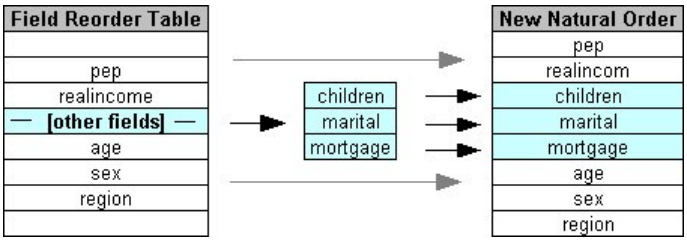
1. Wybierz zmienną z tabeli. Aby wybrać kilka zmiennych, można użyć metody Ctrl+kliknięcie.
2. Aby przenieść węzeł o jeden wiersz w górę lub w dół, można użyć samych przycisków strzałek.

3. Aby przenieść zmienne na dół lub na górę listy, można użyć przycisków z linią i strzałką
4. Określ kolejność zmiennych, które nie zostały tutaj uwzględnione, przesuwając w górę lub w dół wiersz podziału, oznaczony jako [inne zmienne].

Więcej informacji na temat opcji [inne zmienne]

Inne zmienne. Celem wiersza podziału **[inne zmienne]** jest podzielenie tabeli na dwie części.

- Zmienne wyświetlane powyżej węzła podziału zostaną uporządkowane (w sposób, w jaki są wyświetlane w tabeli) nad wszystkimi porządkami naturalnymi użytymi do wyświetlania zmiennych poniżej tego węzła.
- Zmienne wyświetlane poniżej węzła podziału zostaną uporządkowane (w sposób, w jaki są wyświetlane w tabeli) pod wszystkimi porządkami naturalnymi użytymi do wyświetlania zmiennych poniżej tego węzła.



Rysunek 6. Diagram przedstawiający sposób wprowadzania „innych zmiennych" w nowych porządku zmiennych

- Wszystkie pozostałe zmienne, które nie są wyświetlane w tabeli reorganizacji zmiennych, będą wyświetlane pomiędzy zmiennymi „z góry" i „z dołu", zgodnie z ustawieniem węzła podziału.

Można wyróżnić następujące rodzaje ryzyka:

- Zmienne można posortować rosnąco lub malejąco, klikając strzałki nad nagłówkiem poszczególnych kolumn (**Typ**, **Nazwa** i **Składowanie**). W przypadku sortowania według kolumny zmienne, które nie zostaną tutaj określone (wskazane przez wiersz [inne zmienne]), będą sortowanie na końcu w ich porządku naturalnym.
- Kliknięcie przycisku **Wyczyść nieużywane** pozwala usunąć nieużywane zmienne z węzła Reorganizacja. Nieużywane zmienne są wyświetlane w tabeli, jako zapisane czerwoną czcionką. Będzie to oznaczało, że dana zmienna została usunięta podczas wcześniejszych operacji.
- Należy określić kolejność dla wszystkich nowych zmiennych (oznaczonych ikoną błyskawicy jako nowa lub nieokreślona zmienna). Po kliknięciu przycisku **OK** lub **Zastosuj** ikona zniknie.

Uwaga: Jeśli zmienne zostaną dodane we wcześniejszej części strumienia po zastosowaniu porządku użytkownika, nowe zmienne zostaną dołączone u dołu listy użytkownika.

Automatyczne sortowanie

Opcja **Automatyczne sortowanie** służy do określenia parametru sortowania. Opcje w oknie dialogowym zmieniają się w sposób dynamiczny, udostępniając opcje automatycznego sortowania.

Sortuj według. Można wybrać jeden z trzech sposobów sortowania zmiennych wczytywanych do węzła Reorganizacji. Przyciski strzałek wskazują, czy porządek będzie rosnący czy malejący. Należy wybrać jeden z nich, aby dokonać zmiany.

- Nazwa
- Typ
- Składowanie

Zmienne dodane powyżej węzła Reorganizacja po zastosowaniu funkcji automatycznego sortowania zostaną automatycznie umieszczone we właściwej pozycji, zgodnie z wybranym typem sortowania.

Węzeł Przedziały czasowe

Oryginalny węzeł Przedziały czasowe, który był dostępny w programieSPSS Modeler w wersji 17.1 i wcześniejszej, nie jest kompatybilny z serwerem Analytic Server (AS) i został odrzucony w wersji SPSS Modeler 18.0.

Nowy węzeł Przedziały czasowe zawiera szereg zmian w porównaniu do oryginalnego węzła Przedziały czasowe. Nowy węzeł może być używany z produktem Analytic Server lub z produktem SPSS Modeler.

Może on być używany do określenia przedziałów i wyliczenia nowej zmiennej czasu na potrzeby oszacowania lub prognozowania. Obsługiwany jest cały zakres przedziałów czasowych, od sekund po lata.

Węzła należy użyć do wyliczenia nowej zmiennej czasu; typ składowania nowej zmiennej jak taki sam jak wybranej wejściowej zmiennej czasu. Węzeł generuje następujące elementy:

- Zmienną określoną na karcie Zmienne jako **Zmienna czasu**, wraz z wybranym przedrostkiem/przyrostkiem. Domyślny przedrostek to $TI_.
- Zmienne określone na karcie Zmienne jako **Zmienne wymiarów**.
- Zmienne określone na karcie Zmienne jako **Zmienne do agregacji**.

Mogą również zostać wygenerowane zmienne dodatkowe, w zależności od wybranego przedziału lub okresu (np. minuta lub sekunda, w której wykonywany jest pomiar).

Przedział czasowy — opcje zmiennych

Na karcie Zmienne węzła Przedział czasowy można wybrać dane, na podstawie których wyznaczona zostanie nowa zmienna czasu.

Zmienne Wyświetla wszystkie zmienne wejściowe dla węzła, wraz z ikonami typu pomiaru. Wszystkie zmienne czasu mają typ pomiaru „ilościowy”. Należy wybrać zmienną, które będzie używana jako wartość wejściowa.

Zmienna czasu Wyświetla zmienną wejściową, na podstawie której wyznaczany jest nowy przedział czasu; dozwolona jest tylko jedna zmienna ilościowa. Ta zmienna jest używana w węźle Przedziały czasowe jako wartość kluczowa agregacji podczas konwertowania przedziału. Nowa zmienna ma taki sam typ składowania, jak wybrana wejściowa zmienna czasu. Jeśli zostanie wybrana zmienna, które jest liczbą całkowitą, będzie ona traktowana jak indeks czasu.

Zmienne wymiarów Opcjonalnie można tutaj dodać zmienne do utworzenia pojedynczych szeregów czasowych na podstawie wartości zmiennych. Przykładowo, korzystając z danych geoprzestrzennych, można jako wymiaru użyć zmiennej punktowej. W tym przykładzie dane wynikowe węzła Przedział czasowy są sortowane w szeregach czasowych dla każdej wartości punktu ze zmiennej punktowej.

Wymiary stanowią doskonałe rozwiązanie, jeśli używane są spłaszczone dane wielowymiarowe, podobne do tych wygenerowanych przez węzeł TM1 lub do obsługi bardziej złożonych typów danych, takich jak dane geoprzestrzenne. Zasadniczo, można rozważyć użycie opcji **Zmienne wymiarów** jako równoważnika klauzuli **Group By** w zapytaniu SQL lub podobnie opcji **Zmienne grupujące** w węźle Agregacja; jednak opcja **Zmienne wymiarów** jest bardziej wyrafinowania z powodu możliwości obsługi bardziej skomplikowanych struktur danych niż tylko tradycyjne dane z wierszy lub kolumn.

Zmienne do agregacji Należy wybrać zmienne, jakie zostaną zagregowane jako część zmiany okresu zmiennej czasu. Na karcie Budowanie dla tabeli **Ustawienia niestandardowe dla określonych zmiennych** będą wyświetlane tylko zmienne tutaj wybrane. Wszystkie zmienne, które nie zostaną uwzględnione, zostaną odfiltrowane z danych opuszczających węzeł. Oznacza to, że wszystkie zmienne pozostające na liście **Zmienne** zostają wykluczone z danych.

Przedział czasowy — opcje tworzenia

Karta Budowanie umożliwia określenie opcji zmiany przedziału czasu oraz sposobu agregacji zmiennych w danych w oparciu o ich typ pomiaru.

Podczas agregacji danych wszelkie istniejące zmienne daty, czasu i znacznika czasu są zastępowane przez wygenerowane zmienne i usuwane z wyniku. Pozostałe zmienne są agregowane w oparciu o opcje określone przez użytkownika na tej karcie.

Przedział czasowy Należy wybrać przedział i okresowość dla budowania szeregów.

Ustawienie domyślne Należy wybrać domyślną agregację, jaka będzie zastosowana do danych rożnego typu. Ustawienie domyślne jest stosowane na podstawie poziomu pomiaru — na przykład zmienne ciągłe są agregowane z zastosowaniem sumy, a dla zmiennych nominalnych stosowana jest mediana. Można wybrać ustawienia domyślne dla 3 różnych poziomów pomiaru:

- **Ilościowa** Funkcje dostępne dla zmiennych ilościowych to: **Suma**, **Średnia**, **Minimum**, **Maksimum**, **Mediana**, **1. Kwartyl** i **3. Kwartyl**.
- **Nominalna** Dostępne opcje to: **Dominanta**, **Minimum** i **Maksimum**.
- **Flaga** Dostępne opcje to: **Prawda, gdy jakaś jest prawdziwa** lub **Fałsz, jeśli choć jeden fałsz**.

Ustawienia niestandardowe dla określonych zmiennych Można określić wyjątki dla domyślnych ustawień agregacji dla poszczególnych zmiennych. Korzystając z ikon po prawej stronie, można zmienne dodać do lub usunąć z tabeli; można też kliknąć komórkę w odpowiedniej kolumnie, aby zmienić funkcję agregacji używaną dla danej zmiennej. Zmienne bez określonego typu są wykluczane z listy i nie można ich dodać do tabeli.

Nowe rozszerzenie nazwy zmiennej Należy określić **Przedrostek** lub **Przyrostek** zastosowany do wszystkich zmiennych wygenerowanych przez węzeł.

Węzeł Zmiana rzutowania

W przypadku danych geoprzestrzennych lub mapy dwa najbardziej powszechne sposoby identyfikacji współrzędnych to rzutowany układ współrzędnych i układy współrzędnych geograficznych. W programie IBM SPSS Modeler takie składniki oprogramowania, jak funkcje przestrzenne w Konstruktorze wyrażeń, węzeł STP i węzeł Wizualizacja na mapie, używają rzutowanego układu współrzędnych, dlatego wszelkie importowane dane, które zostały zarejestrowane w określonym układzie współrzędnych geograficznych, wymagają zmiany rzutowania. O ile to możliwe, zmienne geoprzestrzenne (wszelkie zmienne z geoprzestrzennym poziomem pomiaru) zostają automatycznie ponownie rzutowane w chwili ich użycia (nie podczas importowania). Jeśli automatyczna zmiana rzutowania zmiennych jest niemożliwa, należy użyć węzła zmiany rzutowania, aby zmienić ich układ współrzędnych. Zmiana rzutowania w taki sposób oznacza, że można poprawić błędy, które są skutkiem użycia niepoprawnego układu współrzędnych.

Na liście poniżej przedstawiono przykładowe sytuacje, w których konieczna może być zmiana rzutowania w celu dokonania zmiany układu współrzędnych:

- *Dołączanie* Jeśli użytkownik próbuje dołączyć dwa zbiory danych z różnymi układami współrzędnych dla zmiennej geoprzestrzennej, program SPSS Modeler wyświetla następujący komunikat o błędzie: Układy współrzędnych zmiennej <*Field1*> i <*Field2*> nie są zgodne. Przeprowadź ponownie projekcję jednej lub obu zmiennych na ten sam układ współrzędnych.

  <*Field1*> i <*Field2*> to nazwy zmiennych geoprzestrzennych, które spowodowały wystąpienie błędu.
- *Wyrażenie if/else* Jeśli używane jest wyrażenie, które zawiera instrukcję if/else razem ze zmiennymi geoprzestrzennymi lub zwracanymi typami w obu częściach wyrażenia, ale układy współrzędnych różnią się, program SPSS Modeler wyświetla następujący komunikat o błędzie: Wyrażenie warunkowe zawiera niezgodne układy współrzędnych: <*arg1*> i <*arg2*>.

  <*arg1*> i <*arg2*> są argumentami „then” lub „else”, które zwracają typ geoprzestrzenny z różnymi układami współrzędnych.
- *Tworzenie listy zmiennych geoprzestrzennych* Aby utworzyć zmienną listy, która składa się z wielu zmiennych geoprzestrzennych, wszystkie argumenty zmiennych geoprzestrzennych wprowadzane do wyrażenia listy muszą należeć do tego samego układu współrzędnych. W przeciwnym razie wyświetlany jest następujący komunikat o błędzie: Układy współrzędnych zmiennej <*Field1*> i <*Field2*> nie są zgodne. Przeprowadź ponownie projekcję jednej lub obu zmiennych na ten sam układ współrzędnych.

Więcej informacji na temat układów współrzędnych zawiera temat Konfigurowanie opcji geoprzestrzennych strumieni w sekcji Praca ze strumieniami w dokumentacji SPSS Modeler — podręcznik użytkownika.

Ustawianie opcji dla węzła Zmiana rzutowania

Zmienne

Zmienne geograficzne

Domyślnie ta lista jest pusta. Można na tę listę przenieść zmienne geoprzestrzenne z listy **Zmienne na nowo odwzorowywane**, aby upewnić się, że dla tych zmiennych rzutowanie nie zostało zmienione.

Zmienne na nowo odwzorowywane

Domyślnie ta lista zawiera wszystkie zmienne geoprzestrzenne stanowiące dane wejściowe dla tego węzła. Dla wszystkich zmiennych z tej listy zostanie przeprowadzona zmiana rzutowania na układ współrzędnych ustawiony w obszarze **Układ współrzędnych**.

Układ współrzędnych

Jak dla strumienia

Należy zaznaczyć tę opcję, aby użyć domyślnego układu współrzędnych.

Określ Jeśli ta opcja zostanie zaznaczona, można za pomocą przycisku **Zmień** wyświetlić okno dialogowe Wybierz układ współrzędnych i wybrać układ, jaki będzie używany do zmiany rzutowania.

Więcej informacji na temat układów współrzędnych zawiera temat Konfigurowanie opcji geoprzestrzennych strumieni w sekcji Praca ze strumieniami w dokumentacji SPSS Modeler — podręcznik użytkownika.

Rozdział 5. Węzły wykresów

Wspólne funkcje węzłów wykresów

W szeregu faz procesu eksploracji danych stosowane są wykresy umożliwiające eksplorację danych przeniesionych do programu IBM SPSS Modeler. Można na przykład podłączyć węzeł Wykresy lub Rozkład do źródła danych w celu uzyskania wglądu w typy danych i rozkłady. Następnie można przeprowadzić rejestrację i manipulacje polami w celu przygotowania danych do operacji modelowania kolejnych węzłów. Innym typowym zastosowaniem wykresów jest sprawdzanie rozkładu i relacji między nowo wyprowadzonymi polami.

Paleta wykresy zawiera następujące węzły:

Węzeł Wizualizacja oferuje wiele różnych typów wykresów w pojedynczym węźle. Korzystając z tego węzła, można wybrać zmienne zawierające dane, dla których ma zostać przeprowadzona eksploracja, a następnie wybrać wykres spośród tych, które zostały udostępnione dla wybranych danych. Węzeł automatycznie filtruje wszelkie typy wykresów, które nie współpracowałyby z wybranymi polami.

Węzeł Rozrzutu przedstawia relacje pomiędzy zmiennymi numerycznymi. Wykres można utworzyć na podstawie dwóch punktów (wykres rozrzutu) lub linii.

Węzeł Rozkładu przedstawia wystąpienia wartości symbolicznych (jakościowych), takich jak typ kredytu lub płeć. Zwykle węzeł rozkładu jest używany do przedstawienia dysproporcji danych, które można później naprawić za pomocą węzła zrównoważenia przed utworzeniem modelu.

Węzeł Histogram pokazuje wystąpienia wartości zmiennych numerycznych. Często używany jest do eksploracji danych przed przystąpieniem do manipulowania i budowy modelu. Podobnie jak w przypadku węzła rozkładu, węzeł histogramu często ujawnia dysproporcje danych.

Węzeł Zbiór przedstawia rozkład wartości dla jednej zmiennej numerycznej względem wartości innej zmiennej. (Tworzy wykresy podobne do histogramów). Jest przydatny do prezentacji zmiennej, której wartości zmieniają się w czasie. Na wykresie 3-W można dodać oś symboliczną odzwierciedlającą rozkład według kategorii.

Węzeł Liniowy tworzy wykres zawierający wiele zmiennych *Y* dla jednej zmiennej *X*. Zmienne *Y* są wykreślane jako kolorowe linie, a każda z nich jest równoważna węzłowi rozrzutu ze stylem ustawionym na wartość **Liniowy** i trybem osi X ustawionym na **Sortuj**. Wykresy wielokrotne są przydatne do zbadania wahań kilku zmiennych w czasie.

Węzeł Sieciowy ilustruje siłę relacji między wartościami co najmniej dwóch zmiennych symbolicznych (jakościowych). Na wykresie linie o różnej szerokości wskazują siłę połączenia. Węzła sieciowego można na przykład użyć do eksploracji relacji między zakupami różnych towarów w witrynie e-sklepu.

Węzeł Sekwencyjny wyświetla co najmniej jeden zbiór danych szeregów czasowych. Zwykle najpierw używany jest węzeł przedziałów czasowych, aby utworzyć zmienną *TimeLabel* (EtykietaCzasu), która będzie używana do oznaczenia osi *x*.

Węzeł Ewaluacyjny pomaga w dokonaniu oceny i porównaniu modeli predykcyjnych. Na wykresie ewaluacyjnym przedstawiane jest, w jakim stopniu modele przewidują określone wyniki. Rekordy sortowane są na podstawie wartości przewidywanej i poziomu ufności predykcji. Rekordy są dzielone na grupy o jednakowej wielkości (**kwantyle**), a następnie tworzone są wykresy wartości wg kryterium biznesowego dla każdego kwantyla, od najwyższego do najniższego. Modele wielokrotne prezentowane są jako osobne linie na wykresie.

Węzeł Wizualizacja na mapie może akceptować wiele połączeń wejściowych i wyświetlać dane geoprzestrzenne na mapie w formie szeregu warstw. Każda warstwa stanowi pojedynczą zmienną geoprzestrzenną; na przykład warstwa podstawowa może być mapą kraju, a nad nią może znajdować się jedna warstwa dróg, jedna warstwa rzek i jedna warstwa miejscowości.

Węzeł Wykres E-plot (Beta) przedstawia relacje pomiędzy zmiennymi liczbowymi. Jest podobny do węzła Wykres, ale oferuje inne opcje, a jego wyniki generowane są za pomocą nowego interfejsu, charakterystycznego dla tego węzła. Zachęcamy do eksperymentowania z nowymi możliwościami tworzenia wykresów, jakie oferuje ten węzeł (mający obecnie status wersji beta).

Stochastyczna metoda porządkowania sąsiadów w oparciu o rozkład t (t-SNE — t-Distributed Stochastic Neighbor Embedding) to narzędzie do wizualizacji danych wysokowymiarowych. Przekształca ono powinowactwa punktów danych w prawdopodobieństwa. Węzeł t-SNE w programie SPSS Modeler został zaimplementowany w języku Python i wymaga biblioteki Python scikit-learn©.

Po dodaniu węzła wykresu do strumienia można dwukrotnie kliknąć węzeł w celu otwarcia okna dialogowego i podania opcji. Większość wykresów zawiera pewną liczbę unikalnych opcji, prezentowanych na jednej lub kilku kartach. Istnieje także kilka opcji kart wspólnych dla wszystkich wykresów. Następujące tematów zawierają więcej informacji na temat tych wspólnych opcji.

Po skonfigurowaniu opcji dla węzła wykresu można uruchomić go w oknie dialogowym lub jako część strumienia. W wygenerowanym oknie wykresu można wygenerować węzły Wyliczanie (Ustaw i Flaga) oraz węzły Wybierz w oparciu o wybór lub obszar danych, w efekcie „podstawiając” dane. Można na przykład użyć tej wydajnej funkcji w celu identyfikacji i wykluczenia wartości skrajnych.

Sposób prezentacji, nakładanie, panele i animacje

Nakładanie i sposób prezentacji

Sposoby prezentacji (i wykresy nakładane) dodają wizualizacjom wymiarowości. Efekt sposobu prezentacji (grupowania, skupiania lub zestawiania) zależy od typu wizualizacji, typu pola (zmiennej) oraz typu i statystyki elementu graficznego. Zmienną jakościową koloru można na przykład wykorzystać do grupowania punktów na wykresie rozrzutu lub tworzenia zestawień na zestawionym wykresie słupkowym. Ciągły przedział liczbowy koloru można także wykorzystać do wskazania wartości przedziału dla każdego punktu na wykresie rozrzutu.

Sposób prezentacji i elementy nakładane, które spełniają potrzeby użytkownika, należy znaleźć drogą eksperymentu. Poniższe opisy mogą pomóc w wyborze właściwych opcji.

Uwaga: Nie wszystkie sposoby prezentacji i elementy nakładane są dostępne dla wszystkich typów wizualizacji.

- **Kolor.** Gdy kolor jest zdefiniowany względem zmiennej jakościowej, rozdziela wizualizację na podstawie poszczególnych kategorii tak, że używany jest jeden poziom koloru dla każdej kategorii. Kiedy kolor jest ciągłym przedziałem liczbowym, to różni się on w zależności od wartości przedziału pola. Jeśli element graficzny (na przykład pasek lub okno) reprezentuje więcej niż jeden rekord/obserwację, a do koloru używane jest pole przedziału, kolor różni się w zależności od wartości *średnie* pola przedziału.
- **Kształt.** Kształt jest definiowany względem zmiennej jakościowej, które rozdziela wizualizację na elementy o różnych kształtach, po jednym dla każdej kategorii.
- **Przezroczystość.** Gdy przezroczystość jest zdefiniowana względem zmiennej jakościowej, rozdziela wizualizację na podstawie poszczególnych kategorii tak, że używany jest jeden poziom przezroczystości dla każdej kategorii. Kiedy przezroczystość jest ciągłym przedziałem liczbowym, to przezroczystość różni się w zależności od wartości przedziału pola. Jeśli element graficzny (na przykład słupek lub prostokąt) reprezentuje więcej niż jeden rekord/obserwację, a do przezroczystości używane jest pole przedziału, kolor różni się w zależności od *wartości średniej* pola przedziału. Dla największej wartości elementy graficzne są w pełni przezroczyste. Dla najmniejszej wartości przezroczystości nie ma w ogóle.
- **Opis danych.** Etykiety danych są zdefiniowane względem dowolnego typu zmiennej, której wartości służą do tworzenia etykiet, które są dołączone do elementów graficznych.
- **Rozmiar.** Gdy rozmiar jest zdefiniowany względem zmiennej jakościowej, rozdziela wizualizację na podstawie poszczególnych kategorii tak, że używany jest jeden poziom rozmiaru dla każdej kategorii. Kiedy rozmiar jest ciągłym przedziałem liczbowym, to rozmiar różni się w zależności od wartości przedziału pola. Jeśli element graficzny (na przykład pasek lub okno) reprezentuje więcej niż jeden rekord/obserwację, a do rozmiaru używane jest pole przedziału, rozmiar różni się w zależności od *wartości średniej* pola przedziału.

Panelowanie i animacje

Panelowanie. Dzielenie na panele, znane również pod nazwą aproksymacji powierzchni ścianami, tworzy tabelę wykresów. Dla każdej kategorii w polach dzielenia na panele generowany jest osobny wykres, ale wszystkie panele są wyświetlane jednocześnie. Dzielenie na panele jest użytecznym narzędziem do sprawdzania, czy wizualizacja podlega warunkom pól dzielenia na panele. Na przykład można podzielić na panele histogram według płci, aby określić, czy częstotliwości rozkładów są takie same wśród osobników płci męskiej i żeńskiej. Dzięki temu można sprawdzić, czy wypłaty różnią się w zależności od płci. Wybierz zmienną jakościową używaną do dzielenia na panele.

Animacje. Animacja przypomina dzielenie na panele, podczas którego wiele wykresów tworzonych jest z wartości pola animacji, ale wykresy te nie są pokazywane razem. Zamiast tego wykorzystuje się elementy sterujące trybu eksploracji, aby animować wynik i przerzucać sekwencję pojedynczych wykresów. Co więcej, w przeciwieństwie do dzielenia na panele, proces animacji nie wymaga zmiennej jakościowej. Można zdefiniować zmienną ciągłą, której wartości są automatycznie rozdzielane do przedziałów. Można zmieniać rozmiar przedziałów za pomocą elementów sterujących animacją, które są dostępne w trybie eksploracji. Nie wszystkie wizualizacje oferują animacje.

Używanie karty Wynik

Dla wszystkich wykresów można określić następujące opcje dotyczące nazwy pliku i wyświetlania wygenerowanych wykresów.

Uwaga: Dla wykresów węzła Rozkład dostępne są dodatkowe ustawienia.

Nazwa wyniku. Określa nazwę utworzonego wykresu po uruchomieniu węzła. **Automatycznie** wybiera nazwę na podstawie węzła, który spowodował wygenerowanie wyniku. Opcjonalnie można wybrać opcję **Użytkownika**, aby określić inną nazwę.

Wynik na ekran. Tę opcję należy wybrać, aby wygenerować i wyświetlić wykres w nowym oknie.

Wynik do pliku. Ta opcja umożliwia zapisanie wyniku w postaci pliku.

- **Wykres wynikowy.** Opcja ta pozwala zaprezentować wynik w formacie graficznym. Dostępna jest tylko do węzłów rozkładu.
- **Tabela wynikowa.** Opcja ta pozwala zaprezentować wynik w postaci tabeli. Dostępna jest tylko do węzłów rozkładu.
- **Nazwa pliku.** Należy określić nazwę pliku używaną dla wygenerowanego wykresu lub tabeli. Przycisk wielokropka **(...)** pozwala określić konkretny plik i lokalizację.
- **Typ pliku.** Należy z listy rozwijanej wybrać typ pliku. Dla wszystkich węzłów wykresów, z wyjątkiem węzła rozkładu z wybraną opcją **Tabela wynikowa**, dostępne są następujące typy plików graficznych.
  - Bitmapa (.bmp)
  - PNG (.png)
  - Obiekt wynikowy (.cou)
  - JPEG (.jpg)
  - HTML (.html)
  - Dokument ViZml (.xml) do użycia z innymi aplikacjami IBM SPSS Statistics.

  W przypadku opcji **Tabela wynikowa** w węźle rozkładu dostępne są następujące typy plików.
  - Dane rozdzielane tabulatorami (.tab)
  - Dane rozdzielane przecinkami (.csv)
  - HTML (.html)
  - Obiekt wynikowy (.cou)

Podział na podstrony. Podczas zapisywania wyniku w formacje HTML ta opcja jest włączona, aby umożliwić kontrolowanie wielkości każdej strony w formacie HTML. (Dotyczy tylko węzła rozkładu).

Wierszy na stronę. Po wybraniu opcji **Podział na podstrony** ta opcja jest aktywowana, aby umożliwić określenie długości każdej strony w formacie HTML. Wartość domyślna to 400 wierszy. (Dotyczy tylko węzła rozkładu).

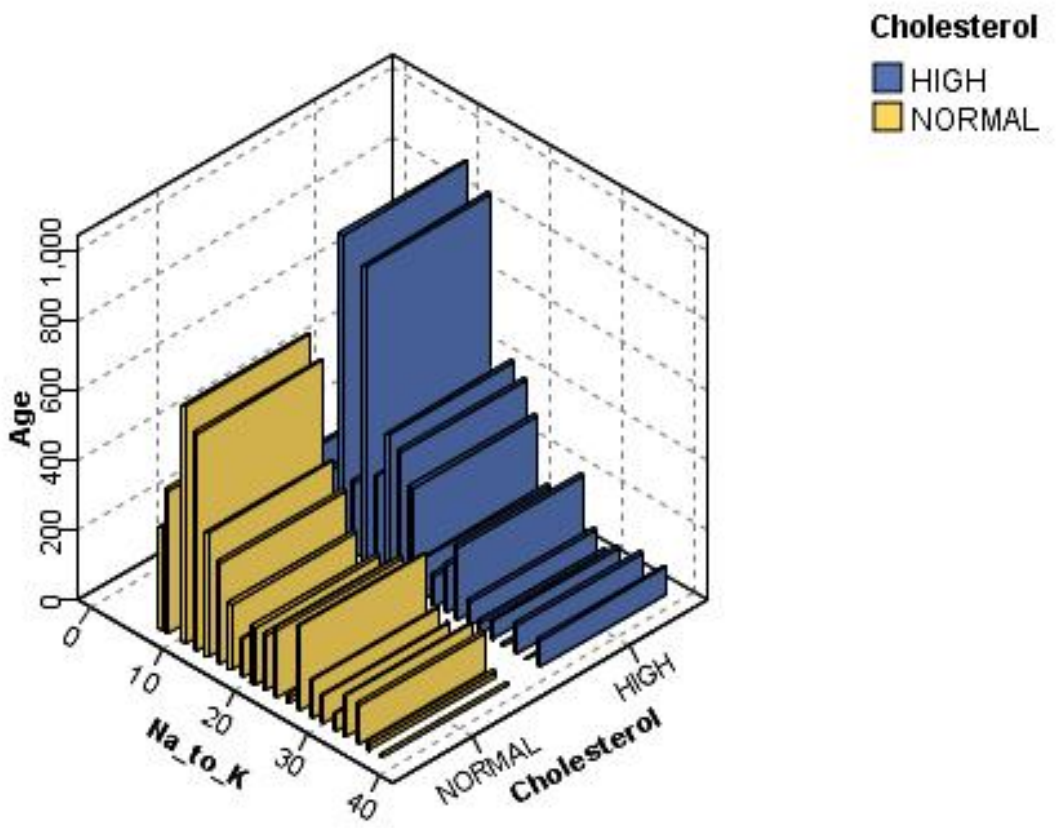
Używanie karty Adnotacje

Ta karta jest używana dla wszystkich węzłów i oferuje opcje zmiany nazwy węzłów, obsługi podpowiedzi użytkownika i zapisywania długich powiadomień.

Wykresy trójwymiarowe

Wykresy i wykresy przedziałowe w programie IBM SPSS Modeler umożliwiają wyświetlanie informacji na trzeciej osi. Zapewnia to dodatkową elastyczność podczas wizualizacji danych w celu wybrania podzbiorów lub wyznaczenia nowych zmiennych do modelowania.

Po utworzeniu wykresu trójwymiarowego można go kliknąć i przeciągnąć mysz, dzięki czemu wykres można obracać i wyświetlać pod dowolnym kątem.



Rysunek 7. Wykres przedziałowy z osiami x, y i z

Możliwe są dwa sposoby tworzenia wykresów trójwymiarowych w programie IBM SPSS Modeler: wykreślanie informacji na trzeciej osi (prawdziwe wykresy trójwymiarowe) i wyświetlanie wykresów z efektami trójwymiarowymi. Obie metody są dostępne dla wykresów i wykresów przedziałowych.

Aby wykreślić informacje na trzeciej osi

1. W oknie dialogowym węzła wykresu należy kliknąć zakładkę **Wykres**.
2. Kliknięcie przycisku 3-W umożliwia aktywowanie opcji dla osi *z*.
3. Za pomocą przycisku selektora zmiennych można wybrać zmienną dla osi *z*. W niektórych przypadkach dozwolone są tutaj tylko zmienne symboliczne. Selektor zmiennych spowoduje wyświetlenie odpowiednich zmiennych.

Aby dodać efekty trójwymiarowe do wykresu

1. Po utworzeniu wykresu należy kliknąć zakładkę **Wykres** w oknie wynikowym.
2. Kliknięcie przycisku 3-W umożliwia przełączenie widoku na wykres trójwymiarowy.

Węzeł Graphboard

Węzeł Graphboard umożliwia wybranie jednego spośród wielu różnych formatów wykresów (wykresy słupkowe, wykresy kołowe, histogramy, wykresy rozrzutu, mapy natężeń itp.) w jednym trybie. W tym celu należy najpierw na pierwszej karcie wybrać pola danych, które mają zostać poddane eksploracji. Za pomocą węzła zostaną wówczas wyświetlone dostępne typy wykresów odpowiadające posiadanym danym. Węzeł automatycznie filtruje wszelkie typy

wykresów, które nie współpracowałyby z wybranymi polami. Istnieje możliwość zdefiniowania szczegółowych lub bardziej zaawansowanych opcji wykresów na karcie Szczegółowe.

Uwaga: W celu edycji węzła lub wyboru typu wykresu konieczne jest podłączenie węzła Graphboard do strumienia danych.

Dostępne są dwa przyciski, które umożliwiają wybór wyświetlanych szablonów (i arkuszy stylów) wizualizacji:

Zarządzaj. Zarządzaj szablonami wizualizacji, arkuszami stylów i mapami na swoim komputerze. Możesz importować, eksportować, zmieniać nazwy i usuwać szablony wizualizacji, arkusze stylów i mapy na swoim komputerze lokalnym. Więcej informacji można znaleźć w temacie “Zarządzanie plikami szablonów, arkuszy stylów oraz map” na stronie 218.

Położenie. Zmień miejsce zapisu szablonów wizualizacji, arkuszy stylów i map. Bieżąca lokalizacja podana jest po prawej stronie przycisku. Więcej informacji można znaleźć w temacie “Ustawianie lokalizacji szablonów, arkuszy stylów i map” na stronie 217.

Karta Opcje podstawowe węzła Graphboard

Jeśli nie masz pewności, jaki typ wizualizacji najlepiej przedstawi Twoje dane, użyj karty Opcje podstawowe. Po wybraniu danych wyświetlony zostanie podzbiór typów wizualizacji, które odpowiadają tym danym. Przykłady można znaleźć w części “Przykłady węzła Graphboard ” na stronie 207.

1. Wybierz co najmniej jedno pole (jedną zmienną) z listy. Aby zaznaczyć wiele elementów pól, należy kliknąć je przy wciśniętym klawiszu Ctrl.

   Uwaga: poziom pomiaru pola determinuje typ dostępnych wizualizacji. Poziom pomiaru można zmienić, klikając prawym klawiszem myszy pole na liście i wybierając odpowiednią opcję. Więcej informacji o dostępnych typach poziomu pomiaru znaleźć można w części “Typy pól (zmiennych)” na stronie 195.
2. Wybierz typ wizualizacji. Opis dostępnych typów można znaleźć w części “Dostępne wbudowane typy wizualizacji Graphboard ” na stronie 199.
3. W przypadku niektórych wizualizacji można wybrać statystykę podsumowującą. Dostępne są różne podzbiory statystyk, zależnie od tego, czy dana statystyka jest oparta na liczności, czy obliczana na podstawie zmiennej ciągłej. Dostępne statystyki zależą także od samego szablonu. Poniżej znajduje się pełna lista statystyk, które mogą być dostępne w następnym kroku.
4. Jeśli chcesz zdefiniować więcej opcji, np. opcjonalne sposoby prezentacji i pola panelowe, kliknij kartę **Opcje szczegółowe**. Więcej informacji można znaleźć w temacie “Karta Opcje szczegółowe węzła Graphboard” na stronie 197.

Statystyki podsumowujące wyliczone na podstawie zmiennej ciągłej

- *Średnia*. Miara tendencji centralnej. Średnia arytmetyczna; suma podzielona przez liczbę obserwacji.
- *Mediana*. Jest to 50. percentyl, czyli taka wartość, że połowa obserwacji ma wartości mniejsze, a druga połowa ma wartości większe od niej. W sytuacji parzystej liczby obserwacji mediana jest średnią dwóch środkowych obserwacji w próbie posortowanej rosnąco lub malejąco. W przeciwieństwie do średniej, na którą wpływ może mieć nawet kilka ekstremalnie dużych lub małych wartości, mediana jest miarą tendencji centralnej niewrażliwą na wartości odstające.
- *Dominanta*. Wartość występująca najczęściej. Jeśli więcej niż jedna wartość występuje z taką samą, największą częstością, każda z nich jest dominantą (wartością modalną).
- *Minimum*. Najmniejsza wartość zmiennej numerycznej.
- *Maksimum*. Największa wartość zmiennej numerycznej.
- *Przedział*. Różnica między wartością minimalną a maksymalną.
- *Środek rozstępu*. Środek rozstępu to wartość, dla której różnica od wartości minimalnej jest równa różnicy od wartości maksymalnej.
- *Suma*. Suma wartości wszystkich obserwacji nieposiadających braków danych.

- *Suma skumulowana*. Skumulowana suma wartości. Każdy element graficzny pokazuje sumę dla jednej podgrupy oraz sumę całkowitą wszystkich poprzednich grup.
- *Suma procentowa*. Procent dla każdej podgrupy bazujący na sumowanym polu w porównaniu do sumy dla wszystkich grup.
- *Skumulowana suma procentowa*. Skumulowany procent dla każdej podgrupy bazujący na sumowanym polu w porównaniu do sumy dla wszystkich grup. Każdy element graficzny pokazuje procent dla jednej podgrupy oraz całkowity procent wszystkich poprzednich grup.
- *Wariancja*. Miara rozproszenia wokół średniej, równa sumie podniesionych do kwadratu odchyleń od średniej, podzielonej przez liczbę obserwacji minus jeden. Wariancja jest mierzona w jednostkach będących kwadratami jednostek miary dla zmiennej, do której wariancja się odnosi.
- *Odchylenie standardowe*. Miara rozproszenia wokół średniej. W przypadku rozkładu normalnego, 68% obserwacji znajduje się w obszarze oddalonym o jedno odchylenie standardowe od średniej, zaś 95% — w przedziale oddalonym o dwa odchylenia standardowe. Na przykład, jeśli średnia wieku osób wynosi 45 lat, a odchylenie standardowe wynosi 10, wówczas 95% rozważanych osób znajduje się w przedziale wiekowym między 25 a 65 lat.
- *Błąd standardowy*. Miara tego, jak bardzo wartość statystyki testowej (sprawdzianu testu) zmienia się pomiędzy próbami. Jest to odchylenie standardowe rozkładu wartości danej statystyki dla poszczególnych prób. Na przykład błąd standardowy średniej to odchylenie standardowe średnich z prób.
- *Kurtoza*. Miara ilości skrajnych wartości odstających. W przypadku rozkładu normalnego wartość statystyki kurtozy wynosi zero. Kurtoza dodatnia oznacza, że w danych jest więcej skrajnych wartości odstających niż w rozkładzie normalnym. Kurtoza ujemna oznacza, że w danych jest mniej skrajnych wartości odstających niż w rozkładzie normalnym.
- *Skośność*. Miara asymetrii rozkładu. Rozkład normalny jest symetryczny, a jego wartość skośności wynosi 0. Rozkład o dużej skośności dodatniej ma długi kraniec z prawej strony. Gdy zaś współczynnik skośności jest ujemny, rozkład ma długi kraniec z lewej strony. Jako wytyczna, wartość skośności przekraczająca dwukrotnie swój błąd standardowy na ogół oznacza odstępstwo od symetrii rozkładu.

Następujące statystyki regionalne mogą wygenerować więcej niż jeden element graficzny dla każdej podgrupy. Używając interwału, obszaru lub elementów obszarów graficznych, statystyką regionalną staje się jeden element graficzny ukazujący rozstęp. Wszystkie inne elementy graficzne stają się dwoma oddzielnymi elementami, jeden ukazujący początek rozstępu i jeden pokazujący jego koniec.

- **Region: Przedział**. Zakres wartości między wartością minimalną a maksymalną.
- **95% przedziału ufności średniej**. Zakres wartości, który ma 95% szans, że zawiera średnią populacji.
- **95% przedziału ufności jednostki**. Przedział liczbowy, który ma 95% prawdopodobieństwa, że będzie zawierać wartość przewidywaną przy założeniu indywidualnej obserwacji.
- **Region: 1 Odchylenie standardowe powyżej/poniżej Średniej**. Zakres wartości od 1 *odchylenia standardowego* powyżej i poniżej *wartości średniej*.
- **Region: 1 Standardowy błąd powyżej/poniżej Średniej**. Zakres wartości od 1 *standardowego błędu* powyżej i poniżej *wartości średniej*.

Statystyka podsumowująca bazująca na liczebności

- **Liczebność**. Liczba wierszy/obserwacji.
- **Liczebność skumulowana**. Skumulowana liczba wierszy/obserwacji. Każdy element graficzny pokazuje liczebność dla jednej podgrupy oraz całkowitą liczebność wszystkich poprzednich grup.
- **Procent liczebności**. Procent wierszy/obserwacji w każdej podgrupie w porównaniu do całkowitej ilości wierszy/obserwacji.
- **Skumulowany procent liczebności**. Skumulowany procent wierszy/obserwacji w każdej podgrupie w porównaniu do całkowitej ilości wierszy/obserwacji. Każdy element graficzny pokazuje procent dla jednej podgrupy oraz całkowity procent wszystkich poprzednich grup.

Typy pól (zmiennych)

Obok pól na liście pól widoczne są ikony określające typ pola i typ danych. Wskazują także zestawy wielokrotnych odpowiedzi.

Tabela 33. Ikony poziomów pomiaru.

Poziom pomiaru	Numeryczna	Łańcuch	Data	Czas
Ilościowy		n/a		
Zbiór uporządkowany				
Ustaw:				

Tabela 34. Ikony zestawów wielokrotnych odpowiedzi.

Typ zestawu wielokrotnych odpowiedzi	Ikona
Zestaw wielokrotnych odpowiedzi, wielokrotne kategorie	
Zestaw wielokrotnych odpowiedzi, wielokrotne dychotomie	

Poziom pomiaru

Poziom pomiaru pola jest istotnym czynnikiem podczas tworzenia wizualizacji. Poniżej przedstawiony jest opis poziomu pomiaru. Klikając prawym przyciskiem myszy pole na liście pól i wybierając opcję, można tymczasowo zmienić poziom pomiaru pola. W większości przypadków trzeba rozważyć tylko dwie najszersze klasyfikacje pól: jakościowe i ilościowe:

Zmienne jakościowe. Dane posiadające ograniczoną liczbę odrębnych wartości lub kategorii (np. płeć czy religia). Zmienne jakościowe mogą być łańcuchami (alfanumerycznymi) lub polami numerycznymi, wykorzystującymi kody liczbowe reprezentujące kategorie (np. 0 = *mężczyzna* i 1 = *kobieta*). Zmienne kategorialne nazywane są również danymi jakościowymi. Zbiory, zbiory uporządkowane i flagi (przełączniki) są zmiennymi jakościowymi.

- *Zbiór*. Zmienna, której wartości reprezentują kategorie bez wewnętrznego rangowania; na przykład wydział, na którym są zatrudnieni pracownicy. Przykładami zmiennych nominalnych są: region, kod pocztowy lub wyznanie. Zmienne tego typu nazywane są także nominalnymi.
- *Zbiór uporządkowany*. Zmienna, której wartości reprezentują kategorie z wewnętrznym rangowaniem, na przykład poziomy zadowolenia z usługi – od bardzo niezadowolonego do bardzo zadowolonego. Przykładami zmiennych uporządkowanych mogą być oceny opinii reprezentujące stopień satysfakcji lub przekonania oraz oceny preferencji. Zmienne tego typu nazywane są także porządkowymi.
- *Flaga*. Zmienna o dwu różnych wartościach, takich jak Tak i Nie lub 1 i 2. Znana również jako zmienna dychotomiczna lub binarna.

Ciągły. Dane mierzone na skali interwałowej lub ilorazowej, których wartości określają zarówno ich porządek, jak i odległość między nimi. Na przykład roczna pensja w wysokości 72 195 PLN jest wyższa niż pensja wynosząca 52 398 PLN, a odległość między tymi dwiema wartościami wynosi 19 797 PLN. Zmienne ilościowe są również zwane danymi ilościowymi, skali lub przedziału liczbowego.

Zmienne jakościowe określają kategorie wizualizacji, zazwyczaj w celu szkicowania osobnych elementów graficznych lub grupowania elementów graficznych. Zmienne ciągłe często są podsumowywane w kategoriach zmiennych jakościowych. Na przykład domyślna wizualizacja dochodu dla kategorii płci pokazuje średni dochód kobiet i mężczyzn. Surowe wartości zmiennych ciągłych można wykreślać, jak na wykresie rozrzutu. Przykładowo wykres rozrzutu dla każdego przypadku może przedstawiać bieżące wynagrodzenia i wynagrodzenia początkowe. Aby pogrupować przypadki według płci, można użyć zmiennej jakościowej.

Typy danych

Poziom pomiaru nie jest jedyną właściwością pola, która określa jego typ. Pole jest także przechowywane jako konkretny typ danych. Do dostępnych typów danych należą łańcuchy (dane inne niż liczbowe, np. litery), wartości liczbowe (liczby rzeczywiste) i daty. W odróżnieniu od poziomu pomiaru typu danych pola nie da się tymczasowo zmienić. Należy zmienić sposób przechowywania danych w oryginalnym zbiorze danych.

Zestawy wielokrotnych odpowiedzi

Niektóre pliki danych obsługują specjalny rodzaj „pól" nazywanych **zestawami wielokrotnych odpowiedzi**. Zestawy wielokrotnych odpowiedzi nie są „polami" w normalnym tego słowa znaczeniu. Zestawy wielokrotnych odpowiedzi wykorzystują wiele pól do rejestracji odpowiedzi na pytania w przypadku, kiedy respondent może udzielić więcej niż jednej odpowiedzi. Są one traktowane podobnie jak zmienne jakościowe i można je, w większości przypadków, poddawać podobnym operacjom.

Zestawy wielokrotnych odpowiedzi muszą być zestawami wielokrotnych dychotomii lub zestawami wielokrotnych kategorii.

Zestaw wielokrotnych dychotomii. Zestaw wielokrotnych dychotomii składa się z wielu pól dychotomii: pól o tylko dwu możliwych wartościach wynoszących tak/nie, występuje/nie występuje, zaznaczone/niezaznaczone. Mimo że pola nie muszą być czysto dychotomiczne, wszystkie pola w zestawie są zakodowane w ten sam sposób.

Przykład: ankieta zawiera pytanie „Z których źródeł wiadomości spośród podanych poniżej korzystasz?" i pięć możliwych odpowiedzi. Respondent może wybrać kilka odpowiedzi, zaznaczając pole wyboru obok każdej z nich. Pięć odpowiedzi staje się pięcioma polami w pliku danych, gdzie 0 oznacza *Nie* (nie zaznaczone) a 1 oznacza *Tak* (zaznaczone).

Zestawy wielokrotnych kategorii. Zestaw wielokrotnych kategorii składa się z wielu pól, zakodowanych w taki sam sposób, często z wieloma kategoriami możliwych odpowiedzi. Na przykład jeden z elementów ankiety jest następujący: „Podaj maksymalnie trzy narodowości, które najlepiej opisują twoje pochodzenie etniczne". Mogą istnieć setki możliwych odpowiedzi, jednak na potrzeby kodowania lista jest ograniczona do 40 najbardziej powszechnych narodowości, a wszystkie pozostałe należą do kategorii "inne". W pliku danych trzy wybrane odpowiedzi stają się trzema zmiennymi, przy czym każda posiada 41 kategorii (40 zakodowanych narodowości i jedna kategoria „inne").

Karta Opcje szczegółowe węzła Graphboard

Jeśli wiesz, jaki typ wizualizacji chcesz utworzyć, lub chcesz dodać do wizualizacji dodatkowe opcje prezentacyjne, panele lub animacje, skorzystaj z karty Opcje szczegółowe. Przykłady można znaleźć w części "Przykłady węzła Graphboard " na stronie 207.

1. Po wybraniu typu wizualizacji na karcie Opcje podstawowe zostanie on wyświetlony. Można też wybrać go z listy rozwijanej. Informacje o dostępnych typach wizualizacji można znaleźć w części "Dostępne wbudowane typy wizualizacji Graphboard " na stronie 199.
2. Bezpośrednio na prawo od miniaturki wizualizacji znajdują się elementy sterujące, umożliwiające określenie wymaganych pól (zmiennych) dla danego typu wizualizacji. Należy określić wszystkie te pola.
3. W przypadku niektórych wizualizacji można wybrać statystykę podsumowującą. W pewnych sytuacjach (np. w przypadku wykresów słupkowych) można przedstawić jedną z opcji podsumowania za pomocą przezroczystości. Opis dostępnych statystyk podsumowujących można znaleźć w części "Karta Opcje podstawowe węzła Graphboard" na stronie 194.

4. Można wybrać jeden lub więcej opcjonalnych sposobów prezentacji. Pozwalają one dodać wymiarowość, umożliwiając uwzględnienie w wizualizacji innych pól. Można na przykład wykorzystać pole do zróżnicowania rozmiaru punktów na wykresie rozrzutu. . Więcej informacji na temat opcjonalnych sposobów prezentacji można znaleźć w sekcji “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190. Uwaga: sposób prezentacji wykorzystujący przezroczystość nie jest obsługiwany przez skrypty.
5. Jeśli tworzysz wizualizację mapy, grupa **Pliki mapy** pokazuje plik lub pliki mapy, które zostaną wykorzystane. Jeśli istnieje domyślny plik mapy, plik ten zostanie wyświetlony. Aby zmienić plik mapy, kliknij **Wybierz plik mapy**, aby wyświetlić okno dialogowe Wybierz mapę. W tym oknie dialogowym możesz także określić domyślny plik mapy. Więcej informacji można znaleźć w temacie “Wybór plików mapy do Wizualizacji mapy”.
6. Można wybrać jedną lub więcej opcji panelowania lub animacji. Więcej informacji na temat opcji panelowania i animacji można znaleźć w części “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190.

Wybór plików mapy do Wizualizacji mapy

Jeśli wybierzesz szablon wizualizacji mapy, będziesz potrzebować pliku mapy, który określa informacje geograficzne potrzebne do narysowania mapy. Jeśli istnieje domyślny plik mapy, zostanie on użyty do stworzenia wizualizacji mapy. Aby wybrać inny plik mapy, kliknij **Wybierz plik mapy** w karcie Szczegółowej, żeby wyświetlić okno dialogowe Wybierz mapę.

Okno dialogowe Wybierz mapy pozwala na wybranie głównego pliku mapy oraz pliku mapy odniesienia. Plik mapy definiuje informacje geograficzne potrzebne do narysowania mapy. Twoja aplikacja została zainstalowana z zestawem standardowych plików mapy. Jeśli masz inny plik kształtu ESRI, który chcesz wykorzystać, musisz najpierw przekonwertować pliki kształtu do formatu pliku SMZ. Więcej informacji można znaleźć w temacie “Konwertowanie i dystrybucja plików kształtu map” na stronie 219. Po przekonwertowaniu mapy kliknij **Zarządzaj...** znajdujące się w oknie dialogowym wyboru szablonu, aby zaimportować mapę do systemu Zarządzania, aby była dostępna w oknie dialogowym Wybierz mapy.

Następnie pojawia się kilka kwestii do rozważenia podczas określania plików mapy:

- Wszystkie szablony mapy wymagają przynajmniej jednego pliku mapy.
- Plik mapy zwykle łączy atrybut klucza mapy z kluczem danych.
- Jeśli szablon nie wymaga klucza mapy podłączonego do klucza danych, wymaga on pliku mapy odniesienia oraz pól określających współrzędne (takie jak długość i szerokość) do rysowania elementów na mapie odniesienia.
- Nakładane szablony mapy wymagają dwóch map: pliku mapy głównej i pliku mapy odniesienia. Najpierw rysowana jest mapa odniesienia, aby znajdowała się pod plikiem mapy głównej.

Więcej informacji na temat terminologii dotyczącej map, tj. atrybutów i właściwości, można znaleźć w części “Główne zagadnienia dotyczące map” na stronie 220.

Plik mapy. Możesz wybrać dowolny plik mapy, znajdujący się w systemie Zarządzania. Znajdują się tam wstępnie zainstalowane pliki mapy oraz pliki mapy zaimportowane przez Ciebie. Aby uzyskać dalsze informacje na temat zarządzania plikami mapy, patrz “Zarządzanie plikami szablonów, arkuszy stylów oraz map” na stronie 218.

Klucz mapy. Określ atrybut, którego chcesz użyć jako klucza łączącego plik mapy z kluczem danych.

Zapisz ten plik mapy oraz ustawienia jako domyślne. Zaznacz to pole wyboru, jeśli chcesz użyć wybranego pliku mapy jako domyślnego. Jeśli określono domyślny plik mapy, nie musisz określać pliku mapy za każdym razem, gdy tworzysz wizualizację mapy.

Klucz danych. Ten element sterujący wyświetla tę samą wartość, która pojawia się w karcie szczegółowej wyboru szablonu. Została ona umieszczona w tym miejscu dla wygody, jeśli zajdzie potrzeba zmiany klucza dla konkretnie wybranego pliku mapy.

Wyświetla wszystkie właściwości mapy w wizualizacji. Gdy opcja ta jest zaznaczona, wszystkie właściwości mapy zostają zrenderowane w wizualizacji, nawet jeśli brak pasującej wartości klucza danych. Jeśli chcesz widzieć tylko

właściwości, dla których posiadasz dane, usuń zaznaczenie tej opcji. Zmienne zidentyfikowane przez klucze mapy ukazane w liście **Niedopasowane klucze mapy** nie zostaną zrenderowane w wizualizacji.

Porównaj wartości mapy i danych. Klucz mapy i klucz danych są ze sobą połączone, aby stworzyć wizualizację mapy. Klucz mapy i klucz danych powinny pochodzić z tej samej domeny (na przykład: kraje i regiony). Kliknij **Porównaj**, aby sprawdzić, czy wartości klucza danych i klucza mapy zgadzają się. Wyświetlana ikona informuje Cię o stanie porównania. Ikony te zostały opisane poniżej. Jeśli porównanie zostało przeprowadzone i istnieją wartości klucza danych niedopasowane do wartości klucza mapy, znajdują się one na liście **Niedopasowane klucze danych**. Na liście **Niedopasowane klucze map** można także zobaczyć, które wartości kluczy map nie mają odpowiadającym im wartości klucza danych. Jeśli opcja **Wyświetlaj wszystkie zmienne mapy w wizualizacji** nie jest zaznaczona, zmienne określane przez te wartości klucza mapy nie zostaną zrenderowane.

Tabela 35. Ikony porównania.

Ikona	Opis
	Nie przeprowadzono żadnych porównań. Jest to stan domyślny zanim klikniesz **Porównaj**. Powinieneś kontynuować ostrożnie, ponieważ nie wiesz, czy wartości klucza danych i klucza mapy się zgadzają.
	Porównanie zostało wykonane i wszystkie wartości klucza danych oraz klucza mapy zgadzają się ze sobą. Dla każdej wartości klucza danych istnieje pasująca właściwość zidentyfikowana przez klucz mapy.
	Porównanie zostało wykonane i niektóre wartości klucza danych oraz klucza mapy nie zgadzają się ze sobą. Dla niektórych wartości klucza danych nie ma pasującej właściwości zidentyfikowanej przez klucz mapy. Powinieneś być ostrożny. Jeśli będziesz kontynuować, wizualizacja mapy nie będzie zawierać wszystkich wartości danych.
	Porównanie zostało wykonane i żadne wartości klucza danych oraz klucza mapy nie zgadzają się ze sobą. Powinieneś wybrać inny klucz danych lub klucz mapy, ponieważ, jeśli będziesz kontynuować, nie zostanie zrenderowana żadna mapa.

Dostępne wbudowane typy wizualizacji Graphboard

Istnieje możliwość utworzenia kilku różnych typów wizualizacji. Wszystkie poniższe wbudowane typy są dostępne na kartach Opcje podstawowe i Opcje szczegółowe. Niektóre opisy szablonów (zwłaszcza szablonów map) identyfikują pola (zmienne) określone w karcie Szczegółowej, używając **tekstu specjalnego**.

Tabela 36. Dostępne typy wykresów.

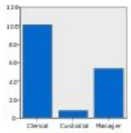
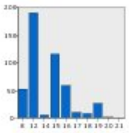
Ikona wykresu	Opis	Ikona wykresu	Opis
	**Wykres słupkowy** Oblicza statystyki podsumowujące dla ciągłego pola liczbowego i wyświetla w postaci słupków wyniki dla każdej kategorii zmiennej jakościowej. *Wymaga*: Zmiennej jakościowej i zmiennej ciągłej.		**Słupkowy Liczebności** Wyświetla w postaci słupków proporcję wierszy/obserwacji w każdej kategorii zmiennej jakościowej. Do utworzenia tego wykresu można także użyć węzła Wykres rozkładu. Węzeł ten udostępnia dodatkowe opcje. Więcej informacji można znaleźć w temacie “Węzeł rozkładu” na stronie 237. *Wymaga*: Pojedynczej zmiennej jakościowej.

Tabela 36. Dostępne typy wykresów (kontynuacja).

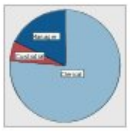
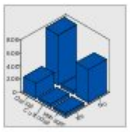
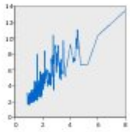
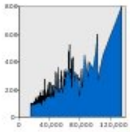
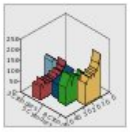
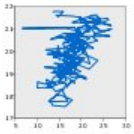
Ikona wykresu	Opis	Ikona wykresu	Opis
	**Wykres kołowy** Oblicza sumę ciągłej zmiennej liczbowej i wyświetla w postaci wycinków koła rozkład części tej sumy w każdej kategorii zmiennej jakościowej. *Wymaga*: Zmiennej jakościowej i zmiennej ciągłej.		**Kołowy Liczebności** Wyświetla w postaci wycinków koła proporcję wierszy/obserwacji w każdej kategorii zmiennej jakościowej. *Wymaga*: Pojedynczej zmiennej jakościowej.
	**Wykres słupkowy trójwymiarowy** Oblicza statystyki podsumowujące dla ciągłej zmiennej liczbowej i wyświetla wyniki dla przecięcia kategorii między dwoma zmiennymi jakościowymi. *Wymaga*: Pary składającej się ze zmiennej jakościowej i zmiennej ciągłej.		**Kołowy 3-W** Taki sam jak wykres kołowy, ale z dodatkowym efektem trójwymiarowości. *Wymaga*: Zmiennej jakościowej i zmiennej ciągłej.
	**Wykres liniowy** Oblicza statystyki podsumowujące dla pola dla każdej wartości kolejnego pola i rysuje linię łączącą wartości. Do utworzenia tego wykresu można także użyć węzła Wykres. Węzeł ten udostępnia dodatkowe opcje. Więcej informacji można znaleźć w temacie “Węzeł Rozrzutu” na stronie 226. *Wymaga*: Pary pól dowolnego typu.		**Warstwowy** Oblicza statystyki podsumowujące dla pola dla każdej wartości kolejnego pola i rysuje obszar łączący wartości. Różnica między wykresem liniowym a warstwowym jest niewielka: warstwa przypomina linię, ale przestrzeń pod nią jest wypełniona kolorem. W razie zastosowania sposobu prezentacji wykorzystującego kolor daje to prosty podział linii i zestawienie warstwy. *Wymaga*: Pary pól dowolnego typu.
	**Warstwowy 3-W** Wyświetla wartości jednego pola wykreślone względem wartości drugiego i podzielone według zmiennej jakościowej. Dla każdej kategorii wykreślany jest element warstwy. *Wymaga*: Zmiennej jakościowej i pary pól dowolnego typu.		**Ścieżkowy** Wyświetla wartości jednego pola wykreślone względem wartości drugiego pola. Wartości są połączone linią w kolejności, w jakiej pojawiają się w oryginalnym zbiorze danych. Uporządkowanie w kolejności jest główną różnicą między wykresem ścieżkowym a liniowym. *Wymaga*: Pary pól dowolnego typu.

Tabela 36. Dostępne typy wykresów (kontynuacja).

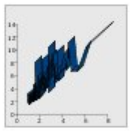
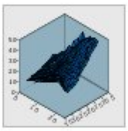
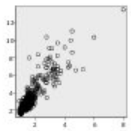
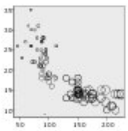
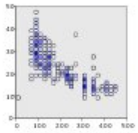
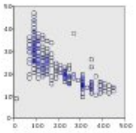
Ikona wykresu	Opis	Ikona wykresu	Opis
	**Wykres wstęgowy** Oblicza statystyki podsumowujące dla pola dla każdej wartości kolejnego pola i rysuje wstęgę łączącą wartości. Wstęga to zasadniczo linia z efektem trójwymiarowości. Nie jest prawdziwym wykresem 3-W. *Wymaga*: Pary pól dowolnego typu.		**Powierzchniowy** Wyświetla wartości trzech ciągłych pól wykreślonych względem wartości jeszcze jednego ciągłego przedziału liczbowego. Wartości są połączone powierzchnią. *Wymaga*: Trzech pól dowolnego typu.
	**Wykres rozrzutu** Wyświetla wartości jednego pola wykreślone względem wartości drugiego pola. Wykres ten może także zaznaczać relacje między polami (jeśli występuje). Do utworzenia tego wykresu rozrzutu można także użyć węzła Wykres. Węzeł ten udostępnia dodatkowe opcje. Więcej informacji można znaleźć w temacie “Węzeł Rozrzutu” na stronie 226. *Wymaga*: Pary pól dowolnego typu.		**Wykres bąbelkowy** Podobnie jak podstawowy wykres rozrzutu, wyświetla wartości jednego pola wykreślone względem wartości drugiego pola. Różnica polega na tym, że wartości trzeciego pola wykorzystuje się do zróżnicowania rozmiaru poszczególnych punktów. *Wymaga*: Trzech pól dowolnego typu.
	**Wykres rozrzutu z kategoryzacją** Podobnie jak podstawowy wykres rozrzutu, wyświetla wartości jednego pola wykreślone względem wartości drugiego pola. Różnica polega na tym, że podobne wartości są umieszczane w przedziałach, w grupach oraz że liczbę obserwacji w każdym kontenerze wskazuje się za pomocą koloru lub rozmiaru. *Wymaga*: Pary zmiennych ciągłych.		**Sześciokątny wykres rozrzutu z kategoryzacją** Patrz opis wykresu rozrzutu z kategoryzacją. Różnica polega na kształcie przedziałów, które mają kształt sześciokątów, nie zaś kół. Powstały sześciokątny wykres rozrzutu z kategoryzacją wygląda podobnie jak wykres rozrzutu z kategoryzacją. Liczba wartości w każdym przedziale będzie jednak różnić się między wykresami ze względu na kształt przedziałów. *Wymaga*: Pary zmiennych ciągłych.

Tabela 36. Dostępne typy wykresów (kontynuacja).

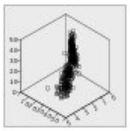
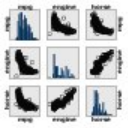
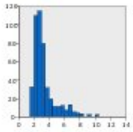
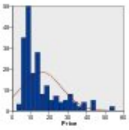
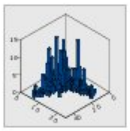
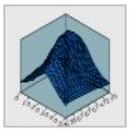
Ikona wykresu	Opis	Ikona wykresu	Opis
	**3-W wykres rozrzutu** Wyświetla wartości trzech pól wykreślonych względem siebie nawzajem. Wykres ten może także zaznaczać relacje między polami (jeśli występuje). Do utworzenia tego wykresu rozrzutu 3-W można także użyć węzła Wykres. Węzeł ten udostępnia dodatkowe opcje. Więcej informacji można znaleźć w temacie "Węzeł Rozrzutu" na stronie 226. *Wymaga*: Trzech pól dowolnego typu.		**Macierz rozrzutu (SPLOM)** Dla każdego pola wyświetla wartości jednego pola wykreślone względem wartości drugiego pola. Macierz rozrzutu przypomina tabelę wykresów rozrzutu. Zawiera także histogram każdego pola. *Wymaga*: Przynajmniej dwóch zmiennych ciągłych.
	**Histogram** Wyświetla rozkład częstości pola. Histogram pomaga określić typ rozkładu i sprawdzić, czy jest on skośny. Do utworzenia tego wykresu można także użyć węzła Wykres histogramu. Węzeł ten udostępnia dodatkowe opcje. Więcej informacji można znaleźć w temacie "Histogram — karta wykresu" na stronie 241. *Wymaga*: Pojedynczego pola dowolnego typu.		**Histogram z krzywą normalną** Wyświetla rozkład częstości zmiennej ciągłej z nałożoną krzywą rozkładu normalnego. *Wymaga*: Pojedynczej zmiennej ciągłej.
	**Histogram 3-W** Wyświetla rozkład częstości pary zmiennych ciągłych. *Wymaga*: Pary zmiennych ciągłych.		**Gęstość 3-W** Wyświetla rozkład częstości pary zmiennych ciągłych. Jest podobny do histogramu 3-W; jedyną różnicą jest przedstawienie rozkładu za pomocą płaszczyzny, nie zaś słupków. *Wymaga*: Pary zmiennych ciągłych.

Tabela 36. Dostępne typy wykresów (kontynuacja).

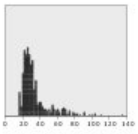
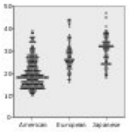
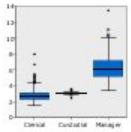
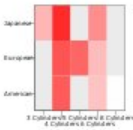
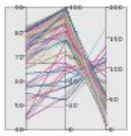
Ikona wykresu	Opis	Ikona wykresu	Opis
	**Wykres punktowy** Wyświetla poszczególne obserwacje/wiersze i zestawia je jako odrębne punkty danych na osi *x*. Wykres ten jest podobny do histogramu pod tym względem, że pokazuje rozkład danych, przedstawia jednak każdą obserwację/każdy wiersz, a nie zagregowaną liczność danego przedziału (przedział wartości). *Wymaga*: Pojedynczego pola dowolnego typu.		**Wykres punktowy 2-W** Wyświetla poszczególne obserwacje/wiersze i zestawia je jako odrębne punkty danych na osi *y* dla każdej kategorii zmiennej jakościowej. *Wymaga*: Zmiennej jakościowej i zmiennej ciągłej.
	**Wykres skrzynkowy** Oblicza pięć statystyk (minimum, pierwszy kwartyl, medianę, trzeci kwartyl i maksimum) dla zmiennej ciągłej dla każdej kategorii zmiennej jakościowej. Wyniki są wyświetlane jako elementy wykresu skrzynkowego/schematu. Wykresy skrzynkowe pomagają sprawdzić, jak rozkład danych ciągłych różni się w poszczególnych kategoriach. *Wymaga*: Zmiennej jakościowej i zmiennej ciągłej.		**Mapa natężeń** Oblicza statystyki podsumowujące dla zmiennej ciągłej dla przecięcia kategorii między dwoma zmiennymi jakościowymi. *Wymaga*: Pary składającej się ze zmiennej jakościowej i zmiennej ciągłej.
	**Równoległy** Tworzy równoległe osie dla każdego pola i wykreśla dla każdego wiersza/każdej obserwacji w danych linię między wartością pola. *Wymaga*: Przynajmniej dwóch zmiennych ciągłych.		**Kartogram liczebności** Oblicza liczebność każdej z kategorii zmiennej jakościowej (**Klucz danych**) i rysuje mapę, na której nasycenie koloru reprezentuje liczebność właściwości mapy odpowiadających kategoriom. *Wymaga*: Zmiennej jakościowej. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.

Tabela 36. Dostępne typy wykresów (kontynuacja).

Ikona wykresu	Opis	Ikona wykresu	Opis
	**Kartogram średnich/median/sum** Oblicza średnią, medianę lub sumę zmiennej ciągłej (**Kolor**) dla każdej kategorii zmiennej jakościowej (**Klucz danych**) i rysuje mapę, używając nasycenia koloru do przedstawienia wyliczonych statystyk we właściwościach mapy odpowiadających kategoriom. *Wymaga*: Zmiennej jakościowej i zmiennej ciągłej. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.		**Kartogram wartości** Rysuje mapę, w której kolor reprezentuje wartości zmiennej jakościowej (**Kolor**) dla właściwości mapy, które odpowiadają wartościom zdefiniowanym przez inną zmienną jakościową (**Klucz danych**). Jeśli dla każdej właściwości jest wiele wartości jakościowych pola Kolor, używana jest wartość modalna. *Wymaga*: Pary zmiennych jakościowych. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.
	**Współrzędne na kartogramie liczebności** Podobne do kartogramu liczebności, poza tym, że istnieją dwie dodatkowe zmienne ciągłe (**Długość** i **Szerokość**) oznaczające współrzędne punktów rysowania kartogramu. *Wymaga*: Zmiennej jakościowej i pary zmiennych ciągłych. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.		**Współrzędne Kartogramu średnich/median/sum** Podobne do Kartogramu średnich/median/sum, poza tym, że istnieją dwie dodatkowe zmienne ciągłe (**Długość** i **Szerokość**) oznaczające współrzędne punktów rysowania kartogramu. *Wymaga*: Zmiennej jakościowej i trzech zmiennych ciągłych. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.
	**Współrzędne na kartogramie wartości** Podobne do kartogramu wartości, poza tym, że istnieją dwie dodatkowe zmienne ciągłe (**Długość** i **Szerokość**) oznaczające współrzędne punktów rysowania kartogramu. *Wymaga*: Pary zmiennych jakościowych i pary zmiennych ciągłych. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.		**Słupki liczebności na mapie** Oblicza proporcję wierszy/obserwacji w każdej kategorii zmiennej jakościowej (**Kategorie**) dla każdej właściwości mapy (**Klucz danych**) i rysuje mapę oraz wykresy słupkowe w środku każdej właściwości mapy. *Wymaga*: Pary zmiennych jakościowych. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.

Tabela 36. Dostępne typy wykresów (kontynuacja).

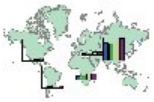
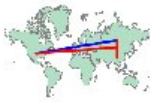
Ikona wykresu	Opis	Ikona wykresu	Opis
	**Słupki na mapie** Oblicza statystykę podsumowującą dla zmiennej ciągłej (**Wartości**) i wyświetla wyniki dla każdej kategorii zmiennej jakościowej (**Kategorie**) dla każdej właściwości mapy (**Klucz danych**) jako wykresy słupkowe umieszczone w środku każdej właściwości mapy. *Wymaga*: Pary składającej się ze zmiennej jakościowej i zmiennej ciągłej. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.		**Koło liczebności mapy** Oblicza proporcję wierszy/obserwacji w każdej kategorii zmiennej jakościowej (**Kategorie**) dla każdej właściwości mapy (**Klucz danych**) i rysuje mapę oraz proporcje jako wycinki koła w środku każdej właściwości mapy. *Wymaga*: Pary zmiennych jakościowych. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.
	**Wykres kołowy na mapie** Oblicza sumę zmiennej ciągłej (**Wartości**) dla każdej kategorii zmiennej jakościowej (**Kategorie**) dla każdej właściwości mapy (**Klucz danych**) i rysuje mapę oraz sumę wycinków wykresu kołowego w środku każdej właściwości mapy. *Wymaga*: Pary składającej się ze zmiennej jakościowej i zmiennej ciągłej. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.		**Wykres liniowy na mapie** Oblicza statystykę podsumowującą dla zmiennej ciągłej (**Y**) dla każdej wartości innej zmiennej (**X**) dla każdej właściwości mapy (**Klucz danych**) i rysuje mapę oraz wykresy liniowe łączące wartości w środku każdej właściwości mapy. *Wymaga*: Zmiennej jakościowej i pary pól dowolnego typu. Plik mapy, którego klucz pasuje do kategorii **Klucza danych**.
	**Współrzędne na mapie odniesienia** Rysuje mapę i punkty, używając zmiennych ciągłych (**Długości** i **Szerokości**), identyfikujących współrzędnych dla punktów. *Wymaga*: Pary pól przedziału. Plik mapy.		**Strzałki na mapie odniesienia** Rysuje mapę i strzałki, używając dla każdej strzałki zmiennych ciągłych identyfikujące punkty początkowe (**Początek dług** i **Początek szer**) oraz punkty końcowe (**Końcowa dług** i **Końcowa szer**). Każdy rekord/obserwacja wyników danych na strzałce na mapie. *Wymaga*: Czterech zmiennych ciągłych. Plik mapy.

Tabela 36. Dostępne typy wykresów (kontynuacja).

Ikona wykresu	Opis	Ikona wykresu	Opis
	**Mapa nakładania się punktów** Rysuje mapę odniesienia i nakłada na nią inną mapę punktową, z właściwościami punktów pokolorowanymi według zmiennej jakościowej (**Kolor**). *Wymaga*: Pary zmiennych jakościowych. Plik mapy punktowej, którego klucz pasuje do kategorii **Klucza danych**. Plik mapy odniesienia.		**Mapa nakładania się wielokątów** Rysuje mapę odniesienia i nakłada na nią inną mapę wielokątną, z właściwościami wielokątów pokolorowanymi według zmiennej jakościowej (**Kolor**). *Wymaga*: Pary zmiennych jakościowych. Plik mapy wielokąta, którego klucz pasuje do kategorii **Klucza danych**. Plik mapy odniesienia.
	**Mapa nakładania się linii** Rysuje mapę odniesienia i nakłada na nią inną mapę liniową, z właściwościami linii pokolorowanymi według zmiennej jakościowej (**Kolor**). *Wymaga*: Pary zmiennych jakościowych. Plik mapy liniowej, którego klucz pasuje do kategorii **Klucza danych**. Plik mapy odniesienia.		

Tworzenie wizualizacji map

W przypadku wielu wizualizacji użytkownik może wybrać tylko spośród dwu opcji: pól (zmiennych) zainteresowania oraz szablonu do wizualizacji tych pól. Nie wymaga się żadnych dodatkowych wyborów ani działań. Wizualizacje map wymagają co najmniej jednego dodatkowego kroku: wyboru pliku map definiującego informacje geograficzne potrzebne do wizualizacji map.

Poniżej znajdują się podstawowe kroki tworzenia prostej mapy:

1. W zakładce Podstawowe wybierz pola zainteresowania. Więcej informacji na temat typu i liczby pól wymaganych dla różnych wizualizacji map można znaleźć w “Dostępne wbudowane typy wizualizacji Graphboard ” na stronie 199.
2. Wybierz szablon mapy.
3. Kliknij kartę Szczegółowe.
4. Sprawdź, czy **Klucz danych** i inne wymagane rozwijane listy są ustawione na poprawne pola.
5. W grupie Pliki map kliknij **Wybierz plik mapy**.
6. Skorzystaj z okna dialogowego Wybierz mapy, aby wybrać plik mapy oraz klucz mapy. Wartości klucza mapy muszą zgadzać się z wartościami dla pola określonego przez **Klucz danych**. Możesz użyć przycisku **Porównaj**, aby porównać te wartości. Jeśli wybierzesz nakładanie szablonów map, będziesz także potrzebować mapy odniesienia. Mapa odniesienia nie jest powiązana z danymi. Jest ona używana jako tło dla mapy głównej. Aby uzyskać więcej informacji na temat okna dialogowego Wybierz mapy, patrz “Wybór plików mapy do Wizualizacji mapy” na stronie 198.
7. Kliknij przycisk **OK**, aby zamknąć okno dialogowe Wybierz mapy.
8. W opcji Wyboru szablonu wizualizacji danych kliknij **Uruchom**, aby stworzyć wizualizację mapy.

Przykłady węzła Graphboard

Ta część zawiera kilka różnych przykładów demonstrujących dostępne opcje. Przykłady te dostarczają także informacje pozwalające zinterpretować powstałe wizualizacje.

Przykłady te wykorzystują strumień o nazwie *graphboard.str*, który odwołuje się do plików danych: *employee_data.sav*, *customer_subset.sav* i *worldsales.sav*. Te pliki są dostępne w folderze *Demos* w instalacji klienta programu IBM SPSS Modeler. Można do niego uzyskać dostęp za pomocą grupy programów IBM SPSS Modeler w menu Start systemu Windows. Plik *graphboard.str* znajduje się w folderze *streams*.

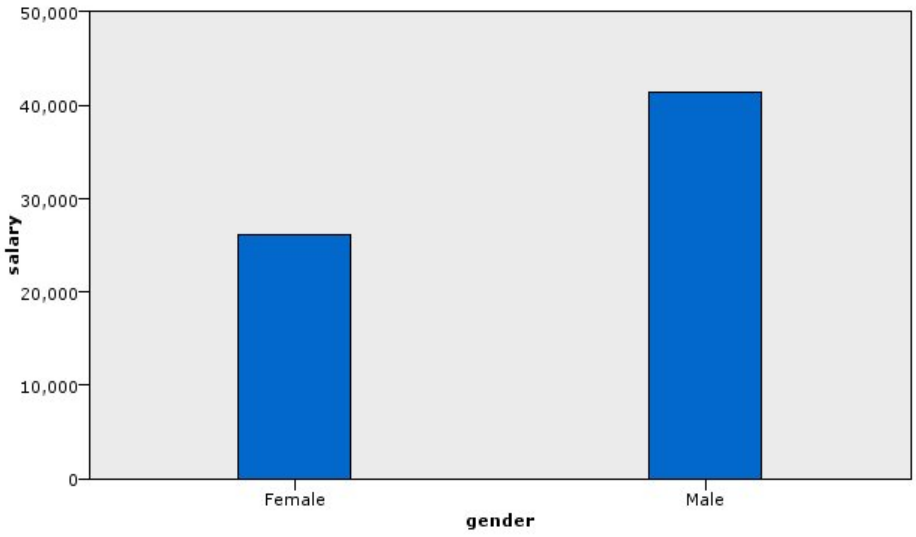
Zaleca się czytanie przykładów w kolejności, w jakiej zostały przedstawione. Kolejne przykłady wykorzystują poprzednie.

Przykład: Wykres słupkowy ze statystyką podsumowującą

Utworzymy wykres słupkowy będący podsumowaniem ciągłego pola liczbowego/zmiennej liczbowej dla każdej kategorii zbioru/zmiennej kategorialnej. Wykres będzie przedstawiać średnie wynagrodzenie mężczyzn i kobiet.

Ten i kilka kolejnych przykładów wykorzystują plik *Employee data (dane pracownika)*, który jest hipotetycznym zbiorem danych zawierającym informacje o pracownikach pewnej firmy.

1. Dodaj węzeł źródłowy Plik Statistics, który wskazuje na plik *employee_data.sav*.
2. Dodaj węzeł Graphboard i otwórz go do edycji.
3. Na karcie Opcje podstawowe wybierz *Płeć* i *Pensja bieżąca* . (Aby zaznaczyć wiele elementów pól/zmiennych, kliknij je, trzymając wciśnięty klawisz Ctrl).
4. Wybierz opcję **Słupkowy**.
5. Z listy rozwijanej Podsumowanie wybierz pozycję **Średnia**.
6. Kliknij przycisk **Uruchom**.
7. Na ekranie z wynikami kliknij przycisk „Wyświetlaj pola i etykiety wartości" znajdujący się na pasku narzędzi (drugi w grupie dwóch przycisków na środku paska narzędzi).



Rysunek 8. Wykres słupkowy ze statystyką podsumowującą

Możemy zauważyć, że:

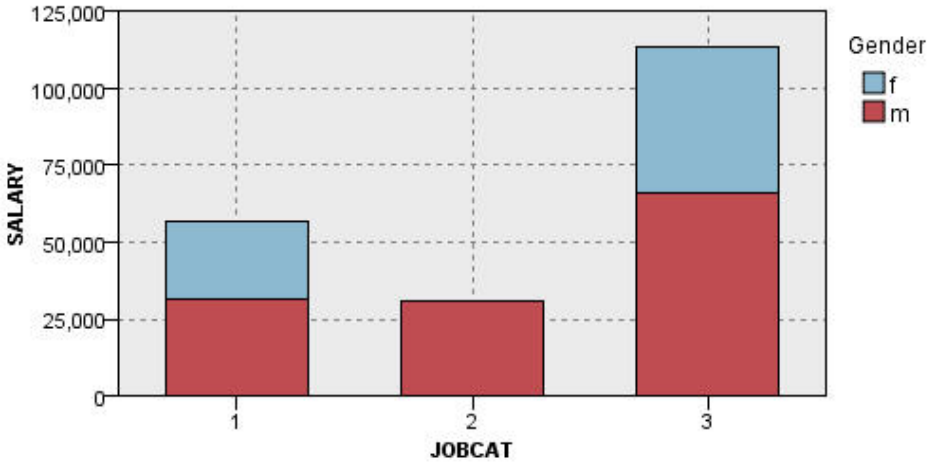
- Na podstawie wysokości słupków widać wyraźnie, że średnie wynagrodzenie mężczyzn jest wyższe od średniego wynagrodzenia kobiet.

Przykład: Zestawiony wykres słupkowy ze statystyką podsumowującą

Utworzymy teraz zestawiony wykres słupkowy, aby sprawdzić, czy różnica w wysokości średniego wynagrodzenia między mężczyznami i kobietami zależy od rodzaju pracy. Być może kobiety średnio częściej niż mężczyźni wykonują określone rodzaje pracy.

Uwaga: Ten przykład korzysta z wartości *Employee data* (Dane pracownika).

1. Dodaj węzeł Graphboard i otwórz go do edycji.
2. Na karcie podstawowej wybierz opcje *Employment Category* (Kategoria zatrudnienia) i *Current Salary* (Bieżące wynagrodzenie). (Aby zaznaczyć wiele elementów pól/zmiennych, kliknij je, trzymając wciśnięty klawisz Ctrl).
3. Wybierz opcję **Słupkowy**.
4. Z listy Podsumowanie wybierz pozycję **Średnia**.
5. Kliknij kartę Szczegółowe. Zwróć uwagę, że są tu uwzględnione opcje wybrane na poprzedniej karcie.
6. W grupie Opcjonalne sposoby prezentacji z listy rozwijanej Kolor wybierz pozycję *płeć*.
7. Kliknij przycisk **Uruchom**.



Rysunek 9. Zestawiony wykres słupkowy

Możemy zauważyć, że:

- Różnica między średnimi wynagrodzeniami dla każdego typu pracy nie jest aż tak duża jak w przypadku wykresu słupkowego, który porównywał średnie wynagrodzenia wszystkich mężczyzn i kobiet. Być może liczba mężczyzn i kobiet w każdej grupie jest różna. Można to sprawdzić, tworząc wykres słupkowy liczebności.
- Bez względu na rodzaj pracy średnie wynagrodzenie mężczyzn zawsze jest większe od średniego wynagrodzenia kobiet.

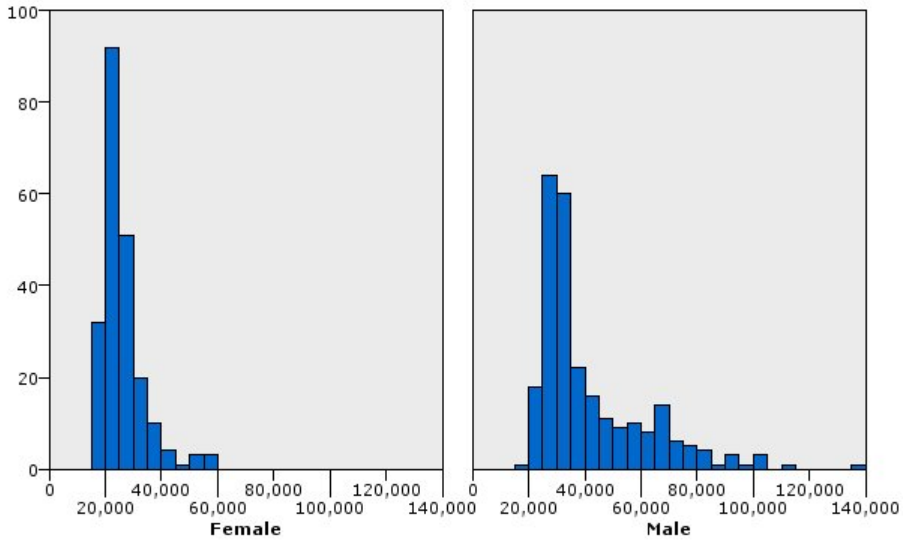
Przykład: Histogram panelowany

Utworzymy histogram panelowany według płci, co pozwoli porównać rozkłady częstości wynagrodzeń mężczyzn i kobiet. Rozkład częstości pokazuje, ile obserwacji/wierszy należy do konkretnych przedziałów wynagrodzenia. Histogram panelowany pomoże nam bliżej przeanalizować różnice wynagrodzeń między płciami.

Uwaga: Ten przykład korzysta z wartości Dane pracownika.

1. Dodaj węzeł Graphboard i otwórz go do edycji.
2. Na karcie Opcje podstawowe wybierz opcje *Pensja bieżąca*.
3. Wybierz pozycję **Histogram**.
4. Kliknij kartę Szczegółowe.
5. W grupie Panele i animacje z listy rozwijanej Paneluj według wybierz pozycję *płeć*.

6. Kliknij przycisk **Uruchom**.



Rysunek 10. Histogram panelowany

Możemy zauważyć, że:

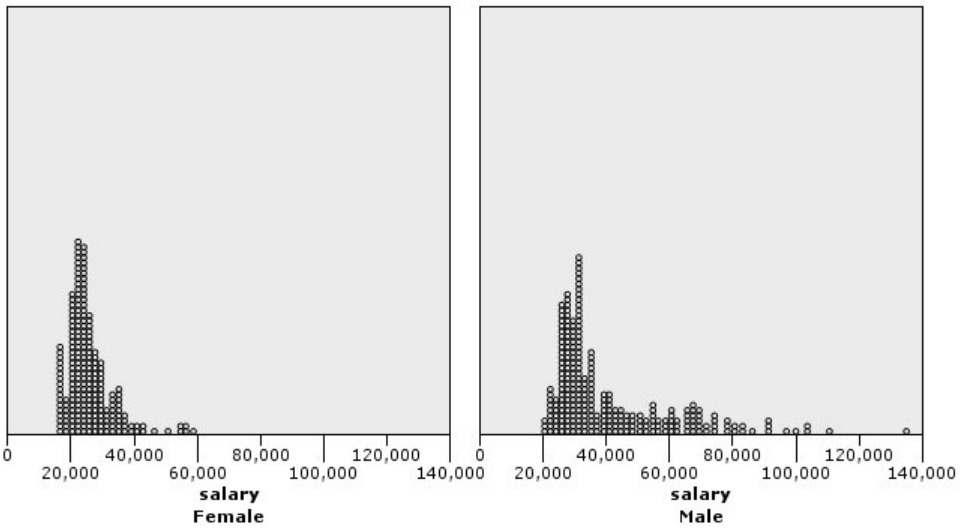
- Żaden z rozkładów częstości nie jest rozkładem normalnym. Oznacza to, że histogramy nie przypominają kształtem dzwonu, jak w przypadku normalnego rozkładu danych.
- Wyższe słupki znajdują się po lewej stronie każdego wykresu. Zarówno w przypadku mężczyzn, jak i kobiet, więcej badanych otrzymuje niższe wynagrodzenie.
- Rozkłady częstości wynagrodzenia mężczyzn i kobiet różnią się. Zwróć uwagę na kształt histogramów. Wśród otrzymujących wyższe wynagrodzenia jest więcej mężczyzn niż kobiet.

Przykład: Panelowany wykres punktowy

Wykres punktowy, podobnie jak histogram, przedstawia rozkład ciągłego przedziału liczbowego. W odróżnieniu od histogramu, który przedstawia liczności przedziałów danych umieszczonych w przedziałach, na wykresie punktowym pokazany jest każdy wiersz/każda obserwacja z danych. Wykres punktowy jest zatem bardziej szczegółowy niż histogram. W rzeczywistości wykorzystanie wykresu punktowego może być najlepszym punktem wyjścia podczas analizowania rozkładów częstości.

Uwaga: Ten przykład korzysta z wartości Dane pracownika.

1. Dodaj węzeł Graphboard i otwórz go do edycji.
2. Na karcie Opcje podstawowe wybierz opcje *Pensja bieżąca*.
3. Wybierz pozycje **Wykres punktowy**.
4. Kliknij kartę Szczegółowe.
5. W grupie Panele i animacje z listy rozwijanej Paneluj według wybierz pozycję *płeć*.
6. Kliknij przycisk **Uruchom**.
7. Zmaksymalizuj wyświetlone okno wyników, aby zobaczyć wyraźniej wykres.



Rysunek 11. Panelowany wykres punktowy

W porównaniu z histogramem (patrz “Przykład: Histogram panelowany” na stronie 208) można zauważyć, że:

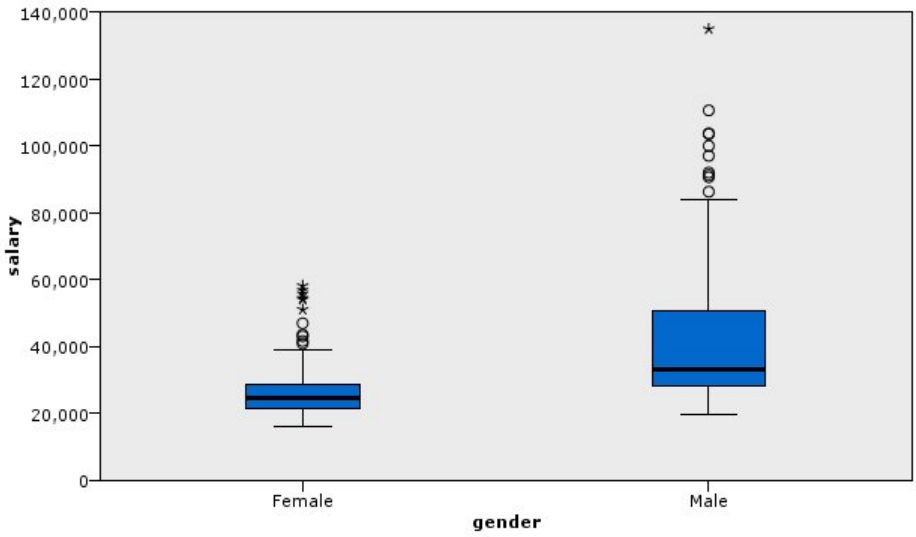
- Szczyt (20 000) pojawiający się na histogramie dotyczącym kobiet wygląda mniej dramatycznie na wykresie punktowym. Wokół tej wartości koncentruje się wiele obserwacji/wierszy, ale większość z nich jest bliższa wartości 25 000. Taki poziom szczegółowości nie jest widoczny na histogramie.
- Choć histogram dotyczący mężczyzn sugeruje, że średnie wynagrodzenie mężczyzn stopniowo spada po 40 000, wykres punktowy pokazuje, że po przekroczeniu tej wartości rozkład jest prawie jednostajny do wartości 80 000. Dla każdej wartości wynagrodzenia w tym przedziale istnieje co najmniej trzech mężczyzn, którzy otrzymują takie wynagrodzenie.

Przykład: Wykres skrzynkowy

Wykres skrzynkowy to kolejny sposób wizualizacji przydatny do sprawdzania rozkładu danych. Wykres skrzynkowy zawiera kilka miar statystycznych, które przeanalizujemy po utworzeniu wizualizacji.

Uwaga: Ten przykład korzysta z wartości Dane pracownika.

1. Dodaj węzeł Graphboard i otwórz go do edycji.
2. Na karcie Opcje podstawowe wybierz *Płeć* i *Pensja bieżąca* . (Aby zaznaczyć wiele elementów pól/zmiennych, kliknij je, trzymając wciśnięty klawisz Ctrl).
3. Wybierz pozycję **Wykres skrzynkowy**.
4. Kliknij przycisk **Uruchom**.



Rysunek 12. Wykres skrzynkowy

Przeanalizujmy różne części wykresu skrzynkowego:

- Ciemna linia pośrodku prostokątów to mediana *wynagrodzenia*. Połowa obserwacji/wierszy ma wartość wyższą niż mediana, druga połowa zaś wartość niższą. Podobnie jak średnia, mediana jest miarą tendencji centralnej. W odróżnieniu od średniej, w mniejszym stopniu wpływają na nią obserwacje/wiersze o wartościach skrajnych. W tym przykładzie mediana jest niższa od średniej (porównaj z “Przykład: Wykres słupkowy ze statystyką podsumowującą” na stronie 207). Różnica między średnią a medianą oznacza, że istnieje kilka obserwacji/wierszy z wartościami skrajnymi, które podnoszą wartość średniej. Oznacza to, że istnieje kilku pracowników, którzy otrzymują wysokie wynagrodzenia.
- Dolna krawędź prostokąta oznacza 25. percentyl. Dwadzieścia pięć procent obserwacji/wierszy ma wartości poniżej 25. percentyla. Górna krawędź prostokąta oznacza 75. percentyl. Dwadzieścia pięć procent obserwacji/wierszy ma wartości powyżej 75. percentyla. Oznacza to, że 50% obserwacji/wierszy leży w obrębie prostokąta. Prostokąt jest znacznie niższy w przypadku kobiet niż w przypadku mężczyzn. Jest to jeden z czynników wskazujących na to, że *wynagrodzenie* kobiet jest mniej zróżnicowane niż wynagrodzenie mężczyzn. Górna i dolna część prostokąta często określane są mianem **zawiasów**.
- Słupki w kształcie litery T, które wystają poza prostokąty, to **ogrodzenia wewnętrzne** lub **wąsy**. Ich wysokość jest równa 1,5 wysokości prostokąta lub, jeśli żadna obserwacja/żaden wiersz nie ma wartości w tym przedziale, wartości minimalnej albo maksymalnej. Jeśli dane mają rozkład normalny, oczekuje się, że około 95% danych znajdzie się między ogrodzeniami wewnętrznymi. W tym przykładzie ogrodzenia wewnętrzne wystają mniej w przypadku kobiet niż mężczyzn, co także wskazuje na mniejsze zróżnicowanie *wynagrodzenia* kobiet niż mężczyzn.
- Punkty to **wartości skrajne**. Definiuje się je jako wartości, które nie mieszczą się w ogrodzeniach zewnętrznych. Wartości skrajne to wartości ekstremalne. Gwiazdki oznaczają **ekstremalne wartości skrajne**. Reprezentują one obserwacje/wiersze, których wartości ponadtrzykrotnie przekraczają wysokość prostokątów. Zarówno w przypadku mężczyzn, jak i kobiet istnieje kilka wartości skrajnych. Należy pamiętać, że średnia jest wyższa od mediany. O wysokości średniej decydują właśnie te wartości skrajne.

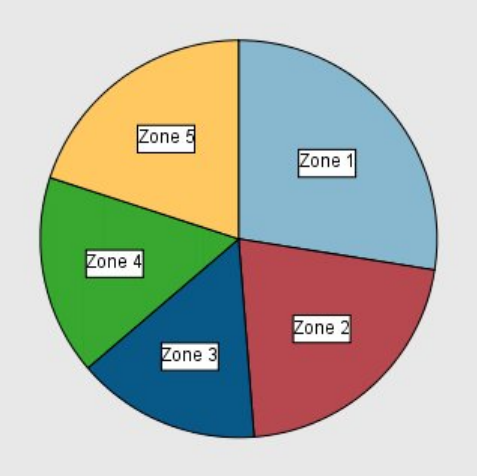
Przykład: Wykres kołowy

Wykorzystamy teraz inny zbiór danych, aby eksplorować parę innych typów wizualizacji. Zbiór danych to plik *customer_subset*, hipotetyczny plik danych, który zawiera informacje o klientach.

Utworzymy najpierw wykres kołowy, aby sprawdzić udziały klientów w różnych regionach geograficznych.

1. Dodaj węzeł źródłowy Plik Statistics, który wskazuje na plik *customer_subset.sav*.
2. Dodaj węzeł Graphboard i otwórz go do edycji.
3. Na karcie Opcje podstawowe wybierz opcję *Wskaźnik geograficzny*.

4. Wybierz pozycję **Wykres kołowy liczności**.
5. Kliknij przycisk **Uruchom**.



Rysunek 13. Wykres kołowy

Możemy zauważyć, że:

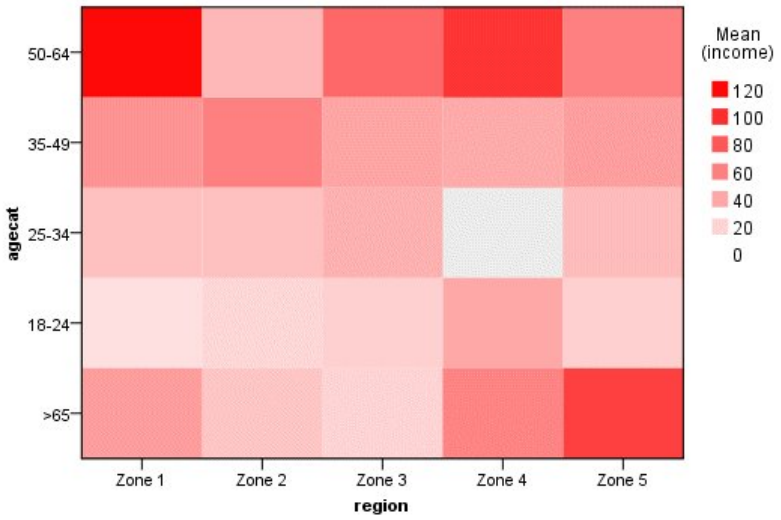
- Strefa 1 ma więcej klientów niż każda inna strefa.
- Dla pozostałych stref klienci są rozdzieleni równomiernie.

Przykład: Mapa natężeń

Stworzymy teraz jakościową mapę natężeń, aby sprawdzić średni przychód dla klientów w różnych regionach geograficznych i różnych grupach wiekowych.

Uwaga: Ten przykład korzysta z *customer_subset*.

1. Dodaj węzeł Graphboard i otwórz go do edycji.
2. Na karcie Opcje podstawowe wybierz w podanej poniżej kolejności: *Wskaźnik geograficzny*, *Kategoria wieku* oraz *Przychód na gospodarstwo podany w tysiącach*. (Aby zaznaczyć wiele elementów pól/zmiennych, kliknij je, trzymając wciśnięty klawisz Ctrl).
3. Wybierz opcję **Mapa natężeń**.
4. Kliknij przycisk **Uruchom**.
5. W oknie z wynikami kliknij przycisk „Wyświetlaj pola i etykiety wartości" znajdujący się na pasku narzędzi (pierwszy z prawej w grupie dwóch przycisków na środku paska narzędzi).



Rysunek 14. Jakościowa mapa natężeń

Możemy zauważyć, że:

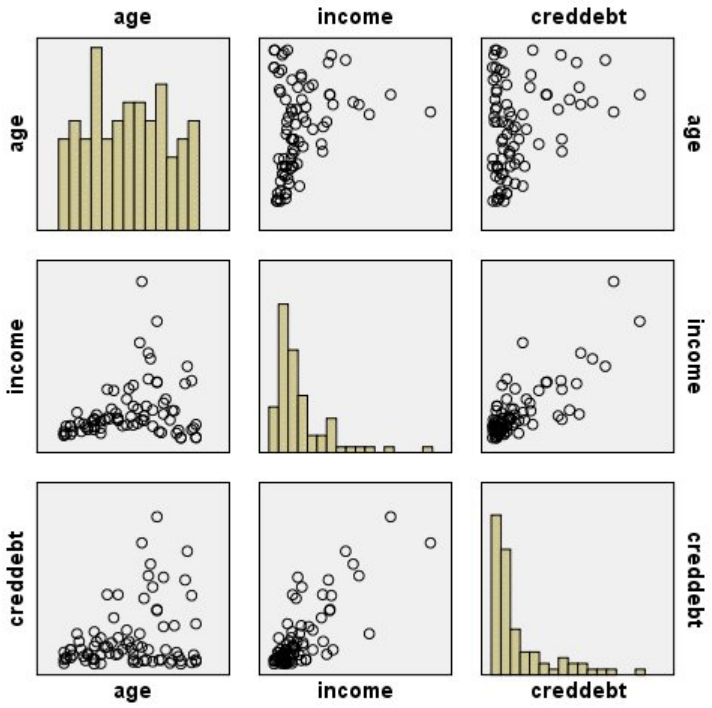
- Mapa natężeń przypomina tabelę, w której do przedstawienia wartości komórek użyto kolorów zamiast liczb. Jasny, głęboki czerwony oznacza najwyższą wartość, natomiast szary – niską wartość. Wartość każdej komórki to średnia pola ciągłego/zmiennej ciągłej dla każdej pary kategorii.
- Poza Strefą 2 i Strefą 5, grupa klientów w wieku od 50 do 64 lat ma większy średni przychód na gospodarstwo domowe niż inne grupy.
- W Strefie 4 nie ma klientów z przedziału wiekowego od 25 do 34 lat.

Przykład: Macierz rozrzutu (SPLOM)

Utworzymy macierz rozrzutu siedmiu różnych zmiennych, aby określić, czy między zmiennymi w zbiorze danych występują jakiekolwiek relacje.

Uwaga: Ten przykład korzysta z *customer_subset*.

1. Dodaj węzeł Graphboard i otwórz go do edycji.
2. Na karcie Opcje podstawowe wybierz *Wiek podany w latach*, *Przychód na gospodarstwo podany w tysiącach* oraz *Dług z kart kredytowych podany w tysiącach* . (Aby zaznaczyć wiele elementów pól/zmiennych, kliknij je, trzymając wciśnięty klawisz Ctrl).
3. Wybierz opcję **SPLOM**.
4. Kliknij przycisk **Uruchom**.
5. Zmaksymalizuj wyświetlone okno wyników, aby zobaczyć wyraźniej macierz.



Rysunek 15. Macierz rozrzutu (SPLOM)

Możemy zauważyć, że:

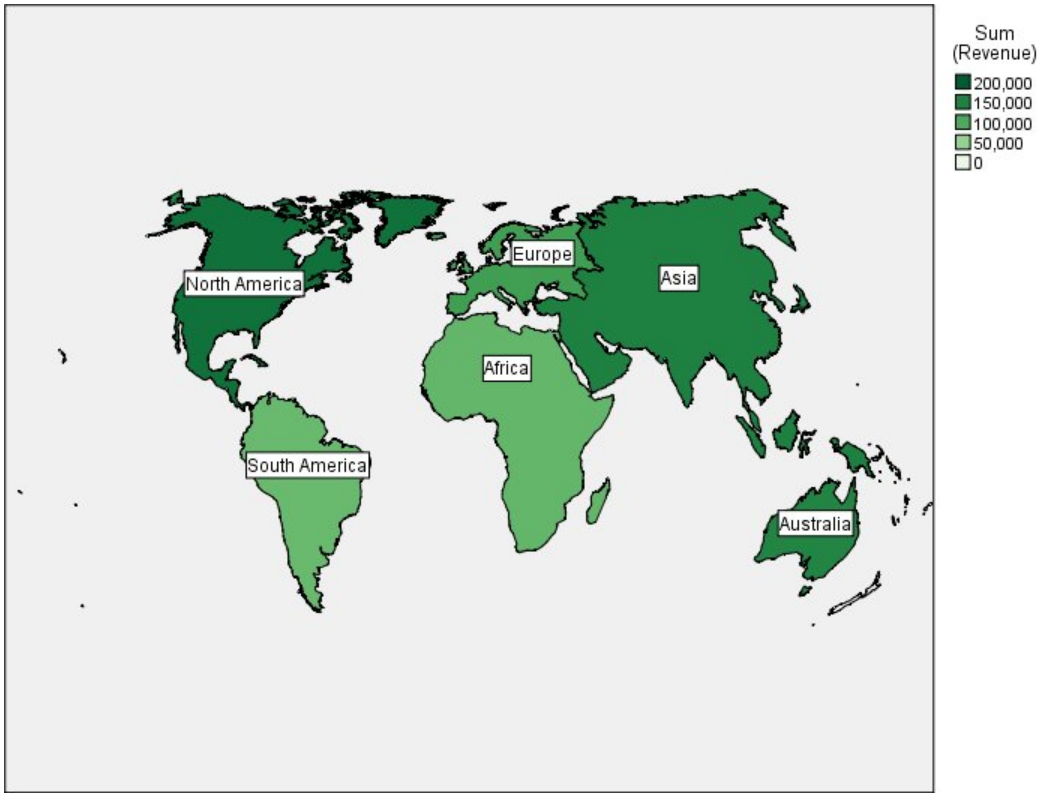
- Histogramy wyświetlone po przekątnej przedstawiają rozkład każdej zmiennej w macierzy rozrzutu. Histogram dotyczący *wieku* jest widoczny w lewej górnej komórce, histogram dotyczący *dochodów* – w środkowej komórce, a dotyczący *zdolności kredytowej* – w prawej dolnej komórce. Wygląda na to, że żadna ze zmiennych nie ma rozkładu normalnego. Żaden z histogramów nie przypomina więc kształtem dzwonu. Można także zauważyć, że histogramy dla *dochodu* i *zdolności kredytowej* są pozytywnie skośne.
- Wydaje się, że nie ma żadnego związku między *wiekiem* a pozostałymi zmiennymi.
- Istnieje liniowa relacja między *dochodem* a *zdolnością kredytową*. *Zdolność kredytowa* wzrasta wraz ze wzrostem *dochodu*. Można utworzyć pojedyncze wykresy rozrzutu tych zmiennych i innych powiązanych z nimi zmiennych, aby dokładniej przeanalizować te relacje.

Przykład: Kartogram (mapa kolorów) sum

Stworzymy teraz wizualizację mapy. Następnie, w kolejnym przykładzie, stworzymy zmienność dla tej wizualizacji. Zbiorem danych jest *worldsales* - plik z hipotetycznymi danymi zawierający przychód ze sprzedaży według kontynentu i produktu.

1. Dodaj węzeł Graphboard i otwórz go do edycji.
2. Na karcie Opcje podstawowe wybierz *Kontynent* i *Przychód*. (Aby zaznaczyć wiele elementów pól/zmiennych, kliknij je, trzymając wciśnięty klawisz Ctrl).
3. Wybierz **Kartogram sum**.
4. Kliknij kartę Szczegółowe.
5. W grupie Opcjonalne sposoby prezentacji z listy rozwijanej Opis danych wybierz pozycję *Kontynent*.
6. W grupie Pliki map kliknij **Wybierz plik mapy**.
7. W oknie dialogowym Wybierz mapy sprawdź, czy dla **Mapy** jest ustawiona wartość *Kontynenty* oraz czy **Klucz mapy** ma wartość *KONTYNENT*.
8. W grupach Mapa porównawcza i Wartości danych kliknij **Porównaj**, aby upewnić się, że klucze mapy zgadzają się z kluczami danych. W tym przykładzie wszystkie wartości kluczy danych mają odpowiadające im klucze mapy i zmienne. Można też zauważyć, że nie ma danych dla Oceanii.
9. W oknie dialogowym Wybierz mapy kliknij **OK**.

10. Kliknij przycisk **Uruchom**.



Rysunek 16. Kartogram sum

Na tej wizualizacji mapy można łatwo dostrzec, że przychód jest najwyższy w Ameryce Północnej, a najniższy w Ameryce Południowej i Afryce. Każdy kontynent posiada etykietę, ponieważ użyliśmy *Kontynentu* jako sposobu prezentacji etykiet danych.

Przykład: Wykres słupkowy na mapie

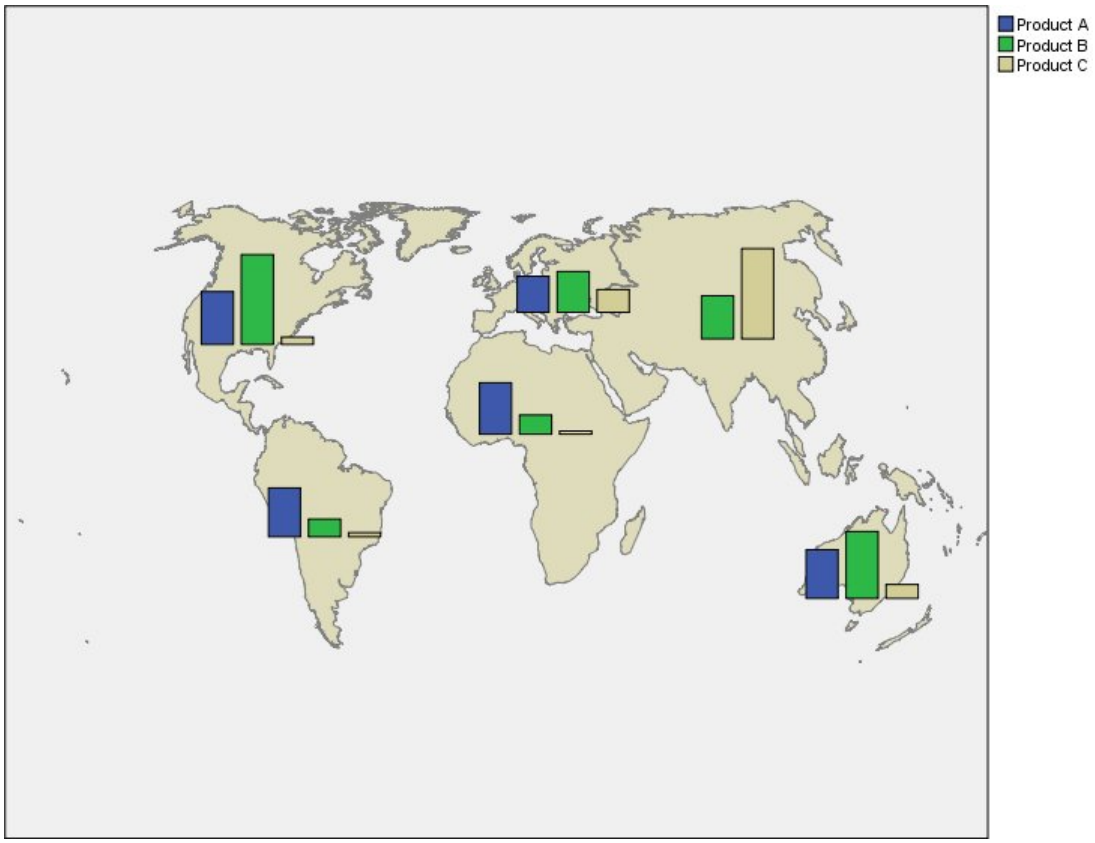
Przykład ten pokazuje, w jaki sposób przychód rozkłada się na produkty dla każdego z kontynentów.

Uwaga: Ten przykład korzysta z *worldsales* (sprzedaż światowa).

1. Dodaj węzeł Graphboard i otwórz go do edycji.
2. Na karcie Opcje podstawowe wybierz *Kontynent*, *Produkt* i *Przychód*. (Aby zaznaczyć wiele elementów pól/zmiennych, kliknij je, trzymając wciśnięty klawisz Ctrl).
3. Wybierz **Słupki na mapie**.
4. Kliknij kartę Szczegółowe.

   Gdy używasz więcej niż jednego pola określonego typu ważne jest, byś sprawdził, czy każde z pól jest przyporządkowane odpowiedniej szczelinie.
5. Z listy rozwijanej Kategorie wybierz pozycję *Produkt*.
6. Z listy rozwijanej Wartości wybierz pozycję *Przychód*.
7. Z listy rozwijanej Klucze danych wybierz pozycję *Kontynent*.
8. Z listy rozwijanej Podsumowanie wybierz pozycję *Suma*.
9. W grupie Pliki map kliknij **Wybierz plik mapy**.
10. W oknie dialogowym Wybierz mapy sprawdź, czy dla **Mapy** jest ustawiona wartość *Kontynenty* oraz czy **Klucz mapy** ma wartość *KONTYNENT*.

11. W grupach Mapa porównawcza i Wartości danych kliknij **Porównaj**, aby upewnić się, że klucze mapy zgadzają się z kluczami danych. W tym przykładzie wszystkie wartości kluczy danych mają odpowiadające im klucze mapy i zmienne. Można też zauważyć, że nie ma danych dla Oceanii.
12. W oknie dialogowym Wybierz mapy kliknij **OK**.
13. Kliknij przycisk **Uruchom**.
14. Zmaksymalizuj wyświetlone okno wyników, aby zobaczyć wyraźniej wyświetlane dane.



Rysunek 17. Wykres słupkowy na mapie

Możemy zauważyć, że:

- Rozkład całkowitego przychodu dla produktów jest bardzo podobny w Ameryce Południowej i Afryce.
- Wszędzie, za wyjątkiem Azji, *Produkt C* generuje najmniejszy przychód.
- W Azji brak przychodu pochodzącego od *Produktu A* lub jest on minimalny.

Karta Wygląd węzła Graphboard

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Opcje wyglądu ogólnego

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

Dobór próby. Podaj metodę dla większych zbiorów danych. Możesz określić maksymalną wielkość zbioru danych lub użyć domyślnej liczby rekordów. Wydajność dla dużych zbiorów danych ulega poprawie po wybraniu opcji **Próba**. Alternatywnie można wybrać wykreślanie wszystkich punktów danych przez wybór opcji **Użyj wszystkich danych**; należy jednak zwrócić uwagę, że może to drastycznie obniżyć wydajność oprogramowania.

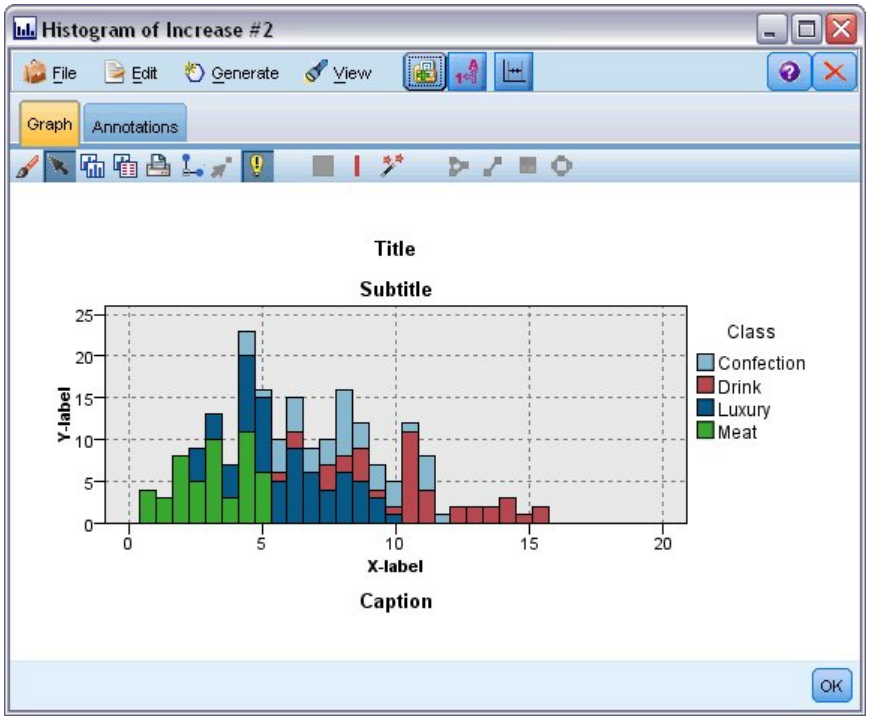
Opcje wyglądu arkusza stylów

Dostępne są dwa przyciski, które umożliwiają wybór wyświetlanych szablonów (i arkuszy stylów) wizualizacji:

Zarządzaj. Zarządzaj szablonami wizualizacji, arkuszami stylów i mapami na swoim komputerze. Możesz importować, eksportować, zmieniać nazwy i usuwać szablony wizualizacji, arkusze stylów i mapy na swoim komputerze lokalnym. Więcej informacji można znaleźć w temacie “Zarządzanie plikami szablonów, arkuszy stylów oraz map” na stronie 218.

Położenie. Zmień miejsce zapisu szablonów wizualizacji, arkuszy stylów i map. Bieżąca lokalizacja podana jest po prawej stronie przycisku. Więcej informacji można znaleźć w temacie “Ustawianie lokalizacji szablonów, arkuszy stylów i map”.

Poniższy przykład przedstawia, gdzie opcje wyglądu znajdują się na wykresie. (Uwaga: wszystkie te opcje znajdują się każdym wykresie).



Rysunek 18. Położenie różnych opcji wyglądu wykresu

Ustawianie lokalizacji szablonów, arkuszy stylów i map

Pliki szablonów wizualizacji, arkuszy stylów wizualizacji oraz map są przechowywane w określonym folderze lokalnym lub w repozytorium IBM SPSS Collaboration and Deployment Services Repository. Przy wyborze szablonów, arkuszy stylów oraz map wyświetlone zostaną tylko te, które są wbudowane w tej lokalizacji. Dzięki przechowywaniu wszystkich szablonów, arkuszy stylów oraz map w jednym miejscu aplikacje IBM SPSS mogą łatwo uzyskać do nich dostęp. Informacje na temat dodawania do tej lokalizacji dodatkowych szablonów, arkuszy stylów oraz map można znaleźć w części “Zarządzanie plikami szablonów, arkuszy stylów oraz map” na stronie 218.

Jak ustawić lokalizację plików szablonów, arkuszy stylów i map

1. W oknie dialogowym szablonu lub arkusza stylów kliknij pozycję **Lokalizacja...**, aby wyświetlić okno dialogowe Szablony, arkusze stylów i mapy.
2. Wybierz opcję dotyczącą domyślnej lokalizacji plików szablonów, arkuszy stylów oraz map:

   Komputer lokalny. Pliki szablonów, arkuszy stylów oraz map znajdują się w określonym folderze na komputerze lokalnym. W systemie Windows XP jest to folder *C:\Documents and Settings\<użytkownik>\Application Data\SPSSInc\Graphboard*. Tego folderu nie można zmienić.

   IBM SPSS Collaboration and Deployment Services Repository. Pliki szablonów, arkuszy stylów oraz map znajdują się we wskazanym przez użytkownika folderze w IBM SPSS Collaboration and Deployment Services Repository. Aby wskazać określony folder, kliknij przycisk **Folder**. Aby uzyskać więcej informacji, zobacz "Używanie IBM SPSS Collaboration and Deployment Services Repository jako lokalizacji plików szablonów, arkuszy stylów oraz map".
3. Kliknij przycisk **OK**.

Używanie IBM SPSS Collaboration and Deployment Services Repository jako lokalizacji plików szablonów, arkuszy stylów oraz map

Szablony i arkusze stylów wizualizacji można przechowywać w repozytorium IBM SPSS Collaboration and Deployment Services Repository. Tą lokalizacją jest określony folder w repozytorium IBM SPSS Collaboration and Deployment Services Repository. Jeśli jest on ustawiony jako lokalizacja domyślna, można wybierać dowolne szablony, arkusze stylów i mapy z tej lokalizacji.

Jak ustawić folder w IBM SPSS Collaboration and Deployment Services Repository jako lokalizację plików szablonów, arkuszy stylów i map

1. W oknie dialogowym zawierającym przycisk Lokalizacja kliknij przycisk **Lokalizacja...**.
2. Zaznacz IBM SPSS Collaboration and Deployment Services Repository.
3. Kliknij przycisk **Folder**.

   Uwaga: Jeśli nie nawiązano wcześniej połączenia z repozytorium IBM SPSS Collaboration and Deployment Services Repository, wyświetli się monit o podanie informacji o połączeniu.
4. W oknie dialogowym Wybieranie folderu wybierz folder, w którym są przechowywane szablony, arkusze stylów i mapy.
5. Opcjonalnie można wybrać etykietę z menu **Pobierz etykietę**. Zostaną wyświetlone tylko pliki szablonów, arkuszy stylów i map posiadające daną etykietę.
6. Jeśli szukasz folderu, który zawiera konkretny szablon lub arkusz stylów, możesz wyszukiwać plików szablonów, arkuszy stylów lub map za pomocą karty Wyszukiwanie. W oknie dialogowym Wybieranie folderu automatycznie zostanie wybrany folder, w którym znajduje się znaleziony plik szablonu, arkusza stylów lub mapy.
7. Kliknij przycisk **Wybierz folder**.

Zarządzanie plikami szablonów, arkuszy stylów oraz map

Szablonami, arkuszami stylów i plikami map można zarządzać lokalnie za pomocą okna dialogowego zarządzania szablonami, arkuszami stylów i mapami. Okno to umożliwia import, eksport, zmianę nazwy i usuwanie szablonów, arkuszy stylów wizualizacji i plików map znajdujących się na komputerze użytkownika.

Kliknij przycisk **Zarządzaj...** w jednym z okien dialogowych, w którym wybiera się szablony, arkusze stylów lub mapy.

Okno dialogowe zarządzania szablonami, arkuszami stylów oraz mapami

Na karcie Szablon wymienione są wszystkie szablony lokalne. Na karcie Arkusz stylów wymienione są wszystkie lokalne arkusze stylów i wyświetlone przykładowe wizualizacje z przykładowymi danymi. Można wybrać jeden z arkuszy stylów, aby zastosować jego style do przykładowych wizualizacji. Więcej informacji można znaleźć w temacie "Stosowanie arkuszy stylów" na stronie 296. Na karcie Mapa wymienione są wszystkie lokalne pliki map. Karta ta wyświetla także klucze mapy, łącznie z wartościami próbnymi, komentarzem informującym, czy zapewniono taką wartość podczas tworzenia mapy, oraz podglądem mapy.

Poniższe przyciski działają bez względu na to, która karta jest obecnie aktywna.

Importuj. Import pliku szablonu wizualizacji, arkusza stylów lub mapy z systemu plików. Import pliku szablonu, arkuszu stylów lub mapy powoduje, że dany element jest dostępny w aplikacji IBM SPSS. Jeśli inny użytkownik wyśle plik szablonu, arkusza stylów lub mapy, należy taki element zaimportować przed wykorzystaniem w aplikacji.

Eksportuj. Eksport pliku szablonu wizualizacji, arkusza stylów lub mapy z systemu plików. Eksportuj plik szablonu, arkusza stylów lub mapy, który ma zostać wysłany do innego użytkownika.

Zmień nazwę. Zmień nazwę wybranego pliku szablonu wizualizacji, arkusza stylów lub mapy. Nie można zmienić nazwy na taką, która jest już używana.

Eksportuj klucz mapy. Eksportuj klucze mapy do pliku z wartościami oddzielonymi przecinkami (CSV). Ten przycisk jest włączony tylko w karcie Mapa.

Usuń. Usuwa wybrany plik szablonu wizualizacji, arkusza stylów lub mapy. Możesz wybrać wiele plików szablonów, arkuszy stylów lub map za pomocą kliknięcia z wciśniętym klawiszem Ctrl. Operacji usuwania nie da się cofnąć, należy więc wykonywać ją ostrożnie.

Konwertowanie i dystrybucja plików kształtu map

Opcja Wyboru szablonu wizualizacji danych pozwala na stworzenie wizualizacji map na podstawie kombinacji szablonu wizualizacji oraz pliku SMZ. Pliki SMZ są podobne do plików kształtu ESRI (format pliku SHP) pod tym względem, że zawierają geograficzne informacje służące do rysowania map (na przykład granice państw), ale są zoptymalizowane na wizualizację map. Opcja Wyboru szablonu wizualizacji danych jest wstępnie zainstalowana z wybraną liczbą plików SMZ. Jeśli posiadasz istniejący plik kształtu ESRI, którego chcesz użyć do wizualizacji mapy, musisz go najpierw przekonwertować do pliku SMZ, używając narzędzia do konwersji map. Narzędzie konwersji map obsługuje pliki kształtu ESRI z punktami, łamanymi lub wielokątami (typami kształtów 1, 3 i 5) zawierające pojedynczą warstwę.

Oprócz konwersji plików kształtu ESRI, Narzędzie konwersji map pozwala na zmodyfikowanie poziomu szczegółowości mapy, zmianę etykiet właściwości, scalanie właściwości, przenoszenie właściwości oraz na wiele innych opcjonalnych zmian. Możesz także wykorzystać Narzędzie konwersji map do zmodyfikowania istniejącego pliku SMZ (włącznie z tymi wstępnie zainstalowanymi).

Edycja zainstalowanych wcześniej plików SMZ

1. Eksportowanie pliku SMZ z systemu Zarządzania. Więcej informacji można znaleźć w temacie “Zarządzanie plikami szablonów, arkuszy stylów oraz map” na stronie 218.
2. Użyj narzędzia do konwersji map, aby otworzyć i edytować wyeksportowany plik SMZ. Zaleca się, by zapisać plik pod inną nazwą. Więcej informacji można znaleźć w temacie “Używanie narzędzia do konwersji map” na stronie 220.
3. Importowanie zmodyfikowanego pliku SMZ do systemu Zarządzania. Więcej informacji można znaleźć w temacie “Zarządzanie plikami szablonów, arkuszy stylów oraz map” na stronie 218.

Dodatkowe zasoby dla plików map

Dane geoprzestrzenne w formacie pliku SHP, które mogą zostać użyte, aby wyjść na przeciw potrzebom mapowania, są dostępne w wielu źródłach prywatnych i publicznych. Jeśli szukasz darmowych danych, sprawdź na stronach lokalnych samorządów. Wiele formatów w niniejszym produkcie oparto na powszechnie dostępnych danych otrzymanych od GeoCommons ( ) oraz od Biura spisu powszechnego Stanów Zjednoczonych (http://www.census.gov).

WAŻNA UWAGA: Informacje dotyczące produktów nienależących do IBM pochodzą od producentów tych produktów, z ich opublikowanych oświadczeń lub innych powszechnie dostępnych źródeł. IBM nie testowało tych produktów i nie może potwierdzić dokładności działania, kompatybilności lub innych informacji związanych z produktami nienależącymi do IBM. Pytania dotyczące możliwości produktów innych podmiotów należy kierować do

dostawców tych produktów. Wszelkie wzmianki w tej publikacji na temat stron internetowych innych podmiotów zostały wprowadzone wyłącznie dla wygody użytkownika i w żadnym wypadku nie stanowią zachęty do ich odwiedzania. Materiały zawarte na tych stronach internetowych nie są częścią materiałów tego programu firmy IBM, chyba że zostało tak określone w pliku Uwag dołączonym do tego programu firmy IBM, i użytkownik korzysta z tych stron z materiałami na własne ryzyko.

Główne zagadnienia dotyczące map

Zrozumienie pewnych podstawowych zagadnień związanych z plikami kształtu pomoże Ci w efektywnym wykorzystaniu narzędzia do konwersji map.

Plik kształtu dostarcza geograficznych informacji do rysowania mapy. Narzędzie konwersji map obsługuje trzy typy plików kształtu:

- **Punkt.** Plik kształtu definiuje lokalizację punktów takich jak miasta.
- **Łamana.** Plik kształtu definiuje ścieżki i ich lokalizacje, jak na przykład rzeki.
- **Wielokąt.** Plik kształtu definiuje obszary ograniczone i ich lokalizacje, jak na przykład kraje.

Najczęściej można spotkać plik kształtu wielokąta. Kartogramy są tworzone z plików kształtu wielokąta. Kartogram używa koloru do przedstawienia wartości wewnątrz osobnych wielokątów (regionów). Pliki kształtu punktu i łamanej są zwykle nakładane dla plik kształtu wielokąta. Przykładem jest plik kształtu punktu miast w Stanach Zjednoczonych nałożony na plik kształtu wielokąta ze stanami.

Plik kształtu składa się z **właściwości**. Właściwości są indywidualnymi jednostkami geograficznymi. Na przykład: właściwościami mogą być kraje, stany, miasta itp. Plik kształtu zawiera także dane o właściwościach. Dane te są przechowywanie w **atrybutach**. Atrybuty są podobne do pól lub zmiennych w pliku danych. Istnieje przynajmniej jeden atrybut, będący dla właściwości **kluczem mapy**. Klucz mapy może być etykietą, jak na przykład państwo lub nazwa stanu. Klucz mapy jest tym, co połączysz do zmiennej/pola w pliku danych, aby stworzyć wizualizację mapy.

Zauważ, że w pliku SMZ będziesz mógł zachować tylko główny atrybut lub atrybuty. Narzędzie konwersji mapy nie obsługuje zapisywania atrybutów dodatkowych. Oznacza to, że jeśli chcesz agregować na różnych poziomach, będziesz musiał stworzyć wiele plików SMZ. Na przykład, jeśli chcemy zagregować stany i regiony USA, potrzebne są osobne pliki SMZ: jeden z kluczem identyfikującym stany i jeden z kluczem identyfikującym regiony.

Używanie narzędzia do konwersji map

Jak uruchomić Narzędzie konwersji map

Z menu wybierz:

Ustawienia > Narzędzie konwersji map

Narzędzie konwersji map posiada cztery główne ekrany (kroki). Jeden z kroków zawiera także kroki podrzędne służące bardziej szczegółowej kontroli nad edycją pliku mapy.

Krok 1 - Wybierz pliki docelowy i źródłowy

Najpierw musisz wybrać plik z mapą źródłową oraz plik docelowy na przekonwertowaną mapę. Dla pliku kształtu będziesz potrzebować zarówno pliku *.shp*, jak i *.dbf*.

Wybierz plik .shp (ESRI) lub .smz do konwersji. Poszukaj na swoim komputerze istniejącego pliku mapy. To jest plik, który przekonwertujesz i zapiszesz jako plik SMZ. Plik *.dbf*, który będzie plikiem kształtu, *musi* być zapisany w tej samej lokalizacji, z taką samą nazwą, jak plik *.shp*. Plik *.dbf* jest wymagany, ponieważ zawiera informacje o atrybutach dla pliku *.shp*.

Ustaw lokalizację docelową oraz nazwę pliku dla przekonwertowanego pliku mapy. Wpisz ścieżkę oraz nazwę pliku dla pliku SMZ, który zostanie utworzony z oryginalnego źródła mapy.

- **Importuj do funkcji Wyboru szablonu.** Oprócz zapisania pliku w systemie plików, możesz też opcjonalnie dodać mapę do listy zarządzania opcji Wyboru szablonu. Jeśli wybierzesz tę opcję, mapa będzie automatycznie dostępna w Wyborze szablonu dla produktów IBM SPSS zainstalowanych na Twoim komputerze. Jeśli nie zaimportujesz teraz do funkcji Wyboru szablonu, będziesz ją musiał zaimportować później ręcznie. Aby uzyskać więcej informacji na temat importowania map do systemu Zarządzania wyborem szablonu, patrz “Zarządzanie plikami szablonów, arkuszy stylów oraz map” na stronie 218.

Krok 2 - Wybierz klucz mapy

Musisz teraz wybrać, które pliki map mają zostać dołączone do pliku SMZ. Następnie możesz zmienić niektóre opcje, które będą miały wpływ na renderowanie mapy. Kolejne kroki Narzędzia do konwersji map zawierają podgląd mapy. Opcje renderowania, które wybierzesz, zostaną użyte do wygenerowania podglądu mapy.

Wybierz główny klucz mapy. Wybierz atrybut, który jest kluczem głównym służącym do identyfikacji i nadawania etykiet właściwościom w mapie. Na przykład, kluczem głównym mapy świata może być atrybut identyfikujący nazwy krajów. Klucz główny połączy także Twoje dane z właściwościami mapy, upewnij się więc, że wartości (etykiety) wybranego atrybutu zgadzają się z wartościami Twoich danych. Podczas wybierania atrybutu wyświetlane są przykładowe etykiety. Jeśli musisz je zmienić, będziesz mógł to zrobić w kolejnym kroku.

Wybierz dodatkowe klucze, które zostaną dołączone. Oprócz głównego klucza mapy, zaznacz wszystkie inne kluczowe atrybuty, które chcesz załączyć w generowanym pliku SMZ. Na przykład niektóre atrybuty mogą zawierać przetłumaczone etykiety. Jeśli oczekujesz danych zakodowanych w innych językach, możesz chcieć zachować te atrybuty. Zwróć uwagę, że możesz wybrać tylko te klucze dodatkowe, które reprezentują te same właściwości, co klucz główny. Przykładowo, jeśli kluczem głównym byłyby pełne nazwy stanów Stanów Zjednoczonych, możesz wybrać tylko te klucze alternatywne, które reprezentują stany Stanów Zjednoczonych, na przykład skróty nazw stanów.

Automatycznie wygładź mapę. Pliki kształtu z łamanymi zwykle zawierają zbyt wiele punktów danych i zbyt wiele szczegółów dla statystycznych wizualizacji map. Nadmiar szczegółu może rozpraszać i negatywnie wpływać na wydajność. Używając wygładzania, możesz zmniejszyć poziom szczegółu i uogólnić mapę. W rezultacie mapa będzie wyglądać ostrzej i będzie się renderować szybciej. Podczas automatycznego wygładzania mapy, maksymalnym kątem jest 15 stopni, a utrzymywanym procentem jest 99. Więcej informacji na temat tych ustawień można znaleźć w “Wygładź mapę” na stronie 222. Zwróć uwagę, że masz możliwość zastosować później, w kolejnym kroku, dodatkowego wygładzania.

W tej samej właściwości usuń granice między stykającymi się wielokątami. Niektóre funkcje mogą zawierać właściwości podrzędne, które posiadają granice wewnętrzne w stosunku do głównych właściwości, które Cię interesują. Na przykład, mapa świata z kontynentami może zawierać wewnętrzne granice krajów znajdujących się na poszczególnych kontynentach. Jeśli wybierzesz tę opcję, wewnętrzne granice nie pojawią się na mapie. W przypadku mapy świata z kontynentami wybór tej opcji usunie granice krajów, ale pozostawi granice kontynentów.

Krok 3 - Edytuj mapę

Teraz, gdy określono już podstawowe opcje mapy, możesz edytować te mniej szczegółowe. Modyfikacje te są opcjonalne. Niniejszy krok z Narzędziem do konwersji map prowadzi Cię przez powiązane zadania oraz wyświetla podgląd mapy tak, byś mógł zweryfikować swoje zmiany. W zależności od typu pliku kształtu (punkt, łamana lub wielokąt) oraz układu współrzędnych niektóre zadania mogą nie być dostępne.

Każde zadanie posiada po lewej stronie Narzędzia do konwersji map następujące często wykorzystywane opcje.

Pokaż etykiety na mapie. Domyślnie etykiety właściwości nie są pokazywane w podglądzie. Możesz zdecydować, czy wyświetlać etykiety. Choć etykiety mogą pomóc zidentyfikować właściwości, mogą też zakłócać bezpośrednie zaznaczenie podglądanej mapy. Włącz tę opcję, gdy jej potrzebujesz, na przykład gdy edytujesz etykiety właściwości.

Zmień kolor podglądu mapy. Domyślnie, podgląd mapy wyświetla obszary w stałych kolorach. Wszystkie właściwości mają ten sam kolor. Możesz wybrać gamę kolorów przydzielonych indywidualnym właściwościom map. Opcja ta może pomóc w rozróżnianiu różnych właściwości na mapie. Będzie szczególnie pomocna, gdy scalasz właściwości i chcesz zobaczyć, w jaki sposób nowe właściwości są reprezentowane w podglądzie.

Każde zadanie posiada po prawej stronie Narzędzia do konwersji map następujące często wykorzystywane opcje.

Cofnij. Kliknij przycisk **Cofnij**, aby wrócić do ostatniego stanu. Możesz cofnąć maksymalnie 100 zmian.

Wygładź mapę: Pliki kształtu z łamanymi zwykle zawierają zbyt wiele punktów danych i zbyt wiele szczegółów dla statystycznych wizualizacji map. Nadmiar szczegółu może rozpraszać i negatywnie wpływać na wydajność. Używając wygładzania, możesz zmniejszyć poziom szczegółu i uogólnić mapę. W rezultacie mapa będzie wyglądać ostrzej i będzie się renderować szybciej. Opcja ta jest niedostępna dla map punktu i łamanej.

Maks. kąt. Maksymalny kąt, którego wartość musi zawierać się w przedziale od 1 do 20, określa tolerancję dla wygładzających zestawów punktów, które są prawie liniowe. Większa wartość daje większą tolerancję dla liniowego wygładzania i pomija większą ilość punktów tworząc bardziej ogólną mapę. Aby zastosować wygładzanie liniowe, Narzędzie konwersji map sprawdza wewnętrzny kąt utworzony między każdym zestawie trzech punktów na mapie. Jeśli różnica 180 oraz kąta jest mniejsza niż określona wartość, Narzędzie konwersji map pominie środkowy punkt. Innymi słowy, Narzędzie konwersji map sprawdza, czy linia stworzona przez trzy punkty jest prawie prosta. Jeśli tak, to Narzędzie do konwersji map traktuje linię jak prostą między punktami końcowymi i pomija punkt środkowy.

Procent do zatrzymania. Procent do zatrzymania, który musi mieć wartość z przedziału od 90 do 100 określa ilość obszaru, do pozostawienia podczas wygładzania mapy. Ta opcja ma wpływ tylko na te funkcje, które mają wiele wielokątów, tak jak w przypadku właściwości zawierającej wyspy. Jeśli całkowity obszar właściwości minus wielokąt jest większy niż określony procent obszaru początkowego, Narzędzie konwersji map pominie wielokąt. Narzędzie konwersji map nigdy nie będzie usuwać wszystkich wielokątów danej właściwości. To jest, bez względu na zastosowaną liczbę wygładzania, zawsze pozostanie przynajmniej jeden wielokąt dla właściwości.

Po wybraniu maksymalnego kąta i procentu, który zostanie zachowany, kliknij **Zastosuj**. Podgląd uaktualni się o zmiany wygładzania. Jeśli konieczne jest ponowne wygładzenie mapy, powtarzaj operację, aż do uzyskanie żądnego poziomu gładkości. Zwróć uwagę, że nie ma limitu dla wygładzania. Jeśli będziesz powtarzać wygładzanie, osiągniesz punkt, w którym nie można zastosować już żadnego dodatkowego wygładzania.

Edytuj etykiety właściwości: Możesz według potrzeb edytować etykiety właściwości (np. aby pasowały do oczekiwanych danych) i zmieniać pozycję etykiet na mapie. Nawet jeśli wydaje Ci się, że nie musisz zmieniać etykiet, powinieneś sprawdzić je przed stworzeniem wizualizacji dla mapy. Ponieważ etykiety nie są domyślnie widoczne w podglądzie, możesz też zaznaczyć opcję **Pokaż etykiety na mapie**, aby je wyświetlić.

Klucze. Wybierz klucz zawierający etykiety właściwości, które chcesz sprawdzić i/lub edytować.

Właściwości. Lista ta wyświetla etykiety właściwości zawarte w wybranym kluczu. Kliknij dwukrotnie etykietę z listy, aby ją edytować. Jeśli etykiety są pokazane na mapie, możesz także dwa razy kliknąć etykietę właściwości bezpośrednio na podglądzie mapy. Jeśli chcesz porównać etykiety z rzeczywistym plikiem danych, kliknij **Porównaj**.

X/Y. Te pola tekstowe zawierają listę aktualnego punktu centralnego wybranej na mapie etykiety właściwości. Jednostki są wyświetlane we współrzędnych mapy. Mogą to być lokalne, współczynniki kartezjańskie (na przykład Krajowy układ współrzędnych) lub współrzędne geograficzne (gdzie **X** jest długością, a **Y** szerokością). Wpisz współrzędne dla nowej pozycji etykiety. Jeśli etykiety są pokazane, możesz także kliknąć i przeciągnąć etykietę na mapę i przesunąć ją. Pola tekstowe zostaną uaktualnione o nową pozycję.

Porównaj. Jeśli masz plik danych, który zawiera wartości danych, które mają pasować do etykiet właściwości dla konkretnego klucza, kliknij **Porównaj**, aby wyświetlić okno dialogowe Porównaj do Zewnętrznego źródła danych. W tym oknie dialogowym będziesz mógł otworzyć plik danych i porównać jego wartości bezpośrednio z tymi w etykietach właściwości klucza mapy.

Okienko dialogowe Porównaj do zewnętrznego źródła danych: Okno dialogowe Porównaj do zewnętrznego źródła danych pozwala CI na otwarcie pliku z wartościami oddzielonymi średnikami (z rozszerzeniem *.txt*) lub pliku z wartościami oddzielonymi przecinkami (z rozszerzeniem *.cvs*) lub pliku danych sformatowanego dla IBM SPSS Statistics (z rozszerzeniem *.sav*). Gdy plik jest otwarty, możesz wybrać pole w pliku danych, aby porównać je do etykiet właściwości w konkretnym kluczu mapy. Następnie możesz poprawić w pliku mapy wszelkie rozbieżności.

Pola w pliku danych. Wybierz pole, którego wartości chcesz porównać do etykiet właściwości. Jeśli pierwszy wiersz pliku *.txt* lub *.csv* zawiera etykiety opisowe dla każdego pola, zaznacz opcję **Użyj pierwszego wiersza jako etykiet kolumn**. W przeciwnym wypadku każde pole zostanie zidentyfikowane na podstawie pozycji zajmowanej w pliku danych (na przykład: „Kolumna 1", „Kolumna 2" itd.).

Klucz do porównania. Wybierz klucz mapy, którego nagłówki właściwości chcesz porównać z wartościami pól pliku danych.

Porównaj. Kliknij, gdy jesteś gotowy do porównania wartości.

Wyniki porównania. Domyślnie tabela Wyników porównania zawiera tylko niedopasowane wartości pól pliku danych. Aplikacja stara się znaleźć powiązaną etykietę właściwości, szukając wstawionych lub brakujących spacji. Kliknij rozwijaną listę w kolumnie *Etykieta mapy*, aby dopasować etykietę właściwości w pliku mapy z wyświetlaną wartością pola. Jeśli w Twoim pliku mapy nie ma odpowiadającej etykiety właściwości, wybierz *Pozostaw niedopasowane*. Jeśli chcesz zobaczyć wartości wszystkich pól, nawet tych, którym już dopasowano etykietę właściwości, usuń zaznaczenie opcji **Wyświetlaj tylko niedopasowane przypadki**. Możesz tak zrobić, jeśli chcesz zastąpić jedno lub więcej dopasowań.

Możesz użyć każdej właściwości tylko raz, aby dopasować ją do wartości pola. Jeśli chcesz dopasować wiele właściwości do pojedynczej wartości pola, możesz scalić właściwości, a następnie dopasować nową, scaloną właściwość do wartości pola. Aby uzyskać dalsze informacje na temat scalania właściwości, patrz “Scal właściwości”.

Scal właściwości: Scalanie właściwości jest przydatne do tworzenia na mapie większych regionów. Na przykład gdybyś konwertował mapę stanów, mógłbyś scalić stany (w tym przypadku – właściwości) do większych regionów: północnego, południowego, wschodniego i zachodniego.

Klucze. Wybierz klucz mapy zawierający etykiety właściwości, które pomogą Ci rozpoznać właściwości, które chcesz scalić.

Właściwości. Kliknij pierwszą właściwość, którą chcesz scalić. Kliknij z wciśniętym klawiszem Ctrl inną właściwość, którą chcesz scalić. Zwróć uwagę, że właściwości zostaną także zaznaczone w podglądzie mapy. Możesz kliknąć i kliknąć z wciśniętym klawiszem Ctrl właściwości bezpośrednio w podglądzie mapy oprócz wybierania ich z listy.

Po zaznaczeniu właściwości, które chcesz scalić, kliknij **Scal**, aby wyświetlić okno dialogowe Nazwij scaloną właściwość, gdzie będziesz mógł wstawić etykietę do nowej właściwości. Po scaleniu właściwości możesz zaznaczyć opcję **Pokoloruj podgląd mapy**, aby upewnić się, że wynik jest taki, jak oczekiwałeś.

Po scaleniu właściwości możesz także przesunąć etykietę do nowej właściwości. Możesz tego dokonać w zadaniu *Edytuj etykiety właściwości*. Więcej informacji można znaleźć w temacie “Edytuj etykiety właściwości” na stronie 222.

Okno dialogowe Nazwij scaloną właściwość: Okno dialogowe Nazwij scaloną właściwość pozwala użytkownikowi na przydzielenie etykiet nowej, scalonej właściwości.

Tabela Etykiety wyświetla informacje o każdym kluczu w pliku mapy i pozwala użytkownikowi na przydzielenie każdemu kluczowi etykiety.

Nowa etykieta. Wpisz nową etykietę dla scalonej wartości, aby przydzielić konkretny klucz mapy.

Klucz. Klucz mapy, któremu przydzielasz nową etykietę.

Stare etykiety. Etykiety dla właściwości, które zostaną scalone w nową właściwość.

Usuń granice między stykającymi się wielokątami. Zaznacz tę opcję, aby usunąć granice między właściwościami, które zostały scalone. Na przykład, jeśli scaliłeś stany w regiony geograficzne, opcja ta usunie granice poszczególnych stanów.

Przesuwaj właściwości: Możesz przesuwać właściwości na mapie. Może to być przydatne, gdy chcesz ułożyć właściwości razem, na przykład stały ląd i oddalone od niego wyspy.

Klucze. Wybierz klucz mapy zawierający etykiety właściwości, które pomogą Ci rozpoznać właściwości, które chcesz przesunąć.

Właściwości. Kliknij pierwszą właściwość, którą chcesz przesunąć. Zwróć uwagę, że właściwość zostanie zaznaczona w podglądzie mapy. Możesz także kliknąć właściwość bezpośrednio w podglądzie mapy.

X/Y. Te pola tekstowe zawierają listę aktualnego punktu centralnego właściwości na mapie etykiety. Jednostki są wyświetlane we współrzędnych mapy. Mogą to być lokalne, współczynniki kartezjańskie (na przykład Krajowy układ współrzędnych) lub współrzędne geograficzne (gdzie **X** jest długością, a **Y** szerokością). Wpisz współrzędne dla nowej pozycji właściwości. Możesz także kliknąć i przeciągnąć właściwość na mapie, aby ją przesunąć. Pola tekstowe zostaną uaktualnione o nową pozycję.

Usuń właściwości: Możesz usunąć z mapy niechciane właściwości. Może to być przydatne, gdy chcesz usunąć śmieci usuwając właściwości, które nie będą interesujące w wizualizacji mapy.

Klucze. Wybierz klucz mapy zawierający etykiety właściwości, które pomogą Ci rozpoznać właściwości, które chcesz usunąć.

Właściwości. Kliknij pierwszą właściwość, którą chcesz usunąć. Jeśli chcesz usunąć wiele właściwości na raz, zaznacz dodatkowe właściwości, klikając z wciśniętym klawiszem Ctrl. Zwróć uwagę, że właściwości zostaną także zaznaczone w podglądzie mapy. Możesz kliknąć i kliknąć z wciśniętym klawiszem Ctrl właściwości bezpośrednio w podglądzie mapy oprócz wybierania ich z listy.

Usuń pojedyncze elementy: Oprócz usuwania całych właściwości, możesz także usunąć niektóre z pojedynczych elementów, składających się na właściwości, jak na przykład jeziora i małe wysepki. Opcja ta jest niedostępna dla map punktu.

Elementy. Kliknij element, który chcesz usunąć. Jeśli chcesz usunąć wiele elementów na raz, zaznacz dodatkowe elementy klikając z wciśniętym klawiszem Ctrl. Zwróć uwagę, że elementy zostaną także zaznaczone w podglądzie mapy. Możesz kliknąć i kliknąć z wciśniętym klawiszem Ctrl elementy bezpośrednio w podglądzie mapy oprócz wybierania ich z listy. Ponieważ lista nazw elementów nie jest opisowa (każdy element ma przydzieloną obrębie właściwości liczbę), powinieneś sprawdzić zaznaczenie na podglądzie mapy, aby upewnić się, że zaznaczono żądane elementy.

Ustaw odwzorowanie:

Odwzorowanie mapy określa sposób, w jaki trójwymiarowa Ziemia jest prezentowana w dwóch wymiarach. Każde odwzorowanie powoduje zniekształcenia. Jednakże niektóre odwzorowania są bardziej odpowiednie w zależności od tego, czy przeglądasz mapę świata, czy też lokalną. Także niektóre odwzorowania zachowają kształt oryginalnych właściwości. Odwzorowania zachowujące kształt są odwzorowaniami wiernokątnymi. Opcja ta jest dostępna tylko dla map ze współrzędnymi geograficznymi (długością i szerokością).

W przeciwieństwie do innych opcji Narzędzia do konwersji mapy, odwzorowanie może zostać zmienione po utworzeniu wizualizacji mapy.

Odwzorowanie. Wybierz odwzorowanie mapy. Jeśli tworzysz mapę całego świata lub półkuli, użyj odwzorowania *Lokalnego*, *Mercatora* lub odwzorowania *Winkela*. Dla mniejszych obszarów użyj odwzorowania *Lokalnego*, *stożkowego wiernokątnego Lamberta* lub *poprzecznego Mercatora*. Wszystkie odwzorowania używają jako daną elipsoidę WGS83.

- Odwzorowanie **Lokalne** jest używane zawsze wtedy, gdy mapa została utworzona za pomocą lokalnego układu współrzędnych, na przykład Krajowego układu współrzędnych. Te układy współrzędnych definiowane są raczej za

pomocą współrzędnych kartezjańskich niż geograficznych (długość i szerokość). W odwzorowaniu Lokalnym linie poziome i pionowe są od siebie równo oddalone w kartezjańskim układzie współrzędnych. Odwzorowanie Lokalne nie jest odwzorowaniem wiernokątnym.

- Odwzorowanie **Mercatora** jest odwzorowaniem wiernokątnym dla map świata. Linie poziome i pionowe są proste i zawsze prostopadłe względem siebie. Zwróć uwagę, że odwzorowanie Mercatora rozszerza się do nieskończoności w miarę jak zbliża się do biegunów północnego i południowego, nie może być więc używane, jeśli Twoja mapa zawiera biegun północny lub południowy. Gdy mapa zbliża się do tych granic, zniekształcenie jest największe.
- Odwzorowanie **Winkela** jest odwzorowaniem niewiernokątnym dla map świata. Pomimo tego, że nie jest ono wiernokątne, zapewnia dobrą równowagę między kształtem i rozmiarem. Wszystkie linie są zakrzywione, poza równikiem i południkiem zero. Jeśli Twoja mapa obejmuje biegun północny lub południowy, to odwzorowanie jest dobrym wyborem.
- Jak sugeruje jego nazwa, odwzorowanie **wiernokątny stożkowy Lamberta** jest odwzorowaniem wiernokątnym i jest używane dla map kontynentów lub mniejszych obszarów lądu, które są dłuższe na linii wschód-zachód niż na linii północ-południe.
- Odwzorowanie **poprzeczne Mercatora** jest innym wiernokątnym odwzorowaniem dla map kontynentów lub mniejszych obszarów lądu. Skorzystaj z tego odwzorowania dla obszarów lądu, które są dłuższe na linii północ-południe niż na linii wschodu-zachodu.

Krok 4 - zakończenie

Na tym etapie możesz dodać uwagę opisującą plik mapy, a także aby z kluczy mapy utworzyć plik z próbnymi danymi.

Klucze mapy. Jeśli w pliku mapy jest wiele kluczy, wybierz klucz mapy, którego etykiety właściwości chcesz wyświetlić w podglądzie. Jeśli stworzysz z mapy plik danych, dla wartości danych zostaną użyte następujące etykiety.

Komentarz. Wpisz komentarz opisujący mapę lub zawierający dodatkowe informację, które mogą okazać się cenne dla użytkowników, na przykład źródła dla oryginalnych plików kształtu. Uwaga pojawi się w systemie zarządzania opcji Wyboru szablonu wizualizacji danych.

Utwórz zbiór danych dla etykiet właściwości. Zaznacz tę opcję, jeśli z wyświetlanych etykiet właściwości chcesz stworzyć plik danych. Gdy klikniesz przycisk **Przeglądaj**, będzie można określić lokalizację i nazwę pliku. Jeśli dodasz rozszerzenie *.txt*, plik zostanie zapisany, jako plik z wartościami oddzielonymi tabulatorami. Jeśli dodasz rozszerzenie *.csv*, plik zostanie zapisany, jako plik z wartościami oddzielonymi przecinkami. Jeśli dodasz rozszerzenie *.sav*, plik zostanie zapisany w formacie programu IBM SPSS Statistics. Format SAV jest formatem domyślnym, jeśli nie określono żadnego rozszerzenia.

Dystrybucja plików map

W pierwszym kroku dla narzędzia do konwersji map wybierasz lokalizację, w której ma zostać zapisany plik SMZ. Możesz także zdecydować się dodać mapę do systemu Zarządzania dla opcji Wyboru szablonu wizualizacji danych. Jeśli wybierzesz zapisanie do systemu Zarządzania, mapa będzie dostępna w każdym produkcie firmy IBM SPSS, który działa na tym samym komputerze.

Aby rozprowadzić mapę wśród innych użytkowników będziesz musiał im wysłać plik SMZ. Ci użytkownicy mogą użyć systemu Zarządzania do zaimportowania mapy. Możesz po prostu wysłać plik, którego lokalizację określono w kroku 1. Jeśli chcesz wysłać plik, który znajduje się w systemie Zarządzania, najpierw musisz go wyeksportować:

1. W oknie Wybór szablonów kliknij **Zarządzaj...**
2. Kliknij kartę Mapa.
3. Wybierz mapę, którą chcesz rozesłać.
4. Kliknij **Eksport...** i wybierz miejsce, w którym chcesz zapisać plik.

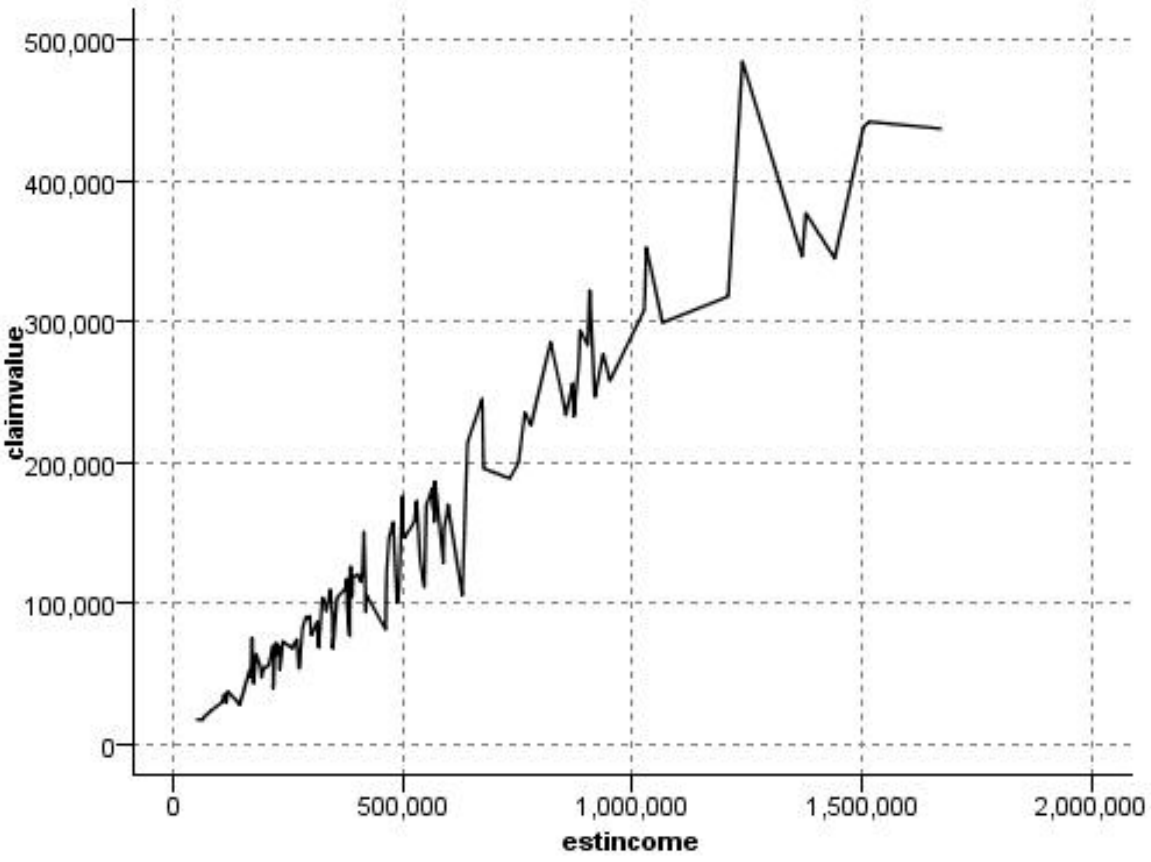
Teraz możesz wysłać innym użytkownikom fizyczny plik mapy. Użytkownicy będą musieli odwrócić ten proces i zaimportować mapę do systemu Zarządzania.

Węzeł Rozrzutu

Węzły wykresu obrazują relacje między zmiennymi liczbowymi. Wykres można utworzyć z punktów (tzw. wykres rozrzutu) albo z linii. Określając w oknie dialogowym tryb osi X, można utworzyć jeden z trzech typów wykresu liniowego.

Tryb osi X = Sortuj

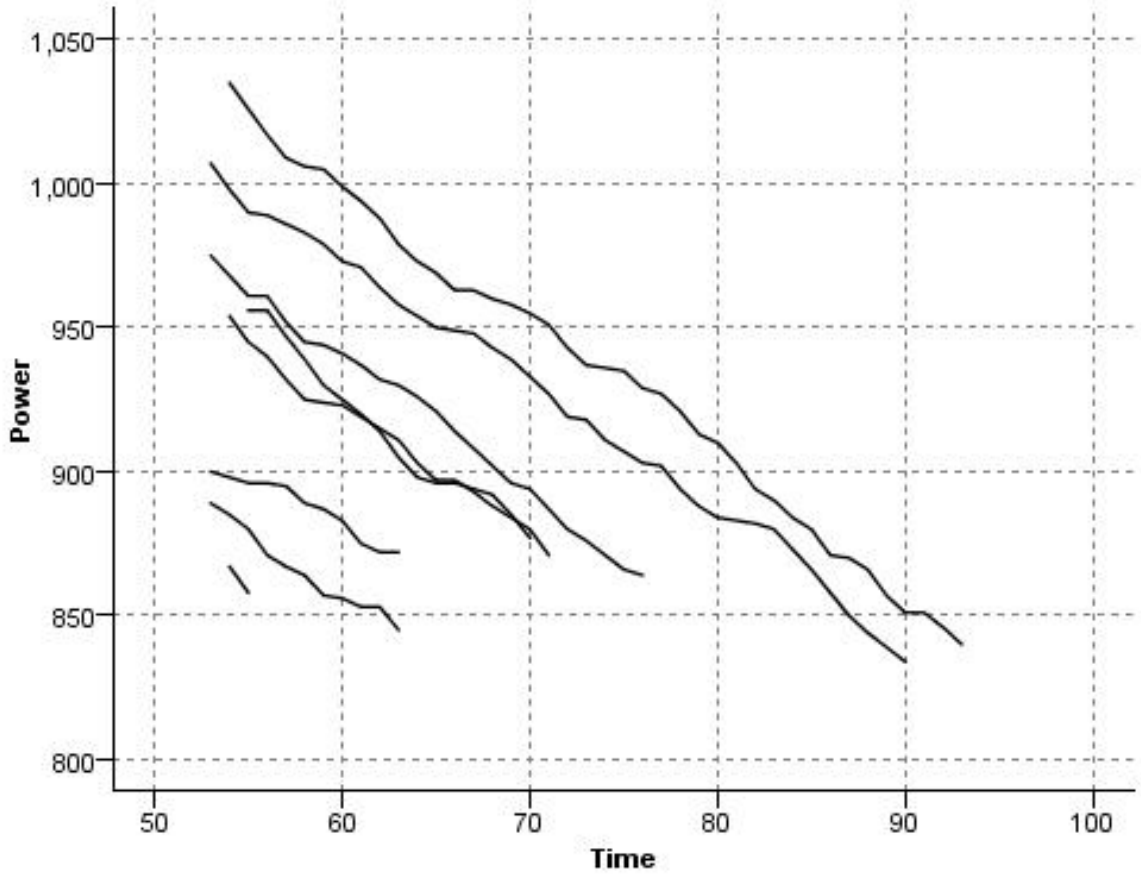
Wybranie trybu osi X **Sortuj** powoduje, że dane będą sortowane według wartości zmiennej wykreślonej na osi *x*. W efekcie na wykresie powstaje pojedyncza linia biegnąca od lewej do prawej strony. Użycie zmiennej nominalnej jako nakładki powoduje narysowanie wielu linii w różnych kolorach, biegnących od lewej do prawej strony wykresu.



Rysunek 19. Wykres liniowy z trybem osi X ustawionym na Sortuj

Tryb osi X = Nakładanie

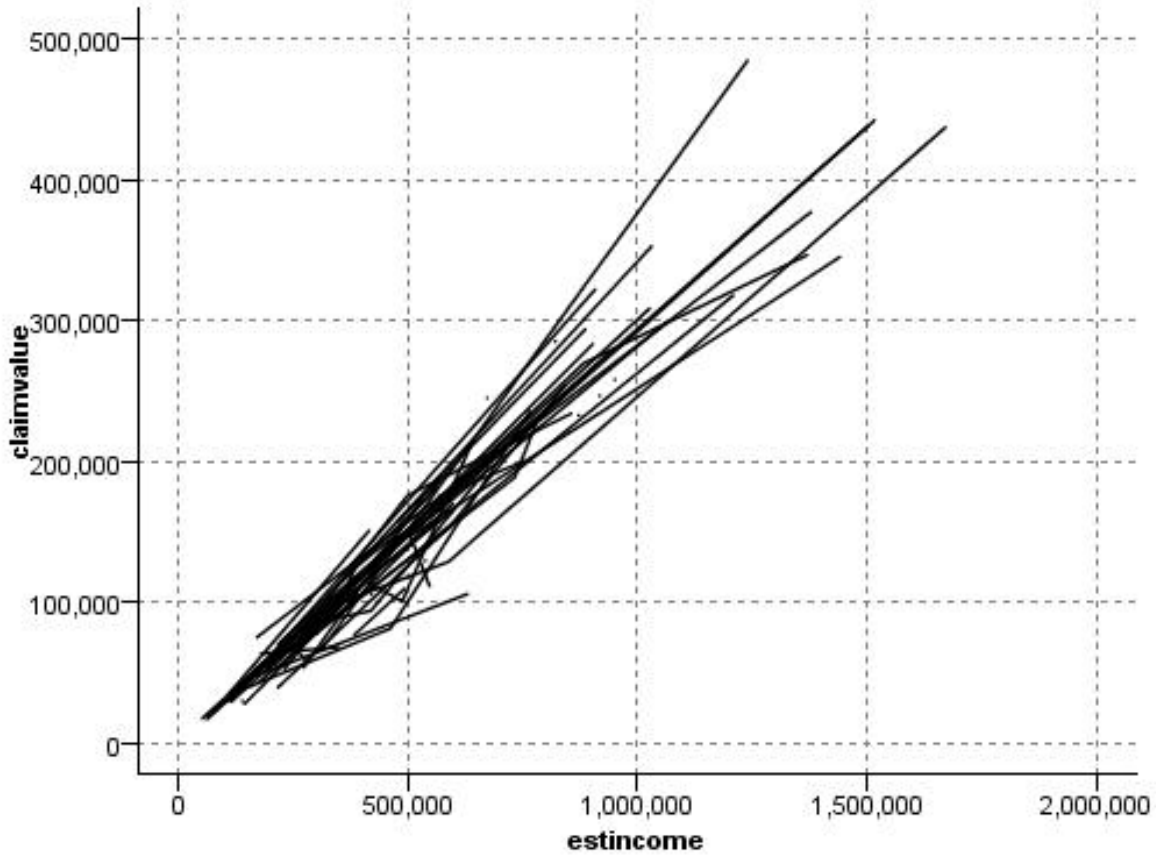
Ustawienie trybu X na **Nakładanie** spowoduje utworzenie wielu wykresów liniowych na tym samym wykresie. Na takim wykresie dane nie są sortowane; kreślone będą jako jedna linia, pod warunkiem że wartości na osi *x* rosną. Jeśli kolejna wartość będzie mniejsza od poprzedniej, rozpoczęta zostanie nowa linia. Na przykład, jeśli wartości na osi *x* rosną od 0 do 100, to wartości *y* będą kreślone jako jedna linia. Gdy *x* spadnie poniżej 100, oprócz pierwszej linii zostanie wykreślona druga. Na gotowym wykresie może znajdować się wiele linii, co ułatwia porównywanie kilku szeregów wartości *y*. Tego typu wykresy są przydatne do analizy danych z okresowym komponentem czasowym, np. zapotrzebowania na energię elektryczną w trakcie kolejnych cykli 24-godzinnych.



Rysunek 20. Wykres liniowy z trybem osi X ustawionym na Nakładanie

Tryb osi X = Według odczytu

Ustawienie trybu osi X na **Według odczytu** powoduje wykreślanie wartości *x* i *y* tak, jak są odczytywane ze źródła danych. Ta opcja jest przydatna w przypadku danych z komponentem szeregu czasowego, gdy chcemy zbadać trendy lub wzorce zależne od kolejności danych. Przed utworzeniem wykresu tego typu konieczne może być posortowanie danych. Celowe może okazać się także porównanie dwóch podobnych wykresów narysowanych w różnych trybach osi X: **Sortuj** oraz **Według odczytu**. Pozwoli to ocenić, na ile wzorzec zależy od sortowania danych.



Rysunek 21. Wykres liniowy przedstawiony wcześniej w trybie sortowania tutaj ponownie narysowany w trybie Według odczytu

Do rysowania wykresów rozrzutu i wykresów liniowych można też używać węzłów wizualizacji. Jednak ten węzeł oferuje do wyboru więcej opcji. Więcej informacji można znaleźć w temacie “Dostępne wbudowane typy wizualizacji Graphboard ” na stronie 199.

Karta węzła wykresu

Na wykresach przedstawiane są wartości zmiennej *Y* w odniesieniu do zmiennej *X*. Często zmienne te odpowiadają odpowiednio zmiennej zależnej i zmiennej niezależnej.

Zmienna X. Z listy należy wybrać zmienną, jaka będzie wyświetlana na poziomej osi *x*.

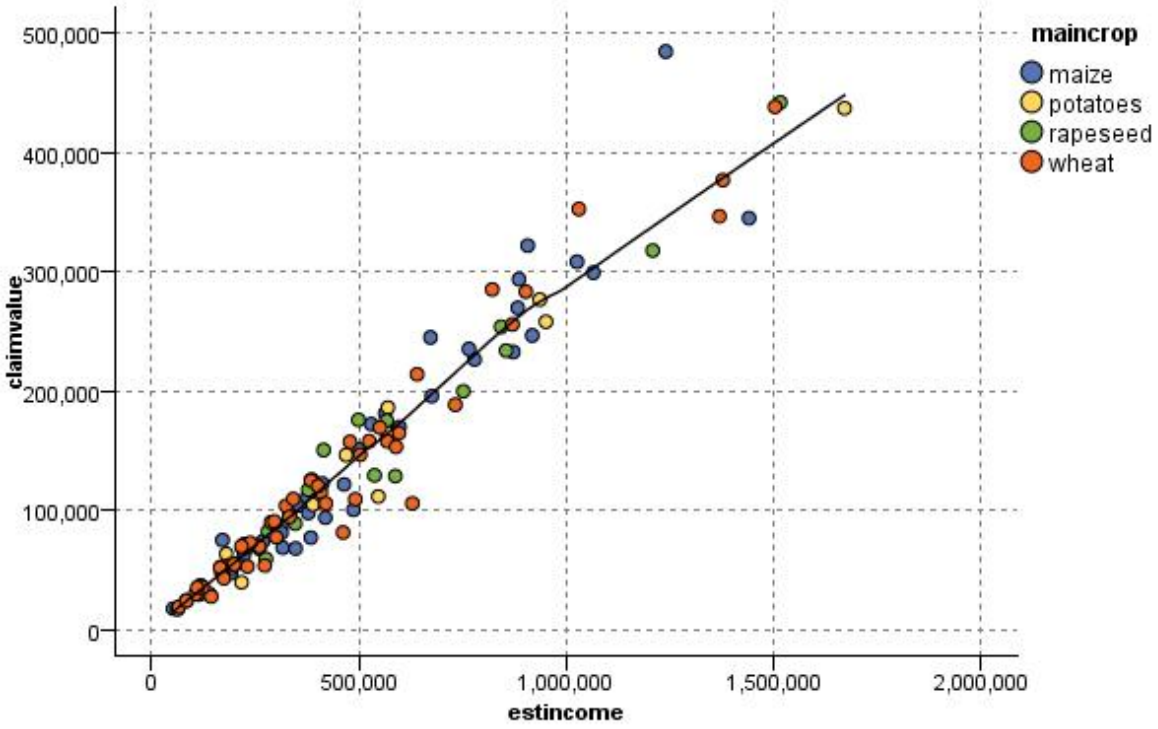
Zmienna Y. Z listy należy wybrać zmienną, jaka będzie wyświetlana na pionowej osi *y*.

Zmienna Z. Po kliknięciu przycisku wykresu 3-W można wybrać zmienną z listy, która będzie wyświetlana na osi *z*.

Nakładanie. Istnieje kilka sposobów ilustrowania kategorii dla wartości danych. Przykładowo można użyć wartości *maincrop* (uprawa główna) dla nałożenia koloru, aby wskazać, że wartości *estincome* (szacowany przychód) i *claimvalue* (wartość roszczenia) dla głównej uprawy rosną wg roszczeń wnioskodawców. Więcej informacji można znaleźć w temacie “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190.

Typ nałożenia. Określa, czy będzie wyświetlana funkcja nałożenia lub wygładzenia. Funkcje wygładzenia i nałożenia są zawsze wyliczane jako funkcja zmiennej y.

- **Brak.** Nie jest wyświetlane żadne nałożenie.
- **Wygładzanie.** Wyświetla wygładzoną linię dopasowania wyliczoną na podstawie regresji lokalnie ważonej metodą cząstkowych najmniejszych kwadratów (LOESS). Ta metoda pozwala na skuteczne obliczenie szeregów regresji, z których każdy jest skoncentrowany na niewielkim obszarze na wykresie. W wyniku tego tworzone są linie „lokalnej" regresji, które są następnie łączone w celu utworzenia krzywej gładkiej.



Rysunek 22. Wykres z nałożeniem wygładzenia typu LOESS

- **Funkcja.** Tę opcję należy wybrać, aby określić znaną funkcję do porównania z wartościami rzeczywistymi. Przykładowo, w celu porównania wartości rzeczywistych do przewidywanych można utworzyć wykres funkcji y = x jako nałożenie. Funkcję dla **y** = należy określić w polu tekstowym. Funkcja domyślna to y = x, ale można określić dowolną funkcję, np. funkcję kwadratową lub dowolne wyrażenie, w odniesieniu do zmiennej x.

  Uwaga: Funkcje nałożenia nie są dostępne na wykresach z panelowaniem lub animacjami.

Po ustawieniu opcji dla wykresu można uruchomić wykres bezpośrednio z okna dialogowego, klikając przycisk **Uruchom**. Można również użyć karty Opcje, aby dokonać dodatkowych specyfikacji, takich jak kategoryzacja, tryb osi X i styl.

Karta opcji wykresu

Styl. Jako styl wykresu należy wybrać **Punkt** lub **Liniowy**. Po wybraniu opcji **Liniowy** aktywowany jest element sterujący **Tryb osi X**. Wybranie opcji **Punkt** spowoduje ustawienie symbolu plus (+) jako domyślny kształt punktu. Po utworzeniu wykresu można zmienić kształt punktu oraz jego wielkość.

Tryb X osi. W przypadku wykresów liniowych należy wybrać tryb osi X, aby zdefiniować styl wykresu liniowego. Można wybrać jedną z opcji **Sortowanie**, **Nakładanie** lub **Jak w danych**. Dla opcji **Nakładanie** lub **Jak w danych** należy określić maksymalną wielkość zbioru danych użytego jako próba dla pierwszych *n* rekordów. W przeciwnym razie zostanie użyta domyślna liczba rekordów, czyli 2000.

Automatyczny przedział X. Zaznaczenie tej opcji umożliwia użycie całego zakresu wartości danych wzdłuż tej osi. Usunięcie zaznaczenia spowoduje użycie jawnego podzbioru wartości na podstawie określonych wartości **Minimum** i **Maksimum**. Wartości można wprowadzić lub można skorzystać ze strzałek. Zakresy automatyczne są zaznaczane domyślnie, aby przyspieszyć tworzenie wykresu.

Automatyczny przedział Y. Zaznaczenie tej opcji umożliwia użycie całego zakresu wartości danych wzdłuż tej osi. Usunięcie zaznaczenia spowoduje użycie jawnego podzbioru wartości na podstawie określonych wartości **Minimum** i **Maksimum**. Wartości można wprowadzić lub można skorzystać ze strzałek. Zakresy automatyczne są zaznaczane domyślnie, aby przyspieszyć tworzenie wykresu.

Automatyczny przedział Z. Tylko wówczas, gdy na karcie Wykres wybrano opcję wykresu 3-W. Zaznaczenie tej opcji umożliwia użycie całego zakresu wartości danych wzdłuż tej osi. Usunięcie zaznaczenia spowoduje użycie jawnego podzbioru wartości na podstawie określonych wartości **Minimum** i **Maksimum**. Wartości można wprowadzić lub można skorzystać ze strzałek. Zakresy automatyczne są zaznaczane domyślnie, aby przyspieszyć tworzenie wykresu.

Rozproszenie. Rozproszenie jest przydatne w przypadku wykresów punktowych dla zbiorów danych, w których powtarzanych jest wiele wartości. Aby rozkład wartości był bardziej czytelny, można użyć funkcji rozproszenia, aby w sposób losowy rozłożyć punkty wokół wartości rzeczywistej.

Uwaga dla użytkowników wcześniejszych wersji oprogramowania IBM SPSS Modeler: Wartość rozproszenia na wykresie w tym wydaniu programu IBM SPSS Modeler korzysta z innej metryki. We wcześniejszych wersjach wartość ta była liczbą rzeczywistą, teraz stanowi proporcję wielkości ramki. Oznacza to, że wartości rozproszenia w starych strumieniach prawdopodobnie będą zbyt duże. Od tego wydania wszystkie niezerowe wartości rozproszenia zostaną przekształcone na wartość 0,2.

Maksymalna liczba rekordów na wykresie. Podaj metodę tworzenia większych zbiorów danych. Można określić maksymalną wielkość zbioru danych lub użyć domyślnej liczby rekordów, wynoszącej 2000. Wydajność dla dużych zbiorów danych ulega poprawie po wybraniu opcji **Kategoria** lub **Próba**. Alternatywnie można wybrać wykreślanie wszystkich punktów danych przez wybór opcji **Użyj wszystkich danych**; należy jednak zwrócić uwagę, że może to drastycznie obniżyć wydajność oprogramowania.

Uwaga: Po ustawieniu trybu osi X na **Nakładanie** lub **Jak w danych** opcje te są wyłączane i użytych jest tylko *n* pierwszych rekordów.

- **Kategoria.** Tę opcję należy wybrać, aby aktywować kategoryzację, jeśli zbiór danych zawiera więcej rekordów niż określona liczba. Kategoryzacja dzieli wykres na mniejsze siatki przed rzeczywistym wykreśleniem go i zlicza liczbę punktów, jakie zostaną wyświetlone w każdej komórce siatki. Na końcowym wykresie w środku ciężkości kategorii wykreślany jest jeden punkt dla każdej komórki (średnia wszystkich lokalizacji punktów w kategorii). Wielkość wykreślanych symboli oznacza liczbę punktów w danym regionie (o ile użyto wielkości jako nałożenia). Użycie środka ciężkości i wielkości do reprezentowania liczby punktów sprawia, że podzielony wykres jest doskonałym sposobem reprezentowania dużych zbiorów danych, ponieważ uniemożliwia wykreślenie zbyt dużej liczby punktów w regionach o dużej gęstości (niemożliwa do rozróżnienia liczba kolorów) oraz redukuje liczbę artefaktów symboli (sztuczne wzory gęstości). Artefakty symboli występują, kiedy określone symbole (głównie symbol plus [+]) kolidują ze sobą, tworząc obszary o dużej gęstości, które nie są dostępne w surowych danych.
- **Przykład.** Tę opcję należy wybrać, aby przeprowadzić próbę losową danych dla rekordów wprowadzonych w polu tekstowym. Domyślną wartością jest 2000.

Karta wyglądu wykresu

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

Etykieta X. Można zaakceptować automatycznie wygenerowaną etykietę osi *x* (pozioma) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

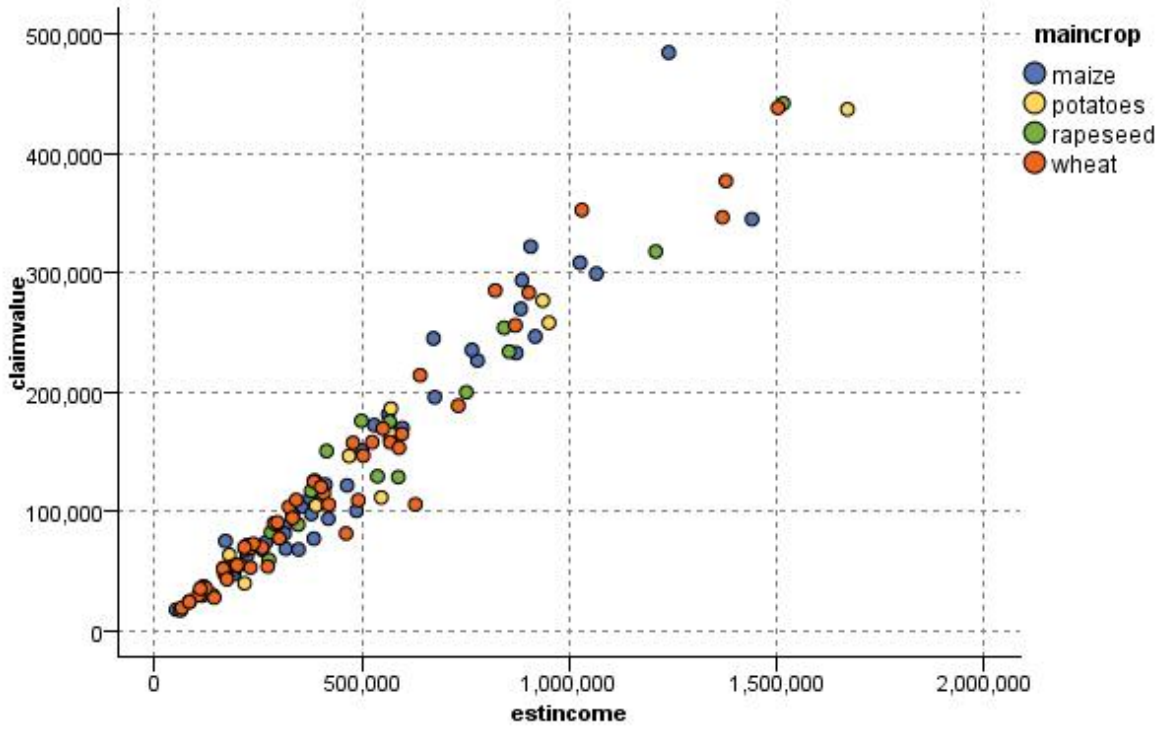
Etykieta Y. Można zaakceptować automatycznie wygenerowaną etykietę osi *y* (pionowa) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Etykieta Z. Opcja dostępna tylko w przypadku wykresów 3-W; można zaakceptować automatycznie wygenerowaną etykietę osi *z* lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Wyświetl linie siatki. Ta opcja jest domyślnie zaznaczona; wyświetla linie siatki za wykresem, dzięki czemu znacznie łatwiej można określić region i punkty odcięcia przedziału. Linie siatki zawsze są wyświetlane na biało, jeśli tło wykresu nie jest białe; wówczas linie są wyświetlane na szaro.

Użycie wykresu

Wykresy i wielokrotne wykresy liniowe są zasadniczo wykresami wartości osi *X* w odniesieniu do osi *Y*. Przykładowo, jeśli przeprowadzana jest eksploracja potencjalnego oszustwa we wnioskach o dotacje rolne, można utworzyć wykres przychodu z wniosku w odniesieniu do przychodu oszacowanego przez sieć neuronową. Użycie nałożenia, takiego jak rodzaj uprawy, pokaże, że istnieje zależność pomiędzy wnioskami (wartość lub liczba) a rodzajem uprawy.



Rysunek 23. Wykres relacji pomiędzy oszacowanym przychodem a wartością wnioskowaną z zastosowaniem nałożenia w postaci rodzaju głównej uprawy

Ponieważ wykresy, wielokrotne wykresy liniowe i wykresy ewaluacyjne są dwuwymiarowym obrazem wartości z osi *Y* w odniesieniu do osi *X*, łatwo można przeprowadzić z nimi interakcję, definiując regiony, oznaczając elementy lub nawet rysując przedziały. Można również wygenerować węzły dla danych reprezentowanych przez dany region, przedział lub element. Więcej informacji można znaleźć w temacie “Eksplorowanie wykresów” na stronie 275.

Węzeł Liniowy

Wykres wielokrotny jest specjalnym typem wykresu, który przedstawia wiele zmiennych *Y* w funkcji jednej zmiennej *X*. Zmienne *Y* są wykreślane jako kolorowe linie, a każda z nich jest równoważna węzłowi wykresu ze stylem ustawionym na **Liniowy** i trybem osi X ustawionym na **Sortuj**. Wykresy wielokrotne są przydatne, gdy mamy dane uporządkowane w czasie i chcemy zbadać fluktuacje kilku zmiennych zachodzące w pewnym przedziale czasu.

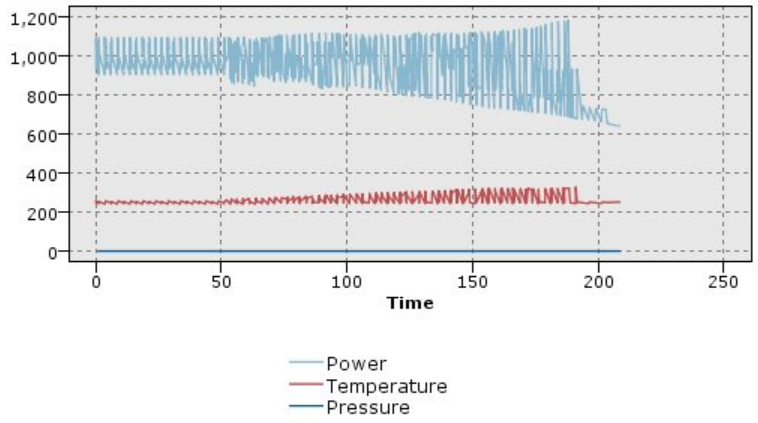
Karta wykresu wielokrotnego

Zmienna X. Z listy należy wybrać zmienną, jaka będzie wyświetlana na poziomej osi *x*.

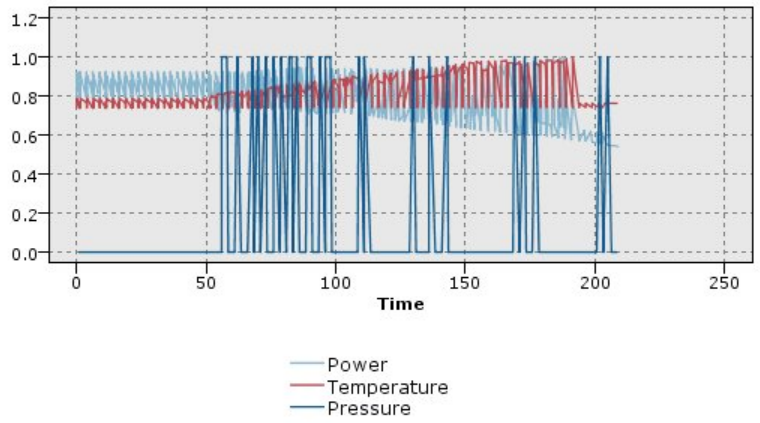
Zmienne Y. Należy wybrać co najmniej jedną zmienną z listy do wyświetlenia w przedziale wartości zmiennej *X*. Przycisk selektora zmiennych umożliwia wybranie wielu zmiennych. Kliknięcie przycisku usuwania umożliwia usunięcie zmiennych z listy.

Nakładanie. Istnieje kilka sposobów ilustrowania kategorii dla wartości danych. Przykładowo można użyć nałożenia animacji w celu wyświetlenia wielu wykresów dla każdej wartości danych. Jest to przydatne w przypadku zestawów składających się z ponad 10 kategorii. W przypadku użycia tej opcji dla zestawów zawierających więcej niż 15 można zauważyć spadek wydajności. Więcej informacji można znaleźć w temacie “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190.

Normalizuj. Tę opcję należy wybrać, aby wyskalować wszystkie wartości *Y* w zakresie od 0 do 1 w celu wyświetlenia na wykresie. Normalizacja ułatwia eksplorację relacji pomiędzy liniami, które w przeciwnym razie mogą zostać zaciemnione z powodu różnic zakresów wartości dla poszczególnych szeregów; jest zalecana w przypadku wykreślania wielu linii na tym samym wykresie lub w przypadku porównywania wykresów w znajdujących się obok siebie panelach. (Normalizacja nie jest konieczna, jeśli wszystkie wartości danych znajdują się w podobnym zakresie).



Rysunek 24. Standardowy wykres wielokrotny przedstawiający wahania źródła energii w czasie (należy zauważyć, że bez normalizacji wyświetlenie wykresu Pressure (Ciśnienie) byłoby niemożliwe)



Rysunek 25. Znormalizowany wykres wielokrotny dla ciśnienia (Pressure)

Funkcja nałożenia. Tę opcję należy wybrać, aby określić znaną funkcję do porównania z wartościami rzeczywistymi. Przykładowo, w celu porównania wartości rzeczywistych do przewidywanych można utworzyć wykres funkcji y = x jako nałożenie. Funkcję dla **y** = należy określić w polu tekstowym. Funkcja domyślna to y = x, ale można określić dowolną funkcję, np. funkcję kwadratową lub dowolne wyrażenie, w odniesieniu do zmiennej x.

Uwaga: Funkcje nałożenia nie są dostępne na wykresach z panelowaniem lub animacjami.

Jeśli liczba rekordów jest większa od. Podaj metodę tworzenia większych zbiorów danych. Można określić maksymalną wielkość zbioru danych lub użyć domyślnej liczby punków, wynoszącej 2000. Wydajność dla dużych zbiorów danych ulega poprawie po wybraniu opcji **Kategoria** lub **Próba**. Alternatywnie można wybrać wykreślanie wszystkich punktów danych przez wybór opcji **Użyj wszystkich danych**; należy jednak zwrócić uwagę, że może to drastycznie obniżyć wydajność oprogramowania.

Uwaga: Po ustawieniu trybu osi X na **Nakładanie** lub **Jak w danych** opcje te są wyłączane i użytych jest tylko *n* pierwszych rekordów.

- **Kategoria.** Tę opcję należy wybrać, aby aktywować kategoryzację, jeśli zbiór danych zawiera więcej rekordów niż określona liczba. Kategoryzacja dzieli wykres na mniejsze siatki przed rzeczywistym wykreśleniem go i zlicza liczbę połączeń, jakie zostaną wyświetlone w każdej komórce siatki. Na końcowym wykresie w środku ciężkości kategorii wykreślane jest jedno połączenie dla każdej komórki (średnia wszystkich punktów połączeń w kategorii).
- **Przykład.** Tę opcję należy wybrać, aby w przeprowadzić próbę losową danych dla określonej liczby rekordów.

Karta wyglądu wykresu wielokrotnego

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

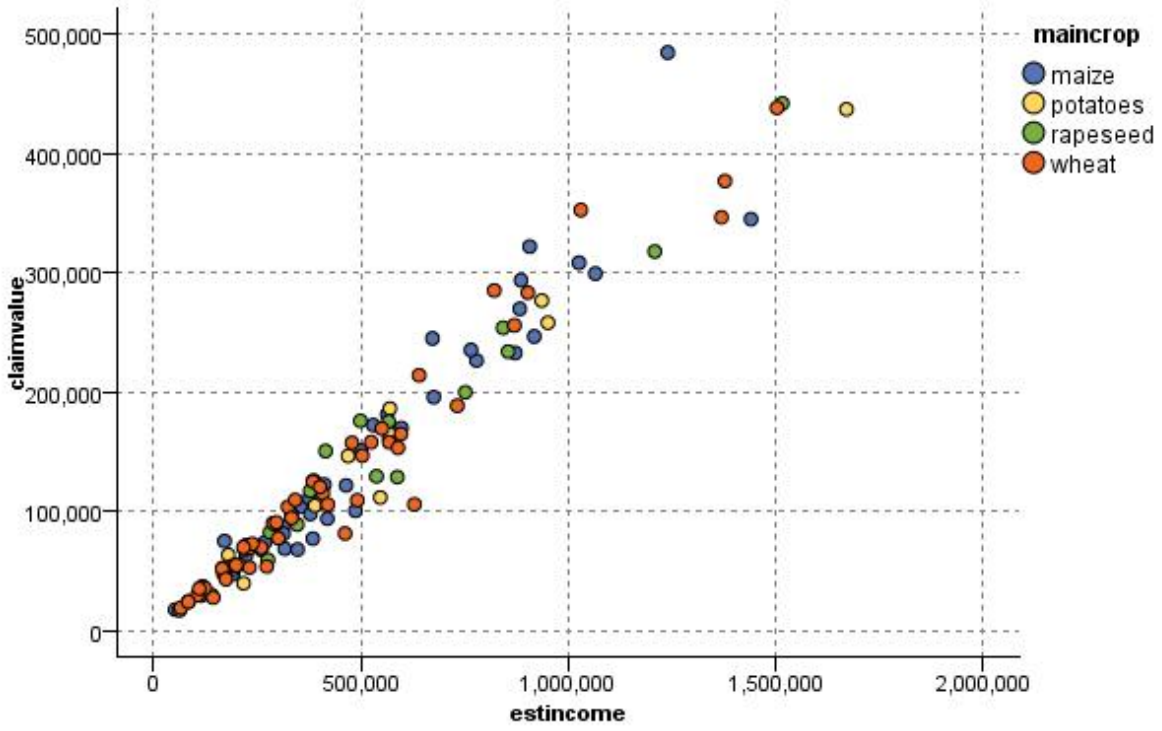
Etykieta X. Można zaakceptować automatycznie wygenerowaną etykietę osi *x* (pozioma) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Etykieta Y. Można zaakceptować automatycznie wygenerowaną etykietę osi *y* (pionowa) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Wyświetl linie siatki. Ta opcja jest domyślnie zaznaczona; wyświetla linie siatki za wykresem, dzięki czemu znacznie łatwiej można określić region i punkty odcięcia przedziału. Linie siatki zawsze są wyświetlane na biało, jeśli tło wykresu nie jest białe; wówczas linie są wyświetlane na szaro.

Korzystanie z wielokrotnego wykresu liniowego

Wykresy i wielokrotne wykresy liniowe są zasadniczo wykresami wartości osi *X* w odniesieniu do osi *Y*. Przykładowo, jeśli przeprowadzana jest eksploracja potencjalnego oszustwa we wnioskach o dotacje rolne, można utworzyć wykres przychodu z wniosku w odniesieniu do przychodu oszacowanego przez sieć neuronową. Użycie nałożenia, takiego jak rodzaj uprawy, pokaże, że istnieje zależność pomiędzy wnioskami (wartość lub liczba) a rodzajem uprawy.



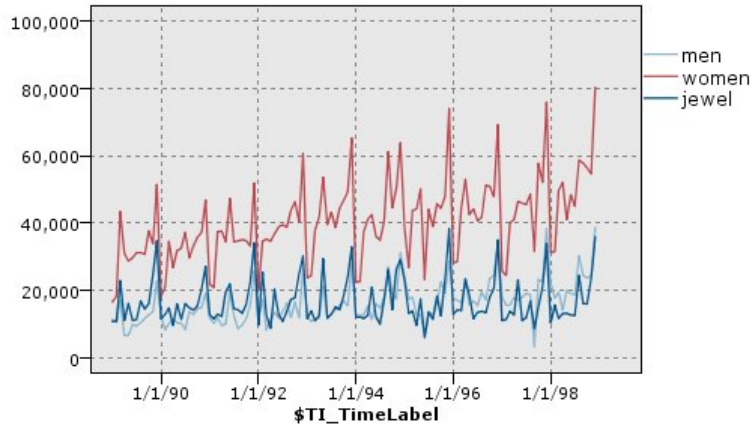
Rysunek 26. Wykres relacji pomiędzy oszacowanym przychodem a wartością wnioskowaną z zastosowaniem nałożenia w postaci rodzaju głównej uprawy

Ponieważ wykresy, wielokrotne wykresy liniowe i wykresy ewaluacyjne są dwuwymiarowym obrazem wartości z osi *Y* w odniesieniu do osi *X*, łatwo można przeprowadzić z nimi interakcję, definiując regiony, oznaczając elementy lub nawet rysując przedziały. Można również wygenerować węzły dla danych reprezentowanych przez dany region, przedział lub element. Więcej informacji można znaleźć w temacie “Eksplorowanie wykresów” na stronie 275.

Węzeł wykresu sekwencyjnego

Węzły wykresu sekwencyjnego umożliwiają wyświetlenie jednego lub większej liczby szeregów czasowych w czasie. Wykreślany szereg musi zawierać wartości numeryczne. Zakłada się ich występowanie w przedziale czasu, w którym okresy są równomierne.

W programie SPSS Modeler wersja 17.1 i wcześniejsze węzeł Przedziały czasu zwykle jest używany przed węzłem Wykres sekwencyjny w celu utworzenia pola *TimeLabel*, które jest używane domyślnie do opatrywania osi *x* na wykresach etykietami.



Rysunek 27. Tworzenie wykresów sprzedaży odzieży i biżuterii męskiej i damskiej w czasie

Tworzenie interwencji i zdarzeń

Istnieje możliwość tworzenia zmiennych Zdarzenie i Interwencja z wykresu czasu poprzez wygenerowanie węzła pochodnego (flagi lub nominalnego) za pomocą menu kontekstowych. Można na przykład utworzyć pole zdarzenia dla strajku kolejarzy, dla którego stan ma wartość Prawda, jeśli zdarzenie to miało miejsce, oraz Fałsz, jeśli nie miało miejsca. W przypadku pola Interwencja, w sytuacji — na przykład — wzrostu cen, można użyć wyliczenia w celu identyfikacji daty wzrostu, przy czym stara cena miałaby wartość 0, zaś nowa — 1. Więcej informacji można znaleźć w temacie “węzeł wyliczeń” na stronie 154.

Karta Wykres sekwencyjny

Wykres. Zapewnia możliwość wyboru sposobu wykreślania danych szeregów czasowych.

- **Wybrane szeregi.** Wyświetla na wykresie wartości wybranych szeregów czasowych. Po wybraniu tej opcji podczas tworzenia wykresu przedziałów ufności należy usunąć zaznaczenie pola wyboru **Normalizuj**.
- **Wybrane modele szeregów czasowych**. Ta opcja użyta razem z modelem szeregów czasowych umożliwia tworzenie wykresów wszystkich powiązanych zmiennych (wartości rzeczywiste i przewidywane, jak również przedziały ufności) dla co najmniej jednego wybranego szeregu czasowego. Wybranie tej opcji powoduje wyłączenie niektórych innych opcji w oknie dialogowym. Jest to opcja preferowana w przypadku tworzenia wykresów przedziałów ufności.

Szeregi. Należy wybrać co najmniej jedną zmienną z danymi szeregów czasowych, dla których ma zostać utworzony wykres. Dane muszą mieć postać numeryczną.

Etykieta osi X. Można wybrać domyślną etykietę lub pojedynczą zmienną, jaka będzie użyta jako etykieta dla osi *x* na wykresach. Jeśli wybrana zostanie opcja domyślna, system użyje zmiennej TimeLabel utworzonej na podstawie poprzedzającego węzła Przedziały czasowe (dla strumieni utworzonych w programie SPSS Modeler w wersji 17.1 lub wcześniejszej) lub sekwencyjnych liczb całkowitych, jeśli węzeł Przedziały czasowe nie został utworzony.

Wyświetl szeregi w oddzielnych panelach. Określa, czy poszczególne szeregi będą wyświetlane w osobnych panelach. Alternatywnie, jeśli nie zostanie wybrane tworzenie paneli, wszystkie szeregi czasowe będą wykreślane na tym samym wykresie, a wygładzanie nie będzie dostępne. W przypadku tworzenia wykresów wszystkich szeregów czasowych na tym samym wykresie każdy szereg będzie reprezentowany przez inny kolor.

Normalizuj. Tę opcję należy wybrać, aby wyskalować wszystkie wartości *Y* w zakresie od 0 do 1 w celu wyświetlenia na wykresie. Normalizacja ułatwia eksplorację relacji pomiędzy liniami, które w przeciwnym razie mogą zostać zaciemnione z powodu różnic zakresów wartości dla poszczególnych szeregów; jest zalecana w przypadku wykreślania wielu linii na tym samym wykresie lub w przypadku porównywania wykresów w znajdujących się obok siebie panelach. (Normalizacja nie jest konieczna, jeśli wszystkie wartości danych znajdują się w podobnym zakresie).

Pokaż. Należy wybrać co najmniej jeden element do wyświetlenia na wykresie. Można wybrać linie, punkty i wygładzanie (LOESS). Wygładzanie jest dostępne tylko w przypadku wyświetlania szeregów w oddzielnych panelach. Domyślnie wybrany jest element linii. Przed uruchomieniem węzła wykresu należy upewnić się, że wybrano co najmniej jeden element wykresu; w przeciwnym razie system zwróci błąd, informujący, że nie wybrano nic do utworzenia wykresu.

Ogranicz rekordy. Tę opcję należy wybrać, aby ograniczyć liczbę wykreślanych rekordów. Należy określić liczbę rekordów, odczytywaną od początku pliku danych, jakie zostaną wykreślone, korzystając z opcji **Maksymalna liczba rekordów na wykresie**. Domyślnie ta liczba jest ustawiona na 2000. Jeśli na wykresie ma znaleźć się ostatnich *n* rekordów z pliku danych, można przed ustawieniem tych rekordów w kolejności malejącej wg czasu użyć węzła Sortowanie.

Karta wyglądu wykresu sekwencyjnego

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

Etykieta X. Można zaakceptować automatycznie wygenerowaną etykietę osi *x* (pozioma) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

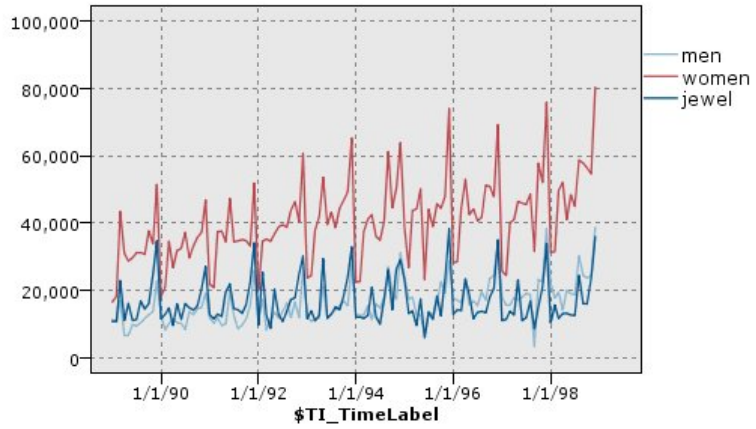
Etykieta Y. Można zaakceptować automatycznie wygenerowaną etykietę osi *y* (pionowa) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Wyświetl linie siatki. Ta opcja jest domyślnie zaznaczona; wyświetla linie siatki za wykresem, dzięki czemu znacznie łatwiej można określić region i punkty odcięcia przedziału. Linie siatki zawsze są wyświetlane na biało, jeśli tło wykresu nie jest białe; wówczas linie są wyświetlane na szaro.

Układ. Tylko w przypadku wykresów sekwencyjnych: można określić, czy wartości czasu będą wykreślane w odniesieniu do osi poziomej czy pionowej.

Użycie wykresu sekwencyjnego

Po utworzeniu wykresu sekwencyjnego dostępnych jest kilka opcji skorygowania obrazu wykresu i wygenerowania węzłów na potrzeby przeprowadzenia dalszej analizy. Więcej informacji można znaleźć w temacie “Eksplorowanie wykresów” na stronie 275.



Rysunek 28. Tworzenie wykresów sprzedaży odzieży i biżuterii męskiej i damskiej w czasie

Po utworzeniu wykresu sekwencyjnego, zdefiniowaniu przedziałów i sprawdzeniu wyników można użyć opcji menu Utwórz oraz menu kontekstowego, aby utworzyć węzły selekcji i wyliczeń. Więcej informacji można znaleźć w temacie “Generowanie węzłów z wykresów” na stronie 282.

Węzeł rozkładu

Wykres lub tabela rozkładu przedstawia wystąpienia wartości symbolicznych (nieliczbowych), takich jak typ kredytu hipotecznego lub płeć, w zbiorze danych. Typowym zastosowaniem węzła rozkładu jest prezentacja nierównowagi w danych, którą można zniwelować, używając węzła ważenia przed przystąpieniem do tworzenia modelu. Węzeł ważenia można wygenerować automatycznie, używając menu Generuj w oknie wykresu lub tabeli rozkładu.

Można także użyć węzła wizualizacji do wygenerowania szeregu wykresów liczebności. Jednak ten węzeł oferuje do wyboru więcej opcji. Więcej informacji można znaleźć w temacie “Dostępne wbudowane typy wizualizacji Graphboard ” na stronie 199.

Uwaga: Do prezentacji wystąpień wartości liczbowych należy używać węzła histogramu.

Karta wykresu rozkładu

Wykres. Należy wybrać typ rozkładu. Opcja **Wybrane zmienne** pozwala wyświetlić rozkład dla wybranej zmiennej. Po wybraniu opcji **Wszystkie flagi (wartości prawda)** wyświetlony zostanie rozkład prawdziwych wartości dla zmiennych typu flaga w danym zbiorze danych.

Zmienna. Należy wybrać zmienną nominalna lub typu flaga, dla których ma zostać wyświetlony rozkład wartości. Na liście wyświetlane są tylko zmienne, które nie zostały w wyraźny sposób ustawione jako numeryczne.

Nakładanie. Należy wybrać zmienną nominalną lub typu flaga, które zostaną użyte jako nałożenie koloru, ilustrując rozkład wartości dla określonej zmiennej. Przykładowo można użyć odpowiedzi na kampanię marketingową (*pep*) jako nałożenie dla liczby dzieci (*children*) w celu zaprezentowania liczby udzielanych odpowiedzi według wielkości rodziny. Więcej informacji można znaleźć w temacie “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190.

Normalizuj według koloru. Tę opcję należy wybrać, aby wyskalować słupki, tak aby wszystkie miały pełną szerokość wykresu. Wartości nałożenia będą równe proporcji każdego słupka, co ułatwi dokonywanie porównań pomiędzy kategoriami.

Sortowanie. Należy wybrać metodę używaną do wyświetlania wartości na wykresach rozkładu. Opcja **Alfabetycznie** umożliwia wyświetlanie wartości w porządku alfabetycznym, a opcja **Wg liczebności** spowoduje wyświetlenie wartości w kolejności malejącej wg wystąpienia.

Skala proporcjonalna. Tę opcję należy wybrać, aby dla rozkładu wartości ustawić taką skalę, zgodnie z którą wartość o największej liczebności będzie wypełniała pełną szerokość wykresu. Wszystkie pozostałe słupki są skalowanie w odniesieniu do tej wartości. Usunięcie zaznaczenia tej opcji spowoduje wyskalowanie słupków w odniesieniu do całkowitej liczebności poszczególnej wartości.

Karta wyglądu wykresu rozkładu

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

Etykieta X. Można zaakceptować automatycznie wygenerowaną etykietę osi *x* (pozioma) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

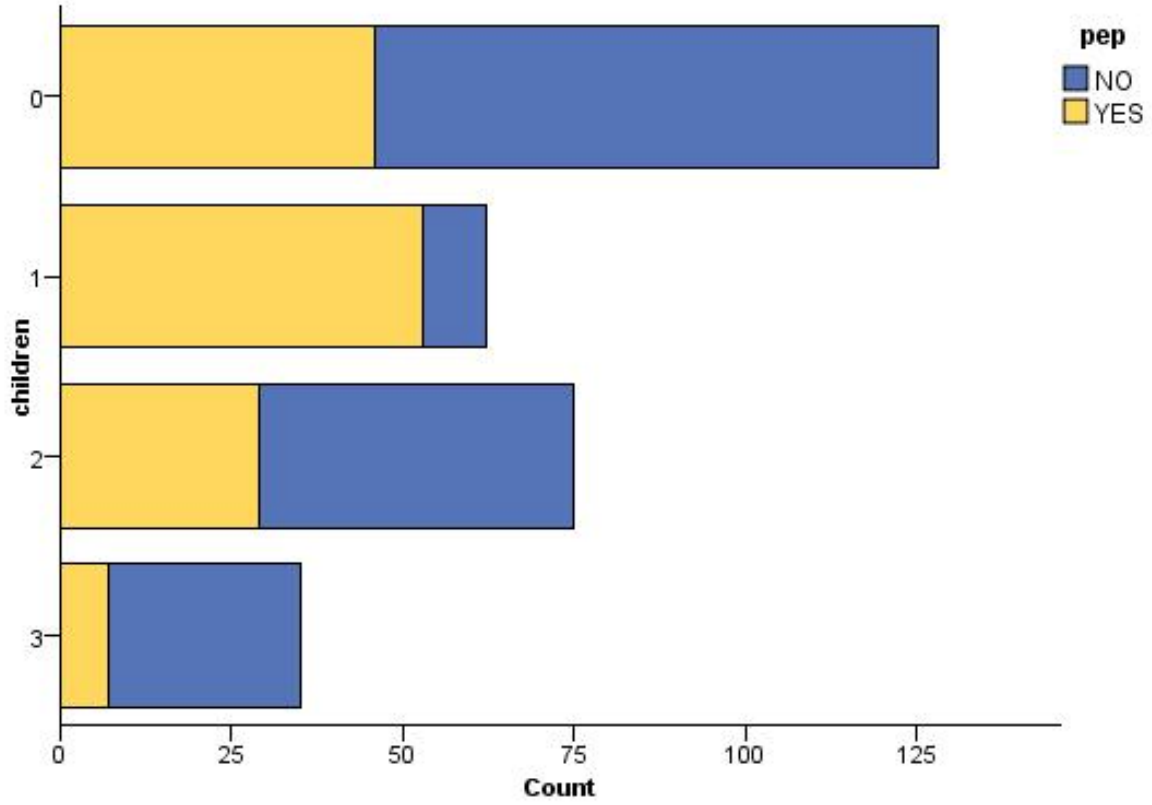
Etykieta Y. Można zaakceptować automatycznie wygenerowaną etykietę osi *y* (pionowa) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Wyświetl linie siatki. Ta opcja jest domyślnie zaznaczona; wyświetla linie siatki za wykresem, dzięki czemu znacznie łatwiej można określić region i punkty odcięcia przedziału. Linie siatki zawsze są wyświetlane na biało, jeśli tło wykresu nie jest białe; wówczas linie są wyświetlane na szaro.

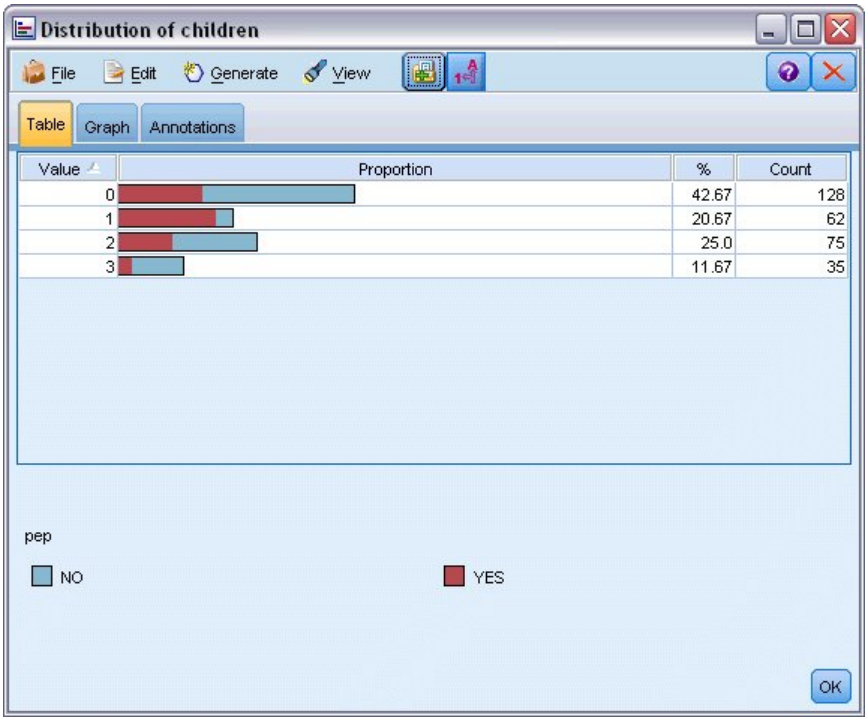
Użycie węzła rozkładu

Węzły rozkładu są używane do wyświetlenia rozkładu wartości symbolicznych w zbiorze danych. Często są używane przed przystąpieniem do manipulacji węzłów w celu przeprowadzenia eksploracji danych i skorygowania dysproporcji. Na przykład, jeśli instancje respondentów bez dzieci występują częściej niż pozostałe typy respondentów, użytkownik może zredukować te instancje, tak aby w późniejszych operacjach eksploracji danych możliwe było wygenerowanie bardziej przydatnej reguły. Węzeł rozkładu ułatwi zbadanie dysproporcji i podjęcie stosownych decyzji.

Odmienność węzła rozkładu polega na tym, że tworzy on zarówno wykres, jak i tabelę do analizy danych.



Rysunek 29. Węzeł rozkładu przedstawiający liczbę osób, które mają lub nie mają dzieci, które odpowiedziały na kampanię marketingową



Rysunek 30. Tabela rozkładu przedstawiająca proporcję osób, które mają lub nie mają dzieci, które odpowiedziały na kampanię marketingową

Po utworzeniu tabeli rozkładu i zbadaniu wyników można użyć opcji z menu do pogrupowania wartości, skopiowania wartości i wygenerowania liczby węzłów w celu przygotowania danych. Ponadto można skopiować lub wyeksportować informacje z wykresu lub tabeli w celu użycia ich w innych aplikacjach, takich jak MS Word lub MS PowerPoint. Więcej informacji można znaleźć w temacie “Drukowanie, zapisywanie, kopiowanie i eksportowanie wykresów” na stronie 297.

Aby wybrać i skopiować wartości z tabeli rozkładu

1. Należy kliknąć i przytrzymać przycisk myszy, przeciągając ją nad wierszami, aby zaznaczyć zbiór wartości. Można również użyć menu Edycja, aby wybrać opcję **Wybierz wszystkie**.
2. Z menu Edycja wybierz opcję **Kopiuj tabelę** lub **Kopiuj tabelę (w tym nazwy zmiennych)**.
3. Wklej do schowka lub do wybranej aplikacji.

   Uwaga: Słupków nie można skopiować bezpośrednio. Zamiast tego kopiowane są wartości tabeli. Oznacza to, że wartości nałożone nie będą wyświetlane w skopiowanej tabeli.

Aby pogrupować wartości z tabeli rozkładu

1. Należy zaznaczyć wartości do pogrupowania, używając metody Ctrl+kliknięcie.
2. Z menu Edycja należy wybrać przycisk **Grupuj**.

Uwaga: podczas grupowania i rozgrupowywania wartości wykres na karcie Wykres jest automatycznie ponownie rysowany, aby odzwierciedlić zmiany.

Można również:

- Rozgrupować wartości, zaznaczając nazwę grupy na liście rozkładu i wybierając opcję **Rozgrupuj** z menu Edycja.
- Edytować grupy, wybierając nazwę grupy na liście rozkładu i wybierając opcję **Edytuj grupę** z menu Edycja. Spowoduje to otwarcie okna dialogowego, w którym można przemieszczać wartości do i z grupy.

Opcje menu Utwórz

Opcje w menu Utwórz umożliwiają wybór podzbioru danych, wyliczanie zmiennej flagi, zmianę grupowania wartości, zmianę klasyfikacji wartości lub równoważenie danych z wykresu lub tabeli. Operacje te spowodują wygenerowanie węzła przygotowania danych i umieszczenie go w obszarze roboczym strumienia. Aby użyć wygenerowanego węzła, należy połączyć go z istniejącym strumieniem. Więcej informacji można znaleźć w temacie “Generowanie węzłów z wykresów” na stronie 282.

Węzeł histogramu

Węzły histogramów przedstawiają występowanie wartości dla pól liczbowych. Są one często używane do eksploracji danych przed przystąpieniem do manipulowania i budowy modelu. Podobnie, jak węzły Rozkład, węzły Histogram są często używane do ujawniania braku równowagi w danych. Choć można używać węzłów Graphboard do tworzenia histogramu, więcej opcji w tym zakresie oferuje niniejszy węzeł. Więcej informacji można znaleźć w temacie “Dostępne wbudowane typy wizualizacji Graphboard ” na stronie 199.

Uwaga: W celu wyświetlenia występowania wartości dla pól symbolicznych należy użyć węzła Rozkład.

Histogram — karta wykresu

Zmienna. Należy wybrać zmienną numeryczną, dla której ma zostać wyświetlony rozkład wartości. Wyświetlone zostaną tylko zmienne, które nie zostały wyraźnie zdefiniowane jako symboliczne (jakościowe).

Nakładanie. Należy wybrać zmienną symboliczną, aby wyświetlić kategorie wartości dla określonej zmiennej. Wybranie zmiennej nałożenia spowoduje przekonwertowanie histogramu na wykres zestawiony, na którym kolory reprezentują różne kategorie zmiennej nałożenia. Po użyciu węzła histogramu dostępne są trzy typy nałożeń: kolor, panel i animacja. Więcej informacji można znaleźć w temacie “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190.

Karta opcji histogramu

Automatyczny przedział X. Zaznaczenie tej opcji umożliwia użycie całego zakresu wartości danych wzdłuż tej osi. Usunięcie zaznaczenia spowoduje użycie jawnego podzbioru wartości na podstawie określonych wartości **Minimum** i **Maksimum**. Wartości można wprowadzić lub można skorzystać ze strzałek. Zakresy automatyczne są zaznaczane domyślnie, aby przyspieszyć tworzenie wykresu.

Przedziały. Należy wybrać opcję **Według liczby** lub **Według szerokości**.

- Opcja **Według liczby** pozwala wyświetlić stałą liczbę słupków, których szerokość zależy od zakresu i liczby określonych przedziałów. Liczbę przedziałów, jakie mają być zastosowane na wykresie, należy określić w opcji **Liczba przedziałów**. Liczbę można skorygować za pomocą strzałek.
- Opcję **Według szerokości** należy wybrać, aby utworzyć wykres ze słupkami o takiej samej szerokości. Liczba przedziałów zależy od określonej szerokości i zakresu wartości. Szerokość słupków należy określić w opcji **Szerokość przedziału**.

Normalizuj według koloru. Tę opcję należy wybrać, aby ustawić wszystkie słupki na takiej samej wysokości, tak aby wyświetlały nałożone wartości jako wartość procentową łącznej liczby obserwacji w każdym słupku.

Przedstaw krzywą normalną. Tę opcję należy wybrać, aby dodać krzywą normalną do wykresu przedstawiającego średnią i wariancję danych.

Oddzielne pasma dla każdego koloru. Ta opcja umożliwia wyświetlenie każdej nałożonej wartości jako osobny przedział na wykresie.

Karta wyglądu histogramu

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

Etykieta X. Można zaakceptować automatycznie wygenerowaną etykietę osi *x* (pozioma) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Etykieta Y. Można zaakceptować automatycznie wygenerowaną etykietę osi *y* (pionowa) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

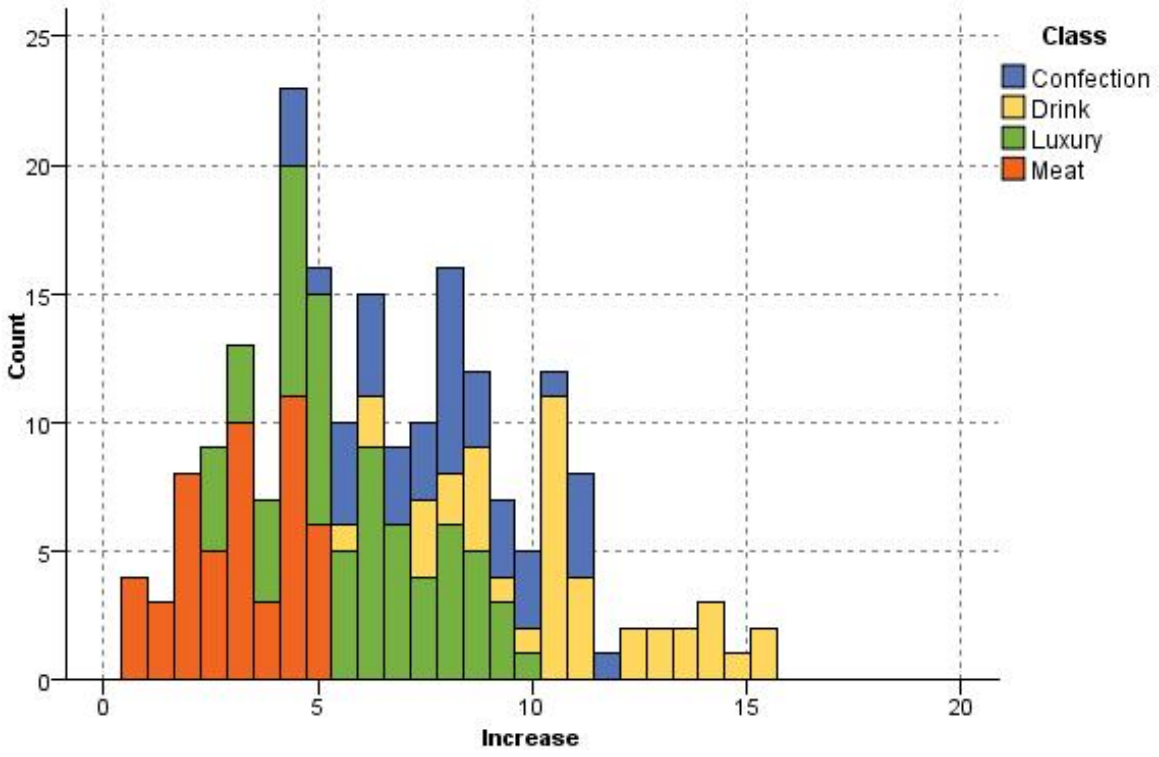
Wyświetl linie siatki. Ta opcja jest domyślnie zaznaczona; wyświetla linie siatki za wykresem, dzięki czemu znacznie łatwiej można określić region i punkty odcięcia przedziału. Linie siatki zawsze są wyświetlane na biało, jeśli tło wykresu nie jest białe; wówczas linie są wyświetlane na szaro.

Używanie histogramów

Histogramy przedstawiają rozkład wartości zmiennej numerycznej, której wartości należą do zakresu wzdłuż osi *x*. Histogramy działają podobnie jak wykresy przedziałowe. Wykresy przedziałowe przedstawiają rozkład wartości jednej zmiennej numerycznej *względem wartości innej zmiennej*, a nie wystąpienia wartości jednej zmiennej.

Po utworzeniu wykresu można zbadać wyniki i zdefiniować przedziały do podzielenia wartości wzdłuż osi *x* lub zdefiniować regiony. Można również oznaczyć elementy na wykresie. Więcej informacji można znaleźć w temacie “Eksplorowanie wykresów” na stronie 275.

Opcje menu Utwórz umożliwiają utworzenie węzłów Zrównoważenie, Selekcja lub Wyliczanie na podstawie danych z wykresu lub dokładniej utworzenie ich wewnątrz przedziałów, regionów lub oznaczonych elementów. Tego typu wykres często jest używany przed rozpoczęciem manipulowania węzłami do eksploracji danych i poprawienia dysproporcji poprzez wygenerowanie węzła zrównoważenia z wykresu i użycie go w strumieniu. Można również wygenerować węzeł wyliczeń typu flaga, aby dodać zmienną pokazującą, do którego przedziału należy dany rekord lub wybrać węzeł selekcji, aby wybrać wszystkie rekordy z określonym zbiorem lub zakresem wartości. Tego typu operacje pomagają skoncentrować się na konkretnym podzbiorze danych w celu przeprowadzenia dalszej eksploracji. Więcej informacji można znaleźć w temacie “Generowanie węzłów z wykresów” na stronie 282.



Rysunek 31. Histogram przedstawiający rozkład wzrostu zakupów wg kategorii po promocji

Węzeł zbioru

Przedziały są podobne do histogramów, z tym że przedstawiają rozkład wartości jednej zmiennej liczbowej względem wartości innej zmiennej, a nie wystąpienia wartości jednej zmiennej. Przedział jest przydatny do prezentacji zmiennej, której wartości zmieniają się w czasie. Na wykresie 3-W można dodać oś symboliczną odzwierciedlającą rozkład według kategorii. Przedziały dwuwymiarowe są prezentowane jako zestawione wykresy słupkowe z nakładkami, o ile są używane. Więcej informacji można znaleźć w temacie “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190.

Karta wykresu przedziałowego

Przedziały. Należy wybrać zmienną, której wartości będą gromadzone i wyświetlane w zakresie wartości dla zmiennej określonej w polu **Względem**. Na liście wyświetlane są zmienne, które nie zostały zdefiniowane jako symboliczne.

Względem. Należy wybrać zmienną, której wartości będą używane do wyświetlania zmiennej określonej w polu **Przedziały**.

Przez. Jeśli ta opcja jest aktywowana podczas tworzenia wykresu 3-W, można wówczas wybrać zmienną nominalną lub flagi używaną do wyświetlania zmiennej przedziałowej wg kategorii.

Operacja. Ta opcja pozwala wybrać, co będzie reprezentował każdy słupek na wykresie. Dostępne opcje to: **Suma**, **Średnia**, **Maksimum**, **Minimum** i **Odchylenie standardowe**.

Nakładanie. Należy wybrać zmienną symboliczną, aby wyświetlić kategorie wartości dla wybranej zmiennej. Wybranie zmiennej nałożenia spowoduje przekształcenie przedziału i utworzenie wielu słupków w różnych kolorach dla każdej kategorii. Ten węzeł udostępnia trzy typy nałożenia: kolor, panel i animacja. Więcej informacji można znaleźć w temacie “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190.

Karta opcji przedziałów

Automatyczny przedział X. Zaznaczenie tej opcji umożliwia użycie całego zakresu wartości danych wzdłuż tej osi. Usunięcie zaznaczenia spowoduje użycie jawnego podzbioru wartości na podstawie określonych wartości **Minimum** i **Maksimum**. Wartości można wprowadzić lub można skorzystać ze strzałek. Zakresy automatyczne są zaznaczane domyślnie, aby przyspieszyć tworzenie wykresu.

Przedziały. Należy wybrać opcję **Według liczby** lub **Według szerokości**.

- Opcja **Według liczby** pozwala wyświetlić stałą liczbę słupków, których szerokość zależy od zakresu i liczby określonych przedziałów. Liczbę przedziałów, jakie mają być zastosowane na wykresie, należy określić w opcji **Liczba przedziałów**. Liczbę można skorygować za pomocą strzałek.
- Opcję **Według szerokości** należy wybrać, aby utworzyć wykres ze słupkami o takiej samej szerokości. Liczba przedziałów zależy od określonej szerokości i zakresu wartości. Szerokość słupków należy określić w opcji **Szerokość przedziału**.

Karta wyglądu wykresu przedziałowego

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

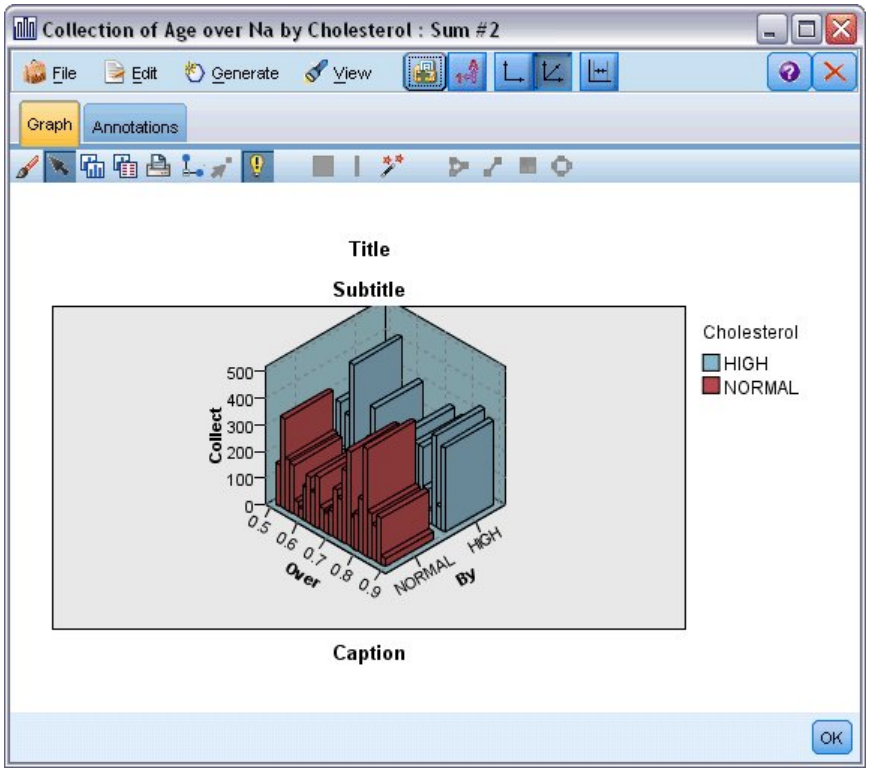
Etykieta **Względem.** Można zaakceptować automatycznie wygenerowaną etykietę lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Etykieta **Przedziały.** Można zaakceptować automatycznie wygenerowaną etykietę lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Etykieta **Przez.** Można zaakceptować automatycznie wygenerowaną etykietę lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Wyświetl linie siatki. Ta opcja jest domyślnie zaznaczona; wyświetla linie siatki za wykresem, dzięki czemu znacznie łatwiej można określić region i punkty odcięcia przedziału. Linie siatki zawsze są wyświetlane na biało, jeśli tło wykresu nie jest białe; wówczas linie są wyświetlane na szaro.

Poniższy przykład przedstawia miejsca, w których znajdują się opcje wyglądu na wykresie w wersji 3-W.



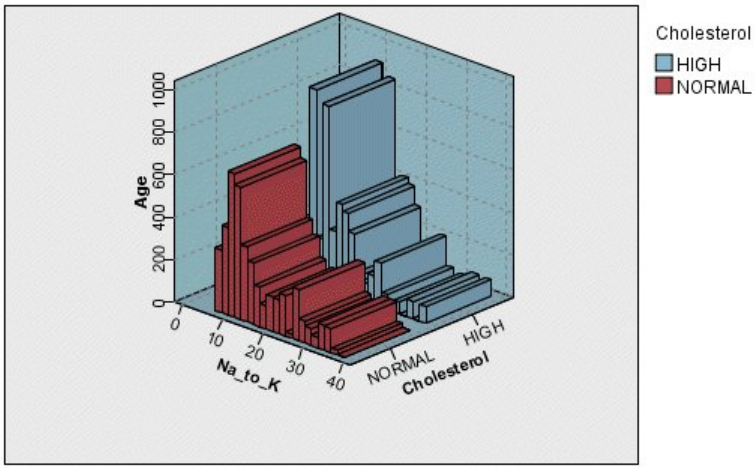
Rysunek 32. Rozmieszczenie opcji wyglądu wykresu na wykresie przedziałowym 3-W

Korzystanie z wykresu przedziałowego

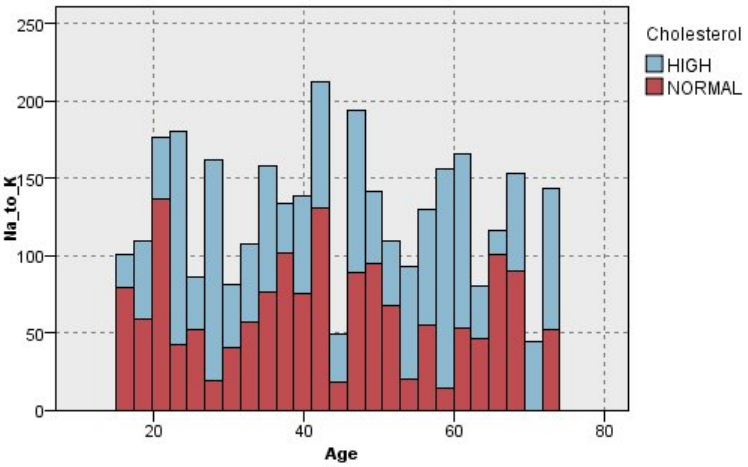
Wykresy przedziałowe przedstawiają rozkład wartości jednej zmiennej numerycznej *względem wartości innej zmiennej*, a nie wystąpienia wartości jednej zmiennej. Histogramy działają podobnie jak wykresy przedziałowe. Histogramy przedstawiają rozkład wartości zmiennej numerycznej, której wartości należą do zakresu wzdłuż osi *x*.

Po utworzeniu wykresu można zbadać wyniki i zdefiniować przedziały do podzielenia wartości wzdłuż osi *x* lub zdefiniować regiony. Można również oznaczyć elementy na wykresie. Więcej informacji można znaleźć w temacie “Eksplorowanie wykresów” na stronie 275.

Opcje menu Utwórz umożliwiają utworzenie węzłów Zrównoważenie, Selekcja lub Wyliczanie na podstawie danych z wykresu lub dokładniej utworzenie ich wewnątrz przedziałów, regionów lub oznaczonych elementów. Tego typu wykres często jest używany przed rozpoczęciem manipulowania węzłami do eksploracji danych i poprawienia dysproporcji poprzez wygenerowanie węzła zrównoważenia z wykresu i użycie go w strumieniu. Można również wygenerować węzeł wyliczeń typu flaga, aby dodać zmienną pokazującą, do którego przedziału należy dany rekord lub wybrać węzeł selekcji, aby wybrać wszystkie rekordy z określonym zbiorem lub zakresem wartości. Tego typu operacje pomagają skoncentrować się na konkretnym podzbiorze danych w celu przeprowadzenia dalszej eksploracji. Więcej informacji można znaleźć w temacie “Generowanie węzłów z wykresów” na stronie 282.



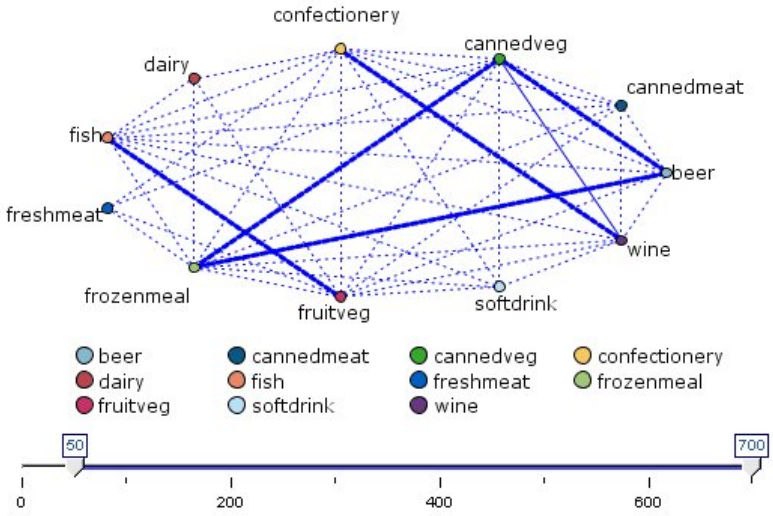
Rysunek 33. Wykresy przedziałowe 3-W przedstawiające sumę Na_to_K względem wartości Age (Wiek) dla normalnego i podwyższonego poziomu cholesterolu



Rysunek 34. Wykres przedziałowy bez osi z nałożeniem koloru dla wartości Cholesterol

Węzeł sieciowy

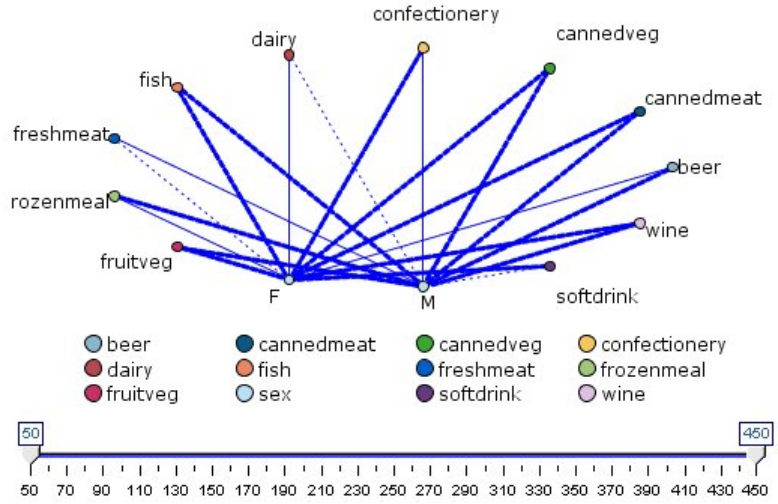
Węzły sieciowe przedstawiają siłę relacji między wartościami dwu lub więcej pól liczbowych. Na wykresie wyświetlane są połączenia z użyciem różnych typów wierszy, wskazujące siłę połączenia. Węzła sieciowego można użyć na przykład do eksploracji relacji między zakupami różnych towarów w witrynie e-sklepu lub w tradycyjnym punkcie sprzedaży detalicznej.



Rysunek 35. Wykres sieciowy przedstawiający relację między zakupami towarów spożywczych

Przekierowane strony WWW

Przekierowane węzły sieciowe są podobne do węzłów sieciowych, ponieważ przedstawiają one siłę relacji między polami symbolicznymi. Przekierowane wykresy sieciowe przedstawiają tylko połączenia z jednego lub więcej pól Od do pojedynczego pola Do. Połączenia są jednokierunkowe, co oznacza, że są połączeniami działającymi w jedną stronę.



Rysunek 36. Przekierowany wykres sieciowy przedstawiający relację między zakupami towarów spożywczych a płcią

Podobnie jak w przypadku węzłów sieciowych, na wykresie wyświetlane są połączenia z użyciem różnych typów wierszy, wskazujące siłę połączenia. Przekierowanego węzła sieciowego można używać na przykład do eksploracji relacji między płcią a skłonnością do zakupu określonych towarów.

Karta wykresu sieciowego

Sieciowy. Tę opcję należy wybrać, aby utworzyć wykres sieciowy ilustrujący siłę relacji między wszystkimi określonymi zmiennymi.

Sieć kierunkowa. Należy wybrać tę opcję, aby utworzyć wykres sieci kierunkowej przedstawiający siłę relacji pomiędzy wieloma zmiennymi i wartościami jednej zmiennej, np. płeć lub religia. Po wybraniu tej opcji aktywowana jest funkcja Do zmiennej, a nazwa elementu sterującego Zmienne poniżej zostaje zmieniona na Od zmiennych w celu zapewnienia dodatkowej przejrzystości.

Do zmiennej (tylko sieć kierunkowa). Należy wybrać zmienną flagi lub nominalną używaną dla sieci kierunkowej. Na liście wyświetlane są tylko zmienne, które nie zostały wyraźnie określone jako numeryczne.

Zmienne/Od zmiennych. Należy wybrać zmienne do utworzenia wykresu sieciowego. Na liście wyświetlane są tylko zmienne, które nie zostały wyraźnie określone jako numeryczne. Przycisk selektora zmiennych umożliwia wybranie kilku zmiennych lub wybranie zmiennych według typu.

Uwaga: W przypadku sieci kierunkowej ten element sterujący jest używany w celu wybrania opcji Od zmiennych.

Pokaż tylko flagi prawdy. Tę opcję należy wybrać, aby dla zmiennej flagi wyświetlić tylko flagi prawdy. Ta opcja upraszcza wyświetlanie wykresu sieciowego i często jest używana w przypadku danych, dla których występowanie wartości dodatnich ma szczególne znaczenie.

Wartości linii. Z listy rozwijanej należy wybrać typ wartości granicznej.

- **Wartość bezwzględna** ustawia wartości graniczne na podstawie liczby rekordów, w których znajduje się każda para wartości.
- **Wartości procentowe ogółem** przedstawia bezwzględną liczbę obserwacji reprezentowaną przez złączenie jako proporcję wszystkich wystąpień każdej pary wartości reprezentowaną na wykresie sieciowym.
- Opcja **Procenty mniejszej zmiennej/wartości** i **Procenty większej zmiennej/wartości** wskazuje, która zmienna/wartość będzie używana do oceny wartości procentowych. Załóżmy na przykład, że w 100 rekordach znajduje się wartość *drugY* (lekY) dla zmiennej *Drug* (Lek) i tylko 10 z nich ma wartość *LOW* (Niski) dla zmiennej *BP*. Jeśli w siedmiu rekordach występują obie wartości *drugY* i *LOW*, ta wartość procentowa będzie wynosiła 70% lub 7%, w zależności od tego, która zmienna stanowi odniesienie, mniejsza (*BP*) czy większa (*Drug*).

Uwaga: W przypadku wykresów sieci kierunkowej trzecia i czwarta opcja powyżej są niedostępne. Można zamiast nich wybrać opcję **Procenty zmiennej/wartości „Do”** oraz **Procenty zmiennej/wartości „Od”**.

Silne łącza są grubsze. Ta opcja jest wybrana domyślnie; jest to standardowy sposób wyświetlania łączy pomiędzy zmiennymi.

Słabe łącza są grubsze. Tę opcję należy wybrać, aby odwrócić znaczenie łączy wyświetlanych jako linie pogrubione. Ta opcja jest często używana do wykrywania oszustw lub badania wartości odstających.

Karta opcji wykresu sieciowego

Karta Opcje dla węzłów sieciowych zawiera dodatkowe opcje umożliwiające dostosowanie wykresu wynikowego.

Liczba łączy. Następujące opcje służą do kontrolowania liczby łączy wyświetlanych na wykresie wynikowym. Niektóre z tych opcji, takie jak **Słabe łącza powyżej** oraz **Silne łącza powyżej**, są również dostępne w oknie wykresu wynikowego. Można również używać suwaka na końcowym wykresie, aby skorygować liczbę wyświetlanych łączy.

- **Maksymalna liczba wyświetlanych łączy.** Określa liczbę wskazującą maksymalną liczbę łączy wyświetlanych na wykresie wynikowym. Wartość można skorygować za pomocą strzałek.
- **Pokaż tylko łącza powyżej.** Określa liczbę wskazującą minimalną wartość, dla której wyświetlane będzie połączenie na wykresie sieciowym. Wartość można skorygować za pomocą strzałek.
- **Pokaż wszystkie łącza.** Określa, że wyświetlane będą wszystkie łącza, niezależnie od wartości minimalnej lub maksymalnej. Jeśli dostępna jest duża liczba zmiennych, wybranie tej opcji może wydłużyć czas przetwarzania.

Odrzuć, gdy zbyt mało rekordów. Tę opcję należy wybrać, aby ignorować połączenia, które są obsługiwane przez zbyt mało rekordów. Dla tej opcji należy ustawić wartość graniczną, wprowadzając liczbę w polu **Minimum rekordów na linię**.

Odrzuć, gdy zbyt dużo rekordów. Tę opcję należy wybrać, aby ignorować silnie obsługiwane połączenia. Należy wprowadzić liczbę w polu **Maksimum rekordów na linię**.

Słabe łącza poniżej. Określa liczbę wskazującą wartość graniczną dla słabych połączeń (linie przerywane) i normalnych połączeń (zwykłe linie). Wszystkie połączenia poniżej tej wartości są uznawane za słabe.

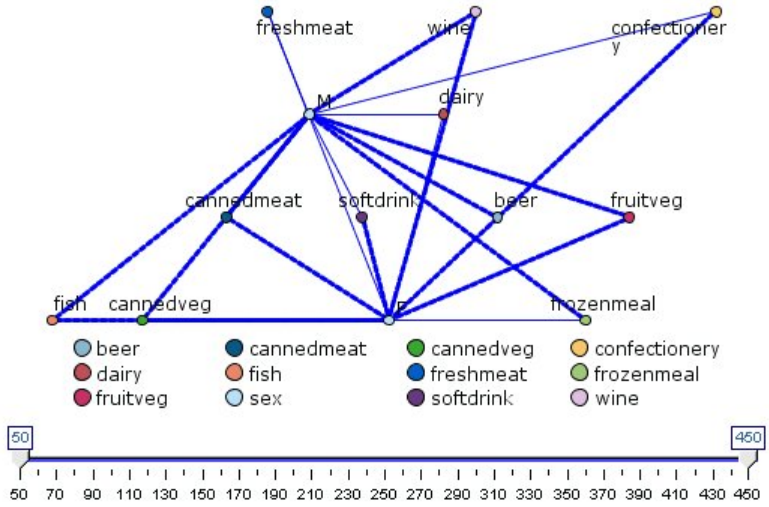
Silne łącza powyżej. Określa wartość graniczną dla silnych połączeń (grube linie) i normalnych połączeń (zwykłe linie). Wszystkie połączenia nad tą wartością są uznawane za silne.

Wielkość łącza. Umożliwia określenie opcji sterowania wielkością łączy:

- **Ciągła zmiana wielkości łączy.** Umożliwia wyświetlenie zakresu wielkości łączy odzwierciedlających różne siły połączeń na podstawie rzeczywistych wartości danych.
- **Rozmiar łącza reprezentuje kategorie Silne/Normalne/Słabe.** Zaznaczenie tej opcji powoduje wyświetlenie trzech sił połączeń — silne, normalne i słabe. Punkty odcięcia dla tych kategorii można określić powyżej oraz na wykresie końcowym.

Wygląd wykresu sieciowego. Należy wybrać typ wyglądu wykresu sieciowego:

- **Układ kołowy.** Po wybraniu tej opcji wyświetlany jest standardowy wykres sieciowy.
- **Układ sieciowy.** Po wybraniu tej opcji używany jest algorytm do grupowania najsilniejszych łączy. Ma to na celu wyróżnienie silnych łączy poprzez ich rozróżnienie przestrzenne oraz przy użyciu linii ważonych.
- **Układ kierunkowy.** Ta opcja umożliwia utworzenie wykresu sieci kierunkowej, na którym opcja **Do zmiennej** z karty wykresu umożliwia skoncentrowanie się na kierunku.
- **Układ siatki.** Ta opcja pozwala utworzyć wykres sieciowy, który znajduje się na równomiernie rozmieszczonej siatce.



Rysunek 37. Wykres sieciowy przedstawiający silne połączenia pomiędzy elementami frozenmeal (produkty mrożone) i cannedveg (warzywa w puszkach) i innymi elementami grocery (artykuły spożywcze)

Uwaga: Podczas filtrowania wyświetlanych łączy (za pomocą suwaka na wykresie sieciowym lub elementu sterującego **Pokaż tylko łącza powyżej** na karcie Opcje węzła sieciowego), może dojść do sytuacji, w której wszystkie nadal wyświetlane łącza są łączami pojedynczej wartości (innymi słowy wszystkie są łączami słabymi, wszystkie są łączami normalnymi lub wszystkie są łączami silnymi, zgodnie z definicją w elementach sterujących **Słabe łącza poniżej** i **Silne łącza powyżej** na karcie Opcje węzła sieciowego). Jeśli tak się stanie, na wynikowym wykresie sieciowym wszystkie łącza będą wyświetlane w postaci linii średniej grubości.

Karta wyglądu wykresu sieciowego

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

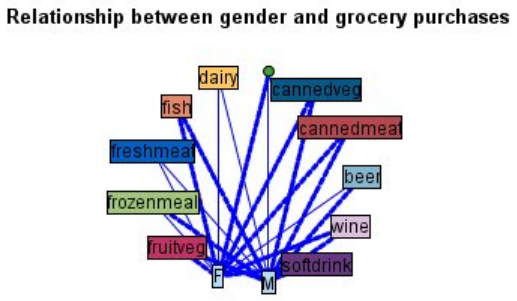
Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

Pokaż legendę. Należy określić, czy legenda będzie wyświetlana. W przypadku wykresów z dużą liczbą zmiennych ukrycie legendy może poprawić wygląd wykresu.

Użyj etykiet w węzłach. Zamiast wyświetlania sąsiadujących etykiet można dołączyć tekst etykiety do każdego węzła. W przypadku wykresów z małą liczbą zmiennych dzięki takiemu rozwiązaniu wykres może być bardziej czytelny.



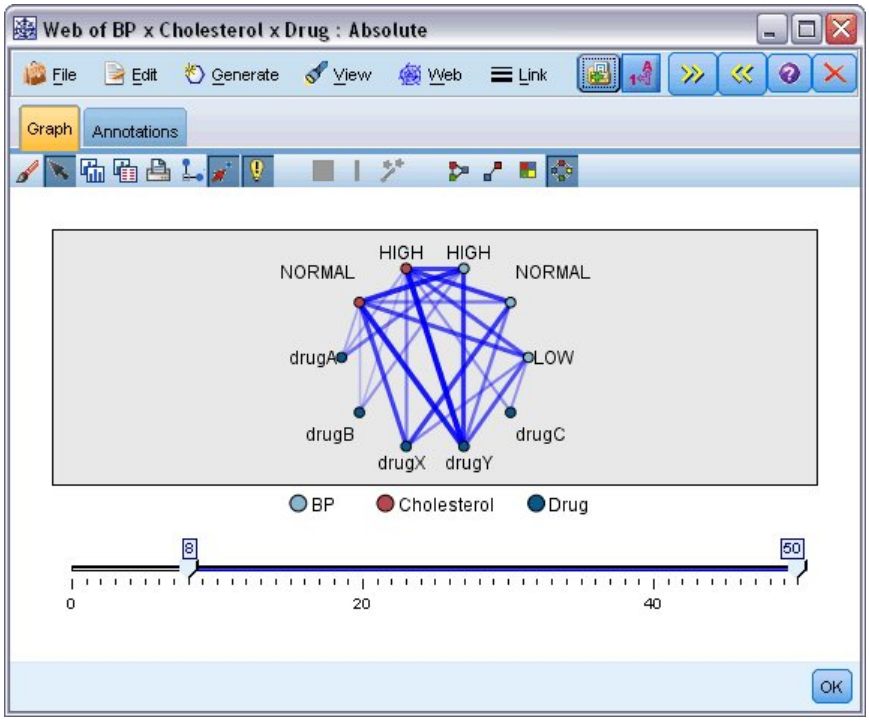
Rysunek 38. Wykres sieciowy przedstawiający etykiety jako węzły

Korzystanie z wykresu sieciowego

Węzły sieciowe służą do przedstawienia siły relacji między wartościami dwu lub więcej pól symbolicznych. Połączenia są wyświetlane na wykresie za pomocą różnych typów linii, wskazujących połączenia o zwiększającej się sile. Węzła sieciowego można na przykład używać do eksploracji relacji pomiędzy poziomami cholesterolu, ciśnieniem krwi i lekiem, który był skuteczny w leczeniu schorzeń pacjentów.

- Silne połączenia są wyświetlane jako gruba linia. Oznacza to, że dwie wartości są silnie powiązane i powinny być dalej eksplorowane.
- Połączenia o średniej sile są wyświetlane jako linia o normalnej grubości.
- Słabe połączenia są wyświetlane jako linia przerywana.
- Jeśli pomiędzy dwiema wartościami nie ma żadnej linii, oznacza to, że te wartości nigdy nie wystąpiły w tym samym rekordzie lub że taka kombinacja występuje w wielu rekordach poniżej wartości granicznej określonej w oknie dialogowym węzła sieciowego.

Po utworzeniu węzła sieciowego dostępnych jest kilka opcji skorygowania obrazu wykresu i wygenerowania węzłów na potrzeby przeprowadzenia dalszej analizy.



Rysunek 39. Wykres sieciowy przedstawiający liczbę silnych relacji, takich jak prawidłowe ciśnienie krwi z lekiem DrugX oraz wysoki poziom cholesterolu z lekiem DrugY

W przypadku obu węzłów, węzła sieciowego i węzła sieci kierunkowej, można:

- Zmienić układ wykresu sieciowego.
- Ukryć punkty w celu uproszczenia wyświetlanego obrazu.
- Zmienić wartości graniczne odpowiadające za style linii.
- Wyróżnić linie pomiędzy wartościami, aby wskazać „wybraną” relację.
- Generować węzeł selekcji dla co najmniej jednego „wybranego” rekordu lub węzeł wyliczeń typu flaga powiązany z co najmniej jedną relacją w sieci.

Aby dostosować punkty

- Punkty można **przesuwać**, klikając je przyciskiem myszy i przeciągając do nowej lokalizacji. Sieć zostanie ponownie narysowana, aby odzwierciedlić nową lokalizację.
- Punkty można **ukryć**, klikając je prawym przyciskiem myszy i wybierając opcję **Ukryj** lub **Ukryj i zaplanuj ponownie** z menu kontekstowego. Opcja **Ukryj** po prostu ukrywa wybrany punkt i linie z nim powiązane. **Ukryj i zaplanuj ponownie** powoduje ponowne narysowanie sieci, tak aby była dostosowana do wprowadzonych zmian. Ręczne przesunięcia zostają cofnięte.
- Wszystkie ukryte punkty można **wyświetlić**, wybierając opcję **Odsłoń wszystko** lub **Odsłoń i zaplanuj ponownie** z menu sieciowego w oknie wykresu. Wybranie opcji **Odsłoń i zaplanuj ponownie** spowoduje ponowne narysowanie sieci, tak aby zawierała wszystkie wcześniej ukryte punkty i ich połączenia.

Aby wybrać lub wyróżnić linie

Wyróżnione linie są wyświetlane na czerwono.

1. Aby wybrać pojedynczą linię, należy ją kliknąć lewym przyciskiem myszy.
2. Aby wybrać wiele linii, należy wykonać jedną z następujących czynności:

- Korzystając z kursora, narysuj okrąg wokół punktów, których linie mają zostać zaznaczone.
- Przytrzymaj klawisz Ctrl i kliknij lewym przyciskiem myszy pojedyncze linie, jakie mają zostać zaznaczone.

Można usunąć zaznaczenie wszystkich wybranych linii, klikając tło wykresu lub wybierając opcję **Wyczyść zaznaczenie** z menu sieciowego w oknie wykresu.

Aby wyświetlić inny układ wykresu sieciowego

W celu zmiany układu wykresu z menu sieciowego należy wybrać opcję **Układ kołowy**, **Układ sieciowy**, **Układ kierunkowy** lub **Układ siatki**.

Aby włączyć lub wyłączyć suwak połączeń

Z menu widoku należy wybrać opcję **Suwak połączeń**.

Aby wybrać lub oznaczyć rekordy dla jednej relacji

1. Kliknij prawym przyciskiem myszy linię reprezentującą wybraną relację.
2. Z menu kontekstowego wybierz opcję **Utwórz węzeł selekcji dla łącza** lub opcję **Utwórz węzeł wyliczeń dla łącza**.

Węzeł selekcji lub wyliczeń jest automatycznie dodawany do obszaru roboczego strumienia z określonymi odpowiednimi opcjami i warunkami:

- Węzeł selekcji wybiera wszystkie rekordy dla danej relacji.
- Węzeł wyliczeń generuje flagę wskazującą, czy wybrana relacja obejmuje warunek typu „prawda” dla rekordów w całym zbiorze danych. Nazwa zmiennej typu flaga jest tworzona poprzez połączenie dwóch wartości tworzących relację z podkreśleniem, np. LOW_drugC lub drugC_LOW.

Aby wybrać lub oznaczyć rekordy dla grupy relacji

1. Wybierz linie na wykresie sieciowym reprezentujące określone relacje.
2. Z menu Utwórz w oknie wykresu wybierz opcję **Węzeł selekcji („And”)**, **Węzeł selekcji („Or”)**, **Węzeł wyliczeń („And”)** lub **Węzeł wyliczeń („Or”)**.

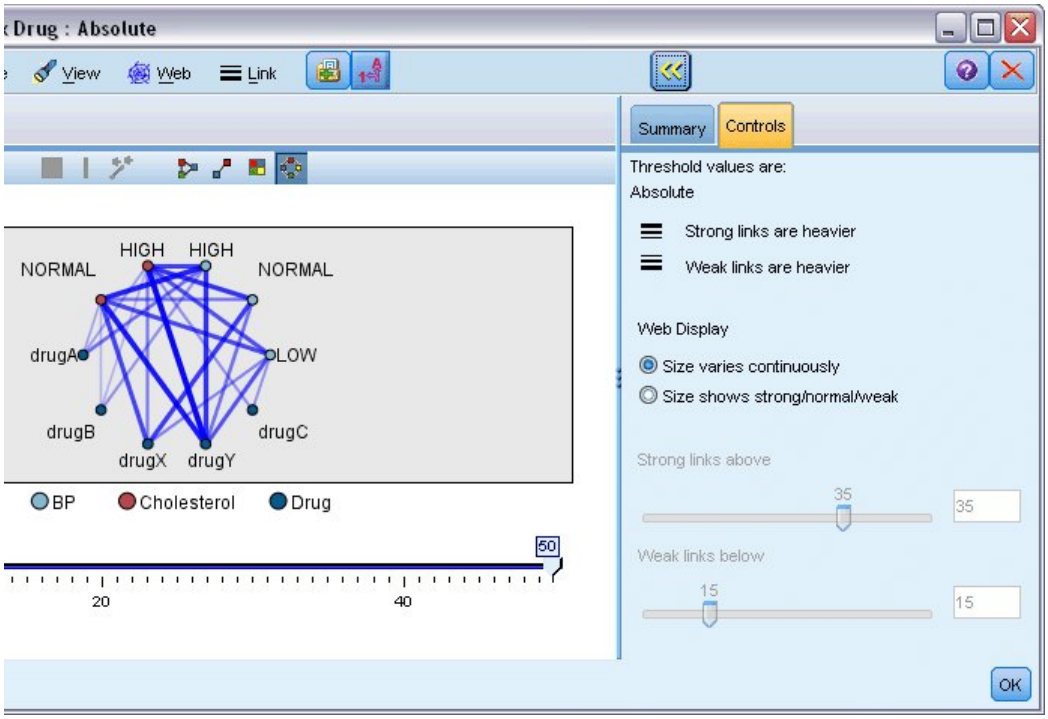
- Węzły „Or” zapewniają alternatywny wybór warunków. Oznacza to, że węzeł będzie miał zastosowanie do rekordów, dla których obowiązują dowolne z wybranych relacji.
- Węzły „And” zapewniają połączenie warunków. Oznacza to, że węzeł będzie miał zastosowanie wyłącznie do rekordów, dla których obowiązują wszystkie wybrane relacje. Jeśli dowolne z wybranych relacji wzajemnie się wykluczają, wystąpi błąd.

Po wyborze węzeł selekcji lub węzeł wyliczeń jest automatycznie dodawany do obszaru roboczego strumienia z określonymi odpowiednimi opcjami i warunkami.

Uwaga: Podczas filtrowania wyświetlanych łączy (za pomocą suwaka na wykresie sieciowym lub elementu sterującego **Pokaż tylko łącza powyżej** na karcie Opcje węzła sieciowego), może dojść do sytuacji, w której wszystkie nadal wyświetlane łącza są łączami pojedynczej wartości (innymi słowy wszystkie są łączami słabymi, wszystkie są łączami normalnymi lub wszystkie są łączami silnymi, zgodnie z definicją w elementach sterujących **Słabe łącza poniżej** i **Silne łącza powyżej** na karcie Opcje węzła sieciowego). Jeśli tak się stanie, na wynikowym wykresie sieciowym wszystkie łącza będą wyświetlane w postaci linii średniej grubości.

Dostosowywanie wartości granicznych dla wykresu sieciowego

Po utworzeniu wykresu sieciowego można dostosować wartości graniczne kontrolujące style linii, używając suwaka na pasku narzędzi w celu zmiany widocznej linii określającej wartości minimalne. Można również wyświetlić dodatkowe opcje wartości granicznych, klikając przycisk żółtej podwójnej strzałki na pasku narzędzi w celu rozwinięcia okna wykresu sieciowego. Następnie należy kliknąć zakładkę **Ustawienia**, aby wyświetlić dodatkowe opcje.



Rysunek 40. Rozwinięte okno z opcjami wyświetlania i wartości granicznych

Wartości graniczne. Wyświetla typ wartości granicznej wybranej podczas tworzenia w oknie dialogowym węzła sieciowego.

Silne łącza są grubsze. Ta opcja jest wybrana domyślnie; jest to standardowy sposób wyświetlania łączy pomiędzy zmiennymi.

Słabe łącza są grubsze. Tę opcję należy wybrać, aby odwrócić znaczenie łączy wyświetlanych jako linie pogrubione. Ta opcja jest często używana do wykrywania oszustw lub badania wartości odstających.

Wygląd wykresu sieciowego. Umożliwia określenie opcji sterowania wielkością łączy w wykresie wynikowym:

- **Ciągła zmiana grubości łączy.** Umożliwia wyświetlenie zakresu wielkości łączy odzwierciedlających różne siły połączeń na podstawie rzeczywistych wartości danych.
- **Styl wyróżnia Silne/Normalne/Słabe łącza.** Zaznaczenie tej opcji powoduje wyświetlenie trzech sił połączeń — silne, normalne i słabe. Punkty odcięcia dla tych kategorii można określić powyżej oraz na wykresie końcowym.

Silne łącza powyżej. Określa wartość graniczną dla silnych połączeń (grube linie) i normalnych połączeń (zwykłe linie). Wszystkie połączenia nad tą wartością są uznawane za silne. Korzystając z suwaka, można skorygować wartość lub można wprowadzić liczbę w polu.

Słabe łącza poniżej. Określa liczbę wskazującą wartość graniczną dla słabych połączeń (linie przerywane) i normalnych połączeń (zwykłe linie). Wszystkie połączenia poniżej tej wartości są uznawane za słabe. Korzystając z suwaka, można skorygować wartość lub można wprowadzić liczbę w polu.

Po skorygowaniu wartości granicznych dla wykresu sieciowego można ponownie zaplanować lub ponownie narysować wykres sieciowy, stosując nowe wartości graniczne za pośrednictwem menu sieciowego znajdującego się na pasku narzędzi wykresu sieciowego. Po znalezieniu ustawień odsłaniających najbardziej istotne wzorce można zaktualizować oryginalne ustawienia w węźle sieciowym (zwanym również nadrzędnym węzłem sieciowym), wybierając opcję **Aktualizuj węzeł nadrzędny** z menu sieciowego w oknie wykresu.

Tworzenie podsumowania dla wykresu sieciowego

Istnieje możliwość utworzenia dokumentu podsumowania wykresu sieciowego, który będzie zawierał listę silnych, średnich i słabych połączeń; w tym celu należy kliknąć żółty przycisk podwójnej strzałki na pasku narzędzi, aby rozwinąć okno wykresu sieciowego. Następnie należy kliknąć kartę **Podsumowanie**, aby wyświetlić tabele dla każdego typu łącza. Tabele można zwijać lub rozwijać za pomocą przycisków przełączania.

Aby wydrukować podsumowanie, należy wybrać następujące opcje z menu w oknie wykresu sieciowego:

Plik > Wydruk podsumowania

---

węzeł ewaluacji

Węzeł ewaluacji umożliwia łatwą ewaluację i porównywanie modeli predykcyjnych w celu wybrania najlepszego modelu do danego zastosowania. Wykresy ewaluacyjne przedstawiają skuteczność modeli w przewidywaniu konkretnych wyników. Ich działanie polega na sortowaniu rekordów na podstawie wartości przewidywanej i ufności przewidywania, dzieleniu rekordów na grupy o równej wielkości (**kwantyle**), a następnie wykreśleniu wartości kryterium biznesowego dla każdego kwantyla, od najwyższego do najniższego. Modele wielokrotne prezentowane są jako osobne linie na wykresie.

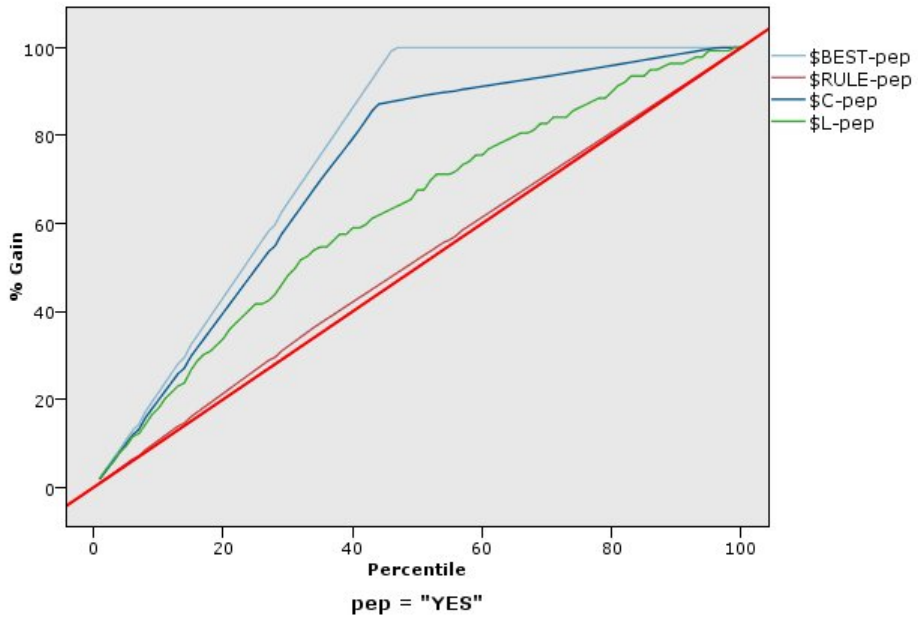
Wyniki uzyskuje się poprzez zdefiniowanie konkretnej wartości lub zakresu wartości jako **trafienia**. Trafienia zwykle oznaczają sukces (np. sprzedaż klientowi) lub zdarzenie będące przedmiotem zainteresowania (np. konkretną diagnozę medyczną). Kryteria trafień można zdefiniować na karcie Opcje okna dialogowego. Można też korzystać z następujących standardowych kryteriów trafień:

- Zmienne wynikowe typu **Flaga** nie wymagają dodatkowych objaśnień: trafienia odpowiadają wartościom *prawda*.
- W przypadku zmiennych wynikowych typu **Nominalne** trafieniem jest pierwsza wartość w zbiorze.
- W przypadku zmiennych wynikowych typu **Ciągłe** trafieniami są wartości większe od połowy zakresu wartości zmiennej.

Istnieje sześć typów wykresów ewaluacyjnych, a na każdym z nich wyróżnione jest inne kryterium ewaluacji.

Wykresy korzyści

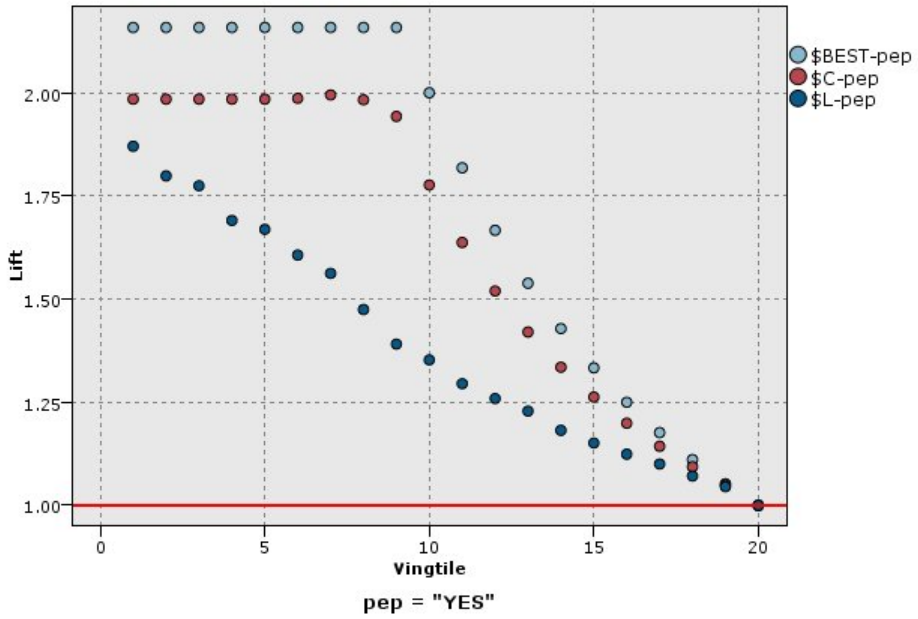
Korzyści zdefiniowane są jako proporcja łącznej liczby trafień w każdym kwantylu. Korzyści oblicza się według wzoru (liczba trafień w kwantylu / łączna liczba trafień) × 100%.



Rysunek 41. Wykresy korzyści (skumulowane) z linią podstawy, najlepszą linią i regułą biznesową.

Wykresy przyrostów

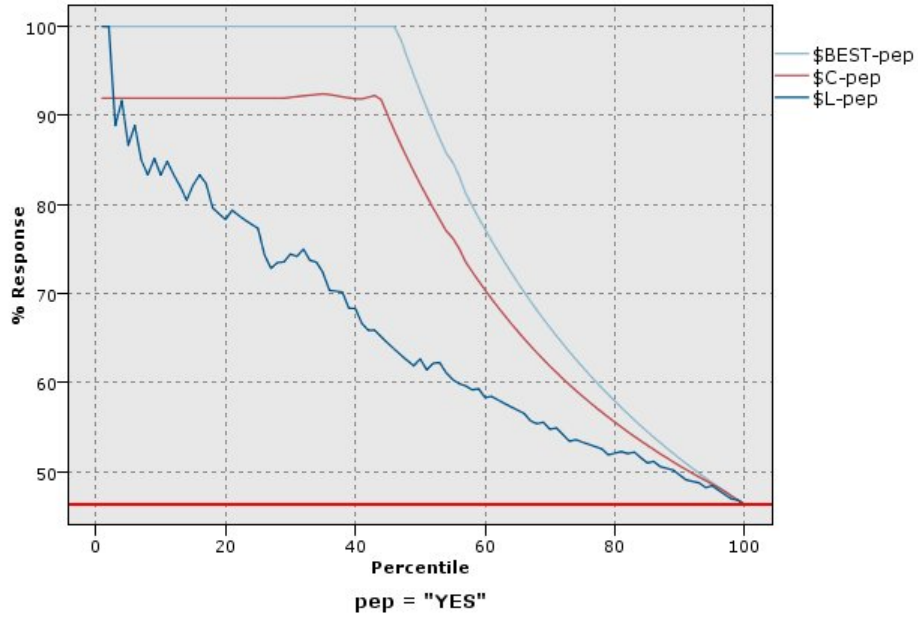
Przyrost porównuje odsetek rekordów w każdym kwantylu będących trafieniami z łącznym odsetkiem trafień w danych uczących. Obliczany jest według wzoru (trafienia w kwantylu / rekordy w kwantylu) / (łącznie trafień / łącznie rekordów).



Rysunek 42. Wykres przyrostu (skumulowany) z punktami i najlepszą linią

Wykresy odpowiedzi

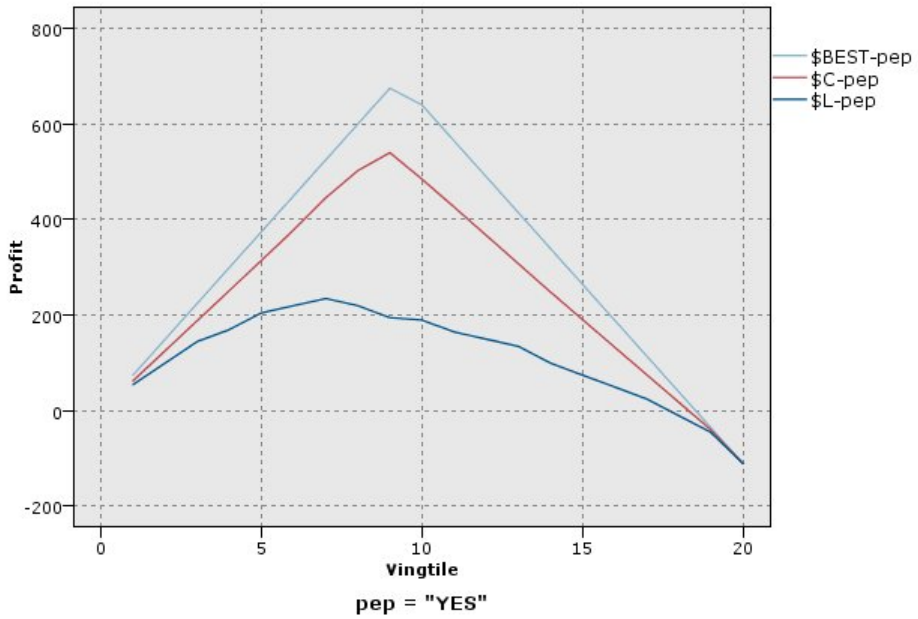
Odpowiedź jest to po prostu odsetek rekordów w kwantylu będących trafieniami. Odpowiedź obliczana jest według wzoru (trafienia w kwantylu / rekordy w kwantylu) × 100%.



Rysunek 43. Wykres odpowiedzi (skumulowany) z najlepszą linią

Wykresy zysków

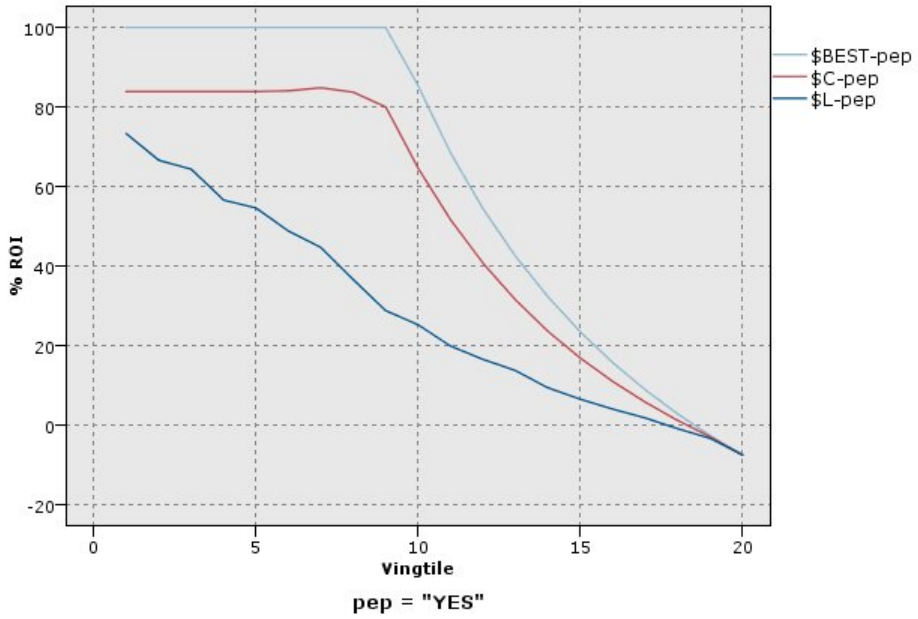
Zysk równy jest **przychodowi** w każdy rekordzie pomniejszonemu o **koszt** w tym rekordzie. Zysk z kwantyla jest po prostu sumą zysków z wszystkich rekordów w tym kwantylu. Przyjmuje się, że przychody mają zastosowanie tylko do trafień, ale koszty — do wszystkich rekordów. Zyski i koszty mogą być stałe lub zdefiniowane przez zmienne w danych. Zyski oblicza się według wzoru (suma przychodów z rekordów w kwantylu − suma kosztów z rekordów w kwantylu).



Rysunek 44. Wykres zysków (skumulowany) z najlepszą linią

Wykresy zwrotu z inwestycji

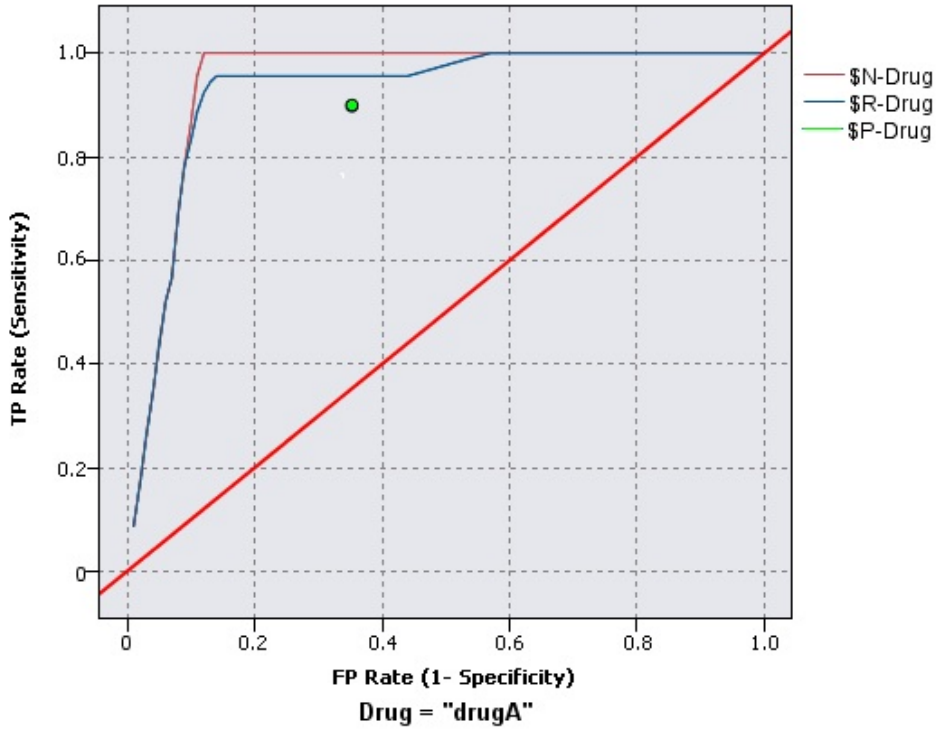
Zwrot z inwestycji (ROI — Return on investment) jest podobny do zysku w tym, że obliczany jest na podstawie przychodów i kosztów. Zwrot z inwestycji jest porównaniem zysków z kosztami z danego kwantyla. Zwrot z inwestycji oblicza się według wzoru (zyski z kwantyla / koszty z kwantyla) × 100%.



Rysunek 45. Wykres zwrotu z inwestycji (skumulowany) z najlepszą linią

Wykresy ROC

Wykresy ROC (oceny poprawności klasyfikatora) mogą być używane tylko w przypadku klasyfikatorów binarnych. Za pomocą wykresów ROC można zwizualizować, uporządkować i wybrać klasyfikatory na podstawie ich skuteczności. Wykres ROC wykreśla stosunek wskaźnika prawdziwie dodatnich (czułości) do wskaźnika fałszywie dodatnich, tzn. jest to wykres czułości klasyfikatora. Wykres ROC przedstawia wzajemną zależność korzyści (obserwacji prawdziwie dodatnich) do kosztów (obserwacji fałszywie dodatnich). Obserwacja prawdziwie dodatnia zachodzi wtedy, gdy zdarzenie jest trafieniem i zostało sklasyfikowane jako trafienie. Wskaźnik prawdziwie dodatnich jest liczony jako iloraz obserwacji prawdziwie dodatnich do liczby wystąpień, które są faktycznie trafieniami. Obserwacja fałszywie dodatnia zachodzi wtedy, gdy zdarzenie jest nietrafione, lecz zostało sklasyfikowane jako trafienie. Wskaźnik fałszywie dodatnich jest liczony jako iloraz obserwacji prawdziwie dodatnich do liczby wystąpień, które nie są trafieniami.



Rysunek 46. Wykres ROC z najlepszą linią

Wykresy ewaluacyjne mogą być także skumulowane, tak aby każdy punkt był równy wartości odpowiedniego kwantyla plus wartości wszystkich wyższych kwantyli. Wykresy skumulowane zwykle lepiej obrazują ogólną wydajność modeli, natomiast wykresy nieskumulowane często lepiej nadają się do ujawniania konkretnych obszarów problemowych w modelach.

Uwaga: Węzeł Ewaluacja nie dopuszcza przecinków w nazwach zmiennych. Jeśli istnieją nazwy zmiennych zawierające przecinki, należy usunąć przecinki albo ująć nazwy zmiennych w cudzysłowy.

Karta wykresu ewaluacyjnego

Typ wykresu. Należy wybrać jeden z następujących typów: **Korzyści**, **Odpowiedź**, **Wzrost**, **Zyski**, **ROI** (zwrot z inwestycji) lub **ROC** (ocena poprawności klasyfikatora).

Wykres skumulowany. Należy wybrać tę opcję, aby utworzyć wykres skumulowany. Wartości na wykresach skumulowanych są wykreślane dla każdego kwantyla plus wszystkie wyższe kwantyle. (Opcja **Wykres skumulowany** jest niedostępna dla wykresów ROC).

Dołącz linię bazową. Tę opcję należy wybrać, aby dołączyć do wykresu linię bazową, wskazującą idealny rozkład losowy trafień, w którym ufność staje się nieistotna. (Opcja **Dołącz linię bazową** jest niedostępna dla wykresów ROI).

Pokaż linię najlepszych wartości. Tę opcję należy wybrać, aby uwzględnić na wykresie linię najlepszych wartości, wskazującą idealną ufność (liczba trafień = 100% obserwacji). (Opcja **Pokaż linię najlepszych wartości**) jest niedostępna dla wykresów ROC).

Użyj kryteriów zysku dla wszystkich typów wykresów. Tę opcję należy zaznaczyć, aby podczas obliczania miar ewaluacyjnych używać kryteriów zysku (koszt, przychód i waga) zamiast normalnej liczby trafień. W przypadku modeli z określonymi zmiennymi przewidywanymi, takimi jak model, który przewiduje przychód uzyskany od klienta w odpowiedzi na ofertę, wartość zmiennej przewidywanej zapewnia lepszą miarę wydajności modelu niż liczba trafień. Zaznaczenie tej opcji aktywuje zmienne **Koszty**, **Przychód** i **Waga** na wykresach zysków, odpowiedzi i przyrostów. Aby użyć kryteriów zysku dla tych trzech typów wykresów, zaleca się ustawienie wartości **Przychód** na zmienną przewidywaną, **Koszt** na 0,0, dzięki czemu zysk będzie równy przychodowi, oraz określenie warunku trafienia zdefiniowanego przez użytkownika dla „prawdy", tak aby wszystkie rekordy były zliczane jako trafienia. (Opcja **Użyj kryteriów zysku dla wszystkich typów wykresów** jest niedostępna dla wykresów ROC).

Znajdź zmienne predykcyjne i przewidywane, wykorzystując. Można wybrać opcję **Metadane zmiennej wyjściowej modelu**, aby wyszukać zmienne predykcyjne na wykresie, używając ich metadanych lub opcję **Format nazwy zmiennej**, aby wyszukać zmienne według nazwy.

Zmienne oceniające. Należy zaznaczyć to pole wyboru, aby aktywować selektor zmiennych oceniających. Następnie należy wybrać co najmniej jedną zmienną zakresu lub ilościową zmienną oceniającą; to znaczy zmienne, które nie są ściśle modelami predykcyjnymi, ale mogą być przydatne do rangowania rekordów z uwzględnieniem skłonności jako trafienia. Węzeł ewaluacji umożliwia porównanie dowolnej kombinacji co najmniej jednej zmiennej oceniającej w jednym lub większej liczbie modeli predykcyjnych. Przykładem może być porównanie kilku zmiennych RFM z najlepszym modelem predykcyjnym.

Zmienna przewidywana. Zmienną przewidywaną należy wybrać za pomocą selektora zmiennych. Należy wybrać dowolną określoną zmienną flagi lub zmienną nominalną z co najmniej dwoma wartościami.

Uwaga: Ta zmienna przewidywana ma zastosowanie tylko do zmiennych oceniających (modele predykcyjne definiują własne przewidywane) i jest ignorowana, jeśli na karcie opcji ustawiono kryterium trafień użytkownika.

Rozdziel na podzbiory. Jeśli do podzielenia rekordów na próbę uczenia, testowania i walidacji używana jest zmienna dzieląca na podzbiory, należy wybrać tę opcję, aby wyświetlić wykres ewaluacyjny osobno dla każdego podzbioru. Więcej informacji można znaleźć w temacie "Węzeł Partycja" na stronie 176.

Uwaga: Podczas podział na podzbiory rekordy zawierające wartości null w zmiennej dzielącej na podzbiory zostaną wykluczone z oceny. Nie będzie to problemem w przypadku użycia węzła Partycja, ponieważ węzły podziału na podzbiory nie generują wartości null.

Wykres. Należy z listy rozwijanej wybrać wielkości kwantyli do wykreślenia na wykresie. Możliwe opcje to: **Kwartyle**, **Kwintyle**, **Decyle**, **Vingtyle**, **Percentyle** i **1000-tyle**. (Opcja **Wykres** jest niedostępna dla wykresów ROC).

Styl. Można wybrać opcje **Liniowy** lub **Punkt**.

Na wszystkich typach wykresów, z wyjątkiem wykresów ROC, dodatkowe elementy sterujące umożliwiają określenie kosztów, przychodu i wagi.

- **Koszty.** Pozwala określić koszty powiązane z poszczególnymi rekordami. Można wybrać przychód **Stały** lub **Zmienny**. W przypadku kosztów stałych należy określić wartość kosztu. W przypadku kosztów zmiennych należy kliknąć przycisk Selektor zmiennych, aby wybrać zmienną stanowiącą zmienną kosztu. (Opcja **Koszty** jest niedostępna dla wykresów ROC).
- **Przychód.** Określa przychód powiązany z poszczególnymi rekordami reprezentującymi trafienie. Można wybrać przychód **Stały** lub **Zmienny**. W przypadku przychodów stałych należy określić wartość przychodu. W przypadku przychodów zmiennych należy kliknąć przycisk Selektor zmiennych, aby wybrać zmienną stanowiącą zmienną przychodu. (Opcja **Przychód** jest niedostępna dla wykresów ROC).

- **Waga.** Jeśli rekordy w danych reprezentują więcej niż jedną jednostkę, można użyć wag częstości, aby skorygować wyniki. Należy określić wagę powiązaną z poszczególnymi rekordami, używając wag **Stała** lub **Zmienna**. W przypadku wag stałych należy określić wartość wagi (liczba jednostek na rekord). W przypadku wag zmiennych należy kliknąć przycisk Selektor zmiennych, aby wybrać zmienną stanowiącą zmienną wagi. (Opcja **Waga** jest niedostępna dla wykresów ROC).

Karta opcji wykresu ewaluacyjnego

Karta Opcje wykresów ewaluacyjnych zapewnia elastyczność definiowania trafień, kryteriów oceniania oraz reguł biznesowych wyświetlanych na wykresie. Można również ustawić opcje eksportowania wyników ewaluacji modelu.

Sukces zdefiniowany przez użytkownika. Tę opcję należy wybrać, aby określić niestandardowy warunek używany do wskazania trafienia. Opcja ta jest przydatna do definiowania wyniku zamiast wyliczania go na podstawie typu zmiennej przewidywanej i kolejności wartości.

- **Warunek.** Jeśli wcześniej wybrano opcję **Sukces zdefiniowany przez użytkownika**, należy określić wyrażenie CLEM dla warunku trafienia. Przykładowo, warunek `@TARGET = "YES"` jest poprawnym warunkiem określającym, że wartość *Yes* (Tak) dla zmiennej przewidywanej zostanie zliczona jako trafienie podczas ewaluacji. Określony warunek zostanie użyty dla wszystkich zmiennych przewidywanych. Aby utworzyć warunek, należy wpisać zmienną lub użyć konstruktora wyrażeń do wygenerowania wyrażenia warunku. Jeśli dane są określone, można wstawić wartości bezpośrednio z konstruktora wyrażeń.

Ocena definiowana przez użytkownika. Tę opcję należy wybrać, aby określić warunek używany do oceniania obserwacji przed przypisaniem ich do kwantyli. Domyślna ocena jest wyznaczana na podstawie wartości przewidywanej i ufności. Zmienna wyrażenia umożliwia określenie niestandardowego wyrażenia oceniania.

- **Wyrażenie.** Należy określić wyrażenie CLEM użyte do oceniania. Przykładowo, jeśli wynik liczbowy w zakresie od 0 do 1 jest uporządkowany tak, że niższe wartości są lepsze niż wyższe, można zdefiniować trafienie jako `@TARGET < 0,5`, a powiązaną ocenę jaki `1 − @PREDICTED`. Wyrażenie oceniania musi dawać wynik w postaci wartości liczbowej. Aby utworzyć warunek, należy wpisać zmienną lub użyć konstruktora wyrażeń do wygenerowania wyrażenia warunku.

Dołącz regułę biznesową. Tę opcję należy wybrać, aby określić warunek reguły odzwierciedlający odpowiednie kryteria. Przykładowo można wyświetlić regułę dla wszystkich obserwacji, gdzie `mortgage = "Y" and income >= 33000` (hipoteka = T, a przychód = 33000). Reguły biznesowe są rysowane na wykresie i oznaczane etykietami w wartościach kluczowych jako *Rule* (Reguła). (Opcja **Dołącz regułę biznesową** jest niedostępna dla wykresów ROC).

- **Warunek.** Należy określić wyrażenie CLEM, jakie zostało użyte do zdefiniowania reguły biznesowej na wykresie wynikowym. Wystarczy wpisać zmienną lub użyć konstruktora wyrażeń do wygenerowania wyrażenia warunku. Jeśli dane są określone, można wstawić wartości bezpośrednio z konstruktora wyrażeń.

Eksportuj wyniki do pliku. Tę opcję należy wybrać, aby wyeksportować wyniki oceny modelu do rozdzielonego pliku tekstowego. Można odczytać ten plik, aby przeprowadzić wyspecjalizowane analizy dla obliczonych wartości. W celu przeprowadzenia eksportu należy ustawić następujące opcje:

- **Nazwa pliku.** Należy wprowadzić nazwę pliku dla pliku wynikowego. Przycisk wielokropka **(...)** umożliwia przejście do odpowiedniego folderu.
- **Separator.** Należy wprowadzić znak, taki jak przecinek lub spacja, aby użyć separatora zmiennych.

Uwzględnij nazwy zmiennych. Tę opcję należy wybrać, aby uwzględnić nazwy zmiennych jako pierwszy wiersz w pliku wynikowym.

Nowy wiersz po każdym rekordzie. Tę opcję należy wybrać, aby każdy rekord rozpoczynał się od nowego wiersza.

Karta wyglądu wykresu ewaluacyjnego

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Tekst. Można zaakceptować automatycznie wygenerowaną etykietę tekstową lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Etykieta X. Można zaakceptować automatycznie wygenerowaną etykietę osi *x* (pozioma) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Etykieta Y. Można zaakceptować automatycznie wygenerowaną etykietę osi *y* (pionowa) lub wybrać opcję **Użytkownika**, aby określić własną etykietę.

Wyświetl linie siatki. Ta opcja jest domyślnie zaznaczona; wyświetla linie siatki za wykresem, dzięki czemu znacznie łatwiej można określić region i punkty odcięcia przedziału. Linie siatki zawsze są wyświetlane na biało, jeśli tło wykresu nie jest białe; wówczas linie są wyświetlane na szaro.

Odczytywanie wyników ewaluacji modelu

Interpretacja wykresu ewaluacyjnego zależy od zakresu dla typu wykresu, jednak niektóre charakterystyki są wspólne dla wszystkich wykresów ewaluacyjnych. W przypadku wykresów skumulowanych wyżej położone linie oznaczają lepsze modele, szczególnie jeśli znajdują się po lewej stronie wykresu. W wielu przypadkach podczas porównywania wielu modeli linie będą się przecinały, tak więc jeden model będzie położony wyżej w jednej części wykresu, a inny będzie znajdował się wyżej w innej jego części. W takim przypadku należy rozważyć, jaka część próby (która definiuje punkt na osi *x*) będzie brana pod uwagę podczas podejmowania decyzji o wyborze modelu.

Większość wykresów nieskumulowanych będzie bardzo podobna. W przypadku dobrych modeli wykresy nieskumulowane powinny unosić się w kierunku lewej strony wykresu i obniżać w prawą jego stronę. (Jeśli wykres nieskumulowany przedstawia wzór piłokształtny, można go wygładzić, redukując liczbę kwantyli i ponownie wykonując wykres). Pochylenia po lewej stronie wykresu lub „wzniesienia” po prawej mogą oznaczać obszary, w których przewidywanie modelu jest słabe. Płaska linia biegnąca przez cały wykres oznacza model, który w istocie nie dostarcza żadnych informacji.

Wykresy korzyści. Skumulowane wykresy korzyści zawsze rozpoczynają się od 0% i kończą na 100% (od lewej do prawej). W dobrym modelu wykres korzyści będzie stopniowo unosił się w kierunku 100%, a następnie przejdzie do linii poziomej. Model, który nie dostarcza żadnych informacji, będzie przebiegał ukośnie od dołu po lewej stronie ku górze po prawej stronie (pokazano na wykresie przy wybranej opcji **Dołącz linię bazową**).

Wykresy wzrostów. Skumulowane wykresy wzrostów mają tendencję do rozpoczynania się powyżej wartości 1,0 i stopniowego opadania do wartości 1,0 (od lewej do prawej). Prawy koniec wykresu reprezentuje cały zbiór danych, dlatego iloraz trafień w skumulowanych kwantylach do trafień w danych wynosi 1,0. W dobrym modelu wzrost powinien rozpoczynać się wyraźnie powyżej wartości 1,0 po lewej stronie, pozostawać płaski w górnej części w kierunku prawej strony, a następnie gwałtownie spadać do 1,0 po prawej stronie. W modelu, który nie dostarcza żadnych informacji, linia będzie utrzymywać się w pobliżu wartości 1,0 na całym wykresie. (Jeśli wybrano opcję **Dołącz linię bazową**, jako odniesienie na wykresie pokazana będzie pozioma linia dla wartości 1,0).

Wykresy odpowiedzi. Skumulowane wykresy odpowiedzi często wyglądają podobnie do wykresów wzrostów, z wyjątkiem skalowania. Wykresy odpowiedzi zwykle rozpoczynają się w pobliżu wartości 100% i stopniowo opadają aż do wartości ogólnego wskaźnika odpowiedzi (łączna liczba trafień/łączna liczba rekordów) na prawym końcu wykresu. W dobrym modelu linia będzie rozpoczynać się w pobliżu lub na wartości 100% po lewej stronie, pozostanie płaska w górnej części w kierunku prawej strony, a następnie będzie gwałtownie spadać w kierunku ogólnego wskaźnika odpowiedzi po prawej stronie wykresu. W modelu, który nie dostarcza żadnych informacji, linia będzie utrzymywać się w pobliżu wartości ogólnego wskaźnika odpowiedzi na całym wykresie. (Jeśli wybrano opcję **Dołącz linię bazową**, jako odniesienie na wykresie pokazana będzie pozioma linia dla wartości ogólnego wskaźnika odpowiedzi).

Wykresy zysków. Skumulowane wykresy zysków przedstawiają sumę zysków przy wzroście wielkości wybranej próby, w kierunku od lewej do prawej. Wykresy zysków zwykle rozpoczynają się w pobliżu 0, stabilnie rosną w kierunku prawej strony aż do osiągnięcia wartości szczytowej lub ustabilizowanego poziomu na środku, po czym

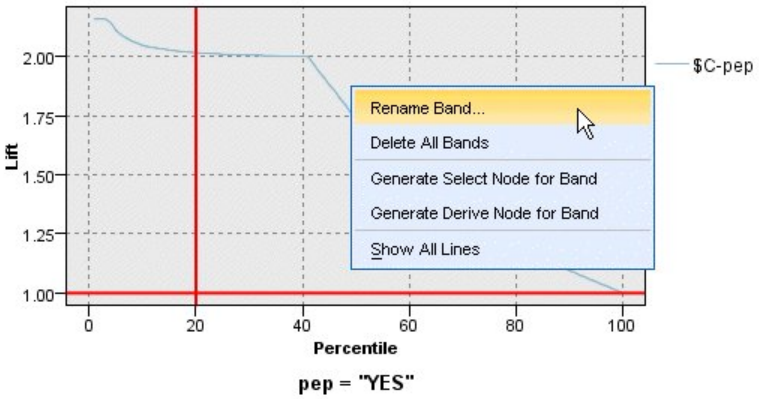
opadają w kierunku prawego końca wykresu. W dobrym modelu zyski będą przedstawiały dobrze zdefiniowaną wartość szczytową w pobliżu środka wykresu. W modelu, który nie dostarcza informacji, linia będzie stosunkowo prosta i może unosić się, opadać lub być pozioma, w zależności od obowiązującej struktury kosztów/przychodów.

Wykresy ROI. Skumulowane wykresy ROI (zwrotu z inwestycji) są podobne do wykresów odpowiedzi i wykresów wzrostów, z wyjątkiem skalowania. Wykresy ROI zwykle rozpoczynają się powyżej wartości 0% i stopniowo opadają, aż do osiągnięcia ogólnego zwrotu z inwestycji dla całego zbioru danych (który może mieć wartość ujemną). W dobrym modelu linia powinna rozpoczynać się wyraźnie powyżej wartości 0%, pozostawać płaska w górnej części w kierunku prawej strony, a następnie dość gwałtownie spadać w kierunku ogólnego zwrotu z inwestycji po prawej stronie wykresu. W modelu, który nie dostarcza żadnych informacji, linia powinna utrzymywać się w pobliżu wartości ogólnego zwrotu z inwestycji.

Wykresy ROC. Krzywe ROC (ocena poprawności klasyfikatora) ogólnie mają kształt skumulowanych wykresów korzyści. Krzywa rozpoczyna się od współrzędnej (0,0) i kończy na współrzędnej (1,1) (od lewej do prawej). Wykres, który unosi się stopniowo w kierunku współrzędnej (0,1), a następnie wyrównuje się oznacza dobry klasyfikator. Model, który klasyfikuje instancje losowo jako trafienia lub braki trafień, będzie przebiegał ukośnie od dołu po lewej stronie ku górze po prawej stronie (wyświetlane na wykresie po wybraniu opcji **Dołącz linię bazową**). Jeśli dla modelu nie podano żadnej zmiennej ufności, model będzie przedstawiony jako pojedynczy punkt. Klasyfikator z optymalną wartością graniczną klasyfikacji znajduje się najbliżej współrzędnej (0,1) lub w lewym górnym rogu wykresu. Lokalizacja ta reprezentuje dużą liczbę instancji, które zostały poprawnie sklasyfikowane jako trafienia, oraz niewielką liczbę instancji, które zostały niepoprawnie sklasyfikowane jako trafienia. Punkty powyżej ukośnej linii reprezentują wyniki dobrej klasyfikacji. Punkty poniżej linii ukośnej reprezentują wyniki słabej klasyfikacji, gorsze niż w przypadku sklasyfikowania instancji w sposób losowy.

Korzystanie z wykresu ewaluacyjnego

Sposób użycia myszy do eksploracji wykresu ewaluacyjnego jest podobny, jak w przypadku histogramu lub wykresu przedziałowego. Oś *x* reprezentuje oceny modelu dla określonych kwantyli, takich jak vingtyle lub decyle.



Rysunek 47. Praca z wykresem ewaluacyjnym

Oś *x* można podzielić na przedziały, podobnie jak w przypadku histogramu, używając ikony rozdzielacza do wyświetlenia opcji automatycznego podziału osi na równe przedziały. Więcej informacji można znaleźć w temacie “Eksplorowanie wykresów” na stronie 275. Można ręcznie edytować granice przedziałów, wybierając z menu Edycja opcję **Przedziały wykresu**.

Po utworzeniu wykresu ewaluacyjnego, zdefiniowaniu przedziałów i sprawdzeniu wyników można użyć opcji menu Utwórz i menu kontekstowego, aby automatycznie utworzyć węzły na podstawie wartości wybranych na wykresie. Więcej informacji można znaleźć w temacie “Generowanie węzłów z wykresów” na stronie 282.

Podczas generowania węzłów na podstawie wykresu ewaluacyjnego zostanie wyświetlony monit o wybranie jednego modelu ze wszystkich modeli dostępnych na wykresie.

Należy wybrać model i kliknąć przycisk **OK**, aby wygenerować nowy węzeł w obszarze roboczym strumienia.

Węzeł Wizualizacja na mapie

Węzeł Wizualizacja na mapie może akceptować wiele połączeń wejściowych i wyświetlać dane geoprzestrzenne na mapie w formie szeregu warstw. Każda warstwa stanowi pojedynczą zmienną geoprzestrzenną; na przykład warstwa podstawowa może być mapą kraju, a nad nią może znajdować się jedna warstwa dróg, jedna warstwa rzek i jedna warstwa miejscowości.

Większość zbiorów danych geoprzestrzennych zwykle zawiera pojedynczą zmienną geoprzestrzenną; jeśli jednak pojedyncze dane wejściowe obejmują wiele zmiennych geoprzestrzennych, można wybrać, które zmienne mają być wyświetlone. W tym samym czasie nie mogą być wyświetlone dwie zmienne z tego samego połączenia wejściowego; można jednak skopiować i wkleić połączenie przychodzące i wyświetlić inną zmienną dla każdego z nich.

Karta wykresu wizualizacji na mapie

Warstwy

W tej tabeli wyświetlane są informacje dotyczące danych wejściowych dla węzła na mapie. Kolejność warstw narzuca kolejność, w jakiej warstwy są wyświetlane w podglądzie mapy oraz w widoku wyjściowym po wykonaniu węzła. Najwyższy wiersz w tabeli oznacza „najwyższą” warstwę, a najniższy wiersz — warstwę „podstawową”; innymi słowy, każda warstwa jest wyświetlana na mapie przed warstwą znajdującą się bezpośrednio pod nią w tabeli.

Uwaga: Jeśli warstwa w tabeli zawiera trójwymiarową zmienną geoprzestrzenną, na wykresie rysowane są tylko osie x i y. Oś z jest ignorowana.

Nazwa

Nazwy są tworzone automatycznie dla każdej warstwy; stosowany jest następujący format: znacznik[węzeł źródłowy:węzeł podłączony]. Domyślnie, znacznik jest wyświetlany w postaci liczby, gdzie 1 reprezentuje pierwsze podłączone źródło danych, 2 drugie itd. W razie potrzeby można nacisnąć przycisk **Edytuj warstwę**, aby zmienić znacznik w oknie dialogowym zmiany opcji warstw mapy. Przykładowo można zmienić znacznik na „roads” (drogi) lub „cities” (miejscowości), aby odzwierciedlał dane wejściowe.

Typ Przedstawia ikonę typu pomiaru zmiennej geoprzestrzennej, jaka została wybrana jako warstwa. Jeśli dane wejściowe zawierają wiele zmiennych z geoprzestrzennym typem pomiaru, domyślnie stosowany jest następujący porządek sortowania:

1. Punkt
2. Łańcuch
3. Wielokąt
4. Multipunkt
5. Multiłańcuch
6. Multiwielokąt

Uwaga: Jeśli dostępne są dwie zmienne z tym samym typem pomiaru, domyślnie wybierana jest pierwsza zmienna (alfabetycznie według nazwy).

Symbol

Uwaga: W tej kolumnie wstawiane są zmienne typu Punkt i Multipunkt.
Przedstawia symbol używany dla zmiennych Punkt lub Multipunkt. W razie potrzeby można nacisnąć przycisk **Edytuj warstwę**, aby zmienić symbol w oknie dialogowym zmiany opcji warstw mapy.

Kolor Przedstawia kolor wybrany do reprezentowania warstwy na mapie. W razie potrzeby można nacisnąć przycisk **Edytuj warstwę**, aby zmienić kolor w oknie dialogowym zmiany opcji warstw mapy. Kolor ma zastosowanie do różnych elementów w zależności od typu pomiaru.

- W przypadku punktów i multipunktów kolor jest stosowany dla symbolu warstwy.
- W przypadku zmiennych typu Łańcuch i Wielokąt kolor jest stosowany dla całego kształtu. Wielokąty zawsze mają czarne kontury; kolor wyświetlany w kolumnie jest kolorem użytym do wypełnienia kształtu.

Podgląd

W tym panelu przedstawiony jest podgląd bieżącego wyboru danych wejściowych w tabeli **Warstwy**. W podglądzie brane są pod uwagę kolejność warstw, symbol, kolor i inne ustawienia wyświetlania powiązane z warstwami oraz, o ile to możliwe, dokonywana jest aktualizacja obrazu po każdej zmianie ustawień. Jeśli w dowolnej części strumienia wprowadzone zostaną zmiany szczegółów, na przykład zmienione zostaną zmienne geoprzestrzenne używane jako warstwy, lub jeśli dokonana zostanie zmiana szczegółów, np. powiązanych funkcji agregujących, konieczne może być kliknięcie przycisku **Odśwież dane** w celu zaktualizowania podglądu.

Opcja **Podgląd** umożliwia wprowadzenie ustawień wyświetlania przed uruchomieniem strumienia. W celu zabezpieczenia przed opóźnieniami w czasie, które mogą wynikać z użycia dużego zbioru danych, podgląd tworzy próby dla każdej warstwy i tworzy obraz na podstawie pierwszych 100 rekordów.

Zmiana warstw na mapie

Okno dialogowe zmiany opcji warstw mapy umożliwia zmianę różnych szczegółów na dowolnej warstwie wyświetlanej na karcie **Wykres** w węźle wizualizacji na mapie.

Szczegóły danych wejściowych

Znacznik

Domyślnie znacznik jest liczbą; można jednak zastąpić tę liczbę bardziej znaczącym znacznikiem, aby ułatwić identyfikację warstwy na mapie. Przykładowo, znacznik może być nazwą danych wejściowych, np. „Cities” (Miejscowości).

Zmienna warstwy

Jeśli dane wejściowe obejmują więcej niż jedną zmienną geoprzestrzenną, należy użyć tej opcji, aby wybrać zmienną, jaka ma być wyświetlana jako warstwa na mapie.

Domyślnie warstwy, z których można dokonać wyboru, ustawione są w następującym porządku sortowania.

- Punkt
- Łańcuch
- Wielokąt
- Multipunkt
- Multiłańcuch
- Multiwielokąt

Wyświetl ustawienia

Użyj grupowania sześciokątnego

Uwaga: Ta opcja wpływa tylko na zmienne typu Punkt i Multipunkt.

Grupowanie sześciokątne (heksagonalne) łączy najbliższe punkty (na podstawie ich współrzędnych x i y) w jeden punkt wyświetlany na mapie. Pojedynczy punkt jest wyświetlany w postaci sześciokąta, ale w rzeczywistości jest renderowany jako wielokąt.

Ponieważ sześciokąt jest renderowany jako wielokąt, wszystkie zmienne punktowe z włączonym grupowaniem sześciokątnym są traktowane jako wielokąty. To oznacza, że jeśli w oknie dialogowym węzła

mapy zostanie wybrana opcja **Porządkuj wg typu**, wszystkie warstwy punktowe z zastosowanym grupowaniem sześciokątnym będą renderowane nad warstwami wielokątów, ale pod warstwami łańcuchów i punktów.

Jeśli grupowanie sześciokątne zostanie zastosowane dla zmiennej typu Multipunkt, zmienna najpierw jest przekształcana na zmienną punktu poprzez kategoryzację wartości multipunktów, co umożliwia obliczenie punktu centralnego. Punkty centralne służą do obliczenia przedziałów sześciokątnych.

Agregacja

Uwaga: Ta kolumna jest dostępna tylko po zaznaczeniu pola wyboru **Użyj grupowania sześciokątnego** oraz opcji **Nałożenie**.

Jeśli zmienna **Nałożenie** zostanie wybrana dla warstwy punktów, dla której zastosowano grupowanie sześciokątne, wszystkie wartości tej zmiennej muszą zostać zagregowane dla wszystkich punktów w sześciokącie. Należy określić funkcję agregującą dla wszystkich zmiennych nałożenia, jakie mają być zastosowane na mapie. Dostępne funkcje agregujące zależą od typu pomiaru.

- Funkcje agregujące dla ilościowego typu pomiaru, z typem składowania Liczba rzeczywista lub Liczba całkowita:
  - Suma
  - Średnia
  - Minimum
  - Maksimum
  - Mediana
  - Pierwszy kwartyl
  - Trzeci kwartyl
- Funkcje agregujące dla ilościowego typu pomiaru, z typem składowania Czas, Data lub Znacznik czasu:
  - Średnia
  - Minimum
  - Maksimum
- Funkcje agregujące dla nominalnego lub jakościowego typu pomiaru:
  - Dominanta
  - Minimum
  - Maksimum
- Funkcje agregujące dla typu pomiaru Flaga:
  - Prawda, gdy jakaś jest prawdziwa
  - Fałsz, jeśli choć jeden fałsz

Kolor

Tej opcji należy użyć, aby wybrać kolor standardowy, który będzie zastosowany dla wszystkich właściwości zmiennej geoprzestrzennej lub zmienną nałożenia, która zastosuje kolory do właściwości na podstawie wartości z innej zmiennej w danych.

Jeśli wybrana zostanie opcja **Standardowy**, można wybrać kolor z palety kolorów, która jest wyświetlana w panelu **Porządek kolorów kategorii na wykresie** na karcie wyświetlania w oknie dialogowym Opcje użytkownika.

W przypadku wybrania opcji **Nałożenie** można wybrać dowolną zmienną z źródła danych, które zawiera zmienną geoprzestrzenną wybraną jako **Zmienna warstwy**.

- W przypadku nominalnych i jakościowych zmiennych nałożenia paleta do wyboru kolorów jest taka sama, jak wyświetlana dla opcji koloru **Standardowy**.
- Dla ilościowych i porządkowych zmiennych nałożenia wyświetlana jest druga lista rozwijana, z której można wybrać kolor. Po wybraniu koloru nałożenie jest stosowane z różnym nasyceniem koloru, w

zależności od wartości w zmiennej ilościowej lub porządkowej. Najwyższa wartość ma kolor wybrany z listy rozwijanej, a dla niższych wartości nasycenie koloru jest odpowiednio mniejsze.

Symbol

Uwaga: Opcja aktywowana tylko dla typów pomiaru Punkt i Multipunkt.

Ta opcja umożliwia wybranie symbolu **Standardowy**, który zostanie zastosowany do wszystkich rekordów zmiennej geoprzestrzennej lub symbolu **Nałożenie**, który zmienia ikonę symbolu dla punktów na podstawie wartości innej zmiennej w danych.

Jeśli wybrana zostanie opcja **Standardowy**, z listy rozwijanej można wybrać jeden z domyślnych symboli, który będzie reprezentował dane punktowe na mapie.

W przypadku wybrania opcji **Nałożenie** można wybrać dowolną zmienną nominalną, porządkową lub jakościową z źródła danych, które zawiera zmienną geoprzestrzenną wybraną jako **Zmienna warstwy**. Dla każdej wartości w zmiennej nałożenia na mapie wyświetlany jest inny symbol.

Przykładowo dane mogą zawierać zmienną punktową, która reprezentuje lokalizacje sklepów; nałożeniem może być zmienna rodzaju sklepu. W tym przykładzie wszystkie sklepy spożywcze mogą być oznaczone na mapie symbolem krzyżyka, a wszystkie sklepy elektroniczne symbolem kwadratu.

Wielkość

Uwaga: Ta opcja jest aktywowana tylko dla typów pomiaru Punkt, Multipunkt, Łańcuch i Multiłańcuch.

Ta opcja umożliwia wybranie wielkości **standardowej**, która zostanie zastosowana do wszystkich rekordów zmiennej geoprzestrzennej lub wielkości **Nałożenie**, która zmienia wielkość ikony symbolu lub grubość linii na podstawie wartości innej zmiennej w danych.

Jeśli zostanie wybrana opcja **Standardowy**, można wybrać wartość szerokości w pikselach. Dostępne opcje to: 1, 2, 3, 4, 5, 10, 20 lub 30.

W przypadku wybrania opcji **Nałożenie** można wybrać dowolną zmienną z źródła danych, które zawiera zmienną geoprzestrzenną wybraną jako **Zmienna warstwy**. Grubość linii lub punktu różni się w zależności od wartości wybranej zmiennej.

Przezroczystość

Ta opcja umożliwia wybranie **standardowej**przezroczystości, która zostanie zastosowana do wszystkich rekordów zmiennej geoprzestrzennej, lub przezroczystości **Nałożenie**, która zmienia przezroczystość symbolu, linii lub wielokąta na podstawie wartości innej zmiennej w danych.

Jeśli wybrana zostanie opcja **Standardowy**, można wybrać jeden z dostępnych poziomów przezroczystości począwszy od 0% (nieprzezroczysty) do 100% (przezroczysty), przy czym przyrost wynosi 10%.

W przypadku wybrania opcji **Nałożenie** można wybrać dowolną zmienną z źródła danych, które zawiera zmienną geoprzestrzenną wybraną jako **Zmienna warstwy**. Dla każdej wartości zmiennej nałożenia wyświetlany jest inny poziom przezroczystości. Przezroczystość jest stosowana do koloru wybranego z listy rozwijanej dla punktu, linii lub wielokąta.

Etykieta danych

Uwaga: Opcja ta jest niedostępna w przypadku zaznaczenia pola wyboru **Użyj grupowania sześciokątnego**.

Tej opcji należy użyć, aby wybrać zmienną, jakie będzie używana jako etykieta danych na mapie. Przykładowo, jeśli zostanie zastosowana dla warstwy wielokątów, etykietą danych może być zmienna nazwy, zawierająca nazwę poszczególnych wielokątów. Jeśli wybrana zostanie zmienna nazwy, nazwy te będą wyświetlane na mapie.

Karta wyglądu wizualizacji na mapie

Przed utworzeniem wykresu możesz określić opcje jego wyglądu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Nagłówek. Należy wprowadzić tekst, jaki będzie używany jako nagłówek wykresu.

Węzeł t-SNE

Stochastyczna metoda porządkowania sąsiadów w oparciu o rozkład t (t-SNE — t-Distributed Stochastic Neighbor Embedding©) to narzędzie do wizualizacji danych wysokowymiarowych. Przekształca ono powinowactwa punktów danych w prawdopodobieństwa. Powinowactwa w przestrzeni pierwotnej są reprezentowane przez gaussowskie prawdopodobieństwa łączne, a powinowactwa w przestrzeni włączanej są reprezentowane przez rozkłady t Studenta. Dzięki temu algorytm t-SNE jest szczególnie czuły na struktury lokalne i ma kilka innych przewag nad wcześniej stosowanymi technikami: [1]

- Ujawnianie struktur w wielu skalach na jednej mapie
- Ujawnianie danych leżących w wielu różnych rozgałęzieniach lub grupach
- Ograniczenie tendencji do skupiania punktów w środku

Węzeł t-SNE w programie SPSS Modeler został zaimplementowany w języku Python i wymaga biblioteki Python scikit-learn©. Aby uzyskać szczegółowe informacje o algorytmie t-SNE i bibliotece scikit-learn, patrz:

- https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html#sklearn.manifold.TSNE
- https://scikit-learn.org/stable/modules/manifold.html#t-sne

Karta Python na palecie węzłów zawiera ten i inne węzły Python. Węzeł t-SNE jest także dostępny na karcie Wykresy.

[1] Piśmiennictwo:

van der Maaten, L.J.P.; Hinton, G. „Visualizing High-Dimensional Data using t-SNE". Journal of Machine Learning Research. 9:2579-2605, 2008.

van der Maaten, L.J.P. „t-Distributed Stochastic Neighbor Embedding".

van der Maaten, L.J.P. „Accelerating t-SNE using Tree-Based Algorithms". Journal of Machine Learning Research. 15(Oct):3221-3245, 2014.

Opcje zaawansowane węzła t-SNE

Wybierz tryb **Prosty** albo **Zaawansowany** w zależności od tego, z których opcji węzła t-SNE chcesz korzystać.

Typ wizualizacji. Wybierz **2W** albo **3W**, aby określić, czy wykres ma być rysowany jako dwuwymiarowy, czy trójwymiarowy.

Metoda. Wybierz metodę **Barnes Hut** albo **Dokładnie**. Domyślnie algorytm obliczania gradientu używa aproksymacji Barnesa-Huta, która jest znacznie szybsza niż metoda dokładna. Aproksymacja Barnesa-Huta umożliwia stosowanie techniki t-SNE do dużych zbiorów danych spotykanych w świecie rzeczywistym. Algorytm dokładny zapewni skuteczniejsze unikanie błędów najbliższego sąsiedztwa.

Pocz. Wybierz metodę inicjowania włączania: **Losowe** albo **PCA** .

Zmienna przewidywana. Wybierz zmienną przewidywaną, która ma zostać przedstawiona jako mapa kolorów na wynikowym wykresie. Jeśli zmienna przewidywana nie zostanie tutaj określona, wykres będzie jednokolorowy.

Optymalizacja

Zmieszanie. Stopień zmieszania związany z liczbą najbliższych sąsiadów używanych w innych algorytmach typu Manifold Learning. Większe zbiory danych zwykle wymagają większego zmieszania. Należy rozważyć wybór wartości z przedziału od **5** do **50**. Wartość domyślna to **30**, a zakres wynosi **2 - 9999999**.

Wstępne wyolbrzymienie. To ustawienie określa, jak ciasno upakowane będą grupy naturalne w przestrzeni włączanej i ile miejsca pozostanie między nimi. Wartość domyślna to **12**, a zakres wynosi **2 - 9999999**.

Współczynnik uczenia. Jeśli współczynnik uczenia będzie za wysoki, dane mogą przypominać „piłkę", a każdy z punktów będzie w przybliżeniu równoodległy od najbliższych sąsiadów. Jeśli Współczynnik uczenia będzie za niski, większość punktów może skupić się w gęstą chmurę z niewielką liczbą punktów odstających. Jeśli funkcja kosztu utknie w nieprawidłowym minimum lokalnym, rozwiązaniem może być zwiększenie współczynnika uczenia. Wartość domyślna to **200**, a zakres wynosi **0 - 9999999**.

Maks. liczba iteracji Maksymalna liczba iteracji optymalizacji. Wartość domyślna to **1000**, a zakres wynosi **250 - 9999999**.

Wielkość kątowa. Wielkość kątowa odległego węzła zmierzona z punktu. Wprowadź wartość z zakresu od **0** do **1**. Wartością domyślną jest **0,5**.

Wartość początkowa

Ustaw wartość początkową generatora liczb losowych. Wybierz tę opcję i kliknij przycisk **Utwórz**, aby wygenerować wartość początkową dla generatora liczb losowych.

Warunek zatrzymania optymalizacji

Maks. liczba iteracji bez postępu. Maksymalna liczba iteracji bez postępu, po której optymalizacja ma zostać zatrzymana. Zaczyna obowiązywać po 250 iteracjach początkowych w przypadku użycia wstępnej nadmiarowości. Należy zwrócić uwagę, że postęp jest sprawdzany co 50 iteracji, zatem ta wartość zostanie zaokrąglona do następnej wielokrotności 50. Wartość domyślna to **300**, a zakres wynosi **0 - 9999999**.

Min. norma gradientu. Jeśli norma gradientu będzie niższa od tego progu, optymalizacja zostanie przerwana. Wartość domyślna to **1.0E-7**.

Metric. Metryka, która ma być stosowana przy obliczaniu odległości między wystąpieniami w macierzy predyktorów. Jeśli metryka jest łańcuchem, musi być jedną z opcji dozwolonych jako parametr metric funkcji scipy.spatial.distance.pdist lub metryką wymienioną w wyliczeniu pairwise.PAIRWISE_DISTANCE_FUNCTIONS. Wybierz jeden z dostępnych typów metryk. Wartością domyślną jest **euclidean**.

Jeśli liczba rekordów jest większa od. Podaj metodę tworzenia większych zbiorów danych. Można określić maksymalną wielkość zbioru danych lub użyć domyślnej liczby punków, wynoszącej 2000. Wydajność dla dużych zbiorów danych ulega poprawie po wybraniu opcji **Kategoria** lub **Próba**. Alternatywnie można wybrać wykreślanie wszystkich punktów danych przez wybór opcji **Użyj wszystkich danych**; należy jednak zwrócić uwagę, że może to drastycznie obniżyć wydajność oprogramowania.

- **Kategoria.** Tę opcję należy wybrać, aby aktywować kategoryzację, jeśli zbiór danych zawiera więcej rekordów niż określona liczba. Kategoryzacja dzieli wykres na mniejsze siatki przed rzeczywistym wykreśleniem go i zlicza liczbę połączeń, jakie zostaną wyświetlone w każdej komórce siatki. Na końcowym wykresie w środku ciężkości kategorii wykreślane jest jedno połączenie dla każdej komórki (średnia wszystkich punktów połączeń w kategorii).
- **Przykład.** Tę opcję należy wybrać, aby w przeprowadzić próbę losową danych dla określonej liczby rekordów.

W poniższej tabeli przedstawiono relację między ustawieniami na karcie Zaawansowane okna dialogowego t-SNE w programie SPSS Modeler a parametrami biblioteki t-SNE w języku Python.

Tabela 37. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr t-SNE w języku Python
Dominanta	mode_type	
Typ wizualizacji	n_components	n_components
Metoda	method	method
Inicjowanie włączania	init	init
Zmienna przewidywana	target_field	target_field
Zmieszanie	perplexity	perplexity
Wstępna nadmiarowość	early_exaggeration	early_exaggeration
Współczynnik uczenia	learning_rate	learning_rate
Maks. liczba iteracji	n_iter	n_iter
Wielkość kątowa	angle	angle
Ustaw wartość początkowa generatora liczb losowych	enable_random_seed	
Wartość początkowa	random_seed	random_state
Maks. liczba iteracji bez postępu	n_iter_without_progress	n_iter_without_progress
Min. norma gradientu	min_grad_norm	min_grad_norm
Wykonaj t-SNE z wieloma poziomami zmieszania	isGridSearch	

Opcje wyników węzła t-SNE

Określ opcje generowania wyników węzła t-SNE na karcie **Wynik**.

Nazwa wyniku. Określa nazwę wyniku uzyskanego po wykonaniu węzła. Wybranie opcji **Automatycznie** powoduje, że nazwa pliku wynikowego będzie określana automatycznie.

Wynik na ekran. Wybierz tę opcję, aby wygenerować i wyświetlić wynik w nowym oknie. Wynik jest także dodawany do Menedżera wyników.

Wynik do pliku. Ta opcja umożliwia zapisanie wyniku w pliku. Wybranie jej powoduje uaktywnienie pól **Nazwa pliku** i **Typ pliku**. Węzeł t-SNE musi mieć dostęp do tego pliku wyników, jeśli w celach porównawczych wykresy mają być tworzone przy użyciu innych zmiennych — lub jeśli wyniki mają być używane jako predyktory w modelach klasyfikacji lub regresji. Model t-SNE tworzy plik wynikowy zmiennych x, y (i z), do którego najłatwiej uzyskać dostęp za pośrednictwem węzła Plik kolumnowy.

Dostęp do danych t-SNE i wykreślanie ich

Jeśli wybrana jest opcja **Wynik do pliku** w celu zapisywania wyników węzła t-SNE w plikach, można w celach porównawczych tworzyć wykresy na podstawie innych zmiennych — lub wykorzystać wyniki jako predyktory w modelach klasyfikacji lub regresji. Model t-SNE tworzy plik wynikowy zmiennych x, y (i z), do którego najłatwiej uzyskać dostęp za pośrednictwem węzła Plik kolumnowy. Ta sekcja zawiera informacje przykładowe.

1. W oknie dialogowym t-SNE otwórz kartę **Wynik**.
2. Wybierz opcję **Wynik do pliku** i wpisz nazwę pliku. Użyj domyślnego typu pliku HTML. Podczas wykonywania modelu wygenerowane zostaną trzy pliki wynikowe w lokalizacji wynikowej:
   - Plik tekstowy (result_xxxxxx.txt)
   - Plik HTML (o podanej nazwie)
   - Plik PNG (tsne_chart_yyyyyy.png)

Plik tekstowy będzie zawierał potrzebne dane, ale ze względów technicznych mogą być one zapisane w formacie standardowym lub naukowym. Jeśli dane są zapisane w formacie naukowym (1.11111111e+01), to konieczne jest utworzenie nowego strumienia, który będzie rozpoznawał ten format:

Dostęp do danych wykresu t-SNE, gdy plik tekstowy zawiera dane liczbowe w formacie naukowym

1. Utwórz nowy strumień (**Plik > Nowy strumień**).
2. Wybierz kolejno opcje **Narzędzia > Właściwości strumienia > Opcje**, wybierz opcję **Formaty liczb**, a następnie jako format wyświetlania liczb wybierz **Naukowy (#.###E+##)**.
3. Dodaj węzeł źródłowy Plik kolumnowy do obszaru roboczego i użyj następujących ustawień na karcie Plik
   - Pomiń wiersze nagłówka: 1
   - Długość rekordu: 54
   - Początek tSNE_x: 3, Długość: 16
   - Początek tSNE_y: 20, Długość: 16
   - Początek tSNE_z : 36, Długość: 16
4. Na karcie Typ liczby powinny być rozpoznane jako rzeczywiste. Kliknij opcję Odczytaj wartości. Powinny pojawić się wartości podobne do poniższych:

Tabela 38. Przykładowe wartości zmiennych

Zmienna	Pomiar	Wartości
tSNE_x	Ilościowy	[-7.07176703,7.14338837]
tSNE_y	Ilościowy	[-9.2188112,8.89647667]
tSNE_x	Ilościowy	[-9.95892882,9.95742482]

5. Dodaj do strumienia węzeł Selekcja, aby możliwe było usunięcie następujących dwóch dolnych wierszy tekstu z pliku (są odczytywane jako wartości null):

   ```
   *****************
   Perform t-SNE (total time 9.5s)
   ```

   Na karcie Ustawienia węzła Selekcja wybierz tryb **Odrzuć** i użyj warunku @NULL(tSNE_x) w celu usunięcia wierszy.
6. Dodaj do strumienia węzeł typu i węzeł eksportu pliku kolumnowego, aby utworzyć zmienną. Węzeł pliku źródłowego, który będzie kopiował i wklejał z powrotem do pierwotnego strumienia.

Dostęp do danych wykresu t-SNE, gdy plik tekstowy zawiera dane liczbowe w formacie standardowym

1. Utwórz nowy strumień (**Plik > Nowy strumień**).
2. Dodaj węzeł źródłowy Plik kolumnowy do obszaru roboczego. Aby uzyskać dostęp do danych t-SNE, wystarczą następujące trzy węzły.

Rysunek 48. Stream for accessing t-SNE plot data in standard numeric format

3. Użyj następujących ustawień na karcie Plik węzła źródłowego Plik kolumnowy.
   - Pomiń wiersze nagłówka: 1

- Długość rekordu: 29
- Początek tSNE_x: 3, Długość: 12
- Początek tSNE_y: 16, Długość: 12

4. Na karcie Filtr można zmienić nazwy zmiennych field1 i field2 na tsneX i tsneY.
5. Dodaj węzeł Łączenie, aby połączyć go ze strumieniem metodą **Porządek**.
6. Można teraz użyć węzła Wykres, aby wykreślić tsneY w funkcji tsneX i oznaczyć kolorem według badanej zmiennej.

Modele użytkowe t-SNE

Modele użytkowe t-SNE zawierają wszystkie informacje zgromadzone przez model t-SNE. Dostępne są następujące karty.

Graph

Na karcie **Wykres** wyświetlany jest wynik węzła t-SNE w formie graficznej. Wykres pyplot rozrzutu przedstawia wynik niskowymiarowy. Jeśli na karcie Zaawansowany węzła t-SNE nie zostanie zaznaczona opcja **Wykonaj t-SNE z wieloma poziomami zmieszania**, to uwzględniony będzie tylko jeden wykres, a nie sześć z różnymi poziomami zmieszania.

Wyniki tekstowe

Na karcie **Wynik tekstowy** wyświetlane są wyniki działania algorytmu t-SNE. Jeśli na karcie Zaawansowany węzła t-SNE wybrano typ wizualizacji **2W**, to wynik będzie przedstawiony jako wartości punktów w dwóch wymiarach. W przypadku wybrania opcji **3W** wynik będzie przedstawiony jako wartości punktów w trzech wymiarach.

Węzeł Wykres E-Plot (Beta)

Węzły Wykres E-Plot (Beta) obrazują relacje między zmiennymi liczbowymi. Wykres E-Plot (Beta) jest podobny do węzła Wykres, ale oferuje inne opcje i nowe funkcje tworzenia wykresów. Zachęcamy do eksperymentowania z nowymi możliwościami tworzenia wykresów, jakie oferuje ten węzeł programu SPSS Modeler.

Węzeł Wykres E-plot (Beta) przedstawia relacje pomiędzy zmiennymi liczbowymi w postaci wykresów rozrzutu, wykresów liniowych i wykresów słupkowych. Nowy interfejs tworzenia wykresów w tym węźle jest intuicyjny i nowoczesny. Charakteryzuje się konfigurowalnością, a tworzone za jego pomocą wykresy danych są interaktywne. Aby uzyskać więcej informacji, patrz “Korzystanie z wykresu E-Plot” na stronie 272.

Wykres E-Plot (Beta): karta Wykres

Na wykresach przedstawiane są wartości zmiennej *Y* w odniesieniu do zmiennej *X*. Często zmienne te odpowiadają odpowiednio zmiennej zależnej i zmiennej niezależnej.

Zmienna X. Z listy należy wybrać zmienną, jaka będzie wyświetlana na poziomej osi *x*.

Zmienna Y. Z listy należy wybrać zmienną, jaka będzie wyświetlana na pionowej osi *y*.

Nakładanie. Istnieją różne sposoby ilustrowania kategorii dla wartości danych. Przykładowo można użyć wartości *maincrop* (uprawa główna) jako kolorowej nakładki, aby wskazać, wartości *estincome* (szacowany przychód) i *claimvalue* (wartość roszczenia) z podziałem na główne uprawy podmiotów zgłaszających roszczenia. Wybierz zmienne, które mają być odwzorowane jako kolory, rozmiary lub kształty w wynikach. Wybierz także wszelkie inne zmienne będące przedmiotem zainteresowania, które mają być uwzględnione w interaktywnych wynikach. Więcej informacji można znaleźć w temacie “Sposób prezentacji, nakładanie, panele i animacje” na stronie 190.

Po ustawieniu opcji dla wykresu E-plot można uruchomić wykres bezpośrednio z okna dialogowego, klikając przycisk **Wykonaj**. Można również użyć karty Opcje, aby dokonać dodatkowych specyfikacji.

Wykres E-Plot (Beta): karta Opcje

Maksymalna liczba rekordów na wykresie. Podaj metodę tworzenia większych zbiorów danych. Można określić maksymalną wielkość zbioru danych lub użyć domyślnej liczby rekordów, wynoszącej 2000. Wydajność dla dużych zbiorów danych ulega poprawie po wybraniu opcji **Próba**. Opcja Losowanie losuje dane dla rekordów wprowadzonych w polu tekstowym. Alternatywnie można wybrać wykreślanie wszystkich punktów danych przez wybór opcji **Użyj wszystkich danych**; należy jednak zwrócić uwagę, że może to drastycznie obniżyć wydajność oprogramowania.

Wykres E-Plot (Beta): karta Wygląd

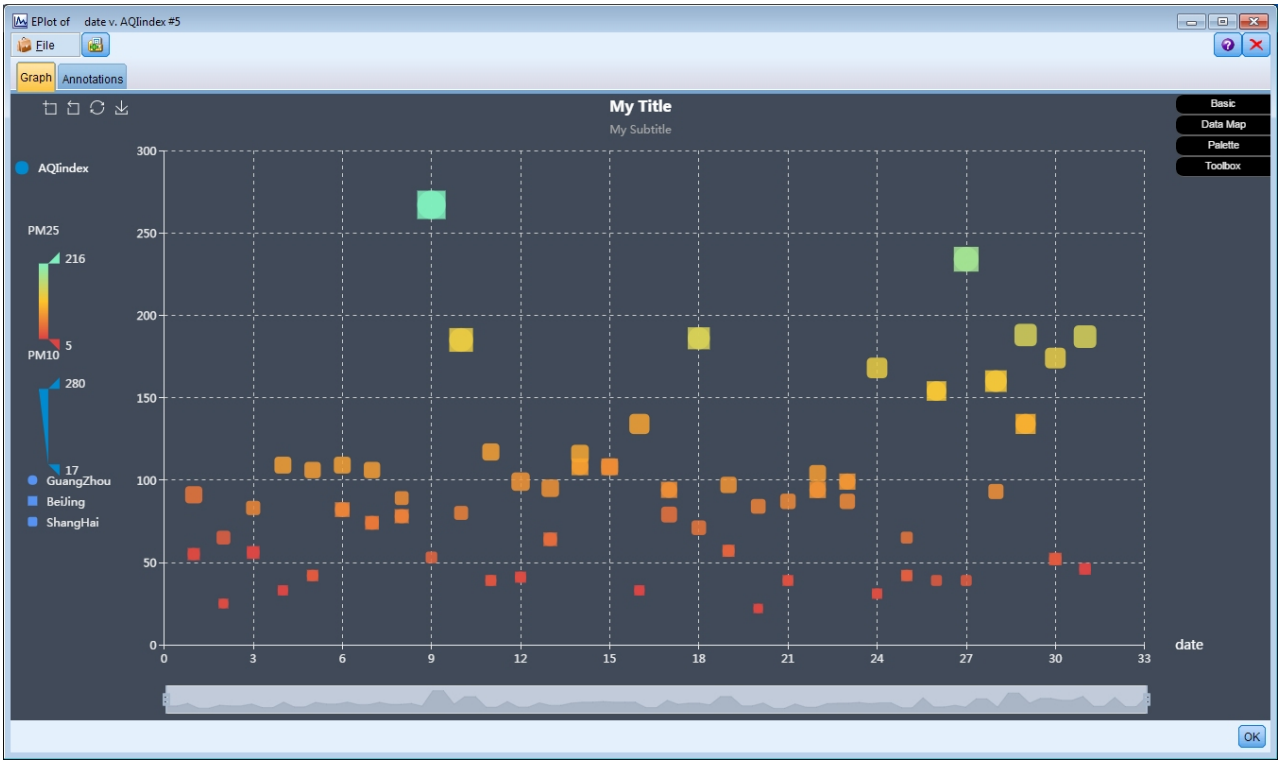
Przed utworzeniem wykresu można w razie potrzeby określić tytuł i podtytuł. Opcje te można też określić lub zmienić po utworzeniu wykresu.

Tytuł. Należy wprowadzić tekst, jaki będzie używany jako tytuł wykresu.

Podtytuł. Należy wprowadzić tekst, jaki będzie używany jako podtytuł wykresu.

Korzystanie z wykresu E-Plot

Węzeł Wykres E-plot (Beta) przedstawia relacje pomiędzy zmiennymi liczbowymi w postaci wykresów rozrzutu, wykresów liniowych i wykresów słupkowych. Nowy interfejs tworzenia wykresów wprowadzony w tym węźle (mającym obecnie status wersji beta) zawiera wiele nowych możliwości i udoskonaleń.



Rysunek 49. E-Plot (Beta) scatterplot graph

W lewym górnym rogu karty Wykres znajduje się pasek narzędzi umożliwiający powiększenie określonej części wykresu, cofnięcie powiększenia, powrót do pierwotnego pełnego widoku lub zapisanie wykresu do wykorzystania w programie zewnętrznym:

Rysunek 50. Toolbar

U dołu okna można użyć suwaka do powiększenia określonej części wykresu. Powiększenie zmienia się, przesuwając małe prostokątne elementy sterujące po prawej i lewej stronie. Aby można było używać suwaka, należy go najpierw włączyć w obszarze opcji Zestaw narzędzi.

Rysunek 51. Zoom slider

Po lewej stronie okna znajdują się elementy do wyboru zakresu prezentowanych wartości. Aby ich użyć, należy najpierw określić opcje w obszarze Mapowanie danych. W poniższym przykładzie zmienna PM25 jest wybrana jako źródło dla mapy kolorów, zmienna PM10 jako źródło dla mapy rozmiarów, a zmienna City jako źródło dla mapy kształtów. Można zatrzymać wskaźnik myszy nad pionowymi paskami kolorów, aby wyróżnić odpowiednie obszary na wykresie, lub przesuwać górny i dolny trójkąt.

Rysunek 52. Range controls

Po prawej stronie okna dostępny jest zestaw rozwijanych opcji, których można użyć do interaktywnej pracy z danymi i zmiany wyglądu wykresu w czasie rzeczywistym.

Rysunek 53. Expandable options

Opcje podstawowe

Rysunek 54. Basic options

Wybierz kompozycję ciemną lub jasną, określ tytuł i podtytuł, wybierz typ wykresu (rozrzutu, liniowy lub słupkowy) i wybierz serię przedstawianą na osi Y. W wypadku wybrania wykresu **Liniowy** wyświetlone zostaną tylko zmienne na osi Y i tylko zmienne z osi Y będą dostępne w opcjach Mapowanie danych dla mapy kolorów i rozmiarów. W wypadku wybrania wykresu **Słupkowy**, w opcjach Mapowanie danych będą dostępne tylko opcje mapy kolorów. Dla serii będą tutaj dostępne wszystkie zmienne **Zainteresowanie** wybrane na karcie Wykres węzła wykresu E-plot.

Opcje Mapowanie danych

Rysunek 55. Data map options

Wybierz zmienną ciągłą lub kategorialną dla mapy **Odwzorowanie kolorów**. W przypadku wybrania zmiennej ciągłej wyświetlane będą kolory od zielonego do czerwonego. Im niższa wartość, tym bliższy czerwieni będzie kolor, a im wyższa wartość, tym kolor będzie bliższy zieleni. W przypadku wybrania zmiennej kategorialnej kolor zmiennej będzie wybierany ze zdefiniowanej palety.
Mapa **Odwzorowanie wielkości** obsługuje tylko zmienne ciągłe. Im mniejsza jest wartość, tym mniejsza będzie na wykresie.
Mapa **Odwzorowanie kształtów** obsługuje tylko zmienne kategorialne. Kształt wyświetlany na mapie jest zdefiniowany względem zmiennej kategorialnej, która rozdziela wizualizację na elementy o różnych kształtach, po jednym dla każdej kategorii.

Opcje Paleta

Rysunek 56. Palette options

Paleta umożliwia zmodyfikowanie kolorów tytułu i serii. Wybierz tytuł lub serię z menu rozwijanego, kliknij opcję **Edytuj kolory predefiniowane**, a następnie kliknij przycisk **Więcej**, aby wybrać kolor. Można dokładnie określić kolor za pomocą pól RGB lub Szesnastkowo.

Opcje Zestaw narzędzi

Rysunek 57. Toolbox options

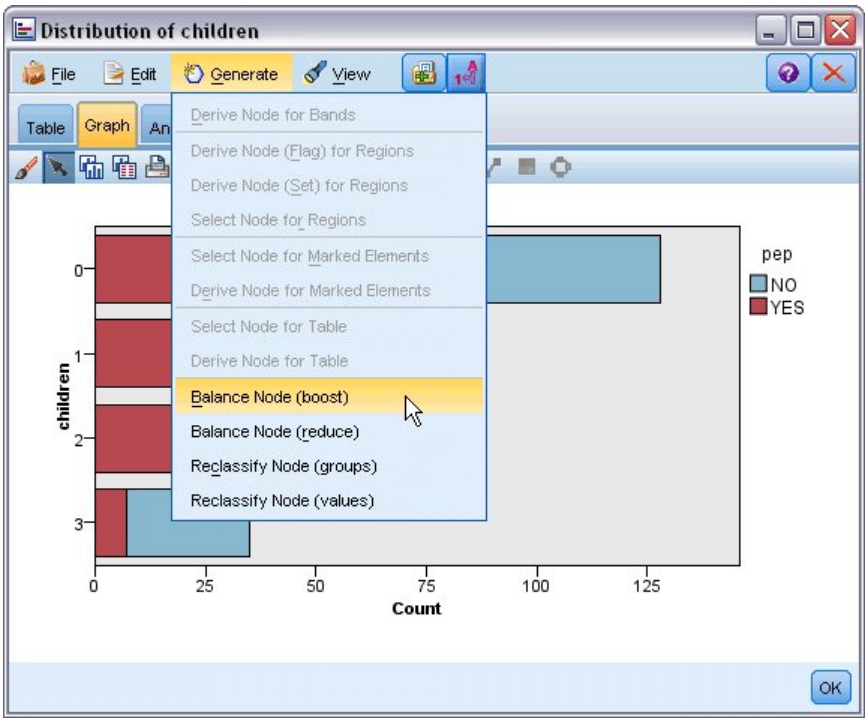
Opcje Zestaw narzędzi umożliwiają włączanie i wyłączanie suwaka powiększenia, określanie właściwości linii siatki oraz włączania i wyłączanie śladu myszy. Opcja Ślad myszy przedstawia dokładne współrzędne po zatrzymaniu wskaźnika nad wykresem.

Eksplorowanie wykresów

Tryb edycji umożliwia edytowanie układu graficznego i wyglądu, natomiast tryb eksploracji umożliwia analityczną eksplorację danych i wartości reprezentowanych przez wykres. Głównym celem eksploracji jest analiza danych, a następnie identyfikacja wartości z użyciem pasm, regionów i oznaczeń w celu wygenerowania węzłów Selekcja, Wyliczanie lub Zrównoważenie. Aby wybrać ten tryb, z menu wybierz **Widok > Tryb eksploracji** (lub kliknij ikonę paska narzędzi).

Podczas gdy w niektórych wykresach można wykorzystać wszystkie narzędzia eksploracji, w innych akceptowane jest tylko jedno. Tryb eksploracji obejmuje:

- definiowanie i edytowanie pasm, które są używane do podziału wartości wzdłuż osi *x* skali. Więcej informacji można znaleźć w temacie “Zastosowanie przedziałów”.
- Definiowanie i edytowanie regionów, które są używane do identyfikacji grup wartości w ramach obszaru prostokątnego. Więcej informacji można znaleźć w temacie “Zastosowanie regionów” na stronie 279.
- Oznaczanie i usuwanie oznaczeń elementów w celu ręcznego wyboru wartości, które mogłyby być używane do generowania węzła Selekcja lub Wyliczanie. Więcej informacji można znaleźć w temacie “Użycie zaznaczonych elementów” na stronie 281.
- Generowanie węzłów za pomocą wartości zidentyfikowanych za pośrednictwem pasm, regionów, oznaczonych elementów i łącz do stron WWW w celu ich użycia w strumieniu. Więcej informacji można znaleźć w temacie “Generowanie węzłów z wykresów” na stronie 282.

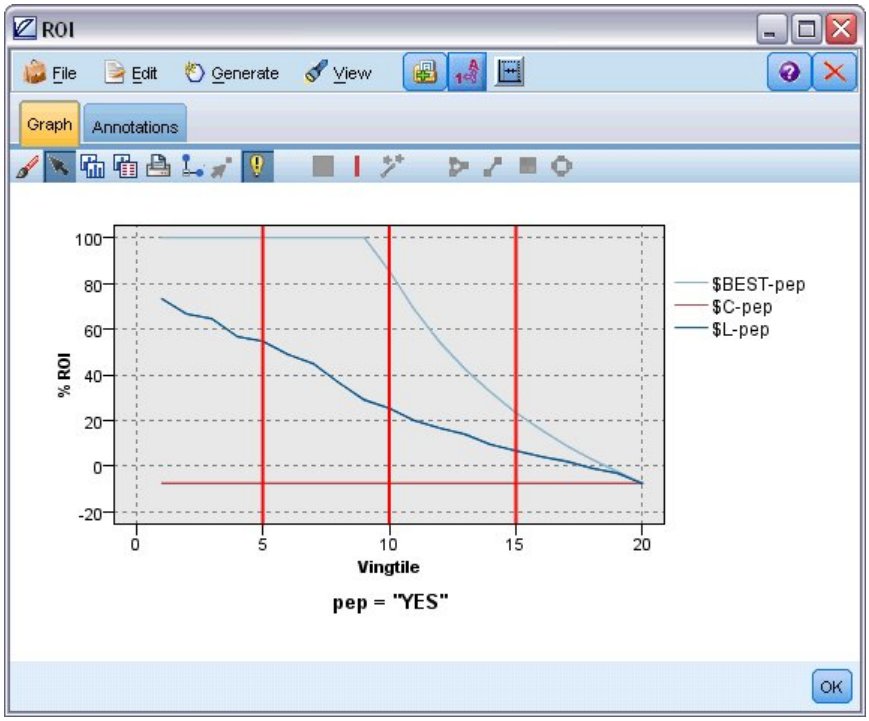


Rysunek 58. Wykres z wyświetlanym menu generowania

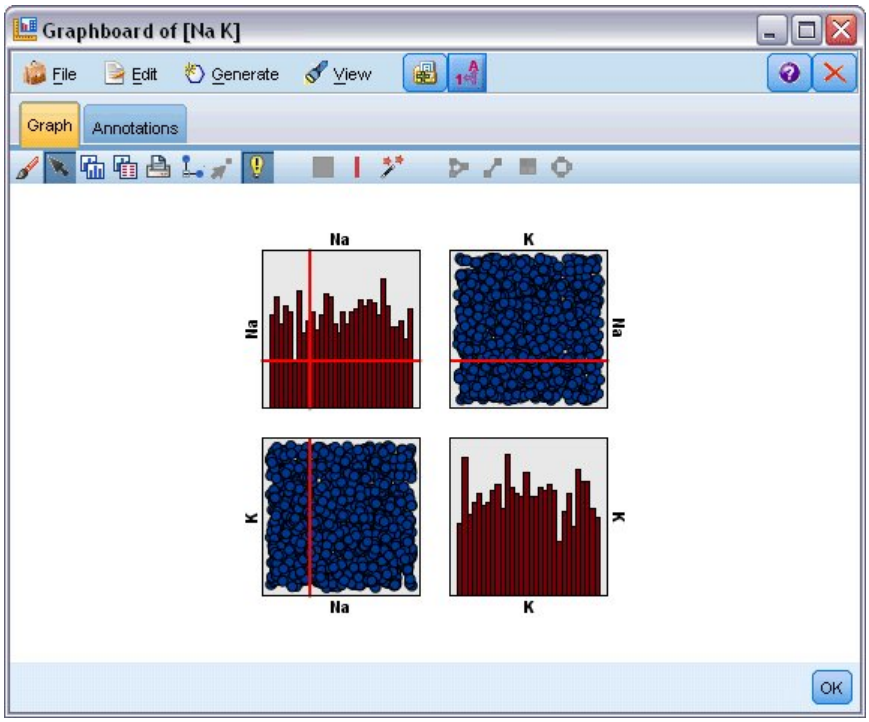
Zastosowanie przedziałów

Na dowolnym wykresie ze zmienną skali na osi *x* można narysować pionowe linie przedziału, które podzielą zakres wartości na osi *x*. Jeśli wykres składa się z wielu paneli, linia przedziału narysowana w jednym z paneli jest również reprezentowana w pozostałych panelach.

Nie we wszystkich wykresach przedziały są akceptowane. Niektóre z wykresów, które umożliwiają zastosowanie przedziałów: histogramy, wykresy słupkowe i rozkłady, wykresy (liniowe, rozrzutu, sekwencyjne itd.), kolekcje i wykresy ewaluacyjne. W przypadku wykresów z panelowaniem przedziały wyświetlane są we wszystkich panelach. W niektórych przypadkach dla macierzowych wykresów rozrzutu (SPLOM) wyświetlana jest pozioma linia przedziału, ponieważ oś, na której narysowano przedział zmiennej, została odwrócona.



Rysunek 59. Wykres z trzema przedziałami



Rysunek 60. Macierzowy wykres rozrzutu (SPLOM) z przedziałami

Definiowanie przedziałów

Na wykresie bez przedziałów dodanie linii przedziału spowoduje podział wykresu na dwa przedziały. Wartość linii przedziału reprezentuje punkt początkowy, nazywany również dolną granicą, drugiego przedziału (przy odczycie

wykresu od lewej do prawej). Podobnie, na wykresie z dwoma przedziałami, dodanie linii przedziału spowoduje podział jednego z tych przedziałów na dwie części, w wyniku czego powstają trzy przedziały. Domyślnie przedziałom nadawana jest nazwa *bandN* (przedziałN), gdzie *N* oznacza liczbę przedziałów od lewej do prawej na osi *x*.

Po zdefiniowaniu przedziału można za pomocą metody przeciągnij i upuść zmienić położenie przedziału na osi *x*. Więcej skrótów dla zadań, takich jak zmiana nazwy, usuwanie lub generowanie węzłów dla określonego przedziału, można wyświetlić, klikając prawym przyciskiem myszy obszar wewnątrz danego przedziału.

Aby zdefiniować przedziały:

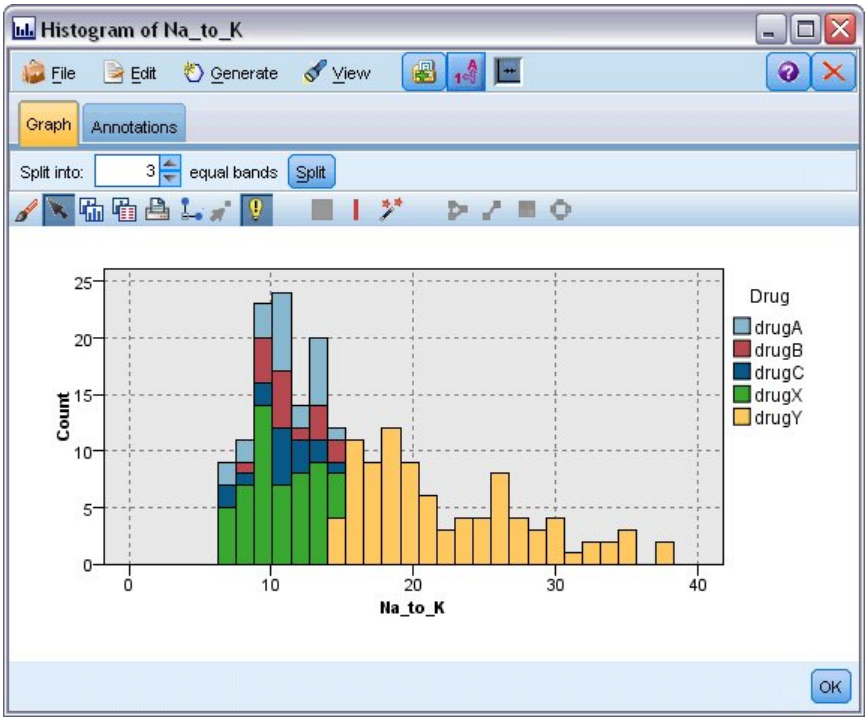
1. Sprawdź, czy jest aktywny tryb eksploracji. Z menu wybierz opcję **Widok> Tryb eksploracji**.
2. Na pasku narzędzi trybu eksploracji kliknij przycisk rysowania przedziału.

Rysunek 61. Przycisk rysowania przedziałów na pasku narzędzi

3. Na wykresie, który akceptuje przedziały, kliknij punkt wartości na osi *x*, w którym ma zostać zdefiniowana linia przedziału.

Uwaga: Alternatywnie można kliknąć ikonę **Podziel wykres na przedziały** na pasku narzędzi i wprowadzić liczbę równych przedziałów, a następnie kliknąć przycisk **Podział**.

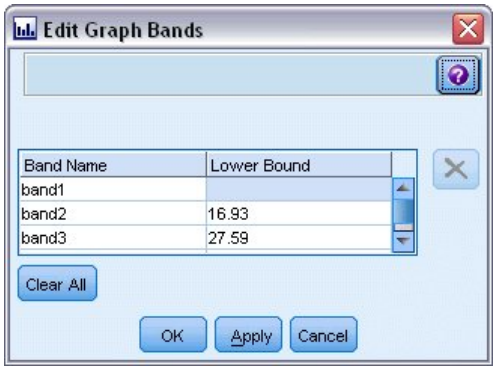
Rysunek 62. Ikona rozdzielacza używanego do rozwijania paska narzędzi z opcjami do podziału na przedziały



Rysunek 63. Pasek narzędzi tworzenia równych przedziałów z aktywnymi przedziałami

Edytowanie, zmiana nazwy i usuwanie przedziałów

Istnieje możliwość edytowania właściwości istniejących przedziałów w oknie dialogowym Edytuj przedziały wykresu lub za pośrednictwem menu kontekstowych na samym wykresie.



Rysunek 64. Okno dialogowe Edytuj przedziały wykresu

Aby edytować przedziały:

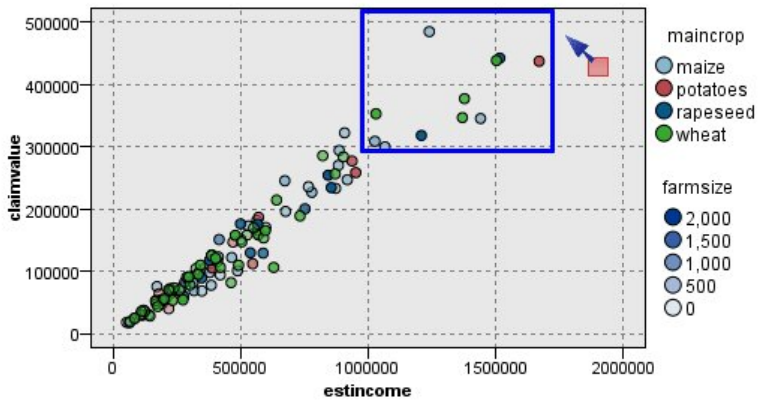
1. Sprawdź, czy jest aktywny tryb eksploracji. Z menu wybierz opcję **Widok> Tryb eksploracji**.
2. Na pasku narzędzi trybu eksploracji kliknij przycisk rysowania przedziału.
3. Z menu wybierz kolejno opcje **Edytuj > Przedziały wykresu**. Zostanie otwarte okno dialogowe Edytuj przedziały wykresu.
4. Jeśli na wykresie znajduje się kilka zmiennych (np. macierzowy wykres rozrzutu (SPLOM)), można wybrać zmienną, jak ma zostać uwzględniona na liście rozwijanej.
5. Dodaj nowy przedział wpisując jego nazwę i dolną granicę. Naciśnij klawisz Enter, aby rozpocząć w nowym wierszu.
6. Przeprowadź edycję granicy przedziału, korygując wartość **Dolna granica**.
7. Zmień nazwę przedziału, wpisując jego nową nazwę.
8. Usuń przedział, wybierając wiersz w tabeli i klikając przycisk Usuń.
9. Kliknij przycisk **OK**, aby zastosować zmiany i zamknąć okno dialogowe.

Uwaga: Alternatywnie można usuwać przedziały i zmieniać ich nazwy bezpośrednio na wykresie, klikając linię przedziału prawym przyciskiem myszy i wybierając odpowiednią opcję z menu kontekstowego.

Zastosowanie regionów

Na dowolnym wykresie z dwoma osiami skali (lub zakresu) można narysować regiony, które umożliwiają pogrupowanie wartości w narysowanym prostokątnym obszarze, który jest nazywany regionem. **Region** to obszar wykresu opisany minimalną i maksymalną wartością *X* i *Y*. Jeśli wykres składa się z wielu paneli, region narysowany w jednym z paneli jest również reprezentowany w pozostałych panelach.

Nie we wszystkich wykresach regiony są akceptowane. Niektóre z wykresów akceptujących regiony to: wykresy (liniowe, rozrzutu, bąbelkowy, sekwencyjny itd.),macierzowy wykres rozrzutu (SPLOM) i wykresy przedziałowe. Regiony te są rysowane w przestrzeni X,Y i dlatego nie mogą być definiowane na wykresach 1-W, 3-W lub wykresach z animacjami. W przypadku wykresów z panelowaniem regiony wyświetlane są we wszystkich panelach. W przypadku macierzowego wykresu rozrzutu (SPLOM) region zostanie wyświetlony na odpowiednim górnym wykresie, ale nie będzie wyświetlany na wykresach diagonalnych, ponieważ przedstawiają on tylko jedną zmienną skali.



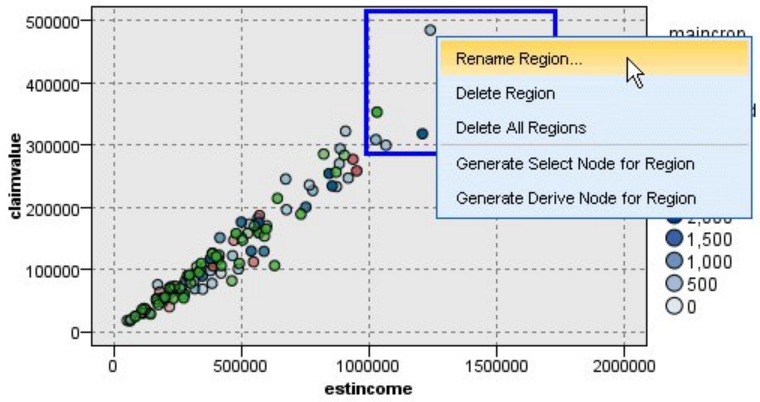
Rysunek 65. Definiowanie regionu dla wysokich wartości wnioskowanych

Definiowanie regionów

Każde zdefiniowanie regionu powoduje utworzenie pogrupowanych wartości. Domyślnie każdy nowy region jest nazywany *Region<N>*, gdzie *N* oznacza liczbę regionów, które zostały już utworzone.

Po zdefiniowaniu regionu można kliknąć linię regionu prawym przyciskiem myszy, aby uzyskać dostęp do podstawowych skrótów. Dodatkowe skróty można jednak wyświetlić, klikając prawym przyciskiem myszy wewnątrz regionu (nie na linii); dotyczą one zadań, takich jak zmiana nazwy, usuwanie lub generowanie węzłów selekcji i wyliczeń dla tego konkretnego regionu.

Można wybrać podzbiory rekordów na podstawie tego, czy są zawarte w konkretnym regionie lub w jednym z kilku regionów. Można również zamieścić informacje o regionie w rekordzie, tworząc węzeł wyliczeń w celu oznaczenia rekordów w oparciu o ich przynależność do regionu. Więcej informacji można znaleźć w temacie “Generowanie węzłów z wykresów” na stronie 282.



Rysunek 66. Eksploracja regionu z wysokimi wartościami wnioskowanymi

Aby zdefiniować regiony:

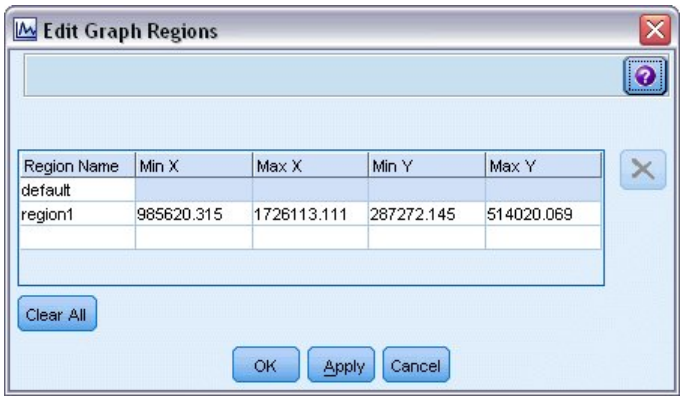
1. Sprawdź, czy jest aktywny tryb eksploracji. Z menu wybierz opcję **Widok> Tryb eksploracji**.
2. Na pasku narzędzi trybu eksploracji kliknij przycisk rysowania regionu.

Rysunek 67. Przycisk rysowania regionów na pasku narzędzi

3. Na wykresie, który akceptuje regiony, kliknij i przeciągnij mysz, aby narysować prostokątny region.

Edytowanie, zmiana nazwy i usuwanie regionów

Istnieje możliwość edytowania właściwości istniejących regionów w oknie dialogowym Edytuj regiony wykresu lub za pośrednictwem menu kontekstowych na samym wykresie.



Rysunek 68. Określanie właściwości zdefiniowanych regionów

Aby edytować regiony:

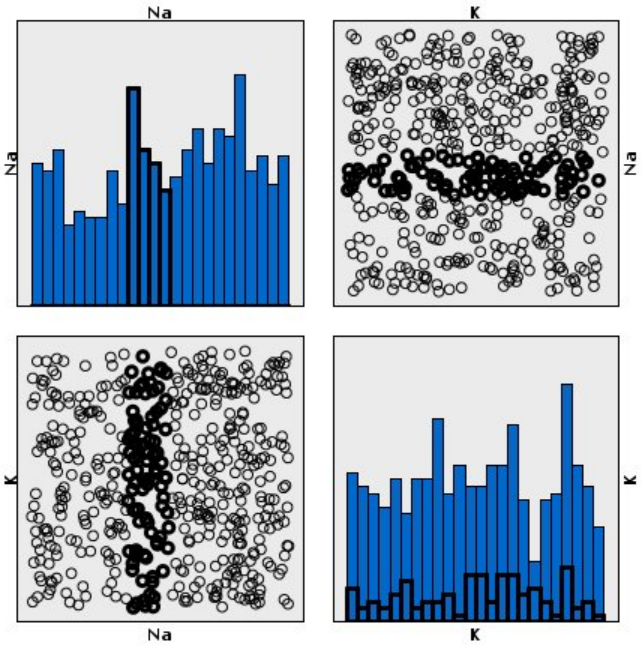
1. Sprawdź, czy jest aktywny tryb eksploracji. Z menu wybierz opcję **Widok> Tryb eksploracji**.
2. Na pasku narzędzi trybu eksploracji kliknij przycisk rysowania regionu.
3. Z menu wybierz kolejno opcje **Edytuj > Regiony wykresu**. Zostanie otwarte okno dialogowe Edytuj regiony wykresu.
4. Jeśli na wykresie znajduje się wiele zmiennych (np. macierzowy wykres rozrzutu (SPLOM)), należy zdefiniować zmienną dla regionu w kolumnach *Field A* (Zmienna A) i *Field B* (Zmienna B).
5. Dodaj nowy region w nowym wierszu, wpisując jego nazwę, wybierając nazwy zmiennych (o ile ma to zastosowanie) i definiując minimalne i maksymalne granice dla każdej zmiennej. Naciśnij klawisz Enter, aby rozpocząć w nowym wierszu.
6. Istniejące granice regionów można edytować, korygując wartości **Min.** i **Maks.** dla zmiennych *A* i *B*.
7. Zmień nazwę regionu, zmieniając jego nazwę w tabeli.
8. Usuń region, wybierając wiersz w tabeli i klikając przycisk Usuń.
9. Kliknij przycisk **OK**, aby zastosować zmiany i zamknąć okno dialogowe.

Uwaga: Alternatywnie można usuwać regiony i zmieniać ich nazwy bezpośrednio na wykresie, klikając linię regionu prawym przyciskiem myszy i wybierając odpowiednią opcję z menu kontekstowego.

Użycie zaznaczonych elementów

Na każdym wykresie można zaznaczać elementy, takie jak słupki, przekroje i punkty. Linie, obszary i powierzchnie można zaznaczać tylko na wykresie sekwencyjnym, wielokrotnym i ewaluacyjnym, ponieważ w tych przypadkach linie odnoszą się do zmiennych. Każde zaznaczenie elementu powoduje wyróżnienie wszystkich danych reprezentowanych przez ten element. Na wszystkich wykresach, na których ta sama obserwacja jest reprezentowana w więcej niż jednym miejscu (np. macierzowe wykresy rozrzutu (SPLOM)), zaznaczanie jest jednoznaczne z wyróżnieniem (ang. brushing). Można zaznaczać elementy na wykresach, a także w przedziałach i regionach. Po każdym zaznaczeniu elementu i

powróceniu do trybu edycji zaznaczenie pozostaje nadal widoczne.



Rysunek 69. Zaznaczanie elementów na macierzowym wykresie rozrzutu (SPLOM)

Elementy można zaznaczać i usuwać ich zaznaczenie, klikając je na wykresie. Po kliknięciu elementu po raz pierwszy w celu zaznaczenia go kolor obramowania elementu zostaje pogrubiony, co oznacza, że został on zaznaczony. Ponowne kliknięcie elementu powoduje, że granica znika i element nie jest już zaznaczony. Aby zaznaczyć wiele elementów, można nacisnąć i przytrzymać klawisz Ctrl, jednocześnie klikając elementy lub można przeciągnąć mysz przez wszystkie elementy, jakie mają zostać zaznaczone, używając „różdżki". Należy pamiętać, że kliknięcie kolejnego obszaru lub elementu bez przytrzymania klawisza Ctrl spowoduje skasowanie zaznaczenia wszystkich wcześniejszych elementów.

Dla zaznaczonych na wykresie elementów można wygenerować węzeł selekcji i wyliczeń. Więcej informacji można znaleźć w temacie "Generowanie węzłów z wykresów".

Aby zaznaczyć elementy:

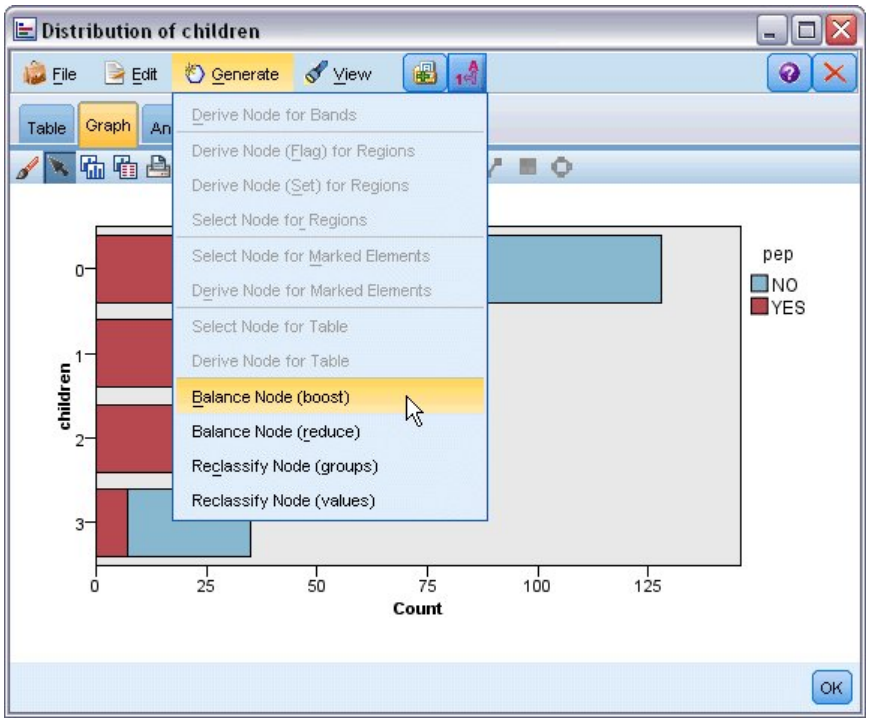
1. Sprawdź, czy jest aktywny tryb eksploracji. Z menu wybierz opcję **Widok> Tryb eksploracji**.
2. Na pasku narzędzi trybu eksploracji kliknij przycisk zaznaczania elementów.
3. Kliknij odpowiedni element lub kliknij i przeciągnij mysz, aby narysować linię wokół regionu zawierającego wiele elementów.

Generowanie węzłów z wykresów

Jedną z najbardziej zaawansowanych właściwości programu IBM SPSS Modeler jest możliwość generowania węzłów z wykresu lub dokonywania wyboru na wykresie. Przykładowo, z wykresu sekwencyjnego można wygenerować węzły wyliczeń i selekcji na podstawie wyboru lub regionu danych, co w rzeczywistości spowoduje utworzenie „podzbiorów" danych. Można na przykład użyć tej wydajnej funkcji w celu identyfikacji i wykluczenia wartości skrajnych.

Jeśli można narysować przedział, można również wygenerować węzeł wyliczeń. Na wykresach z dwoma osiami skali można wygenerować węzły wyliczeń i selekcji z regionów narysowanych na wykresie. Na wykresach z zaznaczonymi elementami z elementów tych można wygenerować węzły wyliczeń i selekcji, a w niektórych przypadkach węzły filtrowania. Generowanie węzła ważenia jest możliwe w przypadku dowolnych wykresów wyświetlających rozkład

liczebności.



Rysunek 70. Wykres z wyświetlanym menu generowania

Po każdym wygenerowaniu węzła jest on umieszczany w obszarze roboczym strumienia, dzięki czemu można połączyć go z istniejącym strumieniem. Z wykresów można wygenerować następujące węzły: Selekcja, Wyliczanie, Zrównoważenie, Filtrowanie i Rekodowanie.

Węzły selekcji

Węzły selekcji mogą zostać wygenerowane w celu przetestowania uwzględnienia rekordów w regionie i wykluczenia wszystkich rekordów znajdujących się poza regionem lub wykluczenia odwrócenia kolejności przetwarzania.

- **Dla przedziałów**. Można wygenerować węzeł selekcji, który będzie uwzględniał lub wykluczał rekordy w danym przedziale. Opcja **Węzeł selekcji tylko dla przedziałów** jest dostępna tylko za pośrednictwem menu kontekstowego, ponieważ należy wybrać, który przedział będzie użyty w węźle selekcji.
- **Dla regionów**. Można wygenerować węzeł selekcji, który będzie uwzględniał lub wykluczał rekordy w danym regionie.
- **Dla zaznaczonych**. Można wygenerować węzły selekcji, aby przechwytywać rekordy odpowiadające zaznaczonym elementom lub łączom wykresu sieciowego.

Węzły wyliczeń

Węzły wyliczeń można wygenerować na podstawie regionów, przedziałów i zaznaczonych elementów. Wszystkie wykresy umożliwiają tworzenie węzłów wyliczeń. W przypadku wykresów ewaluacyjnych wyświetlane jest okno dialogowe umożliwiające wybór modelu. W przypadku węzłów sieciowych możliwe jest użycie opcji **Węzeł wyliczeń („And”)** oraz **Węzeł wyliczeń („Or”)**.

- **Dla przedziałów**. Istnieje możliwość wygenerowania węzła wyliczeń, który będzie tworzył kategorię dla każdego przedziału zaznaczonego na osi, używając nazw przedziałów wymienionych w oknie dialogowym edycji przedziałów jako nazw kategorii.
- **Dla regionów**. Możliwe jest wygenerowanie węzła wyliczeń (**Węzeł wyliczeń (flaga)**), który utworzy zmienną flagi o nazwie *in_region* (w_regionie) z flagami ustawionymi na wartość *T* (Prawda) dla rekordów w każdym regionie

oraz na *F* (Fałsz) dla regionów poza wszystkimi regionami. Można również wygenerować węzeł wyliczeń (**Węzeł wyliczeń (nominalna)**), który utworzy zbiór wartości dla każdego regionu z nową zmienną o nazwie *region* dla każdego rekordu, którego wartością będzie nazwa regionu, w którym znajduje się dany rekord. Rekordy znajdujące poza wszystkimi regionami otrzymują nazwę domyślnego regionu. Wartościami nazw stają się nazwy regionów wyświetlane w oknie dialogowym edycji regionów.

- **Dla zaznaczonych**. Można wygenerować węzeł wyliczeń, który będzie obliczał flagę, która ma wartość *True* (Prawda) dla wszystkich zaznaczonych elementów oraz *False* (Fałsz) dla pozostałych rekordów.

Węzły ważenia

Węzły ważenia można wygenerować w celu skorygowania dysproporcji w danych, np. w celu zredukowania częstotliwości wspólnych wartości (należy użyć opcji menu **Węzeł ważenia (redukcja)**) lub wzmocnienia występowania rzadko występujących wartości (należy użyć opcji **Węzeł ważenia (wzmocnienie)**). Generowanie węzła ważenia jest możliwe w przypadku dowolnych wykresów wyświetlających rozkład liczebności, takich jak histogram, wykres punktowy, wykres przedziałowy, wykres słupkowy liczebności, kołowy liczebności oraz wykres wielokrotny.

Węzły filtrowania

Węzły filtrowania można wygenerować w celu zmiany nazwy lub filtrowania zmiennych na podstawie linii lub węzłów zaznaczonych na wykresie. W przypadku wykresów ewaluacyjnych linia najlepszego dopasowanie nie powoduje wygenerowania węzła filtrowania.

Węzły rekodowania

Węzły rekodowania można wygenerować w celu rekodowania wartości. Ta opcja jest używana w przypadku wykresów rozkładu. Można wygenerować węzeł rekodowania dla **grup** w celu rekodowania określonych wartości wyświetlanej zmiennej w zależności ich uwzględnienia w grupie (grupy należy wybierać metodą Ctrl+kliknięcie na karcie **Tabele**). Można również wygenerować węzeł rekodowania dla **wartości**, aby rekodować dane na istniejący zbiór wartości liczbowych, np. rekodować dane na standardowy zbiór wartości w celu scalenia danych finansowych z różnych firm na potrzeby analizy.

Uwaga: Jeśli wartości są wstępnie zdefiniowane, można je wczytać do programu IBM SPSS Modeler jako plik płaski i użyć rozkładu w celu wyświetlania wszystkich wartości. Następnie można wygenerować węzeł rekodowania (wartości) dla tej zmiennej bezpośrednio z wykresu. W ten sposób wszystkie wartości przewidywane zostaną umieszczone w kolumnie *Nowa wartość* węzła rekodowania (lista rozwijana).

Podczas ustawiania opcji dla węzła Rekodowanie tabela umożliwia usuwanie mapowania ze starych zestawów wartości na nowe wartości określone przez użytkownika:

- **Wartość oryginalna.** W tej kolumnie wyświetlane są istniejące wartości wybranych zmiennych.
- **Nowa wartość.** Ta kolumna służy do wpisania wartości nowej kategorii lub wybrania jej z listy rozwijanej. W przypadku automatycznego generowania węzła rekodowania na podstawie wartości z wykresu rozkładu wartości te są uwzględniane na liście rozwijanej. Dzięki temu można szybko odwzorować istniejące wartości na zbiór znanych wartości. Przykładowo organizacje zajmujące się opieką medyczną niekiedy grupują diagnozy w różny sposób na podstawie sieci lub ustawień regionalnych. Po scaleniu lub zgromadzeniu danych wszystkie strony będą musiały rekodować nowe, a nawet istniejące dane w spójny sposób. Zamiast ręcznego wpisywania z długiej listy każdej wartości przewidywanej można wczytać nadrzędną listę wartości do programu IBM SPSS Modeler, uruchomić wykres rozkładu dla zmiennej *Diagnosis* (Diagnoza) i wygenerować węzeł rekodowania (wartości) dla tej zmiennej bezpośrednio z wykresu. Ten proces udostępni wszystkie wartości przewidywane diagnozy z listy rozwijanej nowych wartości.

Więcej informacji o węźle Rekodowanie zawiera temat “Ustawianie opcji dla węzła Rekodowanie” na stronie 163.

Generowanie węzłów z wykresów

W celu wygenerowania węzłów można użyć menu Utwórz w oknie wyniku graficznego. Wygenerowany węzeł zostanie umieszczony w obszarze roboczym strumienia. Aby użyć węzła, należy połączyć go z istniejącym strumieniem.

Aby wygenerować węzeł z wykresu:

1. Sprawdź, czy jest aktywny tryb eksploracji. Z menu wybierz opcję **Widok> Tryb eksploracji**.
2. Na pasku narzędzi trybu eksploracji kliknij przycisk regionu.
3. Zdefiniuj przedziały, regiony lub wszelkie zaznaczone elementy potrzebne do wygenerowania węzła.
4. Z menu Utwórz wybierz rodzaj węzła, jaki ma zostać utworzony. Aktywne są tylko te węzły, których utworzenie jest możliwe.

Uwaga: Alternatywnie można również wygenerować węzły bezpośrednio z wykresu; w tym celu należy kliknąć prawym przyciskiem myszy i wybrać odpowiednią opcję generowania z menu kontekstowego.

Edycja wizualizacji

Tryb eksploracji umożliwia analityczną eksplorację danych i wartości reprezentowanych przez wizualizację, a tryb edycji umożliwia zmianę wyglądu i układu wizualizacji. Na przykład można zmienić czcionki i kolory tak, aby pasowały do oficjalnej stylistyki danej organizacji. Aby wybrać ten tryb, z menu wybierz **Widok > Tryb edycji** (lub kliknąć ikonę paska narzędzi).

W trybie edycji dostępnych jest kilka pasków narzędzi, które mają wpływ na różne aspekty układu wizualizacji. Jeśli któreś z tych pasków nie są wykorzystywane, można je ukryć. Dzięki temu zwiększa się w oknie dialogowym ilość miejsca dostępnego dla wykresu. Aby zaznaczyć lub odznaczyć paski narzędzi, w menu Widok kliknij odpowiednią nazwę paska narzędzi.

Uwaga: Aby dodać więcej szczegółów do wizualizacji, można skorzystać z tytułu, stopki i etykiet osi. Więcej informacji można znaleźć w temacie “Dodawanie tytułów i stopek” na stronie 295.

W **trybie edycji** dostępnych jest kilka opcji edycji wizualizacji. Można:

- Edytować tekst i go formatować.
- Zmieniać kolor wypełnienia, przezroczystość i desenie ramek oraz elementów graficznych.
- Zmieniać kolor i rysować krawędzie i linie kreską przerywaną.
- Obracać i zmieniać kształt oraz współczynnik proporcji punktów danych.
- Zmieniać rozmiar elementów graficznych (na przykład pasków i punktów).
- Regulować przestrzeń pomiędzy elementami, używając do tego marginesów i wypełnień.
- Określać formatowanie liczb.
- Zmieniać ustawienia osi i skali.
- Sortować, wykluczać i zwijać kategorie na osi kategorii.
- Ustawiać orientację paneli.
- Stosować transformacje układu współrzędnych.
- Zmieniać statystyki, typy elementów graficznych i modyfikatorów konfliktów.
- Zmieniać położenie legendy.
- Stosować arkusze wizualizacji.

Omówione poniżej tematy dotyczą sposobu wykonywania tych różnych zadań. Zaleca się także zapoznanie z ogólnymi regułami edycji wykresów.

Przełączanie się w tryb edycji

Z menu wybierz:

Widok > Tryb edycji

Ogólne reguły edycji wizualizacji

Tryb edycji

Wszystkie zmiany są dokonywane w trybie edycji. Aby włączyć tryb edycji, wybierz z menu:

Widok > Tryb edycji

Uruchom zaznaczony kod

Dostępne opcje edycji zależą od dokonanego wyboru. W zależności od dokonanego wyboru włączane są różne paski narzędzi i okna właściwości. Jedynie włączone elementy mają zastosowanie do bieżącego wyboru. Na przykład, jeżeli wybrana została oś, to w oknie właściwości dostępne są zakładki: Skala, Znaczniki główne i Znaczniki pomocnicze.

Poniżej wypisano wskazówki dotyczące wybierania elementów wizualizacji:

- Kliknij element, aby go zaznaczyć.
- Za pomocą pojedynczego kliknięcia wybierz element graficzny (taki jak punkty wykresu rozrzutu lub słupki wykresu słupkowego). Po dokonaniu wstępnego wyboru kliknij ponownie, aby zawęzić wybór do grupy elementów graficznych lub pojedynczego elementu graficznego.
- Naciśnij klawisz Esc, aby odznaczyć wszystko.

Palety

Kiedy element zostanie wybrany w wizualizacji, następuje aktualizacja różnych palet, aby odzwierciedlić dokonany wybór. Palety zawierają elementy sterujące służące do edycji wybranych pozycji. Paletami mogą być paski narzędzi lub panele z wieloma kontrolkami i zakładkami. Palety mogą być ukryte, aby zapewnić wyświetlanie poprawnej palety do wprowadzania zmian. Sprawdź menu Widok, aby zobaczyć, które z palet są aktualnie wyświetlane.

Można zmieniać pozycję palet poprzez klikanie i przeciąganie pustej przestrzeni palety paska narzędzi lub lewej strony innych palet. Wizualna informacja zwrotna pozwala się upewnić, gdzie paleta ma być umieszczona. W przypadku palet, które nie są paskami narzędzi, można również klikać przycisk zamykania, aby ukryć paletę, oraz przycisk wyłączenia przypięcia, aby wyświetlić paletę w osobnym oknie. Kliknięcie przycisku pomocy wyświetli pomoc dotyczącą określonej palety.

Ustawienia automatyczne

Niektóre ustawienia zawierają opcję **-auto**. Oznacza to, że stosowana jest wartość automatyczna. To, które z ustawień automatycznych są używane, zależy od danej wizualizacji i wartości danych. Można wprowadzać wartości do nadpisania ustawień automatycznych. Jeśli zachodzi potrzeba przywrócenia łańcucha automatycznego, należy skasować bieżącą wartość i nacisnąć klawisz Enter. Ustawienie ponownie wyświetli opcję **-auto-**.

Usuwanie i ukrywanie elementów

Można usuwać lub ukrywać wiele elementów występujących w wizualizacji. Na przykład można ukryć legendę lub etykietę osi. Aby usunąć element, zaznacz go i naciśnij klawisz Delete. Jeśli element nie pozwala się usunąć, nic się nie wydarzy. Jeśli element zostanie przypadkowo usunięty, naciśnij Ctrl+Z, aby cofnąć usunięcie.

Stan

Niektóre paski narzędzi odzwierciedlają stan bieżącego wyboru, inne nie. Paleta właściwości zawsze odzwierciedla stan. Jeśli pasek narzędzi *nie* odzwierciedla stanu, jest o tym wzmianka w tytule opisującym pasek narzędzi.

Edytowanie i formatowanie tekstu

Można edytować tekst i zmieniać formatowanie całego bloku tekstowego. Należy zauważyć, że nie można edytować tekstu, który jest przypisany bezpośrednio do wartości danych. Na przykład nie można edytować etykiety znacznika, ponieważ zawartość etykiety jest dostarczana z powiązanych danych. Jednak można formatować dowolny tekst wizualizacji.

W jaki sposób edytować tekst

1. Dwukrotnie kliknij blok tekstowy. Ta czynność spowoduje zaznaczenie całego tekstu. Wszystkie paski narzędzi będą w tym momencie wyłączone, ponieważ nie można zmieniać żadnej innej części wizualizacji, gdy edytowany jest tekst.
2. Wpisz tekst, który ma zastąpić istniejący tekst. Można również ponownie kliknąć tekst, aby wyświetlić kursor. Umieść kursor w żądanym miejscu i wprowadź dodatkowy tekst.

Formatowanie tekstu

1. Wybierz ramkę zawierającą tekst. Nie klikaj tekstu dwukrotnie.
2. Sformatuj tekst, używając do tego paska narzędzi dotyczącego czcionek. Jeśli pasek narzędzi nie jest włączony, upewnij się, że zaznaczona jest jedynie *ramka* zawierająca tekst. Jeśli zaznaczony jest tekst sam w sobie, pasek narzędzi będzie wyłączony.

Można zmieniać następujące atrybuty czcionki:

- Kolor
- Krój (na przykład Arial lub Verdana)
- Rozmiar (jednostką jest pt, chyba że zostanie wskazana inna jednostka, taka jak pc)
- Grubość linii
- Ustawienie zależne od obramowania tekstu

Formatowanie ma zastosowanie do całego tekstu umieszczonego w ramce. Nie można zmienić formatowania pojedynczych liter ani słów, które znajdują się w danym bloku tekstowym.

Zmiana kolorów, deseni, krawędzi i przezroczystości

Wiele różnych elementów wizualizacji posiada wypełnienie i obramowanie. Najbardziej oczywistym przykładem jest słupek wykresu słupkowego. Kolor słupków jest kolorem wypełnienia. Takie kolory mogą być otoczone ciągłą, czarną ramką.

Istnieją również mniej oczywiste elementy wizualizacji, które mają kolory wypełnienia. Jeśli kolorem wypełnienia jest transparentność, można nie zauważyć, że zastosowano wypełnienie. Jako przykład można rozważyć tekst etykiety osi. Wydaje się, że tekst jest tekstem „unoszącym się”, ale tak naprawdę jest on wyświetlany w ramce, która ma przezroczysty kolor wypełnienia. Ramkę można zobaczyć po zaznaczeniu etykiety osi.

Dowolna ramka wizualizacji może mieć styl wypełnienia i obramowania. Dotyczy to także ramki wokół całej wizualizacji. Ponadto każde wypełnienie ma przypisany poziom przezroczystości/nieprzezroczystości, który można regulować.

Zmiana kolorów, deseni, krawędzi i przezroczystości

1. Wybierz element przeznaczony do formatowania. Zaznacz na przykład słupki wykresu słupkowego lub ramkę zawierającą tekst. Jeśli wizualizacja jest podzielona przez zmienną kategorialną lub pole, można również wybrać grupę, która odpowiada indywidualnej kategorii. Pozwala to na zmianę domyślnego sposobu prezentacji przypisanego do tej grupy. Na przykład można zmienić kolor jednej z grup separacji pionowej na zestawionym wykresie słupkowym.
2. Aby zmienić kolor wypełnienia, kolor ramki lub deseń wypełnienia, należy skorzystać z paska narzędzi dotyczącego koloru.

   Uwaga: Ten pasek narzędzi nie odzwierciedla stanu bieżącego wyboru.

W celu zmiany koloru lub wypełnienia można kliknąć przycisk, aby wybrać wyświetloną opcję, lub kliknąć strzałkę skierowaną w dół, aby wybrać inną opcję. W przypadku kolorów należy zauważyć, że dostępny jest jeden kolor, który wygląda jak biały, ale ma czerwoną przekątną. To kolor przezroczysty. Można go wykorzystać na przykład do tego, aby ukryć obramowania krawędzi histogramu.

- Pierwszy przycisk kontroluje kolor wypełnienia. Jeśli kolor jest powiązany z ciągłym lub porządkowym polem, przycisk ten zmienia kolor wypełnienia na kolor powiązany z najwyższą wartością danych. Możesz użyć karty Kolor znajdującej się na palecie właściwości, by zmienić kolor powiązany z najniższą wartością i brakiem danych. W miarę zwiększania się wartości danych kolor elementów zmienia się przyrostowo z Najmniejszego koloru do Największego koloru.
- Drugi przycisk kontroluje kolor obramowania.
- Trzeci przycisk kontroluje deseń wypełnienia. Styl wypełnienia wykorzystuje kolor obramowania. Z tego powodu wzór wypełnienia widoczny jest tylko wówczas, gdy widoczny jest kolor obramowania.
- Czwarty element sterujący to suwak i pole tekstowe, pozwalające ustawić przezroczystość koloru i wzoru wypełnienia. Mniejsza wartość procentowa oznacza mniejszą nieprzezroczystość i wyższą przezroczystość. 100% to brak przezroczystości.

3. Aby zmienić krawędź obramowania lub linii, należy skorzystać z paska narzędzi dotyczącego linii.

   Uwaga: Ten pasek narzędzi nie odzwierciedla stanu bieżącego wyboru.

   Podobnie jak w przypadku innych pasków narzędzi, można kliknąć przycisk, aby wybrać wyświetloną opcję, lub kliknąć strzałkę skierowaną w dół, aby wybrać inną opcję.

Obracanie i zmienianie kształtu i współczynnik proporcji punktów danych

Można obracać elementy punktów, przypisywać różne wcześniej zdefiniowane kształty lub zmieniać współczynnik proporcji (stosunek szerokości do wysokości).

Modyfikowanie elementów punktów

1. Zaznacz elementy punktów. Nie można obracać ani zmieniać kształtu oraz współczynnika proporcji pojedynczych punktów danych.
2. Należy używać paska narzędzi dotyczącego symboli, aby modyfikować te punkty.
   - Pierwszy przycisk umożliwia zmianę kształtu punktów. Kliknij strzałkę skierowaną w dół i wybierz wcześniej zdefiniowany kształt.
   - Drugi przycisk umożliwia obracanie punktów do określonej pozycji kompasu. Kliknij strzałkę skierowaną w dół, a następnie przeciągnij igłę do żądanej pozycji.
   - Trzeci przycisk umożliwia zmianę współczynnika proporcji. Kliknij strzałkę skierowaną w dół, a następnie kliknij i przeciągnij wyświetlony prostokąt. Kształt prostokąta reprezentuje współczynnik proporcji.

Zmiana rozmiaru elementów graficznych

Można zmieniać rozmiar elementów graficznych używanych w wizualizacji. Obejmuje to m.in. słupki, linie i punkty. Jeśli element graficzny jest rozmiarowany przez zmienną lub pole, określona wartość jest rozmiarem *minimalnym*.

Zmiana rozmiaru elementów graficznych

1. Zaznacz elementy graficzne, których rozmiar chcesz zmienić.
2. Skorzystaj z suwaka lub wprowadź konkretny rozmiar dla opcji dostępnej na pasku narzędzi Symbol. Jednostką są piksele, chyba że zostanie wskazana inna jednostka (poniżej znajduje się pełna lista jednostek i ich skrótów). Można również określić procent (na przykład 30%), co oznacza, że element graficzny używa określonego procentu dostępnego miejsca. Dostępne miejsce zależy od typu elementu graficznego i danej wizualizacji.

Tabela 39. Poprawne skróty jednostek

Skrót	Jednostka
cm	centymetr

Tabela 39. Poprawne skróty jednostek (kontynuacja)

Skrót	Jednostka
w	cal
mm	milimetr
pc	pica
pt	punkt
px	piksel

Definiowanie marginesów i wypełnienia

Jeśli wokół lub wewnątrz ramki jest zbyt mało wolnego miejsca, można zmienić jej ustawienia marginesów lub wypełnienia. **Margines** to ilość miejsca pomiędzy ramką a innymi elementami znajdującymi się wokół niej. **Wypełnienie** to ilość miejsca pomiędzy obramowaniem ramki a jej *zawartością*.

Określanie marginesów i wypełnienia

1. Wybierz ramkę, dla której chcesz określić wartość marginesów i wypełnienia. Może nią być ramka tekstowa, ramka wokół legendy lub nawet ramka danych wyświetlająca elementy graficzne (na przykład słupki i punkty).
2. Skorzystaj z zakładki Marginesy palety właściwości, aby wprowadzić ustawienia. Wszystkie rozmiary są podane w pikselach, chyba że zostanie podana inna jednostka (na przykład cm lub cale).

Formatowanie liczb

Można określić format liczb w etykietach znaczników na osi ciągłej lub etykietach wartości danych zawierających liczby. Można na przykład określić, że liczby wyświetlane w etykietach znaczników są przedstawiane w tysiącach.

Określanie formatów liczb

1. Zaznacz etykiety znaczników osi ciągłej lub etykiety wartości danych, które zawierają liczby.
2. Kliknij zakładkę palety właściwości **Format**.
3. Wybierz żądane opcje formatowania liczb:

   Prefiks. Znak, który będzie wyświetlany przed liczbą. Można na przykład wprowadzić znak dolara ($), jeśli liczby przedstawiają pensje w dolarach amerykańskich.

   Sufiks. Znak, który będzie wyświetlany za liczbą. Można na przykład wprowadzić znak procentu (%), jeśli liczby przedstawiają wartości procentowe.

   Min. cyfr liczby całkowitej. Minimalna liczba cyfr do wyświetlenia w części całkowitej reprezentacji dziesiętnej. Jeśli aktualna wartość nie zawiera minimalnej liczby cyfr, część całkowita tej wartości będzie miała dołączone zera.

   Maks. cyfr liczby całkowitej. Maksymalna liczba cyfr do wyświetlenia w części całkowitej reprezentacji dziesiętnej. Jeśli aktualna wartość przekracza maksymalną liczbę cyfr, część całkowita tej wartości będzie zastąpiona gwiazdkami.

   Min. cyfr liczby dziesiętnej. Minimalna liczba cyfr do wyświetlenia w części dziesiętnej reprezentacji dziesiętnej lub naukowej. Jeśli aktualna wartość nie zawiera minimalnej liczby cyfr, część dziesiętna wartości będzie miała dołączone zera.

   Maks. cyfr liczby dziesiętnej. Maksymalna liczba cyfr do wyświetlenia w części dziesiętnej reprezentacji dziesiętnej lub naukowej. Jeśli aktualna wartość przekroczy maksymalną liczbę cyfr, część dziesiętna zostanie zaokrąglona do odpowiedniej liczby cyfr.

   Naukowy. Określa, czy warto wyświetlać liczby w formacie naukowym. Notacja naukowa jest przydatna w przypadku bardzo dużych lub bardzo małych liczb. Opcja **-auto-** umożliwia aplikacji określenie, czy notacja naukowa jest odpowiednia.

   Skalowanie. Czynnik skalujący, który jest liczbą przez którą dzielona jest wartość pierwotna. Czynnika skalującego należy używać w przypadku liczb o dużych wartościach w celu ograniczenia długości

odpowiadających im etykiet. Po zmianie formatu liczb etykiet znaczników należy dokonać edycji tytułu osi, aby wskazać, jak należy interpretować taką liczbę. Przyjmując na przykład, że na osi ilościowej są wyświetlane pensje i widoczne etykiety 30 000, 50 000 i 70 000, można wprowadzić czynnik skalujący równy 1000, aby były wyświetlane wartości 30, 50 i 70. Następnie należy dokonać edycji tytułu osi ilościowej, wprowadzając tekst w tysiącach.

Nawiasy okrągłe dla -ve. Określa, czy nawiasy okrągłe mają być wyświetlane dla wartości ujemnych.

Grupowanie. Określa, czy wyświetlać znak między grupami cyfr. Bieżąca lokalizacja komputera określa, który znak ma być używany do grupowania cyfr.

Zmiana ustawienia osi i skali

Dostępnych jest kilka opcji modyfikowania osi i skali.

Zmiana ustawień osi i skali

1. Wybierz dowolną część osi (na przykład etykietę osi lub etykiety znaczników).
2. Korzystając z zakładek Skala, Znaczniki główne, Znaczniki pomocnicze palety właściwości, zmień ustawienia osi i skali.

Zakładka Skala

Zakładka Skala nie pojawia się dla wykresów, w których dane są wstępnie agregowane (na przykład w histogramach).

Typ. Definiuje, czy skala ma być liniowa, czy przekształcona. Przekształcenia ilościowe pomagają zrozumieć zależności występujące w danych lub przyjąć założenia niezbędne przy wnioskowaniu statystycznym. Na wykresach rozrzutu można używać przekształconej skali, jeśli zależność pomiędzy zmienną zależną a niezależną lub polami nie jest liniowa. Przekształcenia ilościowe można także stosować do zwiększania symetrii histogramu skośnego, tak aby wyglądem bardziej przypominał rozkład normalny. Należy pamiętać, że przekształcana jest tylko skala, w której wyświetlane są dane; rzeczywiste dane nie są przekształcane.

- **liniowy.** Definiuje liniową, nieprzekształconą skalę.
- **logarytm.** Definiuje skalę przekształconą logarytmiczną o podstawie 10. Aby uwzględnić zero i wartości ujemne, przekształcenie to używa zmodyfikowanej wersji funkcji logarytmu. Funkcja „bezpiecznego logarytmu” (safe log) jest zdefiniowana następująco: sign(x) * log(1 + abs(x)). Stąd safeLog(-99) równa się:

  sign(-99) * log(1 + abs(-99)) = -1 * log(1 + 99) = -1 * 2 = -2
- **potęgowy.** Określa potęgowo przekształconą skalę. Została użyta potęga o wykładniku 0,5. Aby uwzględnić wartości ujemne, przekształcenie to używa zmodyfikowanej wersji funkcji potęgowej. Funkcja „bezpiecznej potęgi” (safe power) jest zdefiniowana następująco: sign(x) * pow(abs(x), 0,5). Stąd safePower(-100) równa się:

  sign(-100) * pow(abs(-100), 0.5) = -1* pow(100, 0.5) = -1 * 10 = -10

Min./Maks./Subtelne minimum/Subtelne maksimum. Określa zakres skali. Wybranie opcji **Subtelne minimum** oraz **Subtelne maksimum** umożliwia aplikacji wybór poprawnej skali w oparciu o występujące dane. Wartości minimalne i maksymalne są „subtelne”, ponieważ zazwyczaj całe wartości są większe lub mniejsze niż maksimum lub minimum wartości danych. Na przykład, jeżeli zakres danych wynosi od 4 do 92, subtelne minimum i subtelne maksimum skali może być 0 i 100 zamiast faktycznego minimum i maksimum zakresu danych. Należy być ostrożnym, aby nie ustawić zakresu, który jest zbyt mały i który ukrywa w ten sposób ważne elementy. Należy zauważyć również, że nie można ustawiać niedwuznacznego minimum i maksimum, jeżeli zostanie wybrana opcja **Uwzględnij zero**.

Margines dolny/Margines górny. Tworzy marginesy w dolnej i/lub górnej części osi. Margines jest prostopadły do wybranej osi. Jednostki są podane w pikselach, chyba że zostanie podana inna jednostka (na przykład cm lub cale). Na przykład, jeżeli opcja **Górny margines** zostanie ustawiona na 5 dla osi pionowej, to poziomy margines o rozmiarze 5 pikseli będzie przebiegał wzdłuż górnej części ramki danych.

Odbicie. Określa, czy skala powinna być odbita.

Uwzględnij zero. Wskazuje, że skala powinna obejmować 0. Opcja ta jest powszechnie używana w przypadku wykresów słupkowych, aby się upewnić, że słupki zaczynają się w zerze, a nie w wartości bliskiej wysokości najmniejszego słupka. Jeśli opcja ta jest wybrana, opcje **Min.** i **Maks.** są wyłączone, ponieważ nie można ustawić własnego minimum i maksimum dla zakresu skali.

Zakładki Znaczniki główne/Znaczniki pomocnicze

Znaczniki lub **znaczniki** są liniami, które są wyświetlane na osi. Określają one wartości o określonych odstępach lub kategoriach. **Znaczniki główne** są znacznikami opisanymi etykietami. Są one dłuższe niż inne znaczniki. **Znaczniki pomocnicze** są znacznikami, które pojawiają się pomiędzy głównymi znacznikami. Niektóre opcje są specyficzne dla typu znacznika, ale większość opcji jest dostępna dla znaczników głównych i pomocniczych.

Pokaż znaczniki. Definiuje, czy na wykresie pokazywane są główne lub pomocnicze znaczniki.

Pokaż linie siatki. Definiuje, czy linie siatki są wyświetlane dla znaczników głównych lub pomocniczych. **Linie siatki** to linie, które przecinają cały wykres od jednej osi, do drugiej.

Położenie. Definiuje położenie znaczników względem osi.

Długość. Definiuje długość znaczników. Jednostki są podane w pikselach, chyba że zostanie podana inna jednostka (na przykład cm lub cale).

Podstawa. *Ma zastosowanie jedynie do znaczników głównych.* Określa wartość, dla której jest wyświetlany pierwszy znacznik główny.

Delta. *Ma zastosowanie jedynie do znaczników głównych.* Definiuje różnice między znacznikami głównymi. Oznacza to, że główne znaczniki będą pojawiać się co *n-tą*wartość, gdzie *n* jest wartością delta.

Przedziały. *Ma zastosowanie jedynie do znaczników pomocniczych.* Definiuje liczbę podziałów znacznikami pomocniczymi pomiędzy znacznikami głównymi. Liczba znaczników pomocniczych jest o jeden mniejsza od liczby podziałek. Na przykład, załóżmy, że w główne znaczniki znajdują się przy wartościach 0 i 100. W przypadku wprowadzenia wartości 2 jako liczby podziałek znaczników pomocniczych uzyskamy *jeden* znacznik pomocniczy na wartości 50, który podzieli zakres 0–100 i utworzy *dwie* podziałki.

Edycja kategorii

Można edytować kategorie wyświetlone na osi kategorii. Można tego dokonywać na kilka sposobów:

- Poprzez zmianę kolejności sortowania wyświetlanych kategorii.
- Poprzez wykluczenie określonych kategorii.
- Poprzez dodawanie kategorii, która nie jest widoczna w zbiorze danych.
- Poprzez zwijanie/łączenie małych kategorii w jedną kategorię.

Zmiana kolejności sortowania kategorii

1. Zaznacz oś kategorii. Paleta Kategorie wyświetli kategorie osi.

   Uwaga: Jeśli paleta nie jest widoczna, należy się upewnić, że została ona włączona. Z menu Widok w opcji IBM SPSS Modeler, wybierz **Kategorie**.
2. Na palecie Kategorie należy wybrać opcję sortowania, wykorzystując do tego listę rozwijaną:

   Użytkownika. Sortowanie kategorii na podstawie kolejności ich występowania na palecie. Należy użyć przycisków strzałek, aby przesuwać kategorie na początek listy, w górę, na dół i na koniec listy.

   Dane. Sortowanie kategorii na podstawie kolejności ich występowania w zbiorze danych.

   Nazwa. Sortowanie kategorii alfabetycznie, za pomocą nazw wyświetlanych na palecie. Może to być albo wartość lub etykieta, w zależności od tego, czy wybrano przycisk paska narzędzi wyświetlający wartości i etykiety.

Wartość. Sortowanie kategorii względem ukrytych wartości danych, przy użyciu wartości wyświetlanych na palecie w nawiasach okrągłych. Z tą opcją są zgodne jedynie źródła danych z metadanymi (takie jak pliki danych IBM SPSS Statistics).

Statystyka. Sortowanie kategorii na podstawie obliczonej statystyki dla każdej kategorii. Przykłady statystyk to liczebności, udziały procentowe i średnie. Opcja ta jest dostępna jedynie wówczas, gdy statystyka jest używana na wykresie.

Dodawanie kategorii

Domyślnie dostępne są tylko kategorie widoczne w zbiorze danych. W razie potrzeby można dodać kategorię do wizualizacji.

1. Zaznacz oś kategorii. Paleta Kategorie wyświetli kategorie osi.

   Uwaga: Jeśli paleta nie jest widoczna, należy się upewnić, że została ona włączona. Z menu Widok w opcji IBM SPSS Modeler, wybierz **Kategorie**.
2. Na palecie Kategorie kliknij przycisk dodawania kategorii:

Rysunek 71. Przycisk Dodaj kategorię.

3. W oknie dialogowym Dodaj nową kategorię wpisz nazwę kategorii.
4. Kliknij przycisk **OK**.

Wykluczanie określonych kategorii

1. Zaznacz oś kategorii. Paleta Kategorie wyświetli kategorie osi.

   Uwaga: Jeśli paleta nie jest widoczna, należy się upewnić, że została ona włączona. Z menu Widok w opcji IBM SPSS Modeler, wybierz **Kategorie**.
2. Na palecie Kategorie należy zaznaczyć nazwę kategorii na liście uwzględniania, a następnie kliknąć przycisk ze znakiem X. Aby z powrotem przywrócić kategorię, należy zaznaczyć jej nazwę na liście Wykluczone, a następnie kliknąć strzałkę znajdującą się z prawej strony listy.

Łączenie/zwijanie małych kategorii

Istnieje możliwość łączenia kategorii, które są tak małe, że nie ma potrzeby wyświetlania ich osobno. Jeśli na przykład masz wykres kołowy z wieloma kategoriami, rozważ zwinięcie kategorii o udziale procentowym mniejszym niż 10. Zwijanie jest dostępne tylko w przypadku statystyk addytywnych. Nie można przykładowo dodawać do siebie średnich, ponieważ średnie nie są addytywne. Łączenie/zwijanie kategorii przy użyciu średniej jest więc niemożliwe.

1. Zaznacz oś kategorii. Paleta Kategorie wyświetli kategorie osi.

   Uwaga: Jeśli paleta nie jest widoczna, należy się upewnić, że została ona włączona. Z menu Widok w opcji IBM SPSS Modeler, wybierz **Kategorie**.
2. Na palecie Kategorie wybierz **Zwiń**, a następnie zdefiniuj wartość procentu. Wszystkie kategorie, których udział procentowy jest mniejszy niż wprowadzona wartość, będą połączone w jedną kategorię. Wartość procentowa jest oparta na statystyce widocznej na wykresie. Zwijanie jest możliwe tylko wówczas, gdy używana jest statystyka oparta na zliczaniu i sumowaniu.

Zmiana orientacji paneli

Jeśli w wizualizacji używane są panele, można zmienić ich orientację.

Zmiana orientacji paneli

1. Wybierz dowolną część wizualizacji.
2. Kliknij zakładkę palety właściwości **Panele**.

3. Z listy **Układ** wybierz opcję:

Tabela. Układa panele w formie tabeli, w której każda wartość ma przypisany wiersz lub kolumnę.

Transponowane. Układa panele w formie tabeli i jednocześnie zamienia miejscami kolumny i wiersze. Opcja ta nie jest tożsama z transponowaniem samego wykresu. Należy zauważyć, że oś *x* i oś *y* po wybraniu tej opcji nie ulegają zmianie.

Lista. Układa panele w formie listy, na której każda komórka reprezentuje kombinację wartości. Kolumny i wiersze nie są dłużej przypisane do indywidualnych wartości. Ta opcja umożliwia zawijanie paneli, jeżeli zachodzi taka potrzeba.

Transformowanie układu współrzędnych

Wiele wizualizacji jest wyświetlanych na płaskim, prostokątnym układzie współrzędnych. Można transformować układ współrzędnych, jeżeli zachodzi taka potrzeba. Można na przykład zastosować do układu współrzędnych transformację biegunową, dodać efekt cienia rzutu ukośnego i transponować osie. Można również cofnąć wszystkie te transformacje, jeżeli są one już zastosowane do bieżącej wizualizacji. Na przykład wówczas, gdy wykres kołowy narysowany jest w układzie współrzędnych biegunowych. Jeśli zachodzi taka potrzeba, można cofnąć transformację biegunową i wyświetlić wykres kołowy w postaci pojedynczego skumulowanego słupka, narysowanego we współrzędnych prostokątnych.

Transformowanie układu współrzędnych

1. Zaznacz układ współrzędny, który ma zostać przetransformowany. Wybór układu współrzędnych odbywa się przez zaznaczenie ramki wokół pojedynczego wykresu.
2. Kliknij zakładkę palety właściwości **Współrzędne**.
3. Zaznacz transformację, która ma zostać zastosowana względem układu współrzędnych. Możesz również odznaczyć transformację, aby cofnąć jej działanie.

   Transponowane. Zmiana orientacji osi nosi nazwę **transponowania**. Jest to operacja podobna do zamiany miejscami osi poziomej i pionowej, dokonywanej w wizualizacji 2-W.

   Biegunowa. Transformacja biegunowa rysuje elementy graficzne pod określonym kątem i w określonej odległości od środka wykresu. Wykres kołowy jest jednowymiarową wizualizacją z transformacją biegunową, która umożliwia rysowanie poszczególnych słupków pod określonymi kątami. Wykres radarowy jest wizualizacją 2-w z transformatą biegunową, która rysuje elementy graficzne o określonych kątach i odległościach od środka wykresu. Wizualizacja 3-W może dodatkowo obejmować głębokość.

   Ukośna. Transformata ukośna dodaje efekt 3-W do elementów graficznych. Ta transformacja dodaje głębię do elementów graficznych, ale jest to efekt czysto dekoracyjny. Nie jest on zależny od żadnej konkretnej wartości danych.

   Taka sama proporcja. Zastosowanie takiej samej proporcji określa, że ta sama odległość na każdej skali odpowiada takiej samej różnicy wartości. Na przykład 2 cm na obu skalach reprezentuje różnicę 1000.

   % odstępu przed transformacją. Jeśli po transformacji osie są skrócone, może zajść potrzeba dodania do wykresu odstępów, zanim zastosuje się transformację. Odstępy skracają wymiary o określony procent przed zastosowaniem względem układu współrzędnych jakichkolwiek transformacji. Możliwe jest kontrolowanie dolnego wymiaru *x*, górnego wymiaru *x*, dolnego wymiaru *y* i górnego wymiaru *y* (w takiej kolejności).

   % odstępu po transformacji. Jeśli zachodzi potrzeba zmiany współczynnika proporcji wykresu, po zastosowaniu transformacji można dodać do wykresu odstępy. Odstępy skracają wymiary o określony procent po zastosowaniu względem układu współrzędnych jakichkolwiek transformacji. Odstępy te mogą być również zastosowane nawet wtedy, gdy względem wykresu nie zastosowano żadnych transformacji. Możliwe jest kontrolowanie dolnego wymiaru *x*, górnego wymiaru *x*, dolnego wymiaru *y* i górnego wymiaru *y* (w takiej kolejności).

Zmiana statystyk i elementów graficznych

Można dokonać konwersję do innego typu, zmienić statystykę używaną do wyrysowania elementu graficznego lub określić modyfikator konfliktu, który decyduje o tym, co się stanie, kiedy elementy graficzne będą na siebie nachodzić.

Konwersja elementów graficznych

1. Zaznacz element graficzny, których chcesz przekonwertować.
2. Kliknij zakładkę palety właściwości **Element**.
3. Wybierz typ elementu graficznego z listy Typ.

Tabela 40. Typy elementów graficznych

Typ elementu graficznego	Opis
Punkt	Znacznik identyfikujący określony punkt danych. Punkt jest używany na wykresach rozrzutu i w innych powiązanych wizualizacjach.
Przedziałowe	Prostokątny kształt rysowany dla określonej wartości danych, który wypełnia przestrzeń między wartością początkową i inną wartością danych. Interwał jest używany na wykresach słupkowych i histogramach.
Wykres liniowy	Linia, która łączy wartości liczbowe.
Ścieżkowy	Linia, która łączy wartości danych w kolejności, w jakiej występują one w zbiorze danych.
Obszar	Linia, która łączy elementy danych z obszarem pomiędzy linią i wypełnioną wartością początkową.
Wielokąt	Kształt wieloboku zamykający okolicę danych. Element ten może być wykorzystany na wykresie rozrzutu z kategoryzacją lub na mapie.
Schemat	Element zawierający pole z wąsami i znacznikami wskazującymi obserwacje odstające. Element schematu jest używany na wykresach skrzynkowych.

Zmiana statystyki

1. Zaznacz element graficzny, którego statystykę chcesz zmienić.
2. Kliknij zakładkę palety właściwości **Element**.

Określanie modyfikatora konfliktu

Modyfikator konfliktu określa, co się dzieje, gdy elementy graficzne na siebie nachodzą.

1. Wybierz element graficzny, dla którego chcesz określić modyfikator konfliktu.
2. Kliknij zakładkę palety właściwości **Element**.
3. Z listy rozwijanej Modyfikator wybierz modyfikator konfliktu. Opcja **-auto-** umożliwia aplikacji określenie, który modyfikator konfliktu jest odpowiedni dla danego elementu graficznego i statystyki.

   Nakładany. Rysowanie elementów graficznych o tej samej wartości tak, aby jeden znajdował się nad drugim.

   Zestawianie. Zestawia elementy graficzne, które normalnie zostałyby na siebie nałożone, jeżeli mają takie same wartości liczbowe.

   Unikanie. Przemieszczanie elementów graficznych obok innych elementów graficznych, które pojawiają się na tej samej wartości, zamiast nakładania ich na siebie. Elementy graficzne są rozmieszczone w sposób symetryczny. Oznacza to, że elementy graficzne są przeniesione do pozycji naprzeciwko siebie względem punktu centralnego. Unikanie jest bardzo podobne do grupowania.

   Sterta. Przemieszczanie elementów graficznych obok innych elementów graficznych, które pojawiają się na tej samej wartości, zamiast nakładania ich na siebie. Elementy graficzne są rozmieszczone w sposób asymetryczny. Oznacza to, że elementy graficzne są umieszczone w stercie jeden na drugim, przy czym element graficzny umieszczony na spodzie zajmuje pozycję w określonym miejscu skali.

   Rozrzut (normalny). Ponowne, losowe rozmieszczanie elementów graficznych o takiej samej wartości danych przy użyciu rozkładu normalnego.

   Rozrzut (stały). Ponowne, losowe rozmieszczanie elementów graficznych o takiej samej wartości danych przy użyciu rozkładu stałego.

Zmiana położenia legendy

Jeśli wykres obejmuje legendę, jest ona zazwyczaj wyświetlana po prawej stronie wykresu. W razie potrzeby można zmienić jej pozycję.

Zmiana pozycji legendy

1. Zaznacz legendę.
2. Kliknij zakładkę palety właściwości **Legenda**.
3. Wybierz położenie legendy.

Kopiowanie wizualizacji i danych wizualizacji

Paleta Ogólne zawiera przyciski służące do kopiowania wizualizacji i jej danych.

Rysunek 72. Przycisk Kopiuj wizualizację

Kopiowanie wizualizacji. Czynność ta spowoduje skopiowanie do schowka wizualizacji w postaci obrazu. Dostępnych jest wiele formatów obrazu. Kiedy obraz jest wklejany do innej aplikacji, można wybrać opcję „wklej specjalnie”, aby wybrać jeden z dostępnych formatów obrazu do wklejenia.

Rysunek 73. Przycisk Kopiuj dane wizualizacji

Kopiowanie danych wizualizacji. Czynność ta powoduje skopiowanie ukrytych danych, które są używane do wyrysowania wizualizacji. Dane są kopiowane do schowka w postaci zwykłego tekstu lub tekstu sformatowanego zgodnie ze standardem HTML. Kiedy dane są wklejane do innej aplikacji, można wybrać opcję „wklej specjalnie”, aby wybrać jeden z dostępnych formatów tekstu do wklejenia.

Skróty klawiaturowe edytora wizualizacji

Tabela 41. Skróty klawiaturowe

Klawisz skrótu	Funkcja
Ctrl+Spacja	Przełącza między trybem eksploracji i edycji
Usuń	Kasuje element wizualizacji
Ctrl+Z	Cofnij
Ctrl+Y	Powtórz
F2	Wyświetla na wykresie zarys wybranych elementów

Dodawanie tytułów i stopek

Dla wszystkich typów wykresów można dodać unikalny tytuł, przypis lub etykiety osi w celu identyfikacji elementów przedstawianych na wykresie.

Dodawanie tytułów do wykresów

1. Z menu wybierz opcje **Edytuj > Dodaj tytuł wykresu**. Ponad wykresem zostanie wyświetlone pole tekstowe zawierające opcję **<TYTUŁ>**.

2. Sprawdź, czy aktywny jest tryb edycji. Z menu wybierz opcję **Widok> Tryb edycji**.
3. Kliknij dwukrotnie tekst **<TYTUŁ>**.
4. Wpisz żądany tytuł i naciśnij klawisz Return.

Dodawanie przypisów do wykresów

1. Z menu wybierz opcje **Edytuj > Dodaj przypis do wykresu**. Pod wykresem zostanie wyświetlone pole tekstowe **<PRZYPIS>**.
2. Sprawdź, czy aktywny jest tryb edycji. Z menu wybierz opcję **Widok> Tryb edycji**.
3. Kliknij dwukrotnie tekst **<PRZYPIS>**.
4. Wpisz żądany tytuł i naciśnij klawisz Return.

Używanie arkuszy stylów

Podstawowe informacje dotyczące wyświetlania wykresu, takie jak kolory, czcionki, symbole i grubości linii, znajdują się w arkuszu stylów. Wraz z programem IBM SPSS Modeler dostarczony jest domyślny arkusz stylów; można jednak w razie potrzeby wprowadzić w nim zmiany. Może na przykład okazać się potrzebny firmowy schemat kolorów do użycia na wykresach, zgodny ze schematem kolorów wykorzystywanym w firmowych prezentacjach. Więcej informacji można znaleźć w temacie “Edycja wizualizacji” na stronie 285.

W węzłach wykresów można użyć trybu edycji w celu wprowadzenia zmian w stylach definiujących wygląd wykresu. Następnie, korzystając z menu **Edycja > Style**, można zapisać zmiany jako arkusz stylów, który będzie obowiązywał do wszystkich wykresów generowanych później z bieżącego węzła wykresu albo stanie się nowym domyślnym arkuszem stylów dla wszystkich wykresów generowanych za pomocą IBM SPSS Modeler.

W opcji **Style** w menu Edycja dostępnych jest pięć opcji arkuszy stylów:

- **Przełącz arkusz stylów.** Wyświetla listę różnych zapisanych arkuszy stylów, które można zmienić w celu zmiany wyglądu wykresów. Więcej informacji można znaleźć w temacie “Stosowanie arkuszy stylów”.
- **Zapisz style w węźle.** Umożliwia zapisanie modyfikacji w wybranych stylach wykresów w celu ich zastosowania do wszelkich przyszłych wykresów utworzonych na podstawie tego samego węzła wykresu w bieżącym strumieniu.
- **Zapisz style jako domyślne.** Umożliwia zapisanie modyfikacji w wybranych stylach wykresów w celu ich zastosowania do wszelkich przyszłych wykresów tworzonych dla dowolnego węzła wykresu w dowolnym strumieniu. Po wybraniu tej opcji można użyć opcji **Zastosuj style domyślne** w celu zmiany wszelkich pozostałych istniejących wykresów tak, aby używały tych samych stylów.
- **Zastosuj style domyślne.** Ta opcja powoduje zmianę wybranych stylów wykresów na obecnie zapisane jako style domyślne.
- **Zastosuj style oryginalne.** Ta opcja powoduje zmianę stylów wykresu z powrotem na oryginalnie dostarczone style domyślne.

Stosowanie arkuszy stylów

Możliwe jest zastosowanie arkusza stylów wizualizacji, który określa właściwości stylistyczne wizualizacji. Na przykład arkusz stylów może definiować między innymi czcionki, styl tekstu i kolory. W pewnym zakresie arkusze stylów skracają edycję, którą należałoby wykonać ręcznie. Należy jednak pamiętać, że arkusz stylów jest ograniczony do zmian *stylu*. Inne zmiany, takie jak położenie legendy lub skala, nie są zapisywane na arkuszu stylów.

Stosowanie arkusza stylów

1. Z menu wybierz:

   Edytuj > Style > Przełącz arkusz stylów
2. Użyj okna dialogowego Przełącz arkusz stylów do wybrania arkusza stylów.
3. Kliknij przycisk **Zastosuj**, aby zastosować arkusz stylów do wizualizacji bez zamykania okna dialogowego. Kliknij przycisk **OK**, aby zastosować arkusz stylów i zamknąć okno dialogowe.

Okno dialogowe Przełącz/wybierz arkusz stylów

W tabeli w górnej części okna dialogowego wymienione są wszystkie dostępne obecnie arkusze stylów wizualizacji. Niektóre arkusze stylów są zainstalowane wstępnie, inne zaś zostały utworzone w programieIBM SPSS Visualization Designer (odrębnym produkcie).

W dolnej części okna dialogowego znajdują się przykładowe wizualizacje z przykładowymi danymi. Wybierz jeden z arkuszy stylów, aby zastosować jego style do przykładowych wizualizacji. Te przykłady pomogą określić, w jaki sposób arkusz stylów wpłynie na rzeczywistą wizualizację.

Okno dialogowe udostępnia także następujące opcje.

Istniejące style. Domyślnie arkusz stylów zastępuje wszystkie style w wizualizacji. To działanie można zmienić.

- **Zastępuj wszystkie style**. Po zastosowaniu arkusza stylów wszystkie style w wizualizacji zostają zastąpione. Dotyczy to także stylów zmodyfikowanych podczas bieżącej sesji edycji.
- **Zachowaj zmodyfikowane style**. Po zastosowaniu arkusza stylów zastąpione zostaną tylko style, których *nie* zmodyfikowano podczas bieżącej sesji edycji. Style zmodyfikowane podczas bieżącej sesji edycji zostają zachowane.

Zarządzaj. Zarządzaj szablonami wizualizacji, arkuszami stylów i mapami na swoim komputerze. Możesz importować, eksportować, zmieniać nazwy i usuwać szablony wizualizacji, arkusze stylów i mapy na swoim komputerze lokalnym. Więcej informacji można znaleźć w temacie “Zarządzanie plikami szablonów, arkuszy stylów oraz map” na stronie 218.

Położenie. Zmień miejsce zapisu szablonów wizualizacji, arkuszy stylów i map. Bieżąca lokalizacja podana jest po prawej stronie przycisku. Więcej informacji można znaleźć w temacie “Ustawianie lokalizacji szablonów, arkuszy stylów i map” na stronie 217.

Drukowanie, zapisywanie, kopiowanie i eksportowanie wykresów

Każdy wykres zawiera szereg opcji umożliwiających zapisywanie lub drukowanie wykresu lub eksportowanie go do innego formatu. Większość z tych opcji jest dostępna z menu Plik. Ponadto z menu Edycja można wybrać opcję kopiowania wykresu, zawartych w nim danych lub obiekt rysunkowy pakietu Microsoft Office na potrzeby innej aplikacji.

Drukowanie

W celu wydrukowania wykresu użyj polecenia menu lub przycisku **Drukuj**. Przed wydrukowaniem możesz użyć opcji **Ustawienia strony** i **Podgląd wydruku** w celu ustawienia opcji wydruku i podglądu wyników. .

Zapisywanie wykresów

W celu zapisania wykresu w pliku wynikowym IBM SPSS Modeler (*.cou) wybierz opcje **Plik > Zapisz** lub **Plik > Zapisz jako** z menu.

lub

Aby zapisać wykres w repozytorium, wybierz opcje **Plik > Zapisz dane wyjściowe** z menu.

Kopiowanie wykresów

Aby skopiować wykres w celu użycia go w innej aplikacji, takiej jak MS Word lubMS PowerPoint, wybierz opcje **Edytuj > Kopiuj wykres** z menu.

Kopiowanie danych

Aby skopiować dane w celu użycia ich w innej aplikacji, takiej jak MS Excel lub MS Word, wybierz opcje **Edytuj > Kopiuj dane** z menu. Domyślnie dane zostaną sformatowane w języku HTML. Użyj opcji **Wklej specjalnie** w innej

aplikacji, aby wyświetlić inne opcje formatowania podczas wklejania.

Kopiowanie obiektu graficznego pakietu Microsoft Office

Istnieje możliwość skopiowania wykresu jako obiektu graficznego pakietu Microsoft Office i użycia go w aplikacjach Microsoft Office, takich jak Excel lub PowerPoint. Aby skopiować wykres, należy wybrać z menu opcje **Edytuj > Kopiuj obiekt graficzny pakietu Microsoft Office**. Zawartość zostanie skopiowana do schowka i domyślnie zapisana zostanie w formacie binarnym. Użyj opcji **Wklej specjalnie** w aplikacji Microsoft Office, aby określić inne opcje formatowania podczas wklejania.

Należy pamiętać, że niektóre treści mogą nie obsługiwać tej funkcji; w takiej sytuacji opcja menu **Kopiuj obiekt graficzny pakietu Microsoft Office** będzie wyłączona. Należy również pamiętać, że wygląd wykresu może zmienić się po wklejeniu do aplikacji Office, ale dane wykresu pozostaną niezmienione.

Istnieje sześć typów wyników graficznych, jakie można skopiować i wkleić do arkusza Excel: Prosty słupkowy, Słupkowy skumulowany, Prosty wykres skrzynkowy, Wykres skrzynkowy grupowany, Prosty rozrzutu i Grupowany rozrzutu. W przypadku użycia opcji Panel i Animacja dla jednego z tych typów wykresów opcja **Kopiuj obiekt graficzny pakietu Microsoft Office** zostanie wyłączona w programie SPSS Modeler. W przypadku innych ustawień, takich jak Opcjonalne sposoby prezentacji lub Nakładanie, opcja jest częściowo obsługiwana. Szczegóły przedstawiono w poniższej tabeli:

Tabela 42. Obsługa funkcji kopiowania obiektu graficznego pakietu Microsoft Office

Szablon wyniku graficznego	Węzeł wykresu w programie Modeler	Typ wykresu w programie Modeler	Ustawienie podstawowe	Opcjonalne sposoby prezentacji	Nałożenie	Obsługa funkcji kopiowania obiektu graficznego pakietu Microsoft Office	Komentarze
Prosty słupkowy	Graphboard	Słupkowy	Tak	Nie	Nie dotyczy	Tak	
		Słupkowy liczebności	Tak	Nie	Nie dotyczy	Tak	
	Rozkład	Słupkowy	Tak	Nie dotyczy	Nie	Tak	
Słupkowy skumulowany	Graphboard	Słupkowy	Tak	Tak	Nie dotyczy	Tak z ograniczeniami	Tak tylko dla zmiennej jakościowej dla opcjonalnych sposobów prezentacji.
		Słupkowy liczebności	Tak	Tak	Nie dotyczy	Tak z ograniczeniami	Tak tylko dla zmiennej jakościowej dla opcjonalnych sposobów prezentacji.
	Rozkład	Słupkowy	Tak	Nie dotyczy	Tak	Tak	
Wykres skrzynkowy	Graphboard	Wykres skrzynkowy	Tak	Nie	Nie dotyczy	Tak z ograniczeniami	Tak tylko dla systemu Windows.
		Wykres skrzynkowy	Tak	Tak	Nie dotyczy	Nie	

Tabela 42. Obsługa funkcji kopiowania obiektu graficznego pakietu Microsoft Office (kontynuacja)

Szablon wyniku graficznego	Węzeł wykresu w programie Modeler	Typ wykresu w programie Modeler	Ustawienie podstawowe	Opcjonalne sposoby prezentacji	Nałożenie	Obsługa funkcji kopiowania obiektu graficznego pakietu Microsoft Office	Komentarze
Wykres skrzynkowy zgrupowany	Graphboard	Wykres skrzynkowy zgrupowany	Tak	Nie	Nie dotyczy	Tak z ograniczeniami	Tak tylko dla systemu Windows.
		Wykres skrzynkowy zgrupowany	Tak	Tak	Nie dotyczy	Nie	
Prosty rozrzutu	Graphboard	Wykres bąbelkowy	Tak	Nie	Nie dotyczy	Tak z ograniczeniami	Tak tylko dla zmiennych ilościowych X i Y oraz zmiennej jakościowej rozmiaru.
		Wykres rozrzutu	Tak	Nie	Nie dotyczy	Tak z ograniczeniami	Tak tylko dla zmiennych ilościowych X i Y.
	Wykresy	Punkt	Tak	Nie dotyczy	Nie	Tak z ograniczeniami	Tak tylko dla zmiennych ilościowych X i Y.
Grupowany rozrzutu	Graphboard	Wykres bąbelkowy	Tak	Tak	Nie dotyczy	Nie	
		Wykres rozrzutu	Tak	Tak	Nie dotyczy	Tak z ograniczeniami	Tak tylko dla zmiennych ilościowych X i Y oraz zmiennych jakościowych dla opcjonalnych sposobów prezentacji.
	Wykresy	Punkt	Tak	Nie dotyczy	Tak	Tak z ograniczeniami	Tak tylko dla zmiennych ilościowych X i Y oraz zmiennej jakościowej w opcji Nakładanie.

Eksportowanie wykresów

Opcja **Eksportuj wykres** umożliwia wyeksportowanie wykresu w jednym z następujących formatów: jako plik bitmapy (.bmp), JPEG (.jpg), PNG (.png), HTML (.html), PDF (.pdf) lub dokument ViZml (.xml) na potrzeby użycia w innych aplikacjach.

Uwaga: Jeśli wybrana zostanie opcja PDF, wykresy są eksportowane jako pliki PDF o wysokiej rozdzielczości, które są obcinane do wielkości grafiki.

W celu wyeksportowania wykresów wybierz opcje **Plik > Eksportuj wykres** z menu, a następnie wybierz format.

Eksportowanie tabel

Opcja **Eksportuj tabele** umożliwia eksportowanie tabel w jednym z następujących formatów: jako plik rozdzielany znakami tabulacji (.tab), rozdzielany przecinkami (.csv) lub jako HTML (.html).

W celu wyeksportowania tabel wybierz opcje **Plik > Eksportuj tabelę** z menu, a następnie wybierz format.

Rozdział 6. Węzły wyników

Przegląd węzłów wyników

Węzły wyników umożliwiają uzyskanie informacji na temat danych i modeli. Udostępniają również mechanizm do eksportowania danych w różnych formatach, tak aby były dostępne za pomocą innych narzędzi.

Dostępne są następujące węzły wyników:

Węzeł Tabela wyświetla dane w formacie tabeli, którą można również zapisać jako plik. Jest to pomocne, kiedy konieczne jest sprawdzenie wartości danych lub wyeksportowanie ich w czytelnej postaci.

Węzeł Macierz tworzy tabelę przedstawiającą relacje pomiędzy zmiennymi. Najczęściej służy do przedstawienia relacji pomiędzy dwiema zmiennymi symbolicznymi, ale również do zaprezentowania relacji pomiędzy zmiennymi flagi lub zmiennymi numerycznymi.

Węzeł Analiza ocenia zdolność modeli predykcyjnych do wygenerowania dokładnych predykcji. Węzły analizy przeprowadzają różne porównania pomiędzy wartościami przewidywanymi a rzeczywistymi dla co najmniej jednego modelu użytkowego. Mogą również porównywać modele predykcyjne pomiędzy sobą.

Węzeł Audyt danych umożliwia kompleksowe spojrzenie na dane, udostępniając statystyki podsumowujące, histogramy i rozkład dla każdej zmiennej, jak również informacje o wartościach odstających, brakujących wartościach i wartościach skrajnych. Wyniki są wyświetlane w postaci czytelnej macierzy, która może zostać posortowana i użyta do wygenerowania wykresów w pełnym rozmiarze oraz węzłów przygotowania danych.

Węzeł Transformacja umożliwia wybór i podgląd wyników transformacji przed zastosowaniem ich w wybranych zmiennych.

Węzeł Statystyki udostępnia informacje podsumowujące na temat zmiennych numerycznych. Oblicza statystyki podsumowujące dla poszczególnych zmiennych oraz korelacje pomiędzy zmiennymi.

Węzeł Średnie porównuje średnie między niezależnymi grupami lub między parami powiązanych zmiennych w celu przetestowania, czy istnieje dla nich znaczna różnica. Na przykład można porównać średnie przychody przed uruchomieniem promocji i po jej zakończeniu lub porównać przychody od klientów, którzy nie otrzymali oferty promocyjnej z tymi, którzy z niej skorzystali.

Węzeł Raport tworzy sformatowane raporty zawierające stały tekst oraz dane i inne wyrażenia wydzielone z danych. Można określić format raportu, korzystając z szablonów tekstowych do zdefiniowania konstrukcji tekstu stałego i danych wyjściowych. Możliwe jest zastosowanie niestandardowego formatowania tekstu poprzez wprowadzenie do szablonu znaczników HTML i ustawienie opcji na karcie Wynik. Wartości danych i inne warunkowe wartości wyjściowe można dołączyć do szablonu za pośrednictwem wyrażeń CLEM.

Węzeł wartości globalnych skanuje dane i oblicza wartości sumaryczne, które mogą zostać użyte w wyrażeniach CLEM. Na przykład można użyć tego węzła do obliczenia statystyk dla zmiennej o nazwie *age* (wiek), a następnie użyć ogólnej średniej dla wartości *age* w wyrażeniach CLEM, wstawiając funkcję @GLOBAL_MEAN(age).

Węzeł Symulacje Dopasowanie sprawdza rozkład statystyczny danych w każdej zmiennej i generuje (lub aktualizuje) węzeł Symulacje Generowanie, stosując rozkład o najlepszym dopasowaniu przypisany do każdej zmiennej. Węzeł Symulacje Generowanie może być wówczas użyty do wygenerowania danych objętych symulacją.

Węzeł Symulacje Wynik przeprowadza ewaluację określonej predykcyjnej zmiennej przewidywanej i prezentuje informacje na temat rozkładu i korelacji zmiennej przewidywanej.

Zarządzanie wynikami

Menedżer wyników wyświetla wykresy i tabele wygenerowane w czasie trwania sesji IBM SPSS Modeler. Każdy wynik można w dowolnej chwili ponownie wyświetlić, klikając go dwukrotnie w menedżerze — bez konieczności uruchamiania odpowiedniego strumienia lub węzła.

Aby wyświetlić menedżera wyników

Należy otworzyć menu Widok i wybrać opcję **Zarządzanie**. Następnie należy kliknąć zakładkę **Wyniki**.

Za pośrednictwem menedżera wyników można:

- Wyświetlić istniejące obiekty wynikowe, takie jak histogramy, wykresy ewaluacyjne i tabele.
- Zmieniać nazwy obiektów wynikowych.
- Zapisywać obiekty wynikowe na dysku lub w repozytorium IBM SPSS Collaboration and Deployment Services Repository (o ile jest dostępne).
- Dodawać pliki wynikowe do bieżącego projektu.
- Usuwać niezapisane obiekty wynikowe z bieżącej sesji.
- Otwierać zapisane obiekty wynikowe lub odtwarzać je z repozytorium IBM SPSS Collaboration and Deployment Services Repository (o ile jest dostępne).

Aby uzyskać dostęp do tych opcji, należy kliknąć prawym przyciskiem myszy w dowolnym miejscu karty wyników.

Wyświetlanie wyników

Wynik jest wyświetlany w oknie przeglądarki wyników. Okno przeglądarki wyników zawiera własny zestaw menu, które umożliwiają drukowanie lub zapisywanie wyniku lub eksportowanie go do innego formatu. Należy pamiętać, że konkretne opcje mogą różnić się w zależności od rodzaju wyniku.

Drukowanie, zapisywanie i eksportowanie danych. Poniżej przedstawiono dodatkowe informacje:

- Aby wydrukować wynik, należy użyć opcji menu lub przycisku **Drukuj**. Przed wydrukowaniem możesz użyć opcji **Ustawienia strony** i **Podgląd wydruku** w celu ustawienia opcji wydruku i podglądu wyników.
- Aby zapisać wynik w pliku wyników programu IBM SPSS Modeler (.cou), należy wybrać opcję **Zapisz** lub **Zapisz jako** z menu Plik.
- Aby zapisać wynik w innym formacie, np. w formacie tekstowym lub HTML, należy wybrać opcję **Eksportuj** z menu Plik. Więcej informacji można znaleźć w temacie “Eksportowanie wyników” na stronie 305.

  Należy pamiętać, że formaty te można wybrać tylko wtedy, gdy wynik zawiera dane, które można w sposób czytelny wyeksportować w dany sposób. Przykładowo, zawartość drzewa decyzyjnego można wyeksportować jako tekst, ale zawartość modelu K-średnich nie będzie miała sensu w postaci tekstowej.
- Aby zapisać wynik we współużytkowanym repozytorium, tak aby inni użytkownicy mogli wyświetlać go za pomocą programu IBM SPSS Collaboration and Deployment Services Deployment Portal, należy wybrać opcję **Publikuj w sieci WWW** z menu Plik. Należy pamiętać, że ta opcja wymaga osobnej licencji programu IBM SPSS Collaboration and Deployment Services.

Wybór komórek i kolumn. Menu Edycja zawiera różne opcje wyboru, usuwania zaznaczenia i kopiowania komórek i kolumn, odpowiednie dla bieżącego typu wyniku. Więcej informacji można znaleźć w “Wybór komórek i kolumn.” na stronie 305.

Generowanie nowych węzłów. Menu Utwórz umożliwia wygenerowanie nowych węzłów na podstawie zawartości przeglądarki wyników. Opcje mogą się różnić w zależności od typu wyniku i elementów wyniku, jakie są aktualnie wybrane. Szczegółowe informacje na temat opcji generowania węzłów dla poszczególnych typów wyników zawiera dokumentacja dotycząca tego wyniku.

Publikowanie w sieci WWW

Funkcja Publikuj w sieci WWW umożliwia publikowanie określonych typów wyników strumienia w centralnym współużytkowanym repozytorium IBM SPSS Collaboration and Deployment Services Repository, które jest elementem podstawowym usług IBM SPSS Collaboration and Deployment Services. Po użyciu tej opcji inni użytkownicy, którzy muszą wyświetlić wynik, będą mogli to zrobić, korzystając z dostępu do Internetu i konta IBM SPSS Collaboration and Deployment Services — program IBM SPSS Modeler nie musi być zainstalowany.

W poniższej tabeli przedstawiono listę węzłów IBM SPSS Modeler, które obsługują funkcję Publikuj w sieci WWW. Wyniki z tych węzłów są zapisywane w repozytorium IBM SPSS Collaboration and Deployment Services Repository w formacie obiektu wynikowego (*.cou*) i można je wyświetlić bezpośrednio w portalu IBM SPSS Collaboration and Deployment Services Deployment Portal.

Pozostałe typy wyników można wyświetlać tylko po zainstalowaniu na komputerze użytkownika odpowiedniej aplikacji (np. IBM SPSS Modeler dla obiektów strumienia).

Tabela 43. Węzły obsługujące funkcję Publikuj w sieci WWW

Typ węzła	Węzeł
Wykresy	Wszystkie
Wynik	Tabela
	Macierz
	Audyt danych
	Transformacja

Tabela 43. Węzły obsługujące funkcję Publikuj w sieci WWW (kontynuacja)

Typ węzła	Węzeł
	Średnie
	Analiza
	Statystyki
	Raport (HTML)
IBM SPSS Statistics	Wynik Statistics

Publikowanie wyników w sieci WWW

Aby opublikować wynik w sieci WWW:

1. W strumieniu IBM SPSS Modeler wykonaj jeden z węzłów wymienionych w tabeli. W wyniku tego utworzony zostanie obiekt wynikowy (na przykład tabela, macierz lub raport) w nowym oknie.
2. W oknie obiektu wynikowego wybierz:

   Plik > Publikuj w sieci WWW

   Uwaga: Jeśli mają zostać wyeksportowane proste pliki HTML, które będą używane za pośrednictwem standardowej przeglądarki WWW, wybierz opcję **Eksportuj** z menu Plik, a następnie opcję **HTML**.
3. Połącz się z repozytorium IBM SPSS Collaboration and Deployment Services Repository.

   Po pomyślnym nawiązaniu połączenia wyświetlane jest okno dialogowe Repozytorium: Składuj, udostępniające wiele opcji składowania.
4. Po wybraniu opcji składowania kliknij przycisk **Składuj**.

Wyświetlanie opublikowanych wyników w sieci WWW

Do korzystania z tej funkcji wymagane jest skonfigurowane konto IBM SPSS Collaboration and Deployment Services. Jeśli na komputerze zainstalowana została odpowiednia aplikacja dla typu obiektu, jaki ma zostać wyświetlony (na przykład IBM SPSS Modeler lub IBM SPSS Statistics), wynik będzie raczej wyświetlany w samej aplikacji, a nie w przeglądarce.

Aby wyświetlić opublikowane wyniki w sieci WWW:

1. Ustaw w przeglądarce: *http://<repos_host>:<repos_port>/peb*

   gdzie *repos_host* i *repos_port* oznaczają nazwę hosta i numer portu dla hosta IBM SPSS Collaboration and Deployment Services.
2. Wprowadź szczegóły logowania do konta IBM SPSS Collaboration and Deployment Services.
3. Kliknij opcję **Repozytorium treści**.
4. Przejdź do obiektu, który chcesz wyświetlić lub go wyszukaj.
5. Kliknij nazwę obiektu. W przypadku niektórych obiektów, takich jak wykresy, renderowanie obiektu w przeglądarce może spowodować opóźnienie.

Wyświetlanie wyników w przeglądarce HTML

Korzystając z karty Zaawansowane modeli użytkowych Liniowy, Regresja logistyczna i Redukcja wymiarów, można wyświetlać informacje w osobnej przeglądarce, takiej jak Internet Explorer. Informacje wyświetlane są w postaci pliku HTML, co umożliwia zapisanie ich i ponowne użycie w dowolnym miejscu, np. w sieci intranet przedsiębiorstwa lub w witrynie internetowej.

Aby wyświetlić informacje w przeglądarce, należy kliknąć przycisk uruchamiania znajdujący się pod ikoną modelu po lewej stronie karty Zaawansowane w modelu użytkowym.

Eksportowanie wyników

W oknie przeglądarki wyników można wybrać opcję wyeksportowania wyniku w innym formacie, np. tekstowym lub HTML. Formaty eksportu różnią się w zależności od typu wyniku, ale ogólnie są podobne do opcji typów plików dostępnych po wybraniu opcji **Zapisz do pliku** w węźle użytym do wygenerowania wyniku.

Uwaga: Formaty te można wybrać tylko wtedy, gdy wynik zawiera dane, które można w sposób czytelny wyeksportować w dany sposób. Przykładowo, zawartość drzewa decyzyjnego można wyeksportować jako tekst, ale zawartość modelu K-średnich nie będzie miała sensu w postaci tekstowej.

Aby wyeksportować wynik:

1. W przeglądarce wyników otwórz menu Plik i wybierz opcję **Eksportuj**. Następnie wybierz typ pliku, jaki zamierzasz utworzyć:
   - **Oddzielone tabulacją (*.tab).** Ta opcja powoduje wygenerowanie sformatowanego pliku tekstowego zawierającego wartości danych. Ten styl jest często przydatny w przypadku generowania informacji w postaci zwykłego tekstu, który można później zaimportować do innych aplikacji. Ta opcja jest dostępna dla węzłów Tabela, Macierz i Średnie.
   - **Oddzielone przecinkami (*.dat).** Ta opcja powoduje wygenerowanie pliku tekstowego rozdzielanego przecinkami zawierającego wartości danych. Ten styl jest często przydatny jako szybki sposób wygenerowania pliku danych, który można zaimportować do arkuszy kalkulacyjnych lub innych aplikacji do analizy danych. Ta opcja jest dostępna dla węzłów Tabela, Macierz i Średnie.
   - **Transponowany separowany tabulacją (*.tab).** Ta opcja jest identyczna jak opcja Oddzielone tabulacją, ale dane są transponowane w taki sposób, że wiersze reprezentują zmienne, a kolumny reprezentują rekordy.
   - **Transponowany rozdzielany przecinkami (*.dat).** Ta opcja jest identyczna jak opcja Oddzielone przecinkami, ale dane są transponowane w taki sposób, że wiersze reprezentują zmienne, a kolumny reprezentują rekordy.
   - **HTML (*.html).** Ta opcja powoduje zapisanie sformatowanego wyniku HTML w pliku lub w plikach.

Wybór komórek i kolumn.

Wiele węzłów, takich jak Tabela, Macierz i Średnie, generuje wyniki w tabelach. Te tabele wynikowe można wyświetlać i można nimi manipulować w podobny sposób; dotyczy to na przykład wyboru komórek, kopiowania całej lub części tabeli do schowka, generowania nowych węzłów na podstawie bieżącego wyboru oraz zapisywania i drukowania tabeli.

Wybór komórek. Aby wybrać komórkę, należy ją kliknąć. Aby wybrać prostokątny zakres komórek, należy kliknąć jeden narożnik wybranego zakresu, przeciągnąć mysz do drugiego narożnika zakresu i zwolnić przycisk myszy. Aby zaznaczyć całą kolumnę, należy kliknąć nagłówek kolumny. Aby zaznaczyć kilka kolumn, należy użyć metody Shift+kliknięcie lub Ctrl+kliknięcie w nagłówkach kolumn.

Po dokonaniu nowego wyboru stary wybór jest kasowany. Przytrzymując wciśnięty klawisz Ctrl podczas dokonywania wyboru można dodawać nowe elementy do już wybranych, bez usuwania wcześniejszych wyborów. Ta metoda umożliwia zaznaczenie wielu nieciągłych obszarów tabeli. Menu Edycja zawiera również opcje **Zaznacz wszystko** i **Wyczyść zaznaczenie**.

Zmiana porządku kolumn. Przeglądarki wyników węzła Tabela i węzła Średnie umożliwiają przenoszenie kolumn w tabeli poprzez kliknięcie nagłówka kolumny i przeciągnięcie go w wybrane miejsce. Kolumny można przemieszczać pojedynczo.

---

węzeł Tabela

Węzeł Tabela umożliwia utworzenie tabeli z listą wartości w danych. Uwzględniane są wszystkie zmienne i wszystkie wartości zawarte w strumieniu, dzięki czemu w prosty sposób można sprawdzać wartości danych lub eksportować je w czytelnej postaci. Opcjonalnie można wyróżnić rekordy, które spełniają określony warunek.

Uwaga: Jeśli używane zbiory danych nie są małe, zaleca się wybranie podzbioru danych, jakie zostaną przekazane do węzła Tabela. Węzeł Tabela nie będzie wyświetlany poprawnie, jeśli liczba rekordów przekracza wielkość, jaka może być zawarta w strukturze wyświetlania (np. 100 milionów wierszy).

Węzeł Tabela — karta Ustawienia

Wyróżnij rekordy spełniające warunek. Można wyróżnić rekordy w tabeli, wprowadzając wyrażenie CLEM, które będzie prawdziwe dla rekordów, które mają zostać wyróżnione. Ta opcja jest aktywna tylko po wybraniu opcji **Wynik na ekran**.

Węzeł Tabela — karta Format

Karta Format zawiera opcje używane do określenia sposobu formatowania dla zmiennej. Ta karta jest współużytkowana przez węzeł Typy. Więcej informacji można znaleźć w temacie “Karta ustawień formatu zmiennej” na stronie 150.

Węzeł Wynik — karta Wynik

W przypadku węzłów generujących wynik w postaci tabeli karta Wynik umożliwia określenie formatu i lokalizacji wyników.

Nazwa wyniku. Określa nazwę uzyskanego wyniku po wykonaniu węzła. **Automatycznie** wybiera nazwę na podstawie węzła, który spowodował wygenerowanie wyniku. Opcjonalnie można wybrać opcję **Użytkownika**, aby określić inną nazwę.

Wynik na ekran (ustawienie domyślne)**.** Tworzy obiekty wynikowe w celu ich wyświetlenia online. Obiekt wynikowy zostanie wyświetlony na kacie Wyniki w oknie menedżera po wykonaniu węzła wynikowego.

Wynik do pliku. Zapisuje wynik w pliku po wykonaniu węzła. Jeśli ta opcja zostanie wybrana, należy wprowadzić nazwę pliku (lub przejść do katalogu i określić nazwę pliku, używając selektora plików) i wybrać typ pliku. Należy pamiętać, że niektóre typy plików mogą być niedostępne dla określonych typów wyników.

Uwaga:

Dane wynikowe są zakodowane zgodnie z poniższymi regułami:

- Podczas wykonywania węzła wyników dla wyniku ustawiona zostanie wartość kodująca strumienia (ustawiona na karcie opcji strumienia).
- Po wygenerowaniu wyniku kodowanie nie zostanie zmienione, nawet jeśli kodowanie strumienia zostanie zmienione.
- Podczas eksportowania węzła wyników plik wynikowy zostaje wyeksportowany z zastosowaniem określonego kodowania strumienia. Po utworzeniu wyników zmiany w kodowaniu strumienia nie wpłyną na wygenerowane wyniki.

Należy pamiętać o następujących wyjątkach od tych reguł:

- Wszystkie wyeksportowane pliki HTML są kodowane w formacie UTF-8.
- Wynik z węzła wyników rozszerzenia jest generowany przez skrypt definiowany przez użytkownika. Dlatego o kodowaniu decyduje skrypt.

Poniżej przedstawiono opcje dostępne podczas zapisywania wyniku do pliku:

- **Dane separowane tabulacją (*.tab).** Ta opcja powoduje wygenerowanie sformatowanego pliku tekstowego zawierającego wartości danych. Ten styl jest często przydatny w przypadku generowania informacji w postaci zwykłego tekstu, który można później zaimportować do innych aplikacji. Ta opcja jest dostępna dla węzłów Tabela, Macierz i Średnie.

- **Dane separowane przecinkami (*.dat).** Ta opcja powoduje wygenerowanie pliku tekstowego rozdzielanego przecinkami zawierającego wartości danych. Ten styl jest często przydatny jako szybki sposób wygenerowania pliku danych, który można zaimportować do arkuszy kalkulacyjnych lub innych aplikacji do analizy danych. Ta opcja jest dostępna dla węzłów Tabela, Macierz i Średnie.
- **HTML (*.html).** Ta opcja powoduje zapisanie sformatowanego wyniku HTML w pliku lub w plikach. W przypadku wyników w tabelach (z węzłów Tabela, Macierz i Średnie) zestaw plików HTML zawiera nazwy zmiennych listy panelu sterowania oraz dane w postaci tabeli HTML. Tabela może być podzielona na kilka plików HTML, jeśli liczba wierszy w tabeli przekracza specyfikację **Wierszy na stronę**. W takim przypadku panel sterowania tworzy odsyłacze do wszystkich stron tabeli i umożliwia nawigowanie w tabeli. W przypadku wyników, które nie mają postaci tabeli, tworzony jest pojedynczy plik HTML, zawierający wyniki dla węzła.

  Uwaga: Jeśli wynik HTML zawiera tylko informacje dotyczące formatowania pierwszej strony, należy wybrać opcję **Podział na podstrony** i skorygować specyfikację **Wierszy na stronę**, aby uwzględnić wszystkie wyniki na jednej stronie. Lub, jeśli szablon wyniku dla węzłów, takich jak Raport, zawiera niestandardowe znaczniki HTML, należy sprawdzić, czy jako typ formatu wybrano opcję **Użytkownika**.
- **Plik tekstowy (*.txt).** Ta opcja generuje plik tekstowy zawierający wynik. Ten styl jest często przydatny podczas generowania wyniku, który można zaimportować do innych aplikacji, takich jak edytory tekstowe lub programowanie do prezentacji. Opcja ta jest niedostępna dla niektórych węzłów.
- **Obiekt wynikowy (*.cou).** Obiekty wynikowe zapisane w tym formacie można otworzyć i wyświetlić w programie IBM SPSS Modeler, dodać do projektów oraz publikować i śledzić za pomocą repozytorium IBM SPSS Collaboration and Deployment Services Repository.

Widok wyników. W przypadku węzła Średnie można określić, czy jako domyślny ma być wyświetlany wynik prosty czy zaawansowany. Należy pamiętać, że podczas przeglądania wygenerowanych wyników widoki te można zmieniać. Więcej informacji można znaleźć w temacie “Węzeł Średnie — przeglądarka wyników” na stronie 327.

Format. W przypadku węzła Raport można wybrać, czy wynik będzie formatowany automatycznie, czy też za pomocą kodu HTML uwzględnionego w szablonie. Aby zezwolić na formatowanie HTML szablonu, należy wybrać opcję **Użytkownika**.

Tytuł. Dla węzła Raport można określić opcjonalny tekst tytułu, jaki będzie wyświetlany u góry wyniku raportu.

Wyróżnij wstawiony tekst. W węźle Raport tę opcję można wybrać, aby wyróżnić tekst wygenerowany za pośrednictwem wyrażeń CLEM w szablonie raportu. Więcej informacji można znaleźć w temacie “Węzeł Raport — karta Szablon” na stronie 329. Ta opcja nie jest zalecana w przypadku użycia formatowania **Użytkownika**.

Wierszy na stronę. Dla węzła Raport należy określić liczbę wierszy, jakie będą zamieszczane na każdej stronie podczas **automatycznego** formatowania raportu wyników.

Transponuj dane. Ta opcja powoduje transponowanie danych przed ich wyeksportowaniem w taki sposób, że wiersze reprezentują zmienne, a kolumny reprezentują rekordy.

Uwaga: W przypadku dużych tabel powyższe opcje mogą być nieco niewydajne, szczególnie w przypadku pracy z użyciem serwera zdalnego. W takim przypadku znacznie lepszą wydajność zapewnia użycie węzła wynikowego Plik. Więcej informacji można znaleźć w temacie “Węzeł eksportu do pliku płaskiego” na stronie 362.

Przeglądarka tabeli

Przeglądarka tabeli wyświetla dane tabelaryczne i umożliwia wykonywanie standardowych operacji, takich jak zaznaczanie i kopiowanie komórek, zmiana porządku kolumn oraz zapisywanie i drukowanie tabeli. Więcej informacji można znaleźć w temacie “Wybór komórek i kolumn.” na stronie 305. Są to takie same operacje, jakie można wykonać podczas wyświetlania podglądu danych w węźle.

Eksportowanie danych z tabeli. Dane z przeglądarki tabeli można wyeksportować, wybierając opcje:

Plik > Eksportuj

Więcej informacji można znaleźć w temacie “Eksportowanie wyników” na stronie 305.

Dane są eksportowane w formacie kodowania domyślnym dla systemu, który jest określany w Panelu sterowania systemu Windows lub, w przypadku pracy w trybie rozproszonym, na serwerze.

Wyszukiwanie w tabeli. Przycisk wyszukiwania (z ikoną lornetki) na głównym pasku narzędzi aktywuje pasek narzędzi wyszukiwania, umożliwiając wyszukiwanie konkretnych wartości w tabeli. Wyszukiwać można do przodu i do tyłu, można wybrać opcję uwzględniania wielkości liter podczas wyszukiwania (przycisk **Aa**), a także można przerywać proces wyszukiwania, używając przycisku Przerwij wyszukiwanie.

Generowanie nowych węzłów. Menu Utwórz zawiera opcje umożliwiające generowanie węzłów.

- **Węzeł selekcji („Rekordy”).** Generuje węzeł Selekcja, który umożliwia wybór rekordów, dla których wybrane są poszczególne komórki w tabeli.
- **Węzeł selekcji („And”).** Generuje węzeł Selekcja, który umożliwia wybór rekordów zawierających *wszystkie* wartości wybrane w tabeli.
- **Węzeł selekcji („Or”).** Generuje węzeł Selekcja, który umożliwia wybór rekordów zawierających *dowolną* wartość wybraną w tabeli.
- **Węzeł wyliczeń („Rekordy”).** Generuje węzeł Wyliczanie w celu utworzenia nowej zmiennej flagi. Zmienna flagi zawiera wartość *T* (Prawda) dla rekordów, dla których zaznaczono w tabeli dowolną komórkę oraz *F* (Fałsz) dla pozostałych rekordów.
- **Węzeł wyliczeń („And”).** Generuje węzeł Wyliczanie w celu utworzenia nowej zmiennej flagi. Zmienna flagi zawiera wartość *T* (Prawda) dla rekordów, dla których *wszystkie* wartości w tabeli zostały zaznaczone oraz *F* (Fałsz) dla pozostałych rekordów.
- **Węzeł wyliczeń („Or”).** Generuje węzeł Wyliczanie w celu utworzenia nowej zmiennej flagi. Zmienna flagi zawiera wartość *T* (Prawda) dla rekordów, dla których w tabeli zaznaczono *dowolną* wartość oraz *F* (Fałsz) dla pozostałych rekordów.

---

węzeł Macierz

Węzeł Macierz umożliwia utworzenie tabeli, która przedstawia relacje pomiędzy zmiennymi. Zazwyczaj jest używany do prezentowania relacji pomiędzy dwiema zmiennymi jakościowymi (flaga, nominalna lub porządkowa), ale może również przedstawiać relacje pomiędzy zmiennymi ilościowymi (zakres liczbowy).

Węzeł Macierz — karta Ustawienia

Karta Ustawienia umożliwia określenie opcji dotyczących struktury macierzy.

Pola. Dostępne są następujące typy wyboru zmiennych:

- **Wybrane.** Ta opcja umożliwia wybranie zmiennej jakościowej dla wierszy i dla kolumn macierzy. Wiersze i kolumny macierzy są definiowane na podstawie listy wartości dla wybranej zmiennej jakościowej. W komórkach macierzy znajdują się statystyki podsumowujące wybrane poniżej.
- **Wszystkie flagi (wartości prawda).** Ta opcja tworzy żądanie utworzenia macierzy zawierającej jeden wiersz i jedną kolumnę dla każdej zmiennej flagi w danych. Komórki macierzy zawierają liczebności podwójnych wartości dodatnich dla każdej kombinacji flag. Innymi słowy, dla wiersza odpowiadającego wartości *bought bread* (kupiony chleb) i kolumny odpowiadającej wartości *bought cheese* (kupiony ser), komórka w miejscu przecięcia tego wiersza i tej kolumny zawiera liczbę rekordów, dla których wartości *bought bread* i *bought cheese* są prawdziwe.
- **Wszystkie numeryczne.** Ta opcja tworzy żądanie utworzenia macierzy zawierającej jeden wiersz i jedną kolumnę dla każdej zmiennej numerycznej. Komórki macierzy reprezentują sumę iloczynów wektorowych dla odpowiedniej pary zmiennych. Innymi słowy, dla każdej komórki w macierzy wartości dla zmiennej wiersza i zmiennej komuny są mnożone dla każdego rekordu, a następnie sumowane pomiędzy rekordami.

Uwzględnij brakujące wartości. Uwzględnia braki danych użytkownika (wartości puste) oraz systemowe braki danych ($null$) w wynikach kolumny i wiersza. Przykładowo, jeśli wartość *N/A* (N/D) została zdefiniowana jako brak danych użytkownika dla wybranej zmiennej kolumny, w tabeli uwzględniona zostanie osobna kolumna z etykietą *N/A*

(przy założeniu, że ta wartość rzeczywiście występuje w danych), tak jak każda inna kategoria. Jeśli zaznaczenie tej opcji będzie usunięte, kolumna *N/A* zostanie wykluczona, niezależnie od tego, jak często występuje.

Uwaga: Opcja uwzględniania wartości braków danych dotyczy tylko sytuacji, w której wybrane zmienne tworzą tabelę krzyżową. Puste wartości są mapowane jako $null$ i są wykluczane z agregacji dla zmiennej funkcji, jeśli jako tryb ustawiona jest opcja **Wybrane**, a zawartość jest ustawiona jako **Funkcja** oraz dla wszystkich zmiennych numerycznych, jeśli tryb jest ustawiony jako **Wszystkie numeryczne**.

Zawartość komórki. Jeśli powyżej wybrano zmienne **Wybrane**, można określić statystyki, jakie będą używane w komórkach macierzy. Należy wybrać statystyki oparte na liczebności lub wybrać zmienną nałożenia, aby zsumować wartości zmiennej numerycznej na podstawie wartości zmiennych wiersza i kolumny.

- **Tabela krzyżowa.** Wartości w komórkach są liczebnościami i/lub wartościami procentowymi określającymi, dla ilu rekordów istnieje odpowiednia kombinacja wartości. Można wskazać, które podsumowania tabeli krzyżowej są wymagane, korzystając z opcji na karcie Wygląd. Globalna wartość chi-kwadrat również jest wyświetlana, razem z istotnością. Więcej informacji można znaleźć w temacie “Przeglądarka wyników węzła Macierz” na stronie 310.
- **Funkcja.** Jeśli wybrana zostanie funkcja podsumowująca, wartości w komórkach będą funkcją wybranej zmiennej nałożenia, o ile wartości kolumn lub wierszy będą odpowiednie. Przykładowo, jeśli zmienna wiersza to *Region*, zmienna kolumny to *Product* (Produkt), a zmienna nałożenia to *Revenue* (Przychód), wówczas komórka w wierszu *Northeast* (Północny Wschód) i w kolumnie *Widget* będzie zawierać sumę (lub wartość średnią, minimalną lub maksymalną) przychodu dla widgetów sprzedanych w regionie północno-wschodnim. Domyślnie, jako funkcja podsumowująca, ustawiona jest **Średnia**. Można wybrać inną funkcję do podsumowania zmiennej funkcji. Dostępne opcje to: **Średnia**, **Suma**, **OdchStd** (odchylenie standardowe), **Maksimum** i **Minimum**.

Węzeł Macierz — karta Wygląd

Karta Wygląd umożliwia kontrolowanie opcji sortowania i wyróżniania dla macierzy, a także statystyk prezentowanych dla macierzy tabeli krzyżowych.

Wiersze i kolumny. Pozwala kontrolować sortowanie nagłówków wierszy i kolumn w macierzy. Wartość domyślna to **Nieposortowane**. Nagłówki kolumn i wierszy można sortować **Rosnąco** lub **Malejąco**.

Nakładanie. Umożliwia wyróżnienie skrajnych wartości w macierzy. Wartości są wyróżniane na podstawie liczebności komórek (w przypadku macierzy tabeli krzyżowych) lub obliczonych wartości (dla macierzy funkcji).

- **Wyróżnij górne.** Konieczne może być wyróżnienie najwyższych wartości w tabeli (na czerwono). Należy określić liczbę wartości do wyróżnienia.
- **Wyróżnij dolne.** Można również wyróżnić najniższe wartości w macierzy (na zielono). Należy określić liczbę wartości do wyróżnienia.

Uwaga: W przypadku dwóch powyższych opcji wyróżniania wiązania mogą powodować, że wyróżnionych będzie więcej wartości, niż określono. Przykładowo, jeśli dostępna jest macierz z sześcioma zerami w komórkach i ustawiona zostanie opcja **Wyróżnij dolne 5**, wyróżnionych zostanie wszystkich sześć zer.

Zawartość komórek tabeli krzyżowej. W przypadku tabel krzyżowych można określić statystyki podsumowujące zawarte w macierzy dla macierzy tabeli krzyżowej. Opcje te nie są dostępne, jeśli na karcie Ustawienia wybrana została opcja **Wszystkie numeryczne** lub **Funkcja**.

- **Liczebności.** Komórki uwzględniają liczbę rekordów zawierających wartość wiersza, której odpowiada wartość kolumny. Jest to domyślna zawartość komórki.
- **Wartości oczekiwane.** Wartość oczekiwana dla liczby rekordów w komórce, przy założeniu, że pomiędzy wierszami i kolumnami nie ma żadnych relacji. Oczekiwane wartości są wyznaczane na podstawie następującej formuły:

  p(*wartość wiersza*) * p(*wartość kolumny*) * *łączna liczba rekordów*
- **Reszty.** Różnica pomiędzy wartością obserwowaną a oczekiwaną.
- **Procent z wiersza.** Procent wszystkich rekordów z wartością wiersza, dla której istnieje odpowiednia wartość kolumny. Suma wartości procentowych w wierszach wynosi 100.

- **Procent z kolumny.** Procent wszystkich rekordów z wartością kolumny, dla której istnieje odpowiednia wartość wiersza. Suma wartości procentowych w kolumnach wynosi 100.
- **Procent z sumy.** Procent wszystkich rekordów, dla których istnieje kombinacja wartości kolumny i wartości wiersza. Suma wartości procentowych dla całej macierzy wynosi 100.
- **Dołącz podsumowania wierszy i kolumn.** Dodaje wiersz i kolumnę do macierzy dla wszystkich sum kolumny i wiersza.
- **Zastosuj ustawienia.** (Tylko przeglądarka wyników) Umożliwia wprowadzanie zmian wyglądu wyników węzła Macierz bez konieczności zamykania i ponownego otwierania przeglądarki wyników. Wystarczy wprowadzić zmiany na tej karcie przeglądarki wyników, kliknąć ten przycisk, a następnie wybrać zakładkę Macierz, aby zobaczyć efekt wprowadzonych zmian.

Przeglądarka wyników węzła Macierz

Przeglądarka wyników wyświetla dane tabeli krzyżowych i umożliwia wykonywanie operacji na macierzy, w tym zaznaczanie komórek, kopiowanie macierzy do schowka (całej lub jej części), generowanie nowych węzłów na podstawie wyboru w macierzy oraz zapisywanie i drukowanie macierzy. Przeglądarki wyników można również używać do wyświetlania wyników z określonych modeli, takich jak modele naiwnego klasyfikatora bayesowskiego (Naive Bayes) z bazy danych Oracle.

Menu Plik i Edycja zawierają standardowe opcje drukowania, zapisywania i eksportowania wyników oraz zaznaczania i kopiowania danych. Więcej informacji można znaleźć w temacie “Wyświetlanie wyników” na stronie 303.

Chi-kwadrat. W przypadku tabeli krzyżowej dwóch zmiennych jakościowych poniżej tabeli wyświetlany jest również ogólny chi-kwadrat Pearsona. Ten test wskazuje prawdopodobieństwo, że dwie zmienne są niepowiązane, w oparciu o różnicę pomiędzy liczebnościami obserwowanymi i liczebnościami, jakie byłyby oczekiwane, jeśli istniałoby powiązanie. Przykładowo, jeśli nie ma powiązania pomiędzy zadowoleniem klienta a lokalizacją sklepu, można oczekiwać, że dla wszystkich sklepów wskaźniki zadowolenia będą podobne. Jeśli jednak dla klientów niektórych sklepów stale notowane są wyższe wskaźniki niż dla pozostałych, można podejrzewać, że nie było to przypadkowe. Im większa różnica, tym mniejsze prawdopodobieństwo, że wynikało to tylko z błędu podczas próbkowania szansy.

- Test chi-kwadrat wskazuje prawdopodobieństwo, że dwie zmienne są niepowiązane, tak więc wszelkie różnice pomiędzy częstościami obserwowanymi i oczekiwanymi wynikają tylko z szansy. Jeśli to prawdopodobieństwo jest bardzo małe — zwykle mniejsze niż 5% — wówczas relacja pomiędzy dwiema zmiennymi będzie istotna.
- Jeśli dostępna jest tylko jedna kolumna lub jeden wiersz (jednokierunkowy test chi-kwadrat), stopnie swobody określa liczba komórek minus jeden. W przypadku dwukierunkowych testów chi-kwadrat stopnie swobody są wyznaczane jako liczba wierszy minus jeden pomnożone przez liczbę kolumn minus jeden.
- Jeśli oczekiwane częstości w komórce są mniejsze niż pięć, podczas interpretowania statystyki chi-kwadrat należy zachować ostrożność.
- Test chi-kwadrat jest dostępny tylko dla tabel krzyżowych dwóch zmiennych. (W przypadku wybrania na karcie Ustawienia opcji **Wszystkie flagi** lub **Wszystkie numeryczne** ten test nie jest wyświetlany).

Menu Utwórz. Menu Utwórz zawiera opcje umożliwiające generowanie węzłów. Operacje te są dostępne tylko dla macierzy tabel krzyżowych i w macierzy musi być wybrana co najmniej jedna komórka.

- **Węzeł wyboru.** Generuje węzeł Selekcja, który umożliwia wybór rekordów zgodnych z dowolną komórką zaznaczoną w macierzy.
- **Węzeł wyliczeń (flaga).** Generuje węzeł Wyliczanie w celu utworzenia nowej zmiennej flagi. Zmienna flagi zawiera wartość *T* (Prawda) dla rekordów, które są zgodne z dowolną komórką wybraną w macierzy oraz wartość *F* dla pozostałych rekordów.
- **Węzeł wyliczeń (nominalne).** Generuje węzeł Wyliczanie w celu utworzenia nowej zmiennej nominalnej. Zmienna nominalna zawiera jedną kategorię dla każdego ciągłego zestawu wybranych komórek w macierzy.

Węzeł Analiza

Węzeł analiza umożliwia określenie możliwości modelu w celu wygenerowania dokładnych predykcji. Węzły analizy przeprowadzają różne porównania pomiędzy wartościami przewidywanymi a rzeczywistymi (zmienna przewidywana) dla co najmniej jednego modelu użytkowego. Węzłów analizy można również używać do porównania modeli predykcyjnych.

Po wykonaniu węzła Analiza podsumowanie wyników analizy jest automatycznie dodawane do sekcji analizy na karcie Podsumowanie dla każdego modelu użytkowego w wykonanym strumieniu. Szczegółowe wyniki analizy są wyświetlane na karcie Wyniki w oknie menedżera lub mogą zostać bezpośrednio zapisane w pliku.

Uwaga: Ponieważ węzły analizy porównują przewidywane wartości z rzeczywistymi, są przydatne tylko w modelach nadzorowanych (czyli w takich, które wymagają zmiennej przewidywanej). W przypadku modeli nienadzorowanych, takich jak algorytmy grupowania, nie ma dostępnych wyników rzeczywistych, jakich można by było użyć do porównania.

Węzeł Analiza — karta Analiza

Karta Analiza umożliwia określenie szczegółów dotyczących analizy.

Macierz zgodności (dla przewidywanych zmiennych symbolicznych lub jakościowych). Przedstawia wzorzec dopasowań pomiędzy każdą wygenerowaną (przewidywaną) zmienną oraz jej zmienną przewidywaną dla przewidywanych zmiennych jakościowych (zmienna typu flaga, nominalna lub porządkowa). W wyświetlanej tabeli wiersze są zdefiniowane przez wartości rzeczywiste, a kolumny przez wartości przewidywane, a liczba rekordów odpowiada liczbie rekordów, w których ten wzorzec znajduje się w każdej komórce. Jest to funkcja przydatna do identyfikowania błędów semantycznych w predykcji. Jeśli istnieje więcej niż jedna wygenerowana zmienna powiązana z tą samą zmienną wynikową, która jednak została utworzona na podstawie innych modeli, zliczane są obserwacje, w których te zmienne są zgodne i niezgodne, a następnie wyświetlane są wartości łączne. W przypadku obserwacji, dla których istnieje zgodność, wyświetlana jest kolejna statystyka typu poprawne/niepoprawne.

Ocena wydajności. Wyświetla statystyki oceny wydajności dla modeli zawierających wyniki jakościowe. Statystyka, utworzona dla każdej kategorii zmiennych wyjściowych, jest miarą średniej możliwej zawartości informacji (w bitach) w modelu dla przewidywania rekordów należących do tej kategorii. Pod uwagę brany jest problem z trudnością sklasyfikowania, dlatego dokładne predykcje dla rzadkich kategorii uzyskają wyższy indeks oceny wydajności niż dokładne predykcje dla często występujących kategorii. Jeśli model jedynie „zgaduje” kategorię, wówczas indeks oceny wydajności dla tej kategorii będzie wynosił 0.

Metryka oceny (AUC i Gini, tylko klasyfikatory binarne). W przypadku klasyfikatorów binarnych ta opcja tworzy raporty dla metryk oceny współczynnika AUC (obszar nad krzywą) i Gini. Obie te metryki oceny są obliczane razem dla każdego modelu binarnego. Wartości metryk są przedstawiane w tabeli w przeglądarce wyników analizy.

Metryka oceny AUC jest obliczana jako obszar pod krzywą ROC (ocena poprawności klasyfikatora) i jest skalarną reprezentacją oczekiwanej wydajności klasyfikatora. Współczynnik AUC zawsze ma wartość z przedziału od 0 do 1, przy czym wyższe wartości reprezentują lepszy klasyfikator. Diagonalna krzywa ROC pomiędzy współrzędnymi (0,0) i (1,1) reprezentuje losowy klasyfikator, a współczynnik AUC wynosi 0,5. Dlatego nie będzie określony realistyczny klasyfikator, a wartość AUC będzie mniejsza niż 0,5.

Metryka oceny współczynnika Gini jest niekiedy używana jako alternatywa dla metryki oceny współczynnika AUC, a obie te miary są ściśle powiązane. Współczynnik Gini jest obliczany jako podwojona powierzchnia pomiędzy krzywą ROC i diagonalną lub jako Gini = 2AUC - 1. Współczynnik Gini zawsze ma wartość z przedziału od 0 do 1, przy czym wyższe wartości reprezentują lepszy klasyfikator. Współczynnik Gini jest ujemny w mało prawdopodobnym przypadku, kiedy krzywa ROC znajduje się poniżej diagonalnej.

Wartości ufności (jeżeli dostępne). W przypadku modeli generujących zmienną ufności ta opcja tworzy raporty dla statystyk wartości ufności i ich relacji z predykcjami. Dla tej opcji dostępne są dwa ustawienia:

- **Wartość graniczna dla.** Informuje o poziomie ufności, powyżej którego dokładność będzie określoną wartością procentową.
- **Poprawa dokładności.** Informuje o poziomie ufności, powyżej którego dokładność jest zwiększana przez określony czynnik. Przykładowo, jeśli ogólna dokładność wynosi 90%, a ta opcja zostanie ustawiona na wartość 2,0, zgłoszona wartość będzie ufnością wymaganą dla 95-procentowej dokładności.

Znajdź zmienne predykcyjne i przewidywane, wykorzystując. Określa, w jaki sposób zmienne predykcyjne są dopasowywane do oryginalnej zmiennej przewidywanej.

- **Metadane zmiennej wyjściowej modelu.** Dopasowuje zmienne predykcyjne do zmiennej przewidywanej na podstawie informacji o zmiennej modelu, zezwalając na dopasowanie, nawet jeśli nazwa zmiennej predykcyjnej została zmieniona. Informacje na temat zmiennej modelu można również uzyskać dla dowolnej zmiennej predykcyjnej za pośrednictwem okna dialogowego Wartości w węźle Typy. Więcej informacji można znaleźć w temacie “Użycie okna dialogowego Wartości” na stronie 144.
- **Format nazwy zmiennej.** Dopasowuje zmienne na podstawie konwencji tworzenia nazw. Przykładowo, wartości predykcyjne wygenerowane przez model użytkowy C5.0 dla zmiennej przewidywanej o nazwie *response* (odpowiedź) muszą znajdować się w zmiennej o nazwie *$C-response* ($C-odpowiedź).

Rozdziel na podzbiory. Jeśli do podzielenia rekordów na próbę uczenia, testowania i walidacji używana jest zmienna dzieląca na podzbiory, należy wybrać tę opcję, aby wyświetlić wyniki osobno dla każdego podzbioru. Więcej informacji można znaleźć w temacie “Węzeł Partycja” na stronie 176.

Uwaga: Podczas rozdzielania przez podział na podzbiory rekordy zawierające wartości w zmiennej dzielącej na podzbiory zostaną wykluczone z analizy. Nie będzie to problemem w przypadku użycia węzła Partycja, ponieważ węzły podziału na podzbiory nie generują wartości null.

Analiza definiowana przez użytkownika. Można określić własne obliczenia dla analizy, jakie będą używane podczas przeprowadzania oceny modeli. Wyrażenia CLEM umożliwiają określenie, co powinno zostać obliczone dla każdego rekordu oraz w jaki sposób połączyć oceny z poziomu rekordu, aby uzyskać ocenę ogólną. Korzystając z funkcji **@TARGET** i **@PREDICTED** można odpowiednio utworzyć odniesienie do wartości przewidywanej (rzeczywisty wynik) i wartości predykcyjnej.

- **Jeżeli.** Należy określić wyrażenie warunkowe, jeżeli konieczne jest użycie różnych obliczeń w zależności od niektórych warunków.
- **To.** Należy określić obliczenie, jakie zostanie wykonane, o ile warunek Jeżeli jest prawdziwy.
- **Inaczej.** Należy określić obliczenie, jakie zostanie wykonane, o ile warunek Jeżeli jest fałszywy.
- **Wykorzystanie.** Należy wybrać statystyki do obliczenia ogólnej oceny w oparciu o oceny indywidualne.

Podziel analizę według wartości zmiennych. Wyświetla zmienne jakościowe, jakich można użyć do podziału analizy. Oprócz ogólnej analizy zgłoszone zostaną osobne analizy dla każdej kategorii każdej zmiennej podziału.

Przeglądarka wyników analizy

Przeglądarka wyników analizy wyświetla wyniki uzyskiwane po wykonaniu węzła Analiza. Typowe opcje zapisywania, eksportowania i drukowania są dostępne w menu Plik. Więcej informacji można znaleźć w temacie “Wyświetlanie wyników” na stronie 303.

Podczas przeglądania wyników analizy po raz pierwszy wyniki zostają rozwinięte. Aby ukryć wyniki po ich przejrzeniu, należy użyć rozszerzanego elementu sterującego znajdującego się po lewej stronie pozycji w celu zwinięcia określonych wartości lub kliknąć przycisk **Zwiń wszystko**, aby zwinąć wszystkie wyniki. Aby ponownie wyświetlić wyniki po ich zwinięciu, należy użyć rozszerzanego elementu sterującego po lewej stronie pozycji w celu wyświetlenia wyników lub kliknąć przycisk **Rozwiń wszystko**, aby wyświetlić wszystkie wyniki.

Wyniki dla zmiennej wyjściowej. Wyniki analizy składają się z sekcji dla każdej zmiennej wyjściowej, dla której istnieje odpowiednia zmienna predykcyjna utworzona przez wygenerowany model.

Porównywanie. W sekcji zmiennej wyjściowej znajduje się sekcja podrzędna dla każdej zmiennej predykcyjnej powiązanej z tą zmienną wyjściową. W przypadku jakościowych zmiennych wyjściowych na najwyższym poziome tej sekcji znajduje się tabela z liczbą i procentem poprawnych i niepoprawnych predykcji oraz łączną liczbą rekordów w strumieniu. W przypadku numerycznych zmiennych wyjściowych w tej sekcji wyświetlane są następujące informacje:

- **Błąd minimalny.** Wyświetla minimalny błąd (różnica pomiędzy wartościami obserwowanymi a przewidywanymi).
- **Błąd maksymalny.** Wyświetla maksymalny błąd.
- **Błąd średni.** Wyświetla wartość średnią dla błędów ze wszystkich rekordów. Wskazuje, gdzie znajduje się **odchylenie** systematyczne (silniejsza tendencja do przeszacowania niż do niedoszacowania lub odwrotnie) w modelu.
- **Średni błąd bezwzględny.** Wyświetla średnią dla wartości bezwzględnych błędów we wszystkich rekordach. Wskazuje średnią wielkość błędu, niezależnie od kierunku.
- **Odchylenie standardowe.** Wyświetla odchylenie standardowe dla błędów.
- **Korelacja liniowa.** Wyświetla liniową korelację pomiędzy wartościami przewidywanymi a rzeczywistymi. Statystyki mają wartości od –1,0 do 1,0. Wartości zbliżone do +1,0 oznaczają silny, dodatni związek, czyli że wysokie wartości przewidywane są związane z wysokimi wartościami rzeczywistymi, a niskie wartości przewidywane są związane z niskimi wartościami rzeczywistymi. Wartości zbliżone do –1,0 oznaczają silny, ujemny związek, czyli że wysokie wartości przewidywane są związane z niskimi wartościami rzeczywistymi i odwrotnie. Wartości zbliżone do 0,0 oznaczają słaby związek, czyli że wartości przewidywane są mniej lub bardziej niezależne w odniesieniu do wartości rzeczywistych. *Uwaga*: Pusta wartość w tym miejscu oznacza, że w danym przypadku nie można obliczyć liniowej korelacji, ponieważ wartości rzeczywiste lub przewidywane są stałe.
- **Wystąpienia.** Wyświetla liczbę rekordów użytych w analizie.

Macierz zbieżności. Jeśli dla jakościowych zmiennych wyjściowych w opcjach analizy utworzono żądanie dla macierzy zbieżności, wyświetlana jest tutaj dodatkowa sekcja zawierająca macierz. Wiersze reprezentują rzeczywiste obserwowane wartości, a kolumny wartości przewidywane. Komórka w tabeli wskazuje liczbę rekordów dla każdej kombinacji wartości przewidywanych i rzeczywistych.

Ocena wydajności. Jeśli dla jakościowych zmiennych wyjściowych w opcjach analizy utworzono żądanie dla statystyki oceny wydajności, wyświetlane są tutaj wyniki oceny wydajności. Wyświetlana jest każda kategoria wyników wraz z jej statystyką oceny wydajności.

Raport wartości ufności. Jeśli dla jakościowych zmiennych wyjściowych w opcjach analizy utworzono żądanie dla wartości ufności, wartości te są tutaj wyświetlane. Dla wartości ufności modelu zgłaszane są następujące statystyki:

- **Zakres.** Wyświetla zakres (najmniejsze i największe wartości) wartości ufności dla rekordów w danych strumienia.
- **Średnia ufność dla poprawnych.** Wyświetla średnią ufność dla rekordów, które zostały poprawnie sklasyfikowane.
- **Średnia ufność dla niepoprawnych.** Wyświetla średnią ufność dla rekordów, które zostały nieprawidłowo sklasyfikowane.
- **Zawsze poprawne powyżej progu.** Wyświetla wartość progową ufności, powyżej której predykcje są zawsze poprawne oraz procent obserwacji, które spełniają to kryterium.
- **Zawsze niepoprawne poniżej progu.** Wyświetla wartość progową ufności, poniżej której predykcje są zawsze niepoprawne oraz procent obserwacji, które spełniają to kryterium.
- **Dokładność powyżej X%.** Wyświetla poziom ufności, przy którym dokładność wynosi *X%*. *X* jest przybliżoną wartością określoną dla opcji **Wartość graniczna dla** w opcjach analizy. W przypadku niektórych modeli i zbiorów danych nie ma możliwości wybrania wartości ufności, która zapewni dokładnie taką wartość graniczną, jak określono w opcjach (zwykle z powodu grup podobnych obserwacji z taką samą wartością ufności zbliżoną do wartości granicznej). Zgłaszana wartość graniczna jest wartością najbardziej zbliżoną do określonego kryterium dokładności, jaką można uzyskać dla pojedynczej wartości granicznej ufności.
- **Powyżej X-krotnej poprawności.** Wyświetla wartość ufności, przy której dokładność jest *X* razy lepsza niż dla całego zbioru danych. *X* jest wartością określoną dla opcji **Poprawa dokładności** w opcjach analizy.

Zgodność pomiędzy. Jeśli w strumieniu uwzględnione są co najmniej dwa wygenerowane modele, które przewidują tą samą zmienną wyjściową, wyświetlone zostaną również statystyki **zgodności** pomiędzy predykcjami wygenerowanymi

przez te modele. Obejmują one liczbę i procent rekordów, dla których predykcje są zgodne (dla jakościowych zmiennych wyjściowych) lub statystyki podsumowania błędów (dla ilościowych zmiennych wyjściowych). W przypadku zmiennych jakościowych obejmuje analizę predykcji w porównaniu do wartości rzeczywistych dla podzbioru rekordów, w którym modele są zgodne (generują taką samą wartość przewidywaną).

Metryka ewaluacyjna. Jeśli dla klasyfikatorów binarnych w opcjach analizy utworzono żądanie dla metryki ewaluacyjnej, wartości metryki oceny współczynnika AUC i Gini są wyświetlane w tabeli w tej sekcji. Tabela zawiera jeden wiersz dla każdego modelu klasyfikatora binarnego. Tabela metryki ewaluacyjnej jest raczej wyświetlana dla każdej zmiennej wyjściowej, a nie dla poszczególnych modeli.

węzeł Audyt danych

Węzeł Audyt danych umożliwia kompleksowe spojrzenie na dane wprowadzone do programu IBM SPSS Modeler, które są zaprezentowane w postaci czytelnej macierzy, która może zostać posortowana i użyta do wygenerowania wykresów w pełnym rozmiarze oraz węzłów przygotowania danych.

- Na karcie Audyt wyświetlany jest raport zawierający statystyki podsumowujące, histogramy i wykresy rozkładu, które mogą pomóc we wstępnym zrozumieniu danych. Raport zawiera również ikonę składowania, która znajduje się przed nazwą zmiennej.
- Na karcie Jakość w raporcie z audytu wyświetlane są informacje na temat wartości odstających, skrajnych i brakujących oraz udostępniane narzędzia umożliwiające obsługę tych wartości.

Korzystanie z węzła Audyt danych

Węzeł Audyt danych może zostać dołączony bezpośrednio do węzła źródłowego lub poniżej określonego węzła Typy. Można również wygenerować dowolną liczbę węzłów przygotowania danych na podstawie wyników. Można na przykład wygenerować węzeł filtrowania, który będzie wykluczał zmienne ze zbyt dużą liczbą braków danych, by były użyteczne w modelowaniu, i wygenerować Superwęzeł, który będzie podstawiał braki danych dla dowolnych lub wszystkich pozostałych zmiennych. Właśnie w takich sytuacjach przejawia się prawdziwy potencjał audytu danych — pozwala on bowiem nie tylko oceniać obecny stan danych, lecz także podejmować działania na podstawie wyników tej oceny.

Monitorowanie lub próbkowanie danych. Ponieważ wstępny audyt jest najbardziej efektywny w przypadku wielkich zbiorów danych (big data), w celu zredukowania czasu przetwarzania podczas początkowej eksploracji można utworzyć węzeł próby, który umożliwi wybór tylko określonego podzbioru rekordów. Węzeł Audyt danych może być również użyty w połączeniu z węzłami, takimi jak Dobór predyktorów oraz Wykrywanie anomalii, w badawczych etapach analizy.

Węzeł Audyt danych — karta Ustawienia

Karta Ustawienia umożliwia określenie podstawowych parametrów dla audytu.

Domyślny. Wystarczy dołączyć węzeł do strumienia i kliknąć przycisk **Uruchom**, aby wygenerować raport z audytu dla wszystkich zmiennych w oparciu o ustawienia domyślne, zgodnie z informacjami poniżej:

- Jeśli nie określono ustawień dla węzła typu, wszystkie zmienne są uwzględniane w raporcie.
- Jeśli wprowadzono ustawienia dla węzła typu (niezależnie od tego, czy zostały określone), wyświetlane są wszystkie zmienne typu *wejściowa*, *przewidywana* i *łącznie*. Jeśli wystąpi pojedyncza zmienna *przewidywana*, należy użyć zmiennej nałożenia. Jeśli określona zostanie więcej niż jedna zmienna *przewidywana*, nie można zastosować domyślnego nałożenia.

Użyj ustawień użytkownika. Tę opcję należy zaznaczyć, aby ręcznie wybrać zmienne. Przycisk selektora zmiennych po prawej stronie umożliwia wybranie zmiennych pojedynczo lub według typu.

Zmienna nałożenia. Zmienna nałożenia jest używana podczas rysowania miniaturowych wykresów wyświetlanych w raporcie z audytu. W przypadku zmiennej ilościowej (zakres numeryczny) obliczane są również statystyki dla dwóch

zmiennych (kowariancja i korelacja). Jeśli obecna jest pojedyncza zmienna *przewidywana* oparta na ustawieniach węzła Typy, pełni ona funkcję domyślnej zmiennej nałożenia, zgodnie z opisem powyżej. Alternatywnie można wybrać opcję **Użyj ustawień użytkownika**, aby określić nałożenie.

Pokaż. Umożliwia określenie, czy w wyniku będą dostępne wykresy oraz wybranie statystyk, które będą wyświetlane domyślnie.

- **Wykresy.** Wyświetla wykres dla każdej wybranej zmiennej; w zależności od danych może to być wykres rozkładu (słupkowy), histogram lub wykres rozrzutu. Wykresy są wyświetlane jako miniatury we wstępnym raporcie, ale wykresy w pełnym wymiarze i węzły wykresów również mogą zostać wygenerowane. Więcej informacji można znaleźć w temacie “Audyt danych — przeglądarka wyników” na stronie 316.
- **Podstawowe statystyki/Statystyki zaawansowane.** Określa poziom statystyk wyświetlanych domyślnie w wynikach. To ustawienie dotyczy wstępnego wyświetlania, jednak w wyniku dostępne są wszystkie statystyki, niezależnie od tego ustawienia. Więcej informacji można znaleźć w temacie “Wyświetl statystyki” na stronie 317.

Mediana i dominanta. Oblicza medianę i dominantę dla wszystkich zmiennych w raporcie. Należy pamiętać, że w przypadku dużych zbiorów danych te statystyki mogą wydłużyć czas przetwarzania, ponieważ ich obliczenie trwa dłużej od pozostałych. Tylko w przypadku mediany: w niektórych przypadkach zgłoszona wartość może być wyznaczona na podstawie przykładowych 2000 rekordów (a nie dla pełnego zbioru danych). Próbkowanie odbywa się na podstawie zmiennej w przypadku obserwacji, dla których w przeciwnym razie dojdzie do przekroczenia limitów pamięci. Po zakończeniu próbkowania wyniki zostaną opatrzone etykietami, takimi jak w wyniku (raczej *Sample Median* (Przykładowa mediana) niż po prostu *Median* (Mediana)). Wszystkie pozostałe statystyki (oprócz mediany) są zawsze używane z zastosowaniem pełnego zbioru danych.

Zmienne puste lub bez typu. W przypadku użycia z określonymi danymi zmienne bez określonego typu nie są uwzględniane w raporcie z audytu. Aby uwzględnić zmienne bez typu (w tym puste zmienne), należy wybrać opcję **Wyczyść wszystkie** w dowolnym wcześniejszym węźle Typy. Dzięki temu dane nie będą określone, co spowoduje, że wszystkie zmienne będą uwzględnione w raporcie. Może to być na przykład przydatne, jeśli użytkownik zamierza uzyskać pełną listę wszystkich zmiennych lub wygenerować węzeł Filtr, który będzie wykluczał puste zmienne. Więcej informacji można znaleźć w temacie “Filtrowanie zmiennych z brakami danych” na stronie 319.

Audyt danych — karta Jakość

Karta Jakość w węźle Audyt danych udostępnia opcje obsługi braków danych, wartości odstających oraz wartości skrajnych.

Braki danych

- **Liczebność rekordów z ważnymi wartościami.** Tę opcję należy zaznaczyć, aby wyświetlić liczbę rekordów z poprawnymi wartościami dla każdej zmiennej poddanej ocenie. Należy pamiętać, że wartości null (niezdefiniowane), wartości puste, białe znaki i puste łańcuchy zawsze traktowane są jako wartości niepoprawne.
- **Podzielone liczebności rekordów z nieważnymi wartościami.** Tę opcję należy zaznaczyć, aby wyświetlić liczbę rekordów z niepoprawną wartością każdego typu dla poszczególnych zmiennych.

Wartości skrajne i odstające

Metoda wykrywania wartości odstających lub skrajnych. Obsługiwane są dwie metody:

Odchylenie standardowe od średniej. Wykrywa wartości odstające i wartości skrajne na podstawie liczby standardowych odchyleń od średniej. Przykładowo, jeśli dostępna jest zmienna o wartości średniej wynoszącej 100 i z odchyleniem standardowym wynoszącym 10, można wprowadzić wartość 3,0, aby wskazać, że wszystkie wartości poniżej 70 lub powyżej 130 powinny być traktowane jako wartości odstające.

Rozstęp ćwiartkowy. Wykrywa wartości odstające i skrajne na podstawie rozstępu ćwiartkowego, który jest zakresem zawierającym dwa centralne kwartyle (pomiędzy 25. i 75. percentylem). Przykładowo, w oparciu o ustawienie domyślne wynoszące 1,5 dolna wartość graniczna dla wartości odstających będzie wynosiła Q1 – 1,5 * IQR, a górna

wartość graniczna będzie wynosiła Q3 + 1,5*IQR. Należy pamiętać, że w przypadku dużych zbiorów danych użycie tej opcji może zmniejszyć wydajność.

Audyt danych — przeglądarka wyników

Przeglądarka audytu danych to wszechstronne narzędzie umożliwiające przegląd danych. Na karcie Audyt wyświetlane są miniaturowe wykresy, ikony składowania oraz statystyki dla wszystkich zmiennych, natomiast na karcie Jakość wyświetlane są informacje dotyczące wartości odstających, wartości skrajnych oraz braków danych. Na podstawie wstępnych wykresów i statystyk podsumowujących można podjąć decyzję, czy konieczne jest przekodowanie zmiennych numerycznych, wyliczenie nowej zmiennej czy też rekodowanie wartości zmiennej nominalnej. Można również przeprowadzić dalszą eksplorację, używając bardziej złożonych wizualizacji. Jest to możliwe bezpośrednio z przeglądarki raportów z audytu za pośrednictwem menu Utwórz, które pozwala na utworzenie dowolnej liczby węzłów, jakie mogą następnie zostać użyte do transformacji lub wizualizacji danych.

- Można sortować kolumny, klikając ich nagłówek, lub zmienić porządek kolumn, używając funkcji „przeciągnij i upuść". Obsługiwana jest również większość standardowych operacji związanych z wynikami. Więcej informacji można znaleźć w temacie "Wyświetlanie wyników" na stronie 303.
- Wartości i przedziały dla zmiennych można wyświetlać, dwukrotnie klikając zmienną w kolumnach Poziom pomiaru lub Unikalne.
- Pasek narzędzi lub menu Edycja umożliwiają wyświetlanie lub ukrywanie etykiet wartości lub wybieranie statystyk, jakie mają zostać wyświetlone. Więcej informacji można znaleźć w temacie "Wyświetl statystyki" na stronie 317.
- Można sprawdzić ikony składowania znajdujące się po lewej stronie nazw zmiennych. Składowanie to sposób przechowywania danych w zmiennej. Przykładowo w zmiennej zawierającej wartości 1 i 0 składowane są dane w postaci liczb całkowitych. Różni się to od poziomu pomiaru, który opisuje użycie danych i nie wpływa na składowanie. Więcej informacji można znaleźć w temacie "Ustawienia składowania i formatowania zmiennej" na stronie 9.

Wyświetlanie i generowanie wykresów

Jeśli nie wybrano żadnego nałożenia, karta Audyt umożliwia wyświetlenie wykresów słupkowych (zmienne nominalne i zmienne typu flaga) lub histogramów (zmienne ilościowe).

W przypadku nałożenia dla zmiennej nominalnej lub flagi wykresy są kolorowane zgodnie z wartościami nałożenia.

W przypadku nałożenia dla zmiennej ilościowej generowane są raczej dwuwymiarowe wykresy rozrzutu, a nie jednowymiarowe wykresy słupkowe i histogramy. W takim przypadku oś *x* odwzorowuje zmienną nałożenia, umożliwiając wyświetlanie takiej samej skali na wszystkich osiach *x*.

- Zmienne flagi lub nominalne: należy ustawić kursor myszy nad słupkiem, aby wyświetlić odpowiednią wartość lub etykietę w okienku podpowiedzi.
- W przypadku zmiennych flagi i nominalnych w celu zmiany orientacji miniatur wykresów z poziomej na pionową należy użyć paska narzędzi.
- Aby z dowolnej miniatury wygenerować pełnowymiarowy wykres, należy dwukrotnie kliknąć miniaturę lub zaznaczyć miniaturę i wybrać z menu Utwórz opcję **Wynik graficzny**. *Uwaga*: Jeśli miniaturowy wykres został utworzony na podstawie danych z próby, wygenerowany wykres będzie zawierał wszystkie obserwacje, o ile oryginalny strumień danych jest nadal otwarty.

  Wykres można wygenerować tylko wtedy, gdy węzeł Audyt danych, który utworzył wynik, jest połączony ze strumieniem.
- Aby wygenerować pasujący węzeł wykresu, należy wybrać co najmniej jedną zmienną na karcie Audyt i z menu Utwórz wybrać opcję **Węzeł wykresu**. Węzeł wynikowy jest dodawany do obszaru roboczego strumienia i może być użyty do ponownego utworzenia wykresu za każdym razem, kiedy węzeł jest uruchomiony.
- Jeśli zestaw nałożenia zawiera więcej niż 100 wartości, generowane jest ostrzeżenie i nałożenie nie jest uwzględniane.

Wyświetl statystyki

Okno dialogowe Wyświetl statystyki umożliwia wybór statystyk wyświetlanych na karcie Audyt. Ustawienia początkowe są określane w węźle Audyt danych. Więcej informacji można znaleźć w temacie “Węzeł Audyt danych — karta Ustawienia” na stronie 314.

Minimum. Najmniejsza wartość zmiennej numerycznej.

Maksimum. Największa wartość zmiennej numerycznej.

Suma. Suma wartości wszystkich obserwacji nieposiadających braków danych.

Przedział. Różnica między największą a najmniejszą wartością zmiennej numerycznej; maksimum minus minimum.

Średnia. Miara tendencji centralnej. Średnia arytmetyczna; suma podzielona przez liczbę obserwacji.

Błąd standardowy średniej. Miara tego, jak bardzo wartość średniej może się zmieniać dla różnych prób losowanych z tego samego rozkładu. Może być wykorzystywana do pobieżnego porównania rzeczywistej wartości średniej z wartością hipotetyczną (tj. można sądzić, że te dwie wartości są różne, jeśli iloraz różnicy i błędu standardowego jest mniejszy od -2 lub większy od +2).

Odchylenie standardowe. Miara rozproszenia wokół wartości średniej, równa pierwiastkowi z wariancji. Odchylenie standardowe mierzy się w tych samych jednostkach co pierwotną wartość.

Wariancja. Miara rozproszenia wokół średniej, równa sumie podniesionych do kwadratu odchyleń od średniej, podzielonej przez liczbę obserwacji minus jeden. Wariancja jest mierzona w jednostkach będących kwadratami jednostek miary dla zmiennej, do której wariancja się odnosi.

Skośność. Miara asymetrii rozkładu. Rozkład normalny jest symetryczny, a jego wartość skośności wynosi 0. Rozkład o dużej skośności dodatniej ma długi kraniec z prawej strony. Gdy zaś współczynnik skośności jest ujemny, rozkład ma długi kraniec z lewej strony. Jako wytyczna, wartość skośności przekraczająca dwukrotnie swój błąd standardowy na ogół oznacza odstępstwo od symetrii rozkładu.

Błąd standardowy skośności. Iloraz skośności i jej błędu standardowego; może być wykorzystywany jako test normalności (tzn. można odrzucić założenie o normalności, jeśli wartość ilorazu jest mniejsza od -2 lub większa od +2). Wysoka dodatnia wartość dla skośności wskazuje na długi prawy kraniec; skrajnie ujemna wartość wskazuje na długi lewy kraniec.

Kurtoza. Miara ilości skrajnych wartości odstających. W przypadku rozkładu normalnego wartość statystyki kurtozy wynosi zero. Kurtoza dodatnia oznacza, że w danych jest więcej skrajnych wartości odstających niż w rozkładzie normalnym. Kurtoza ujemna oznacza, że w danych jest mniej skrajnych wartości odstających niż w rozkładzie normalnym.

Błąd standardowy kurtozy. Iloraz kurtozy i jej błędu standardowego; może być wykorzystywany jako test normalności (tzn. można odrzucić założenie normalności, jeśli wartość ilorazu jest mniejsza od -2 lub większa od +2). Wysoka dodatnia wartość dla kurtozy wskazuje na to, iż krańce rozkładu są dłuższe niż te dla rozkładu normalnego; ujemna wartość dla kurtozy wskazuje na krótsze krańce (podobnie jak w rozkładach prostokątnych).

Swoista. Oszacowuje wielkość wszystkich efektów równocześnie, korygując każdy z nich ze względu na wszystkie inne efekty każdego typu.

Ważne. Poprawne obserwacje nieposiadające systemowych ani zdefiniowanych przez użytkownika braków danych. Należy pamiętać, że wartości null (niezdefiniowane), wartości puste, białe znaki i puste łańcuchy zawsze traktowane są jako wartości niepoprawne.

Mediana. Jest to 50. percentyl, czyli taka wartość, że połowa obserwacji ma wartości mniejsze, a druga połowa ma wartości większe od niej. W sytuacji parzystej liczby obserwacji mediana jest średnią dwóch środkowych obserwacji w

próbie posortowanej rosnąco lub malejąco. W przeciwieństwie do średniej, na którą wpływ może mieć nawet kilka ekstremalnie dużych lub małych wartości, mediana jest miarą tendencji centralnej niewrażliwą na wartości odstające.

Dominanta. Wartość występująca najczęściej. Jeśli więcej niż jedna wartość występuje z taką samą, największą częstością, każda z nich jest dominantą (wartością modalną).

Należy również pamiętać, że mediana i dominanta są domyślnie usuwane, aby zwiększyć wydajność, ale można je wybrać na karcie Ustawienia w węźle Audyt danych. Więcej informacji można znaleźć w temacie “Węzeł Audyt danych — karta Ustawienia” na stronie 314.

Statistics for Overlays

Jeśli używana jest ilościowa (zakres liczbowy) zmienna nałożenia, dostępne są również następujące statystyki:

Kowariancja. Nieustandaryzowana miara powiązania dwóch zmiennych, równa sumie iloczynów wektorowych odchyleń wartości tych zmiennych od ich średnich podzielonej przez N-1.

Przeglądarka audytu danych — karta Jakość

Karta Jakość w przeglądarce audytu danych wyświetla wyniki analizy jakości danych i umożliwia określenie sposobu postępowania z wartościami odstającymi, skrajnymi i z brakami danych.

Podstawianie brakujących wartości: W raportach z audytu wyświetlana jest lista wartości procentowych dla kompletnych rekordów dla każdej zmiennej, wraz z liczbą poprawnych wartości, wartości null oraz pustych wartości. Można podstawić braki danych odpowiednio dla konkretnych zmiennych, a następnie wygenerować Superwęzeł w celu zastosowania tych transformacji.

1. W kolumnie **Podstawianie braków** określ typ wartości, jakie mają zostać podstawione, o ile takie istnieją. Można podstawić puste wartości i/lub wartości null lub określić niestandardowy warunek lub wyrażenie, na podstawie którego wybierane będą wartości do podstawienia.

   Istnieje kilka typów braków danych rozpoznawanych przez program IBM SPSS Modeler:

   - **Null lub systemowe braki danych.** Są to wartości niełańcuchowe pozostawione jako puste w bazie danych lub w pliku źródłowym i niezdefiniowane jako „brakujące” w węźle źródłowym ani w węźle wprowadzania danych. Systemowe braki danych są wyświetlane jako wartości **$null$**. Należy zwrócić uwagę, że puste łańcuchy tekstowe nie są traktowane jako wartości null w programie IBM SPSS Modeler, mimo że mogą być one traktowane jako wartości null przez niektóre bazy danych.
   - **Puste łańcuchy i białe znaki.** Puste łańcuchy tekstowe i białe znaki (łańcuchy bez widocznych znaków) są traktowane odmiennie niż wartości null. Puste łańcuchy są w większości przypadków traktowane jako równoważne białym znakom. Na przykład po wybraniu opcji traktowania białych znaków jako pustych w węźle źródłowym albo w węźle Typy to ustawienie ma zastosowanie także do pustych łańcuchów.
   - **Puste lub zdefiniowane przez użytkownika braki danych.** Są to wartości takie jak nieznane, 99 czy –1, zdefiniowane jawnie w węźle źródłowym lub w węźle Typy jako braki danych. Opcjonalnie można także wybrać traktowanie wartości null i białych znaków jako wartości pustej, co pozwala oznaczyć je z myślą o specjalnym ich traktowaniu i wykluczeniu ich z większości obliczeń. Można na przykład użyć funkcji @BLANK do traktowania tych wartości, wraz z brakami danych innego typu, jako wartości pustej.

2. W kolumnie **Metoda** należy wybierz metodę, jaka ma zostać użyta.

   Wprowadzanie braków danych umożliwiają następujące metody:

   Stała. Zastępuje stałą wartość (średnią dla zmiennej, środek zakresu lub wskazaną stałą).

   Losowa. Podstawia wartość losową w oparciu o rozkład normalny lub jednostajny.

   Wyrażenie. Umożliwia określenie wyrażenia niestandardowego. Można na przykład zastąpić wartości zmienną globalną utworzoną za pośrednictwem węzła wartości globalnych.

   Algorytm. Podstawia wartość przewidywaną przez model w oparciu o algorytm C&RT. Dla każdej zmiennej wprowadzonej tą metodą dostępny będzie osobny model C&RT oraz węzeł wypełniania zastępujący wartości puste i wartości null wartością przewidywaną przez model. Następnie stosowany jest węzeł filtrowania umożliwiający usuwanie zmiennych predykcyjnych wygenerowanych przez model.

3. Aby wygenerować Superwęzeł braków danych, z menu wybierz:

   Utwórz > Superwęzeł braków danych

   Zostanie wyświetlone okno dialogowe Superwęzeł braków danych.
4. Wybierz opcję **Wszystkie zmienne** lub **Tylko wybrane zmienne** i w razie potrzeby określ wielkość próby. (Określona próba jest wartością procentową; domyślnie próbkowanych jest 10% wszystkich rekordów).
5. Kliknij przycisk **OK**, aby dodać wygenerowany Superwęzeł do obszaru roboczego strumienia.
6. Dołącz Superwęzeł do strumienia, aby zastosować transformacje.

W ramach Superwęzła używana jest odpowiednia kombinacja modelu użytkowego oraz węzłów Wypełnianie i Filtrowanie. Aby zrozumieć sposób działania, można przeprowadzić edycję Superwęzła i kliknąć przycisk **Wejdź do środka**; można dodawać, edytować lub usuwać konkretne węzły Superwęzła, aby doprecyzować jego działanie.

Traktowanie wartości odstających i wartości skrajnych: W raporcie z audytu zamieszczana jest lista wartości odstających i skrajnych dla każdej zmiennej w oparciu o opcje wykrywania określone w węźle Audyt danych. Więcej informacji można znaleźć w temacie “Audyt danych — karta Jakość” na stronie 315. Dla konkretnych zmiennych wartości można odpowiednio wymusić, odrzucić lub wyzerować, a następnie wygenerować Superwęzeł, który pozwoli zastosować transformacje.

1. W kolumnie **Postępowanie** określ sposób działania w przypadku wartości odstających i skrajnych dla konkretnych zmiennych.

   Następujące sposoby postępowania z wartościami odstającymi i skrajnymi:

   - **Wymuś.** Zastępuje wartości odstające i skrajne najbliższą wartością, która nie jest uznawana jako skrajna. Przykładowo, jeśli wartość odstająca jest zdefiniowana jako każda wartość powyżej lub poniżej trzech standardowych odchyleń, wówczas wszystkie wartości odstające zostaną zastąpione najwyższą lub najniższą wartością z tego zakresu.
   - **Odrzuć.** Odrzuca rekordy z wartościami odstającymi lub skrajnymi dla konkretnej zmiennej.
   - **Wyzeruj.** Zastępuje wartości odstające i skrajne wartością null lub systemowym brakiem danych.
   - **Wymuś odstające/odrzuć skrajne.** Odrzuca tylko wartości skrajne.
   - **Wymuś odstające/wyzeruj skrajne.** Zeruje tylko wartości skrajne.
2. Aby wygenerować Superwęzeł, z menu wybierz:

   Utwórz > Superwęzeł odstających i skrajnych

   Zostanie wyświetlone okno dialogowe Superwęzeł wartości odstających.
3. Wybierz opcje **Wszystkie zmienne** lub **Tylko wybrane zmienne**, a następnie kliknij przycisk **OK**, aby dodać wygenerowany Superwęzeł do obszaru roboczego strumienia.
4. Dołącz Superwęzeł do strumienia, aby zastosować transformacje.

Opcjonalnie można przeprowadzić edycję Superwęzła i wejść do środka, aby przeglądać jego zawartość lub dokonywać zmian. W ramach Superwęzła wartości można odrzucać, wymuszać lub wyzerować, używając serii węzłów Selekcja i/lub Wypełnianie.

Filtrowanie zmiennych z brakami danych: Korzystając z przeglądarki audytu danych można utworzyć nowy węzeł filtrowania w oparciu o wyniki analizy jakości; umożliwia to opcja Generuj filtr w oknie dialogowym Jakość.

Dominanta. Należy wybrać odpowiednią operację dla określonych zmiennych: **Uwzględnij** lub **Wyklucz**.

- **Wybrane zmienne.** Węzeł filtrowania będzie uwzględniał zmienne wybrane na karcie Jakość lub będzie je wykluczał. Można na przykład posortować tabelę w kolumnie **Ukończono %**, za pomocą metody Shift+kliknięcie wybrać najmniejsze wypełnione zmienne, a następnie wygenerować węzeł filtrowania który te zmienne wyklucza.
- **Zmienne z procentową jakością wyższą niż.** Węzeł filtrowania będzie uwzględniał/wykluczał zmienne, w których wartość procentowa kompletnych rekordów jest większa niż określona wartość graniczna. Domyślną wartością progową jest 50%.

Filtrowanie pustych zmiennych lub zmiennych bez określonego typu

Należy pamiętać, że po określeniu wartości danych zmienne bez typu lub puste zmienne zostają wykluczone z wyników audytu oraz z większości pozostałych wyników w programie IBM SPSS Modeler. Zmienne te są ignorowane podczas modelowania, ale mogą powodować niepotrzebne przeciążenie i zamieszanie. W takim przypadku należy użyć przeglądarki audytu danych, aby wygenerować węzeł filtrowania, który usunie te zmienne ze strumienia.

1. Aby upewnić się, czy wszystkie zmienne zostały uwzględnione w audycie, w tym również zmienne puste lub bez określonego typu, kliknij opcję **Wyczyść wszystkie** we wcześniejszym węźle źródłowym lub typu lub ustaw wartości na <*Przepuść*> dla wszystkich zmiennych.
2. W przeglądarce audytu danych przeprowadź sortowanie w kolumnie **Ukończono %**, wybierz zmienne, które nie mają w ogóle poprawnych wartości (lub zastosuj inną wartość graniczną), i korzystając z menu Utwórz, utwórz węzeł filtrowania, który może zostać dodany do strumienia.

Wybór rekordów z brakami danych: Za pośrednictwem przeglądarki audytu danych można utworzyć nowy węzeł Selekcja, korzystając z wyników analizy jakości.

1. W przeglądarce audytu danych wybierz zakładkę Jakość.
2. Z menu wybierz opcję:

   Utwórz > Węzeł selekcji braków danych

   Wyświetlane jest okno dialogowe Utwórz węzeł selekcji.

Wybierz rekord, gdy jest on. Określ, czy rekordy powinny być zachowywane jeśli są **Prawidłowe** czy **Nieprawidłowe**.

Wyszukiwanie nieprawidłowych wartości. Należy określić, co należy sprawdzić, aby wyszukać nieprawidłowe wartości.

- **Wszystkie zmienne.** Węzeł selekcji sprawdzi wszystkie zmienne w celu wyszukania nieprawidłowych wartości.
- **Zmienne zaznaczone w tabeli.** Węzeł selekcji sprawdzi tylko zmienne aktualnie wybrane w tabeli wyników węzła Jakość.
- **Zmienne z procentową jakością wyższą niż.** Węzeł selekcji sprawdzi wszystkie zmienne, w których wartość procentowa kompletnych rekordów jest większa niż określona wartość graniczna. Domyślną wartością progową jest 50%.

Uważaj rekord za nieprawidłowy, jeżeli znaleziono niepoprawną wartość w. Należy określić warunek, na podstawie którego rekord zostanie uznany jako nieprawidłowy.

- **Dowolna z powyższych zmiennych.** Węzeł selekcji uzna rekord jako nieprawidłowy, jeśli *dowolna* z określonych powyżej zmiennych zawiera niepoprawną wartość dla tego rekordu.
- **Wszystkie z powyższych zmiennych.** Węzeł selekcji uzna rekord jako nieprawidłowy, jeśli *wszystkie* zmienne określone powyżej zawierają niepoprawną wartość dla tego rekordu.

Generowanie innych węzłów w celu przygotowania danych

Różne węzły używane w procesie przygotowywania danych można wygenerować bezpośrednio z przeglądarki audytu danych, w tym węzły Rekodowanie, Kategoryzacja oraz Wyliczanie. Na przykład:

- Nową zmienną można wyliczyć na podstawie wartości *claimvalue* (wartość roszczenia) oraz *farmincome* (przychód gospodarstwa), wybierając obie te wartości w raporcie z audytu oraz opcję **Wyliczanie** z menu Utwórz. Nowy węzeł jest dodawany do obszaru roboczego strumienia.
- Podobnie na podstawie wyników audytu można określić, czy rekodowanie wartości *farmincome* na przedziały percentylowe zapewni bardziej przejrzystą analizę. Aby wygenerować węzeł kategoryzacji, należy zaznaczyć wiersz zmiennej i wybrać opcję **Kategoryzacja** z menu Utwórz.

Po wygenerowaniu węzła i dodaniu go do obszaru roboczego strumienia należy dołączyć węzeł do strumienia i otworzyć go, aby określić opcje dla wybranych zmiennych.

Węzeł Transformacja

Normalizacja zmiennych wejściowych jest istotnym etapem przed zastosowaniem tradycyjnych technik oceniania, takich jak regresja, regresja logistyczna czy analiza dyskryminacyjna. Techniki dokonują założeń na temat rozkładów normalnych danych, które mogą nie być prawdziwe dla wielu zmiennych danych surowych. Jednym z rozwiązań, które pozwala poradzić sobie z danymi rzeczywistymi, jest zastosowanie transformacji (przekształceń), które zbliżą element danych surowych do rozkładu normalnego. Ponadto, zmienne znormalizowane można w prosty sposób porównać z każdą inną zmienną — na przykład, przychód i wiek mają całkiem inne skale w zmiennych danych surowych, ale po znormalizowaniu rzeczywisty wpływ każdej ze zmiennych można w łatwy sposób zinterpretować.

Węzeł Transformacja udostępnia przeglądarkę wyników, dzięki której można szybko dokonać oceny wzrokowej i określić, jakie przekształcenie przeprowadzić. Wystarczy rzut oka, aby stwierdzić, czy zmienne mają rozkład normalny i w razie konieczności wybrać przekształcenie i zastosować je. Można wybrać wiele zmiennych i wykonać jedno przekształcenie dla zmiennej.

Po wybraniu preferowanych przekształceń dla zmiennych można wygenerować węzły Wyliczanie lub Wypełnianie, które przeprowadzają przekształcenia i dołączyć je do strumienia. Węzeł wyliczeń tworzy nowe zmienne, podczas gdy węzeł wypełniania wykonuje przekształcenie istniejących. Więcej informacji można znaleźć w temacie “Tworzenie wykresów” na stronie 323.

Węzeł Transformacja — karta Zmienne

Na zakładce Zmienne można określić, które zmienne danych mają zostać użyte do wyświetlania możliwych przekształceń i zastosowania ich. Przekształcaniu można poddać tylko zmienne numeryczne. Należy kliknąć przycisk Selektor zmiennych i wybrać co najmniej jedną zmienną numeryczną z wyświetlonej listy.

Węzeł Transformacja — karta Opcje

Karta Opcje umożliwia określenie typu transformacji, jakie mają zostać uwzględnione. Można uwzględnić wszystkie dostępne przekształcenia lub wybrać przekształcenia pojedynczo.

W drugim przypadku można również wprowadzić liczbę określającą przesunięcie danych dla przekształcenia odwrotnego i logarytmicznego. Jest to przydatne, jeśli duża ilość zer w danych będzie powodowała odchylenie wyników dla średniej i odchylenia standardowego.

Załóżmy na przykład, że dostępna jest zmienna o nazwie *BALANCE* (Saldo), która zawiera pewne wartości zerowe, a użytkownik zamierza poddać ją przekształceniu odwrotnemu. Aby uniknąć niepożądanych odchyleń, można wybrać opcję **Odwrotność (1/x)** i wprowadzić wartość 1 w zmiennej **Zastosuj przesunięcie danych**. (Należy pamiętać, że przesunięcie to nie jest powiązane z przesunięciem wykonywanym za pośrednictwem funkcji @OFFSET w programie IBM SPSS Modeler.)

Wszystkie formuły. Wskazuje, że wszystkie dostępne przekształcenia powinny zostać obliczone i wyświetlone w wynikach.

Wybrane formuły. Umożliwia wybranie różnych przekształceń, jakie zostaną obliczone i wyświetlone w wynikach.

- **Odwrotność (1/x).** Wskazuje, że w wynikach powinno być wyświetlone przekształcenie odwrotne.
- **Logarytm naturalny (log n).** Wskazuje, że w wynikach powinno być wyświetlone przekształcenie $\log_n$.
- **Logarytm dziesiętny (log 10).** Wskazuje, że w wynikach powinno być wyświetlone przekształcenie $\log_{10}$.
- **Wykładniczo.** Wskazuje, że w wynikach powinno być wyświetlone przekształcenie wykładnicze (e^x).
- **Pierwiastek kwadratowy.** Wskazuje, że w wynikach powinno być wyświetlone przekształcenie pierwiastka kwadratowego.

Węzeł Transformacja — karta Wynik

Karta Wynik umożliwia określenie formatu wynikowego i lokalizacji wyniku. Wyniki mogą być również wyświetlane na ekranie lub można je wysłać do jednego ze standardowych typów plików. Więcej informacji można znaleźć w temacie “Węzeł Wynik — karta Wynik” na stronie 306.

Węzeł Transformacja — przeglądarka wyników

Przeglądarka wyników umożliwia wyświetlanie wyników wykonywania węzła Transformacja. Przeglądarka jest wszechstronnym narzędziem, umożliwiającym wyświetlanie wielu przekształceń dla zmiennej w postaci miniatur, co pozwala na szybkie porównanie zmiennych. Korzystając z opcji dostępnych w menu Plik, można zapisywać, eksportować lub drukować wynik. Więcej informacji można znaleźć w temacie “Wyświetlanie wyników” na stronie 303.

Dla każdej transformacji (innej niż Wybrana transformacja) poniżej wyświetlana jest legenda w następującym formacie:

Średnia (Odchylenie standardowe)

Generowanie węzłów dla transformacji

Przeglądarka wyników stanowi pomocny punkt startowy dla przygotowania danych. Przykładowo, użytkownik zamierza znormalizować zmienną *AGE* (Wiek), aby można było zastosować technikę ocenienia (np. regresja logistyczna lub analiza dyskryminacyjna), która zakłada, że rozkład jest normalny. Na podstawie wstępnych wykresów i statystyk podsumowujących można zdecydować, czy zmienna *AGE* ma zostać przekształcona zgodnie z określonym rozkładem (np. logarytmicznym). Po wybraniu rozkładu można wygenerować węzeł wyliczania, w którym do oceniania zastosowana będzie standaryzacja.

Za pośrednictwem przeglądarki wyników można wygenerować następujące węzły operacji na zmiennych:

- Wyliczanie
- Wypełnianie

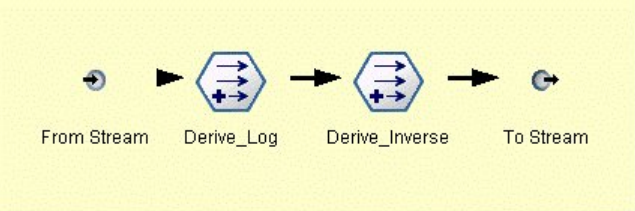
Węzeł wyliczeń tworzy nowe zmienne z zastosowaniem odpowiednich transformacji, podczas gdy węzeł filtrowania przekształca istniejące zmienne. Węzły są umieszczane w obszarach roboczych w postaci Superwęzła.

Jeśli to samo przekształcenie zostanie wybrane dla różnych zmiennych, węzeł wypełniania lub wyliczania będzie zawierał formuły dla tego typu przekształcenia dla wszystkich zmiennych, których to przekształcenie dotyczy. Załóżmy na przykład, że wybrano zmienne i przekształcenia, przedstawione w poniższej tabeli, aby wygenerować węzeł Wyliczanie.

Tabela 44. Przykład generowania węzła Wyliczanie

Zmienna	**Transformacja**
AGE	Bieżący rozkład
INCOME	Logarytm
OPEN_BAL	Odwrotność
BALANCE	Odwrotność

Superwęzeł zawiera następujące węzły:



Rysunek 74. Superwęzeł w obszarze roboczym

W tym przykładzie węzeł Derive_Log zawiera formułę logarytmiczną dla zmiennej *INCOME* (Przychód), a węzeł Derive_Inverse zawiera formuły odwrotności dla zmiennych *OPEN_BAL* (Saldo otwarcia) i *BALANCE* (Saldo).

Aby wygenerować węzeł

1. Dla każdej zmiennej w przeglądarce wyników wybierz wymagane przekształcenie.
2. Z menu Utwórz wybierz odpowiednio opcje **Węzeł wyliczania** lub **Węzeł wypełniania**.

   Spowoduje to otworzenie odpowiednio okna dialogowego węzła wyliczania lub wypełniania.

   Wybierz opcję **Niestandaryzowana transformacja** lub **Standaryzacja (Statystyka z)**, w zależności od potrzeb. Druga opcja powoduje zastosowanie podczas transformacji statystyki *z*; statystyka *z* reprezentuje wartości w postaci funkcji odległości od średniej dla zmiennej w odchyleniach standardowych. Przykładowo, jeśli dla zmiennej *AGE* zastosowana zostanie transformacja logarytmiczna i wybrana będzie opcja standaryzacji, końcowe równanie dla generowanego węzła będzie miało następującą postać:

   (log(AGE)-Mean)/SD

Po wygenerowaniu węzła i wyświetlaniu w obszarze roboczym strumienia:

1. Dołącz go do strumienia.
2. W przypadku Superwęzła opcjonalnie kliknij dwukrotnie węzeł, aby wyświetlić jego zawartość.
3. Opcjonalnie, kliknij dwukrotnie węzeł wyliczania lub wypełniania, aby zmodyfikować opcje dla wybranych zmiennych.

Tworzenie wykresów: W przeglądarce wyników na podstawie miniatury histogramu można wygenerować pełnowymiarowy histogram z wynikami.

Aby wygenerować wykres

1. Kliknij dwukrotnie miniaturę wykresu w przeglądarce wyników.

   lub

   Wybierz miniaturę wykresu w przeglądarce wyników.
2. Z menu Utwórz wybierz opcję **Wynik graficzny**.

Zostanie wyświetlony histogram z nałożoną krzywą rozkładu normalnego. Umożliwia to porównanie bliskości dopasowania poszczególnych dostępnych przekształceń z rozkładem normalnym.

Uwaga: Wykres można wygenerować tylko wtedy, gdy węzeł Transformacja, który utworzył wynik, jest połączony ze strumieniem.

Pozostałe operacje: Korzystając z przeglądarki wyników, można również:

- Sortować siatkę wynikową według kolumny Zmienna.
- Eksportować wynik do pliku HTML. Więcej informacji można znaleźć w temacie “Eksportowanie wyników” na stronie 305.

Węzeł Statystyki

Węzeł Statystyki udostępnia informacje podsumowujące na temat zmiennych numerycznych. Można uzyskać podsumowujące statystyki dla poszczególnych zmiennych i korelacji pomiędzy zmiennymi.

Węzeł Statystyki — karta Ustawienia

Zbadaj. Należy wybrać zmienną lub zmienne, dla których potrzebne są indywidualne statystyki podsumowujące. Można wybrać kilka zmiennych.

Statystyki. Należy wybrać statystyki, jakie znajdą się w raporcie. Dostępne opcje to: **Liczebność**, **Średnia**, **Suma**, **Min.**, **Maks.**, **Rozstęp**, **Wariancja**, **Odch. Std.**, **Błąd standardowy średniej**, **Mediana** i **Dominanta**.

Korelacja. Należy wybrać zmienną lub zmienne, jakie mają zostać skorelowane. Można wybrać kilka zmiennych. Po wybraniu zmiennych korelacji jako wynik wyświetlone zostaną korelacje pomiędzy każdą zmienną Zbadaj a zmiennymi korelacji.

Ustawienia korelacji. Można określić opcje wyświetlania siły korelacji w wyniku.

Ustawienia korelacji

IBM SPSS Modeler może charakteryzować korelacje z etykietami opisowymi, aby podkreślić istotne relacje. **Korelacja** mierzy siłę relacji pomiędzy dwiema zmiennymi ilościowymi (zakres liczbowy). Może mieć wartości od –1,0 do 1,0. Wartości zbliżone do +1,0 oznaczają silny dodatni związek, czyli wysokie wartości jednej zmiennej są związane z wysokimi wartościami innej zmiennej, a niskie wartości są związane z niskimi. Wartości zbliżone do –1,0 oznaczają silny ujemny związek, czyli wysokie wartości jednej zmiennej są związane z niskimi wartościami innej zmiennej i odwrotnie. Wartości zbliżone do 0,0 oznaczają słaby związek, czyli wartości dwóch zmiennych są mniej lub bardziej niezależne.

Korzystając z okna dialogowego Ustawienia korelacji, można sterować wyświetlaniem etykiet korelacji, zmieniać wartości progowe definiujące kategorie, a także zmieniać etykiety stosowane dla każdego zakresu. Ponieważ sposób charakteryzowania wartości korelacji zależy w znacznym stopniu od obszaru problemu, można dostosować zakresy i etykiety, tak aby dopasować je do konkretnej sytuacji.

Pokaż w wyniku etykiety siły korelacji. Ta opcja jest wybrana domyślnie. Usunięcie zaznaczenia tej opcji spowoduje pominięcie etykiet opisowych w wyniku.

Siła korelacji. Dostępne są dwie opcje służące do definiowania i dodawania etykiet siły korelacji:

- **Przedstaw siłę korelacji według ważności (1-p).** Etykiety dla korelacji są tworzone na podstawie ważności, zdefiniowanej jako 1 minus istotność lub 1 minus prawdopodobieństwo, że różnica średnich może być wyjaśniona tylko przez szansę. Im wartość bardziej zbliżona do 1, tym większa szansa, że dwie zmienne *nie* są niezależne — innymi słowy, że istnieje pomiędzy nimi pewna relacja. Tworzenie etykiet korelacji na podstawie ważności jest ogólnie bardziej zalecane niż na podstawie wartości bezwzględnej, ponieważ tłumaczy zmienność danych — na przykład, współczynnik wynoszący 0,6 może wskazywać duże znaczenie w jednym zbiorze danych i brak znaczenia w innym. Domyślnie, wartości istotności z przedziału od 0,0 do 0,9 są opatrzone etykietą *Słaba*, te z przedziału od 0,9 do 0,95 etykietą *Średnia*, a te z przedziału od 0,95 do 1,0 etykietą *Silna*.
- **Przedstaw siłę korelacji według wartości bezwzględnej.** Etykiety korelacji są tworzone na podstawie wartości bezwzględnej współczynnika korelacji Pearsona, który mieści się w zakresie od –1 do 1, jak opisano powyżej. Im bardziej wartość bezwzględna miary zbliżona do 1, tym silniejsza korelacja. Domyślnie, korelacje z zakresu od 0,0 do 0,3333 (jako wartość bezwzględna) są opatrzone etykietą *Słaba*, te z zakresu od 0,3333 do 0,6666 etykietą *Średnia*, a te z przedziału od 0,6666 do 1,0 etykietą *Silna*. Należy jednak pamiętać, że istotność każdej wartości jest trudna do uogólnienia pomiędzy różnymi zbiorami danych; z tego względu w większości przypadków zalecane jest definiowanie korelacji na podstawie prawdopodobieństwa, a nie wartości bezwzględnej.

Węzeł Statystyki — przeglądarka wyników

Przeglądarka wyników węzła Statistics wyświetla wyniki analizy statystycznej i umożliwia wykonywanie różnych operacji, takich jak wybór zmiennych, generowanie nowych węzłów na podstawie dokonanego wyboru oraz zapisywanie i drukowanie wyników. W menu Plik dostępne są standardowe opcje zapisywania, eksportowania i drukowania, a menu Edycja udostępnia standardowe opcje edytowania. Więcej informacji można znaleźć w temacie “Wyświetlanie wyników” na stronie 303.

Podczas przeglądania wyników węzła Statistics po raz pierwszy wyniki zostają rozwinięte. Aby ukryć wyniki po ich przejrzeniu, należy użyć rozszerzanego elementu sterującego znajdującego się po lewej stronie pozycji w celu zwinięcia określonych wartości lub kliknąć przycisk **Zwiń wszystko**, aby zwinąć wszystkie wyniki. Aby ponownie wyświetlić wyniki po ich zwinięciu, należy użyć rozszerzanego elementu sterującego po lewej stronie pozycji w celu wyświetlenia wyników lub kliknąć przycisk **Rozwiń wszystko**, aby wyświetlić wszystkie wyniki.

Wynik zawiera sekcję dla każdej zmiennej *Zbadaj*, w której znajduje się tabela z żądanymi statystykami.

- **Liczebności.** Liczba rekordów z poprawnymi wartościami dla zmiennej.
- **Średnia.** Średnia wartość dla zmiennej we wszystkich rekordach.
- **Suma.** Suma wartości dla zmiennej we wszystkich rekordach.
- **Min.** Minimalna wartość dla zmiennej.
- **Maks.** Maksymalna wartość dla zmiennej.
- **Zakres.** Różnica między wartością minimalną a maksymalną.
- **Wariancja.** Miara zmienności wartości zmiennej. Obliczana jest poprzez określenie różnicy pomiędzy każdą wartością a ogólną średnią, podniesienie jej do kwadratu, zsumowanie dla wszystkich wartości i podzielenie przez liczbę rekordów.
- **Odchylenie standardowe.** Kolejna miara zmienności wartości zmiennej, obliczana jako pierwiastek kwadratowy wariancji.
- **Błąd standardowy średniej.** Miara niepewności oszacowania średniej dla zmiennej przy założeniu, że średnia jest stosowana dla nowych danych.
- **Mediana.** Wartość „środkowa” zmiennej; czyli wartość, która oddziela górną połowę danych od dolnej (na podstawie wartości zmiennej).
- **Dominanta.** Najbardziej powszechna wartość osobliwa w danych.

Korelacje. Jeśli określono zmienne korelacji, wynik również będzie zawierał sekcję z listą korelacji Pearsona pomiędzy zmienną Zbadaj i każdą zmienną korelacji oraz opcjonalnie etykiety opisowe dla wartości korelacji. Więcej informacji można znaleźć w temacie “Ustawienia korelacji” na stronie 324.

Menu Utwórz. Menu Utwórz zawiera opcje umożliwiające generowanie węzłów.

- **Filtrowanie.** Generuje węzeł filtrowania, umożliwiający odfiltrowanie zmiennych, które nie są skorelowane lub są słabo skorelowane z innymi zmiennymi.

Generowanie węzła filtrowania z węzła Statistics

Węzeł filtrowania wygenerowany za pośrednictwem przeglądarki wyników węzła Statistics umożliwia filtrowanie zmiennych na podstawie ich korelacji z innymi zmiennymi. Korelacje są sortowane w kolejności wartości bezwzględnej, z uwzględnieniem najwyższych korelacji (zgodnie z kryterium ustawionym w obszarze Generuj filtr w oknie dialogowym Statystyka); następnie tworzony jest filtr, który obejmuje wszystkie zmienne wyświetlane dla wysokich korelacji.

Dominanta. Należy zdecydować, w jaki sposób korelacje mają być wybierane. Opcja **Uwzględnij** spowoduje zachowanie zmiennych wyświetlanych w określonych korelacjach. Opcja **Wyklucz** spowoduje odfiltrowanie zmiennych.

Uwzględnij/Wyklucz zmienne występujące w. Należy zdefiniować kryterium wyboru korelacji.

- **Maksymalna liczba korelacji.** Umożliwia wybór określonej liczby korelacji i uwzględnia/wyklucza zmienne wyświetlane dla dowolnej z tych korelacji.
- **Maksymalny procent korelacji (%).** Umożliwia wybór określonego procenta (*n%*) korelacji i uwzględnia/wyklucza zmienne wyświetlane dla dowolnej z tych korelacji.
- **Korelacje większe niż.** Umożliwia wybór korelacji, których wartość bezwzględna jest większa od określonej wartości granicznej.

Węzeł Średnie

Węzeł Średnie porównuje średnie między niezależnymi grupami lub między parami powiązanych zmiennych w celu przetestowania, czy istnieje dla nich znaczna różnica. Na przykład można porównać średnie przychody przed uruchomieniem promocji i po jej zakończeniu lub porównać przychody od klientów, którzy nie otrzymali oferty promocyjnej, z przychodami od tych, którzy z niej skorzystali.

Średnie można porównać na dwa różne sposoby, w zależności od dostępnych danych:

- **Między grupami w zmiennej.** Aby porównać niezależne grupy, należy wybrać zmienną testową i zmienną grupującą. Przykładowo podczas wysyłania materiałów promocyjnych można wykluczyć próbę „wstrzymanych" klientów i porównać średnie przychody dla grupy wstrzymanej z pozostałymi. W takim przypadku należy określić pojedynczą zmienną testową, która będzie określała przychód dla każdego klienta oraz zmienną flagi lub nominalną, która będzie określała, czy klienci ci otrzymali ofertę. Próby są niezależne, to znaczy każdy rekord jest przypisywany do jednej lub drugiej grupy i nie ma możliwości powiązania określonego obiektu jednej grupy z określonym obiektem z drugiej. Można również określić zmienną nominalną zawierającą więcej niż dwie wartości w celu porównania średnich dla wielu grup. Podczas wykonywania węzeł wykonuje obliczenia testu jednoczynnikowej analizy wariancji (ANOVA) dla wybranych zmiennych. Jeśli istnieją tylko dwie grupy zmiennych, wyniki jednoczynnikowej analizy wariancji są w zasadzie takie same jak dla testu *t* dla prób niezależnych. Więcej informacji można znaleźć w temacie "Porównywanie średnich dla grup niezależnych".
- **Pomiędzy parami zmiennych.** Podczas porównywania średnich dwóch powiązanych zmiennych grupy muszą być połączone w pary, aby wyniki były znaczące. Można na przykład porównać średnie przychody z tej samej grupy klientów przed rozpoczęciem promocji i po jej rozpoczęciu lub porównać wskaźniki użycia dla usług pomiędzy parami mąż-żona, aby sprawdzić, czy się różnią. Każdy rekord zawiera dwie osobne, ale powiązane miary, które można porównać w znaczący sposób. Podczas wykonywania węzeł wykonuje obliczenia testu *t* dla prób zależnych dla każdej wybranej pary zmiennych. Więcej informacji można znaleźć w temacie "Porównywanie średnich pomiędzy parami zmiennych".

Porównywanie średnich dla grup niezależnych

W celu porównania średniej dla co najmniej dwóch niezależnych grup należy wybrać opcję **Między grupami w zmiennej** w węźle Średnie.

Zmienna grupująca. Należy wybrać numeryczną zmienną flagi lub zmienną nominalną z co najmniej dwoma odmiennymi wartościami, która dzieli rekordy na grupy, jakie mają zostać porównane, takie jak: kto otrzymał ofertę a kto nie. Niezależnie od liczby zmiennych testowych można wybrać tylko jedną zmienną grupującą.

Zmienne testowe. Należy wybrać co najmniej jedną zmienną numeryczną, która zawiera miary, jakie mają zostać przetestowane. Dla każdej wybranej zmiennej wykonany zostanie osobny test. Przykładowo można przetestować wpływ danej promocji na użycie, przychód i poziom odejścia.

Porównywanie średnich pomiędzy parami zmiennych

Aby porównać średnie pomiędzy parami zmiennych, należy wybrać opcję **Pomiędzy parami zmiennych** w węźle Średnie. Zmienne muszą być powiązane, aby wyniki były znaczące, np. przychody przed akcją promocyjną i po jej zakończeniu. Można również wybrać kilka par zmiennych.

Zmienna 1. Należy wybrać zmienną numeryczną, która zawiera pierwszą miarę do porównania. W badaniu przed i po będzie to zmienna „przed".

Zmienna 2. Należy wybrać drugą zmienną do porównania.

Dodaj. Dodaje wybraną parę zmiennych do listy Pary zmiennych testowych.

W razie potrzeby należy powtórzyć wybór zmiennych i dodać do listy wiele par.

Ustawienia korelacji. Umożliwia określenie opcji tworzenia etykiet siły korelacji. Więcej informacji można znaleźć w temacie “Ustawienia korelacji” na stronie 324.

Opcje węzła Średnie

Na karcie Opcje można ustawić wartości graniczne *p*, na podstawie których tworzone są etykiety Ważne, Brzegowe lub Nieważne. Można również edytować etykietę dla poszczególnych rang. Ważność jest mierzona w skali procentowej i można ją ogólnie zdefiniować jako 1 minus prawdopodobieństwo uzyskania danego wyniku (np. różnica średnich pomiędzy dwiema zmiennymi) jako wartość skrajna lub bardziej skrajna niż obserwowany wynik dla jednej szansy. Przykładowo wartość *p* większa niż 0,95 oznacza mniej niż 5% szans, że wynik może zostać objaśniony tylko przez jedną szansę.

Etykiety ważności. Istnieje możliwość edytowania etykiet używanych dla poszczególnych par lub grup w wyniku. Etykiety domyślne to: *ważne*, *brzegowe* i *nieważne*.

Wartości odcięcia. Określa wartość graniczną dla każdej rangi. Zwykle wartości *p* większe niż 0,95 uzyskają rangę Ważne, a wartości niższe niż 0,9 będą określone jako Nieważne, jednak te wartości graniczne można skorygować odpowiednio do potrzeb.

Uwaga: Miary ważności są dostępne w wielu węzłach. Konkretne obliczenia zależą od węzła oraz od typu zmiennej przewidywanej i wejściowej, ale wartości nadal można porównywać, ponieważ wszystkie są mierzone w skali procentowej.

Węzeł Średnie — przeglądarka wyników

Przeglądarka wyników węzła Średnie wyświetla dane tabel krzyżowych i umożliwia wykonywanie standardowych operacji, takich jak zaznaczanie i kopiowanie pojedynczych wierszy tabeli, sortowanie według kolumny oraz zapisywanie i drukowanie tabeli. Więcej informacji można znaleźć w temacie “Wyświetlanie wyników” na stronie 303.

Konkretne informacje w tabeli zależą od typu porównania (grupy w zmiennej lub osobne zmienne).

Sortuj według. Umożliwia sortowanie wyniku wg konkretnej kolumny. Należy kliknąć strzałkę w górę lub w dół, aby zmienić kierunek sortowania. Alternatywnie można kliknąć dowolny nagłówek kolumny, aby przeprowadzić sortowanie według tej kolumny. (Aby zmienić kierunek sortowania w kolumnie, należy kliknąć jeszcze raz).

Widok. Aby ustawić poziom szczegółów, można wybrać **Prosty** lub **Zaawansowany**. Widok zaawansowany obejmuje wszystkie informacje z widoku prostego oraz dodatkowe szczegóły.

Wynik średniej — porównywanie grup w zmiennej

Podczas porównywania grup w zmiennej nazwa zmiennej grupującej jest wyświetlana nad tabelą wyników, a średnie i powiązane statystyki są zgłaszane osobno dla każdej grupy. Tabela zawiera osobny wiersz dla każdej zmiennej testowej.

Wyświetlane są następujące kolumny:

- **Zmienna.** Zawiera listę nazw wybranych zmiennych testowych.
- **Średnie według grupy.** Wyświetla średnią dla każdej kategorii zmiennej grupującej. Przykładowo można porównać osoby, które otrzymały ofertę specjalną (*New Promotion* — Nowa promocja), z osobami, które takiej oferty nie otrzymały (*Standard*). W widoku zaawansowanym wyświetlane są również odchylenie standardowe, błąd standardowy oraz liczebność.

- **Ważność.** Wyświetla wartość i etykietę ważności. Więcej informacji można znaleźć w temacie “Opcje węzła Średnie” na stronie 327.

Zaawansowane wyniki

W widoku zaawansowanym wyświetlane są następujące dodatkowe kolumny.

- **Test F.** Ten test opiera się na współczynniku wariancji pomiędzy grupami i wariancji w każdej grupie. Jeśli średnie są takie same dla wszystkich grup, współczynnik *F* będzie zbliżony do 1, ponieważ oba oszacowania obejmują tę samą wariancję populacji. Im wyższy współczynnik, tym większa wariancja pomiędzy grupami i większa szansa, że istnieje znaczna różnica.
- **df.** Wyświetla stopnie swobody.

Wynik średniej — porównywanie par zmiennych

W przypadku porównywania osobnych zmiennych tabela wyników zawiera wiersz dla każdej wybranej pary zmiennych.

- **Zmienna 1/2.** Wyświetla nazwę pierwszej i drugiej zmiennej w każdej parze. W widoku zaawansowanym wyświetlane są również odchylenie standardowe, błąd standardowy oraz liczebność.
- **Średnia 1/2.** Wyświetla średnią dla każdej zmiennej (odpowiednio).
- **Korelacja.** Mierzy siłę relacji pomiędzy dwiema zmiennymi ilościowymi (zakres liczbowy). Wartości zbliżone do +1,0 oznaczają silny, dodatni związek, a wartości zbliżone do –1,0 oznaczają silny, ujemny związek. Więcej informacji można znaleźć w temacie “Ustawienia korelacji” na stronie 324.
- **Różnica średnich.** Wyświetla różnicę pomiędzy dwiema średnimi zmiennych.
- **Ważność.** Wyświetla wartość i etykietę ważności. Więcej informacji można znaleźć w temacie “Opcje węzła Średnie” na stronie 327.

Zaawansowane wyniki

W wynikach zaawansowanych wyświetlane są następujące dodatkowe kolumny:

95% przedział ufności. Dolna i górna granica zakresu, w którym rzeczywista średnia prawdopodobnie znajdzie się w 95% wszystkich możliwych prób tej wielkości w danej populacji.

Test T. Statystyka *t* jest wyznaczana przez podzielenie różnicy średnich przez błąd standardowy. Im wyższa wartość bezwzględna statystyki, tym większe prawdopodobieństwo, że średnie nie są takie same.

df. Wyświetla stopnie swobody dla statystyki.

Węzeł Raport

Węzeł Raport umożliwia tworzenie sformatowanych raportów zawierających stały tekst oraz dane i inne wyrażenia wydzielone z jednych danych. Można określić format raportu, korzystając z szablonów tekstowych do zdefiniowania konstrukcji tekstu stałego i danych wyjściowych. Możliwe jest zastosowanie niestandardowego formatowania tekstu poprzez wprowadzenie do szablonu znaczników HTML i ustawienie opcji na karcie Wynik. Wartości danych i inne wyniki warunkowe są zamieszczane w raporcie za pośrednictwem wyrażeń CLEM zamieszczanych w szablonie.

Rozwiązania alternatywne dla węzła Raport

Węzeł Raport jest najczęściej stosowany do wyświetlania listy rekordów lub wyniku obserwacji ze strumienia, takich jak wszystkie rekordy spełniające określony warunek. Z tego względu może być traktowany jako mniej ustrukturyzowana alternatywa dla węzła Tabela.

- Jeśli potrzebny jest raport zawierający listę informacji o zmiennej lub informacje o tym, co zostało zdefiniowane w strumieniu, a nie same dane (np. definicje zmiennych określone w węźle Typy), wówczas można użyć skryptu.

- Aby wygenerować raport zawierający wiele obiektów wynikowych (np. zbiór modeli, tabel i wykresów wygenerowanych na podstawie jednego lub kilku strumieni), dla którego wyniki mogą być zapisane w wielu formatach (takich jak tekst, HTML i Microsoft Word/Office), można użyć projektu IBM SPSS Modeler.
- W celu utworzenia listy nazw zmiennych bez używania skryptów można przed węzłem Tabela zamieścić węzeł Losowanie, który odrzuci wszystkie rekordy. W wyniku tego powstanie tabela bez wierszy, którą można będzie transponować podczas eksportu, aby powstała lista nazw zmiennych w postaci pojedynczej kolumny. (W tym celu należy wybrać opcję **Transponuj dane** na karcie Wynik w węźle Tabela).

Węzeł Raport — karta Szablon

Tworzenie szablonu. Aby zdefiniować zawartość raportu, w węźle raportu na karcie Szablon można utworzyć szablon. Szablon składa się z wierszy tekstu (każdy wiersz określa pewną zawartość raportu) oraz z wierszy specjalnych znaczników służących do określania zakresu wierszy zawartości. W każdym wierszu zawartości przed wysłaniem wiersza do raportu obliczane są wyrażenia CLEM ujęte w nawiasy kwadratowe ([]). Dla wiersza w szablonie istnieją trzy możliwe zakresy:

Stała. Wiersze nie, które nie są oznaczone, są uznawane jako stałe. Wiersze stałe są kopiowane do raportu tylko raz, po obliczeniu dowolnego wyrażenia, jakie zawierają. Na przykład wiersz

```
This is my report, printed on [@TODAY]
```

spowoduje skopiowanie do raportu pojedynczego wiersza zawierającego tekst i bieżącą datę.

Globalne (powtórz ALL). Wiersze zawarte pomiędzy specjalnymi znacznikami #ALL i # są kopiowane do raportu raz dla każdego rekordu danych wejściowych. Wyrażenia CLEM (ujęte w nawiasach) są obliczane na podstawie bieżącego rekordu dla każdego wiersza wynikowego. Na przykład wiersze

```
#ALL
For record [@INDEX], the value of AGE is [AGE]
#
```

będą obejmowały jeden wiersz dla każdego rekordu określającego numer rekordu i wiek.

Aby wygenerować listę wszystkich rekordów:

```
#ALL
[Age] [Sex] [Cholesterol]  [BP]
#
```

Warunkowe (powtórz WHERE). Wiersze zawarte pomiędzy specjalnymi znacznikami #WHERE <*warunek*> i # są kopiowane do raportu raz dla każdego rekordu, w którym podany warunek jest prawdziwy. Warunek jest wyrażeniem CLEM. (W przypadku warunku WHERE nawiasy są opcjonalne). Na przykład wiersze

```
#WHERE [SEX = 'M']
Male at record no. [@INDEX] has age [AGE].
#
```

spowodują zapisanie jednego wiersza w pliku dla każdego rekordu zawierającego dla płci wartość *M* (kobieta). Kompletny raport będzie zawierał wiersze stałe, globalne i warunkowe zdefiniowane poprzez zastosowanie szablon do danych wejściowych.

Opcje wyświetlania lub zapisywania wyników wspólne dla różnych typów węzłów wyników można określić za pomocą karty Wynik. Więcej informacji można znaleźć w temacie “Węzeł Wynik — karta Wynik” na stronie 306.

Zapisywanie wyników danych w formacie HTML lub XML

Zamieszczenie znaczników HTML lub XML bezpośrednio w szablonie pozwala na zapisywanie raportów w jednym z tych formatów. Przykładowo, następujący szablon utworzy tabelę HTML.

```
This report is written in HTML.
Only records where Age is above 60 are included.

<HTML>
  <TABLE border="2">
   <TR>
     <TD>Age</TD>
     <TD>BP</TD>
     <TD>Cholesterol</TD>
     <TD>Drug</TD>
   </TR>

  #WHERE Age > 60
    <TR>
       <TD>[Age]</TD>
       <TD>[BP]</TD>
       <TD>[Cholesterol]</TD>
       <TD>[Drug]</TD>
     </TR>
  #
  </TABLE>
</HTML>
```

Węzeł Raport — przeglądarka wyników

Przeglądarka raportów wyświetla zawartość wygenerowanego raportu. W menu Plik dostępne są standardowe opcje zapisywania, eksportowania i drukowania, a menu Edycja udostępnia standardowe opcje edytowania. Więcej informacji można znaleźć w temacie “Wyświetlanie wyników” na stronie 303.

Węzeł Globalne

Węzeł wartości globalnych skanuje dane i oblicza wartości sumaryczne, które mogą zostać użyte w wyrażeniach CLEM. Na przykład węzła wartości globalnych można użyć do obliczenia statystyk dla zmiennej o nazwie *age* (wiek), a następnie użyć ogólnej średniej dla wartości *age* w wyrażeniach CLEM, wstawiając funkcję @GLOBAL_MEAN(age).

Węzeł Globalne — karta Ustawienia

Tworzone wartości globalne. Należy wybrać zmienną lub zmienne, dla których wartości globalne mają być dostępne. Można wybrać kilka zmiennych. Dla każdej zmiennej należy określić statystyki, jakie będą obliczane, upewniając się, że statystyki te zostały wybrane w kolumnach obok nazwy zmiennej.

- **ŚREDNIA.** Średnia wartość dla zmiennej we wszystkich rekordach.
- **SUMA.** Suma wartości dla zmiennej we wszystkich rekordach.
- **MIN.** Minimalna wartość dla zmiennej.
- **MAKS.** Maksymalna wartość dla zmiennej.
- **ODCHSTD.** Odchylenie standardowe, które jest miarą zmienności wartości zmiennej i jest obliczane jako pierwiastek kwadratowy wariancji.

Domyślne operacje. Wybrane tutaj opcje będę stosowane po dodaniu nowych zmiennych do listy wartości globalnych powyżej. Aby zmienić domyślnie ustawione statystyki, należy zaznaczyć lub usunąć zaznaczenie odpowiednich statystyk. Można również użyć przycisku **Zastosuj**, aby zastosować operacje domyślne do wszystkich zmiennych z listy.

Uwaga: Niektóre operacje nie mają zastosowania w przypadku zmiennych nienumerycznych (np. Suma dla zmiennej data/czas). Operacje, których nie można użyć dla wybranej zmiennej, są wyłączone.

Wyczyść wszystkie wartości globalne przed wykonaniem. Tę opcje należy zaznaczyć, aby przed obliczeniem nowych wartości usunąć wszystkie wartości globalne. Jeśli ta opcja nie zostanie zaznaczona, nowo obliczone wartości zastąpią stare, ale wartości globalne, które nie zostaną ponownie obliczone, pozostaną dostępne.

Wyświetl podgląd globalnych utworzonych po wykonaniu. Jeśli ta opcja zostanie zaznaczona, po wykonaniu w oknie dialogowym właściwości strumienia wyświetlana będzie karta Globalne, zawierająca obliczone wartości globalne.

Węzeł Symulacje Dopasowanie

Węzeł Symulacje Dopasowanie dopasowuje zestaw potencjalnych rozkładów statystycznych do poszczególnych zmiennych w danych. Dopasowanie każdego rozkładu do zmiennej jest oceniane na podstawie kryterium dobroci dopasowania. Po wykonaniu węzła Symulacje Dopasowanie tworzony jest węzeł Symulacje Generowanie (lub następuje aktualizacja istniejącego węzła). Do każdej zmiennej przypisany jest rozkład i najlepszym dopasowaniu. Węzeł Symulacje Generowanie może być wówczas użyty do wygenerowania danych objętych symulacją dla każdej zmiennej.

Chociaż węzeł Symulacje Dopasowanie jest węzłem końcowym, nie powoduje dodania modelu do palety wygenerowanych modeli, dodaje natomiast wynik lub wykres do karty wyników lub eksportuje dane.

Uwaga: Jeśli dane historyczne są rzadkie (to znaczy zawierają dużo braków danych), składnik odpowiedzialny za dopasowanie może mieć trudności w wyszukaniu wystarczającej liczby poprawnych wartości, aby dopasować rozkłady do danych. W przypadku kiedy dane są rzadkie, przed dopasowaniem należy usunąć rzadkie zmienne, o ile nie są one wymagane, lub podstawić braki danych. Korzystając z opcji na karcie **Jakość** w węźle audytu danych, można wyświetlić liczbę kompletnych rekordów, określić, które zmienne są rzadkie i wybrać metodę podstawiania. Jeśli liczba rekordów jest niewystarczająca dla dopasowania rozkładu, można użyć węzła równoważenia, aby zwiększyć liczbę rekordów.

Użycie węzła Symulacje Dopasowanie do automatycznego tworzenia węzła Symulacje Generowanie

Po wykonaniu węzła Symulacje Dopasowanie po raz pierwszy tworzony jest węzeł Symulacje Generowanie zawierający łącze aktualizacji dla węzła Symulacje Dopasowanie. Jeśli węzeł Symulacje Dopasowanie zostanie wykonany ponownie, nowy węzeł Symulacje Generowanie zostanie utworzony tylko w przypadku usunięcia łącza aktualizacji. Węzeł Symulacje Dopasowanie umożliwia również zaktualizowanie połączonego węzła Symulacje Generowanie. Wynik zależy od tego, czy w obu węzłach znajdują się takie same zmienne oraz czy zmienne nie zostały zablokowane w węźle Symulacje Generowanie. Więcej informacji można znaleźć w temacie “Węzeł Symulacje Generowanie” na stronie 52.

Węzeł Symulacje Dopasowanie umożliwia również zaktualizowanie połączonego węzła Symulacje Generowanie. Aby zdefiniować łącze aktualizacji dla węzła Symulacje Generowanie, należy wykonać następujące kroki:

1. Kliknij prawym przyciskiem myszy węzeł Symulacje Dopasowanie.
2. Z menu wybierz opcję **Zdefiniuj łącze aktualizacji**.
3. Kliknij węzeł Symulacje Generowanie, dla którego ma zostać zdefiniowane łącze aktualizacji.

Aby usunąć łącze aktualizacji pomiędzy węzłem Symulacje Dopasowanie a węzłem Symulacje Generowanie, kliknij łącze aktualizacji prawym przyciskiem myszy i wybierz opcję **Usuń łącze**.

Dopasowywanie rozkładu

Rozkład statystyczny to teoretyczna częstość wystąpienia wartości, jakie mogą istnieć dla zmiennej. W węźle Symulacje Dopasowanie zestaw rozkładów statystycznych jest porównywany z danymi każdej zmiennej. Rozkłady dostępne dla dopasowania zostały opisane w temacie “Rozkłady” na stronie 61. Parametry teoretycznego rozkładu są korygowane, tak aby zapewnić najlepsze dopasowanie do danych zgodnie z pomiarem dobroci dopasowania; stosowane jest kryterium Anderson-Darling lub kryterium Kolmogorov-Smirnov. Wyniki dopasowania rozkładu za pośrednictwem węzła Symulacje Dopasowanie pokazują, które rozkłady zostały dopasowane, najlepsze oszacowania parametrów dla poszczególnych rozkładów oraz stopień dopasowania każdego rozkładu do danych. W czasie

dopasowywania rozkładu obliczane są również korelacje pomiędzy zmiennymi z liczbowym typem składowania oraz kontyngencje pomiędzy zmiennymi z rozkładem jakościowym. Wyniki dopasowywania rozkładu służą do utworzenia węzła Symulacje Generowanie.

Przed dopasowaniem rozkładów do danych w pierwszym 1000 rekordów przeprowadzane jest sprawdzenie, czy nie ma w nich braków danych. Jeśli braków danych jest zbyt wiele, dopasowanie rozkładu jest niemożliwe. W takiej sytuacji należy zdecydować, czy odpowiednie będzie użycie jednej z następujących opcji:

- Użycie poprzedzającego węzła w celu usunięcia rekordów zawierających braki danych.
- Użycie węzła poprzedzającego w celu wprowadzenia wartości do braków danych.

Podczas dopasowywania rozkładu braki danych nie są wykluczane. Jeśli w danych występują braki danych zdefiniowane przez użytkownika i wartości te mają zostać wykluczone z dopasowywania rozkładu, wówczas należy ustawić te wartości jako systemowe braki danych.

Podczas dopasowywania rozkładu rola zmiennej nie jest brana pod uwagę. Przykładowo zmienne z rolą **przewidywana** są traktowane tak samo, jak zmienne z rolami **wejściowa**, **brak**, **oba elementy**, **podział**, **separacja**, **częstość** i **Id**.

Zmienne traktowane są odmiennie w czasie dopasowywania rozkładu w zależności od ich typu składowania oraz poziomu pomiaru. Sposób traktowania zmiennych w czasie dopasowywania rozkładu został opisany w poniższej tabeli.

Tabela 45. Dopasowywanie rozkładu w zależności od typu składowania i poziomu pomiaru zmiennych

<table>
<tr><th rowspan="2">Typ składowania</th><th colspan="6">Poziom pomiaru</th></tr>
<tr><th>Ilościowy</th><th>Jakościowy</th><th>Flaga</th><th>Nominalny</th><th>Porządkowy</th><th>Nieokreślony</th></tr>
<tr><td>Łańcuch</td><td>Niemożliwe</td><td colspan="3">Jakościowy, dopasowywane są rozkłady Dice'a i stały</td><td rowspan="8">Zmienna jest ignorowana i nie jest przekazywana do węzła Symulacje Generowanie.</td></tr>
<tr><td>Liczba całkowita</td><td rowspan="6">Dopasowywane są wszystkie rozkłady. Obliczane są korelacje i kontyngencje.</td><td colspan="3" rowspan="6">Dopasowywany jest rozkład jakościowy. Korelacje nie są obliczane.</td><td rowspan="6">Dopasowywane są rozkłady dwumianowy, ujemny dwumianowy oraz Poissona; obliczane są korelacje.</td></tr>
<tr><td>Liczba rzeczywista</td></tr>
<tr><td>Czas</td></tr>
<tr><td>Data</td></tr>
<tr><td>Znacznik czasu</td></tr>
<tr><td>Nieznane</td><td colspan="5">Na podstawie danych określany jest odpowiedni typ składowania.</td></tr>
</table>

Zmienne z porządkowym typem pomiaru są traktowane jak zmienne ilościowe i są uwzględniane w tabeli korelacji w węźle Symulacje Generowanie. Jeśli konieczne jest dopasowanie do zmiennej porządkowej rozkładu innego niż dwumianowy, ujemny dwumianowy lub Poissona, należy zmienić poziom pomiaru zmiennej na ilościowy. Jeśli wcześniej zdefiniowano etykietę dla każdej wartości zmiennej porządkowej, a następnie poziom pomiaru zostanie zmieniony na ilościowy, etykiety zostaną utracone.

Zmienne z pojedynczymi wartościami nie są podczas dopasowywania rozkładu traktowane inaczej niż zmienne z wieloma wartościami. Zmienne z typem składowania czas, data lub znacznik czasu są traktowane jako numeryczne.

Dopasowywanie rozkładów do zmiennych podziału

Jeśli dane zawierają zmienną podziału, a dopasowywanie rozkładu ma zostać wykonane osobno dla każdego podziału, konieczne jest przeprowadzenie transformacji danych za pomocą węzła Restrukturyzacja. Korzystając z węzła restrukturyzacji, należy wygenerować nową zmienną dla każdej wartości zmiennej podziału. Restrukturyzowane dane mogą być następnie użyte do dopasowywania rozkładu w węźle dopasowania symulacji.

Węzeł Symulacje Dopasowanie — karta Ustawienia

Nazwa węzła źródłowego. Nazwę wygenerowanego (lub zaktualizowanego) węzła Symulacje Generowanie można utworzyć automatycznie, wybierając opcję **Automatycznie**. Wygenerowana automatycznie nazwa jest nazwą określonego węzła Symulacje Dopasowanie, jeśli określono nazwę niestandardową (lub jest to Sim Gen, jeśli w węźle dopasowania symulacji nie określono nazwy niestandardowej). Opcję **Użytkownika** należy wybrać, aby określić nazwę niestandardową w sąsiednim polu tekstowym. Jeśli pole tekstowe nie będzie edytowane, domyślną nazwą niestandardową pozostanie Sim Gen.

Opcje dopasowania Korzystając z tych opcji, można określić sposób dopasowania rozkładów oraz sposób oceny dopasowania rozkładów.

- **Liczba losowanych obserwacji.** Ta opcja określa liczbę obserwacji użytych podczas dopasowywania rozkładów do zmiennych w zbiorze danych. Opcja **Wszystkie obserwacje** umożliwia dopasowanie rozkładów do wszystkich rekordów w danych. Jeśli zbiór danych jest bardzo duży, warto rozważyć ograniczenie liczby obserwacji używanych do dopasowywania rozkładu. Aby użyć tylko N pierwszych obserwacji, należy wybrać opcję **Ogranicz do pierwszych N obserwacji**. Liczbę obserwacji, jakie mają zostać użyte, można określić, klikając strzałki. Alternatywnie można użyć węzła poprzedzającego, aby wybrać próbę losową rekordów do dopasowania rozkładu.
- **Kryterium dobroci dopasowania (ilościowe).** W przypadku zmiennych ilościowych należy wybrać test Anderson-Darling lub test dobroci dopasowania Kolmogorov-Smirnoff, aby porangować rozkłady podczas dopasowywania rozkładów do zmiennych. Test Anderson-Darling jest wybrany domyślnie i jest on szczególnie zalecany, aby zapewnić najlepsze możliwe dopasowanie w regionach końcowych. Obie statystyki są obliczane dla każdego potencjalnego rozkładu, ale tylko wybrana statystyka zostanie użyta do określenia kolejności rozkładów i ustalenia najlepiej dopasowanego rozkładu.
- **Przedziały (tylko rozkład empiryczny).** W przypadku zmiennych ilościowych rozkład Empiryczny jest funkcją skumulowanego rozkładu danych historycznych. Jest to prawdopodobieństwo dla każdej wartości lub zakresu wartości i jest wyliczane bezpośrednio na podstawie danych. Liczbę przedziałów, jakie będą użyte do obliczenia rozkładu empirycznego dla zmiennych ilościowych, można określić, klikając strzałki. Domyślną wartością jest 100, a maksymalną 1000.
- **Zmienna ważąca (opcjonalnie).** Jeśli zbiór danych zawiera zmienną ważącą, należy kliknąć ikonę wybierania zmiennych i wybrać zmienną ważącą z listy. Zmienna ważąca zostanie wykluczona z procesu dopasowywania rozkładu. Na liście wyświetlane są wszystkie zmienne w zbiorze danych, dla których wybrano ilościowy poziom pomiaru. Można wybrać tylko jedną zmienną ważącą.

Węzeł Symulacje Wynik

Węzeł Symulacje Wynik jest węzłem końcowym, który umożliwia ocenę określonej zmiennej, udostępnia rozkład dla zmiennej oraz tworzy wykresy rozkładu i korelacji. Ten węzeł jest głównie używany do oceny zmiennych ilościowych. Wypełnia wykres ewaluacyjny wygenerowany przez węzeł ewaluacyjny i jest przydatny do oceny zmiennych dyskretnych. Kolejna różnica polega na tym, że węzeł Symulacje Wynik ocenia pojedynczą predykcję dla kilku iteracji, podczas gdy węzeł ewaluacyjny ocenia wiele predykcji, każdą z pojedynczą iteracją. Iteracje są generowane, jeśli dla parametru rozkładu w węźle Symulacje Generowanie określono więcej niż jedną wartość. Więcej informacji można znaleźć w temacie “Iteracje” na stronie 61.

Węzeł oceny symulacji jest przeznaczony do użycia z danymi, które zostały uzyskane z węzłów dopasowania symulacji i generowania symulacji. Może jednak zostać użyty z każdym innym węzłem. Pomiędzy węzłem generowania symulacji i węzłem oceny symulacji można zamieścić dowolną liczbę kroków przetwarzania.

Ważne: Węzeł oceny symulacji wymaga co najmniej 1000 rekordów z poprawnymi wartościami dla zmiennej przewidywanej.

Węzeł Symulacje Wynik — karta Ustawienia

Na karcie Ustawienia węzła Symulacje Wynik można określić rolę każdej zmiennej w zbiorze danych oraz dostosować wyniki generowane po wykonaniu symulacji.

Wybierz element. Umożliwia przełączanie pomiędzy trzema widokami węzła Symulacje Wynik: Zmienne, Funkcje gęstości oraz Wyniki.

Widok Zmienne

Zmienna przewidywana. Jest to wymagana zmienna. Należy kliknąć strzałkę, aby z listy rozwianej wybrać zmienną przewidywaną zbioru danych. Wybrana zmienna może mieć poziom pomiaru ilościowy, porządkowy lub nominalny, ale nie może mieć poziomu pomiaru data lub nieokreślonego.

Zmienna iteracyjna (opcjonalnie). Jeśli w danych występuje zmienna iteracyjna wskazująca, do której iteracji przynależą poszczególne rekordy w danych, należy ją tutaj wybrać. Oznacza to, że każda iteracja będzie oceniana osobno. Wybrać można tylko zmienne z poziomem pomiaru: ilościowy, porządkowy lub nominalny.

Dane wejściowe są już posortowane zgodnie ze zmienną iteracyjną. Ta opcja jest aktywna, jeśli w polu **Zmienna iteracyjna (opcjonalnie)** określono zmienną iteracyjną. Tę opcję należy wybrać wyłącznie po uzyskaniu pewności, że dane wejściowe są już posortowane według zmiennej iteracyjnej w polu **Zmienna iteracyjna (opcjonalnie)**.

Maksymalna liczba iteracji do wykreślenia. Ta opcja jest aktywna, jeśli w polu **Zmienna iteracyjna (opcjonalnie)** określono zmienną iteracyjną. Liczbę iteracji, jakie mają zostać wykreślone, można określić, klikając strzałki. Określenie tej liczby pozwoli uniknąć podjęcia próby wykreślenia zbyt wielu iteracji na jednym wykresie, co mogłoby sprawić, że wykres będzie zbyt trudny do zinterpretowania. Najniższy poziom maksymalnej liczby iteracji, jaki można ustawić, to 2; poziom najwyższy to 50. Maksymalna liczba iteracji do wykreślenia jest początkowo ustawiona na wartość 10.

Zmienne wejściowe na wykresach tornado. Wykres korelacji tornado jest wykresem słupkowym, który wyświetla współczynniki korelacji pomiędzy określoną zmienną przewidywaną i wszystkimi określonymi danymi wejściowymi. Należy kliknąć ikonę wybierania zmiennych, aby z listy dostępnych zmiennych symulowanych wybrać zmienne wejściowe, jakie będą uwzględnione na wykresie tornado. Wybrane mogą zostać tylko zmienne wejściowe z ilościowym i porządkowym typem pomiaru. Zmienne wejściowe nominalne, bez określonego typu i typu data są niedostępne na liście i nie można ich wybrać.

Widok Funkcje gęstości

Za pomocą opcji w tym widoku można dostosować wyniki dla funkcji gęstości prawdopodobieństwa i funkcji skumulowanego rozkładu dla docelowej wartości ilościowej, a także wykresów słupkowych wartości przewidywanych dla przewidywanych zmiennych jakościowych.

Funkcje gęstości. Funkcje gęstości są głównymi średnimi sondowania zbioru wyników Twojej symulacji.

- **Funkcja gęstości prawdopodobieństwa (PDF).** Tę opcję należy wybrać, aby utworzyć funkcję gęstości prawdopodobieństwa dla zmiennej przewidywanej. Funkcja gęstości prawdopodobieństwa wyświetla rozkład wartości zmiennych przewidywanych. Funkcji tej można użyć do określenia prawdopodobieństwa, że zmienna przewidywana znajduje się w konkretnym regionie. Dla przewidywanych zmiennych jakościowych (zmiennych przewidywanych z poziomem pomiaru nominalnym lub porządkowym) generowany jest wykres słupkowy, który wyświetla procent obserwacji przypadających na każdą z kategorii zmiennej przewidywanej.

- **Funkcja skumulowanego rozkładu (CDF).** Tę opcję należy wybrać, aby utworzyć funkcję skumulowanego rozkładu dla zmiennej przewidywanej. Skumulowana funkcja gęstości wyświetla prawdopodobieństwo, że wartość zmiennej przewidywanej jest mniejsza lub równa określonej wartości. Jest ona dostępna tylko dla ciągłych zmiennych przewidywanych.

Linie odniesienia (ilościowe). Te opcje są dostępne, jeśli wybrano opcje **Funkcja gęstości prawdopodobieństwa (PDF)** i/lub **Funkcja skumulowanego rozkładu (CDF)**. Korzystając z tych opcji, do funkcji gęstości prawdopodobieństwa i funkcji skumulowanego rozkładu można dodać wiele stałych pionowych linii odniesienia.

- **Średnia.** Tę opcję należy wybrać, aby dodać linię odniesienia dla wartości średniej zmiennej przewidywanej.
- **Mediana.** Tę opcję należy wybrać, aby dodać linię odniesienia dla wartości mediany zmiennej przewidywanej.
- **Odchylenia standardowe.** Tę opcję należy wybrać, aby dodać linie odniesienia w miejscu wyznaczonym przez dodanie lub odjęcie określonej liczby standardowych odchyleń od średniej zmiennej przewidywanej. Zaznaczenie tej opcji powoduje aktywowanie sąsiedniego pola **Liczba**. Aby wybrać liczbę standardowych odchyleń, należy kliknąć przyciski strzałek. Minimalna liczba standardowych odchyleń wynosi 1, a maksymalna 10. Liczba standardowych odchyleń jest początkowo ustawiona na 3.
- **Percentyle.** Tę opcję należy wybrać, aby dodać linie odniesienia dla dwóch wartości percentyla rozkładu zmiennej przewidywanej. Po wybraniu tej opcji aktywowane są sąsiednie pola tekstowe **Dolny** i **Górny**. Przykładowo, wprowadzenie wartości 90 w polu tekstowym **Górny** spowoduje dodanie linii odniesienia dla 90. percentyla zmiennej przewidywanej, co oznacza wartość, poniżej której przypada 90% obserwacji. Podobnie, wartość 10 w polu **Dolny** reprezentuje dziesiąty percentyl zmiennej przewidywanej, co oznacza wartość, poniżej której przypada 10% obserwacji.
- **Niestandardowe linie referencyjne.** Tę opcję należy wybrać, aby dodać linie odniesienia dla określonej wartości na osi poziomej. Wybranie tej opcji powoduje aktywowanie sąsiedniej tabeli **Wartości**. Każde wprowadzenie poprawnej wartości w tabeli **Wartości** powoduje dodanie nowego pustego wiersza u dołu tabeli. *Poprawną* liczbą jest liczba z zakresu wartości zmiennej przewidywanej.

Uwaga: Jeśli na jednym wykresie wyświetlanych jest wiele funkcji gęstości lub funkcji rozkładu (z wielu iteracji), linie odniesienia (inne niż linie niestandardowe) są osobno stosowane dla każdej funkcji.

Przewidywana zmienna jakościowa (tylko PDF). Opcje te są dostępne tylko po zaznaczeniu opcji **Funkcja gęstości prawdopodobieństwa (PDF)**.

- **Wartości kategorii do zaraportowania.** W przypadku modeli z jakościowymi zmiennymi przewidywanymi wynik dla modelu jest zbiorem prawdopodobieństw — po jednym dla każdej kategorii w taki sposób, że wartość zmiennej przewidywanej pojawia się w każdej kategorii. Kategoria z najwyższym prawdopodobieństwem jest uznawana za przewidywaną kategorię i służy do generowania wykresu słupkowego dla funkcji gęstości prawdopodobieństwa. Aby wygenerować wykres słupkowy, należy wybrać opcję **Przewidywana kategoria**. Wybranie opcji **Przewidywane prawdopodobieństwa** umożliwia wygenerowanie histogramów rozkładu przewidywanych prawdopodobieństw dla każdej kategorii zmiennej przewidywanej. Można również wybrać opcję **Łącznie**, aby wygenerować oba typy wykresów.
- **Grupowanie w analizie czułości.** Symulacje, które obejmują iteracje analizy czułości generują niezależną zmienną przewidywaną (lub predykcyjną zmienną przewidywaną z modelu) dla każdej iteracji zdefiniowanej przez analizę. Na każdą różniącą się wartość parametru rozkładu przypada jedna iteracja. Jeśli istnieją iteracje, wykres słupkowy kategorii przewidywanej dla jakościowej zmiennej przewidywanej jest wyświetlany jako zgrupowany wykres słupkowy, zawierający wyniki dla wszystkich iteracji. Należy wybrać opcję **Grupuj razem kategorie** lub opcję **Grupuj razem iteracje**.

Widok Wyniki

Wartości percentyli dla rozkładów zmiennych przewidywanych. Te opcje umożliwiają utworzenie tabeli wartości percentyli dla rozkładów zmiennych przewidywanych oraz określenie percentyli do wyświetlenia.

Stwórz tabelę wartości percentyli. W przypadku docelowych wartości ilościowych tę opcję należy wybrać, aby utworzyć tabelę określonych percentyli dla rozkładów zmiennych przewidywanych. Aby określić percentyle, należy wybrać jedną z następujących opcji:

- **Kwartyle.** Kwartyle to 25., 50. i 75. percentyle rozkładu zmiennej przewidywanej. Obserwacje są podzielone na grupy jednakowej wielkości.
- **Przedziały.** Jeśli konieczne jest utworzenie równej liczby grup innej niż cztery, należy wybrać opcję **Przedziały**. Zaznaczenie tej opcji powoduje aktywowanie sąsiedniego pola **Liczba**. Kliknięcie strzałek umożliwia określenie liczby przedziałów. Minimalna liczba przedziałów to 2, a maksymalna to 100. Liczba przedziałów jest początkowo ustawiona na 10.
- **Wskazane percentyle.** Zaznaczenie opcji **Wskazane percentyle** umożliwia określenie pojedynczych percentyli, na przykład 99. percentyl. Wybranie tej opcji powoduje aktywowanie sąsiedniej tabeli **Wartości**. Każde wprowadzenie poprawnej wartości (od 1 do 100) w tabeli **Wartości** powoduje dodanie nowego pustego wiersza u dołu tabeli.

Wynik węzła oceny symulacji

Po wykonaniu węzła Symulacje Wynik wynik jest dodawany do menedżera wyników. Przeglądarka wyników węzła Symulacje Wynik wyświetla wyniki wykonania węzła Symulacje Wynik. W menu **Plik** dostępne są standardowe opcje zapisywania, eksportowania i drukowania, a menu **Edycja** udostępnia standardowe opcje edytowania. Więcej informacji można znaleźć w temacie “Wyświetlanie wyników” na stronie 303. Menu **Widok** jest aktywowane tylko w przypadku wybrania jednego z wykresów. Nie jest wyświetlane dla tabeli rozkładu lub wyników informacyjnych. Z menu **Widok** można wybrać opcję **Tryb edycji**, aby zmienić układ i wygląd wykresu, lub opcję **Tryb eksploracji**, aby przeprowadzić eksplorację danych i wartości reprezentowanych przez wykres. Tryb statyczny mocuje linie odniesienia wykresu (i suwaki) w ich bieżących położeniach, tak że nie można ich ruszyć. Tryb statyczny jest jedynym trybem, w którym można kopiować, eksportować lub drukować wykres wraz z liniami odniesienia. Aby wybrać ten tryb, należy kliknąć opcję **Tryb statyczny** w menu **Widok**.

Okno przeglądarki wyników węzła Symulacje Wynik składa się z dwóch paneli. Po lewej stronie okna znajduje się panel nawigacji, w którym wyświetlane są miniatury wykresów wygenerowane po wykonaniu węzła Symulacje Wynik. Po wybraniu miniatury wykres wynikowy jest wyświetlany na panelu po prawej stronie okna.

Panel nawigacji

Panel nawigacji przeglądarki wyników zawiera miniatury wykresów wygenerowane w wyniku symulacji. Miniatury wyświetlane na panelu nawigacji zależą od poziomu pomiaru zmiennej przewidywanej oraz od opcji wybranych w oknie dialogowym węzła Symulacje Wynik. Opisy miniatur zamieszczono w poniższej tabeli.

Tabela 46. Miniatury w panelu nawigacji

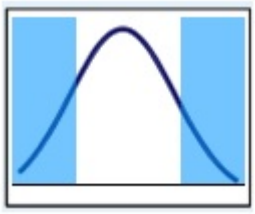
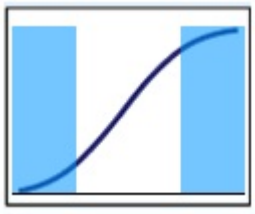
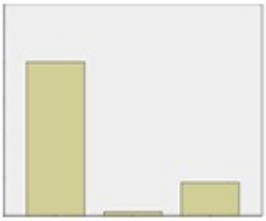
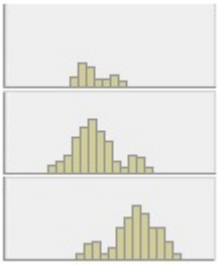
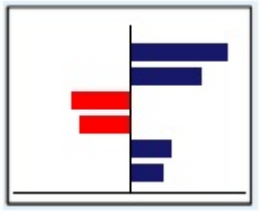
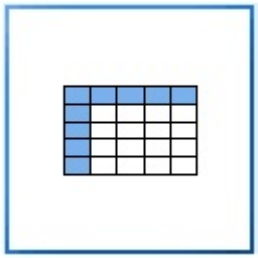
Miniatura	Opis	Komentarze
	Funkcja gęstości prawdopodobieństwa	Ta miniatura jest wyświetlana tylko wtedy, gdy poziom pomiaru zmiennej przewidywanej jest ilościowy, a w widoku funkcji gęstości w oknie dialogowym węzła Symulacje Wynik wybrano opcję **Funkcja gęstości prawdopodobieństwa (PDF)**. Jeśli poziom pomiaru zmiennej przewidywanej jest jakościowy, ta miniatura nie jest wyświetlana.
	Funkcja skumulowanego rozkładu	Ta miniatura jest wyświetlana tylko wtedy, gdy poziom pomiaru zmiennej przewidywanej jest ilościowy, a w widoku funkcji gęstości w oknie dialogowym węzła Symulacje Wynik wybrano opcję **Funkcja skumulowanego rozkładu (CDF)**. Jeśli poziom pomiaru zmiennej przewidywanej jest jakościowy, ta miniatura nie jest wyświetlana.

Tabela 46. Miniatury w panelu nawigacji (kontynuacja)

Miniatura	Opis	Komentarze
	Przewidywane wartości kategorii	Ta miniatura jest wyświetlana tylko wtedy, gdy poziom pomiaru zmiennej przewidywanej jest jakościowy, w widoku funkcji gęstości w oknie dialogowym węzła Symulacje Wynik wybrano opcję **Funkcja gęstości prawdopodobieństwa (PDF)**, a w obszarze **Wartości kategorii do raportu** wybrano opcję **Przewidywana kategoria** lub **Łącznie**. Jeśli poziom pomiaru zmiennej przewidywanej jest ilościowy, ta miniatura nie jest wyświetlana.
	Przewidywane prawdopodobieństwa kategorii	Ta miniatura jest wyświetlana tylko wtedy, gdy poziom pomiaru zmiennej przewidywanej jest jakościowy, w widoku funkcji gęstości w oknie dialogowym węzła Symulacje Wynik wybrano opcję **Funkcja gęstości prawdopodobieństwa (PDF)**, a w obszarze **Wartości kategorii do raportu** wybrano opcję **Przewidywane prawdopodobieństwa** lub **Łącznie**. Jeśli poziom pomiaru zmiennej przewidywanej jest ilościowy, ta miniatura nie jest wyświetlana.
	Wykresy tornado	Ta miniatura jest wyświetlana, jeśli w widoku zmiennych w oknie dialogowym węzła Symulacje Wynik w polu **Zmienne wejściowe na wykresach tornado** wybrano co najmniej jedną zmienną wejściową.
	Tabela rozkładu	Ta miniatura jest wyświetlana tylko wtedy, gdy poziom pomiaru zmiennej przewidywanej jest ilościowy, a w widoku wyników w oknie dialogowym węzła Symulacje Wynik wybrano opcję **Utwórz tabelę wartości percentyli**. Dla tego wykresu menu **Widok** jest wyłączone. Jeśli poziom pomiaru zmiennej przewidywanej jest jakościowy, ta miniatura nie jest wyświetlana.
	Informacje	Ta miniatura jest zawsze wyświetlana. Dla tego wyniku menu **Widok** jest wyłączone.

Wynik w postaci wykresu

Dostępne typy wykresów wynikowych zależą od poziomu pomiaru zmiennej przewidywanej, od tego, czy użyta została zmienna iteracji oraz od opcji wybranych w oknie dialogowym węzła Symulacje Wynik. Wiele z wykresów wygenerowanych przez symulację udostępnia interaktywne funkcje pozwalające na dostosowanie sposobu wyświetlania. Funkcje interaktywne są dostępne po kliknięciu opcji **Opcje wykresu**. Wszystkie wykresy symulacji są wizualizacjami danych.

Wykresy funkcji gęstości prawdopodobieństwa dla ciągłych zmiennych przewidywanych. Na tym wykresie przedstawiane jest prawdopodobieństwo i częstość, przy czym skala prawdopodobieństwa znajduje się na lewej osi pionowej, a częstotliwości na prawej osi pionowej. Na wykresie dostępne są dwie przesuwające się pionowe linie odniesienia dzielące wykres na osobne regiony. W tabeli poniżej wykresu wyświetlany jest procent rozkładu w każdym z regionów. Jeśli na jednym wykresie wyświetlanych jest kilka funkcji gęstości (ze względu na iteracje), w tabeli znajduje się osobny wiersz dla prawdopodobieństw powiązanych z każdą funkcją gęstości oraz dodatkową kolumnę, która zawiera nazwę iteracji i kolor powiązany z każdą funkcją gęstości. Iteracje są wyświetlane w tabeli w kolejności alfabetycznej, zgodnie z etykietą iteracji. Jeśli nie jest dostępna żadna etykieta iteracji, zamiast niej używana jest wartość iteracji. Tej tabeli nie można edytować.

Każda linia odniesienia ma suwak (odwrócony trójkąt), który umożliwia łatwe przesuwanie linii. Każdy suwak jest opatrzony etykietą wskazującą jego aktualne położenie. Domyślnie suwaki są ustawione na 5. i 95. percentylu rozkładu. W przypadku wielu iteracji suwaki są ustawione na 5. i 95. percentylu pierwszej iteracji z tabeli. Nie można przesuwać linii, tak aby krzyżowały się ze sobą.

Po kliknięciu opcji **Opcje wykresu** dostępne są dodatkowe funkcje. W szczególności, można wyraźnie ustawić pozycje suwaków, dodać stałe linie odniesienia i zmienić widok wykresu z krzywej ciągłej na histogram. Więcej informacji można znaleźć w temacie “Opcje: Wykresy” na stronie 339. Aby skopiować lub wyeksportować wykres, należy kliknąć go prawym przyciskiem myszy.

Wykresy funkcji skumulowanego rozkładu dla ciągłych zmiennych przewidywanych. Wykres ten ma takie same dwie przesuwalne pionowe linie odniesienia oraz powiązaną tabelę, które zostały opisane dla wykresu funkcji gęstości prawdopodobieństwa. W przypadku wielu iteracji elementy sterujące suwaka i tabela działają tak samo, jak funkcja gęstości prawdopodobieństwa. Dla funkcji rozkładu używane są te same kolory, jakie zostały użyte do określenia, która funkcja gęstości należy do poszczególnych iteracji.

Ten wykres zapewnia również dostęp do okna dialogowego Opcje wykresu, które umożliwia na bezpośrednie ustawienie pozycji suwaków, dodanie stałych linii odniesienia oraz określenie, czy funkcja skumulowanego rozkładu jest wyświetlana jako funkcja rosnąca (domyślnie) lub malejąca. Więcej informacji można znaleźć w temacie “Opcje: Wykresy” na stronie 339. Aby skopiować, wyeksportować lub edytować wykres, należy kliknąć go prawym przyciskiem myszy. Wybranie opcji **Edycja** otwiera wykres w ruchomym oknie edytora wizualizacji.

Wykres przewidywanych wartości kategorii dla przewidywanych zmiennych jakościowych. Dla przewidywanych zmiennych jakościowych na wykresie słupkowym wyświetlane są wartości przewidywane. Wartości przewidywane wyświetlane są jako procent zmiennej przewidywanej, dla której przewiduje się, że będzie należała w danej kategorii. W przypadku przewidywanych zmiennych jakościowych z iteracjami analizy czułości wyniki dla przewidywanej kategorii docelowej są wyświetlane jako zgrupowany wykres słupkowy obejmujący wyniki dla wszystkich iteracji. Wykres jest pogrupowany według kategorii lub według iteracji, w zależności od opcji, jaka została wybrana w obszarze **Grupowanie do analizy czułości** w widoku funkcji gęstości w oknie dialogowym węzła Symulacje Wynik. Aby skopiować, wyeksportować lub edytować wykres, należy kliknąć go prawym przyciskiem myszy. Wybranie opcji **Edycja** otwiera wykres w ruchomym oknie edytora wizualizacji.

Wykres przewidywanego prawdopodobieństwa kategorii dla przewidywanych zmiennych jakościowych. W przypadku przewidywanych zmiennych jakościowych histogram przedstawia rozkład przewidywanych prawdopodobieństw dla każdej kategorii zmiennej. Dla przewidywanych zmiennych jakościowych z iteracjami analizy czułości histogramy są wyświetlane według kategorii lub według iteracji, w zależności od tego, która opcja została wybrana w obszarze **Grupowanie do analizy czułości** w widoku funkcji gęstości w oknie dialogowym węzła Symulacje Wynik. Jeśli histogramy są pogrupowane według kategorii, lista rozwijana zawierająca etykiety iteracji umożliwia wybranie iteracji do wyświetlenia. Iteracje do wyświetlenia można również wybrać, klikając wykres

prawym przyciskiem myszy i wybierając odpowiednią iterację z podmenu **Iteracja**. Jeśli histogramy są pogrupowane według iteracji, lista rozwijana zawierająca nazwy kategorii umożliwia wybranie kategorii do wyświetlenia. Kategorie do wyświetlenia można również wybrać, klikając wykres prawym przyciskiem myszy i wybierając odpowiednią kategorię z podmenu **Kategoria**.

Ten wykres jest dostępny tylko dla podzbioru modeli i po wybraniu opcji generowania prawdopodobieństw wszystkich grup w modelu użytkowym. Przykładowo, w modelu użytkowym Regresja logistyczna należy wybrać opcję **Dołącz wszystkie prawdopodobieństwa**. Opcja ta jest dostępna w następujących modelach użytkowych:

- regresja logistyczna, SVM, Bayesa, sieć neuronowa i KNN
- modele eksploracji w bazie danych Db2/ISW dla regresji logistycznej, drzew decyzyjnych i Naïve Bayes

Domyślnie opcja generowania prawdopodobieństw wszystkich grup w tych modelach użytkowych nie jest zaznaczona.

Wykresy tornado. Wykres tornado to wykres słupkowy przedstawiający czułość zmiennej przewidywanej dla każdej określonej zmiennej wejściowej. Czułość jest mierzona na podstawie korelacji zmiennej przewidywanej z poszczególnymi zmiennymi wejściowymi. W tytule wykresu znajduje się nazwa zmiennej przewidywanej. Każdy słupek na wykresie reprezentuje korelacje pomiędzy zmienną przewidywaną a zmienną wejściową. Symulowane zmienne wejściowe uwzględniane na wykresie są zmiennymi wejściowymi wybranymi w polu **Zmienne wejściowe na wykresach tornado** w widoku funkcji gęstości w oknie dialogowym węzła Symulacje Wynik. Każdy słupek jest opatrzony etykietą z wartością korelacji. Słupki są uporządkowane według wartości bezwzględnej korelacji, od wartości największej do najmniejszej. Jeśli dostępne są iteracje, dla każdej z nich generowany jest osobny wykres. Każdy wykres zawiera podtytuł, w którym zapisana jest nazwa iteracji.

Tabela rozkładu. Ta tabela zawiera wartość zmiennej przewidywanej, poniżej której znajduje się określony procent obserwacji. Tabela zawiera wiersz dla każdej wartości percentyla, jaka została określona w widoku wyników w oknie dialogowym węzła Symulacje Wynik. Wartościami percentyla mogą być kwartyle, różna liczba równo rozmieszczonych percentyli lub indywidualnie określone percentyle. W tabeli rozkładu znajduje się kolumna dla każdej iteracji.

Informacje. W tej sekcji znajduje się ogólne podsumowanie zmiennych i rekordów użytych do ewaluacji. Zawiera ona również zmienne wejściowe oraz liczebności rekordów, rozbite dla każdej iteracji.

Opcje: Wykresy

Okno dialogowe Opcje wykresu pozwala na dostosowanie sposobu wyświetlania aktywnych wykresów funkcji gęstości prawdopodobieństwa i funkcji skumulowanego rozkładu wygenerowanych za pomocą symulacji.

Widok. Lista rozwijana **Widok** jest stosowana tylko do wykresu funkcji gęstości prawdopodobieństwa. Pozwala na przełączanie widoku wykresu między krzywą ciągłą a histogramem. Jeśli na jednym wykresie wyświetlanych jest kilka funkcji gęstości (dla kilku iteracji), ta funkcja jest wyłączona. W przypadku kilku funkcji gęstości funkcje te mogą być wyświetlane tylko jako krzywe ciągłe.

Porządek. Lista rozwijana **Porządek** jest stosowana tylko do wykresu funkcji skumulowanego rozkładu. Określa ona, czy funkcja skumulowanego rozkładu jest wyświetlana jako funkcja rosnąca (domyślnie) czy malejąca. Jeśli wyświetlana jest ona jako funkcja malejąca, wartość funkcji w danym punkcie na osi poziomej jest prawdopodobieństwem, że zmienna przewidywana znajduje się na prawo od tego punktu.

Pozycje suwaków. Pole tekstowe **Górna** zawiera bieżące położenie prawej przesuwanej linii odniesienia. Pole tekstowe **Dolna** zawiera bieżące położenie lewej przesuwanej linii odniesienia. Można bezpośrednio ustawić pozycje suwaków, wprowadzając wartości w polach tekstowych **Górna** i **Dolna**. Wartość w polu tekstowym **Dolna** musi być ściśle mniejsza niż wartość w polu tekstowym **Górna**. Można usunąć lewą linię odniesienia, wybierając opcję **-Nieskończoność**, co spowoduje ustawienie położenia linii jako minus nieskończoność. To działanie powoduje wyłączenie pola tekstowego **Dolna**. Można usunąć prawą linię odniesienia, wybierając opcję **Nieskończoność**, co spowoduje ustawienie położenia linii w nieskończoności. To działanie powoduje wyłączenie pola tekstowego **Górna**. Nie można usunąć obu linii odniesienia; wybranie opcji **-Nieskończoność** wyłącza pole wyboru **Nieskończoność** i odwrotnie.

Linie referencyjne. Do funkcji gęstości prawdopodobieństwa i funkcji skumulowanego rozkładu możesz dodać wiele stałych pionowych linii odniesienia.

- **Średnia.** Można dodać linię odniesienia w miejscu wartości średniej zmiennej przewidywanej.
- **Mediana.** Można dodać linię odniesienia w miejscu mediany zmiennej przewidywanej.
- **Odchylenia standardowe.** Linie odniesienia można dodać w położeniu plus i minus określonej liczby odchyleń standardowych od średniej zmiennej przewidywanej. Można wprowadzić liczbę standardowych odchyleń, jaka będzie użyta w sąsiednim polu tekstowym. Minimalna liczba standardowych odchyleń wynosi 1, a maksymalna 10. Liczba standardowych odchyleń jest początkowo ustawiona na 3.
- **Percentyle.** Można dodać linie odniesienia dla jednej lub dwóch wartości percentyli rozkładu dla zmiennej przewidywanej, wprowadzając wartości w polach tekstowych **Dolny** i **Górny**. Na przykład: wartość 95 w polu tekstowym **Górny** reprezentuje 95. percentyl, co jest wartością, poniżej której znajduje się 95% obserwacji. Podobnie, wartość 5 w polu tekstowym **Dolny** reprezentuje 5. percentyl, co jest wartością, poniżej której znajduje się 5% obserwacji. Dla pola tekstowego **Dolny** minimalna wartość percentyla wynosi 0, a maksymalna 49. Dla pola tekstowego **Górny** minimalna wartość percentyla wynosi 50, a maksymalna 100.
- **Pozycje niestandardowe.** Możesz dodać linie odniesienia o określonych wartościach wzdłuż osi poziomej. Można usunąć niestandardowe linie odniesienia, usuwając wpis z siatki.

Po kliknięciu przycisku **OK** suwaki, etykiety nad suwakami, linie odniesienia i tabela pod wykresem zostaną zaktualizowane, tak aby odzwierciedlały opcje wybrane w oknie dialogowym Opcje wykresu. Kliknięcie przycisku **Anuluj** spowoduje zamknięcie okna dialogowego bez wprowadzania zmian. Linie odniesienia można usunąć, usuwając zaznaczenie powiązanego wyboru dokonanego w oknie dialogowym Opcje wykresu i klikając przycisk **OK**.

Uwaga: Jeśli na jednym wykresie wyświetlanych jest wiele funkcji gęstości lub rozkładu (ze względu na wyniki z iteracji analiz czułości), linie odniesienia (inne niż linie niestandardowe) są osobno stosowane dla każdej funkcji. Wyświetlane są tylko linie odniesienia dla pierwszej iteracji. Etykiety linii odniesienia zawierają etykietę iteracji. Etykieta iteracji pochodzi z poprzedzającego węzła, zwykle z węzła Symulacja lub Utwórz. Jeśli nie jest dostępna żadna etykieta iteracji, zamiast niej używana jest wartość iteracji. Opcje **Średnia**, **Mediana**, **Odchylenia standardowe** i **Percentyle** są wyłączone dla funkcji skumulowanego rozkładu z wieloma iteracjami.

---

węzeł importowania przez rozszerzenie

Jeśli wybrano opcję **Wynik na ekran** na karcie **Wynik** okna dialogowego Rozszerzenie Wynik, wyniki będą wyświetlane na ekranie w oknie przeglądarki wyników. Wynik jest także dodawany do Menedżera wyników. Okno przeglądarki wyników zawiera własny zestaw menu, które umożliwiają drukowanie lub zapisywanie wyniku lub eksportowanie go do innego formatu. Menu **Edycja** zawiera tylko opcję **Kopiuj**. Przeglądarka wyników węzła Rozszerzenie Wynik ma dwie karty: kartę **Wynik tekstowy**, na której wyświetlane są wyniki tekstowe, i kartę **Wynik graficzny**, na której wyświetlane są wykresy.

Jeśli wybrano opcję **Wynik do pliku** na karcie **Wynik** okna dialogowego Rozszerzenie Wynik, przeglądarka wyników nie będzie wyświetlana po pomyślnym wykonaniu węzła Rozszerzenie Wynik.

Węzeł Rozszerzenie Wynik — karta Polecenia

Wybierz język poleceń: **R** albo **Python for Spark**. Więcej informacji można znaleźć w następujących sekcjach. Gdy polecenia będą gotowe, można kliknąć przycisk **Wykonaj**, aby wykonać węzeł Rozszerzenie Wynik. Obiekty wynikowe są dodawane do Menedżera wyników lub, opcjonalnie, do pliku określonego w polu **Nazwa pliku** na karcie **Wynik**.

Polecenia R

Polecenia R. Do tego pola można wpisać lub wkleić własny skrypt R służący do analizy danych.

Konwertuj zmienne typu flaga. Określa sposób traktowania zmiennych typu flaga. Dostępne są dwie opcje: **Łańcuchy na czynnik, liczby całkowite i rzeczywiste na liczby typu double** oraz **Wartości logiczne (Prawda,**

Fałsz). W przypadku wybrania opcji **Wartości logiczne (Prawda, Fałsz)** pierwotne wartości zmiennych typu flaga zostaną utracone. Na przykład, jeśli zmienna ma wartości Mężczyzna i Kobieta, to zostaną zamienione na Prawda i Fałsz.

Konwertuj brakujące wartości na wartość niedostępności danych (NA) pakietu R. Gdy ta opcja jest wybrana, wszelkie brakujące wartości są przekształcane w wartość NA w języku R. W języku R wartość NA oznacza brakujące wartości. Niektóre funkcje R przyjmują argument sterujący zachowaniem funkcji w przypadku, gdy dane zawierają wartość NA. Na przykład funkcja może oferować opcję automatycznego wykluczania rekordów zawierających wartość NA. Jeśli ta opcja nie będzie wybrana, wszelkie brakujące wartości będą przekazywane do skryptu R bez zmian, co może powodować błędy podczas jego wykonywania.

Konwertuj zmienne daty/czasu na klasy pakietu R ze specjalną kontrolą stref czasowych. Gdy ta opcja jest wybrana, zmienna typu data lub data/czas są przekształcane w obiekty date/time języka R. Należy wybrać jedną z następujących opcji:

- **R POSIXct.** Zmienne typu data lub data/czas są przekształcane w obiekty POSIXct języka R.
- **R POSIXlt (lista).** Zmienne typu data lub data/czas są przekształcane w obiekty POSIXlt języka R.

Uwaga: Formaty POSIX są opcjami zaawansowanymi. Opcji tych należy używać tylko wtedy, gdy w skrypcie R nakazano traktowanie zmiennych daty/czasu w sposób wymagający zastosowania tych formatów. Formaty POSIX nie mają zastosowania względem zmiennych z formatami czasu.

Polecenia Python

Polecenia Python. Do tego pola można wpisać lub wkleić własny skrypt Python służący do analizy danych. Aby uzyskać więcej informacji na temat języka Python for Spark, patrz Python for Spark i Pisanie skryptów w języku Python for Spark.

Węzeł Rozszerzenie Wynik — karta Wynik z konsoli

Karta **Wynik z konsoli** zawiera wszelkie wyniki odbierane podczas wykonywania skryptu w języku R lub Python for Spark na karcie Polecenia (na przykład, jeśli używany jest skrypt R, to na tej karcie wyświetlane są wyniki odbierane z konsoli R podczas wykonywania skryptu z pola **Polecenia R** na karcie **Polecenie**). Wyniki te mogą zawierać komunikaty o błędach lub ostrzeżenia generowane podczas wykonywania skryptu w języku R lub Python. Wyniki można wykorzystać przede wszystkim do debugowania skryptu. Karta **Wynik z konsoli** zawiera także skrypt z pola **Polecenia R** lub **Polecenia Python**.

Po każdym wykonaniu skryptu Rozszerzenie Wynik zawartość karty **Wynik z konsoli** jest nadpisywana wynikami z konsoli R lub środowiska Python for Spark. Wyników nie można edytować.

Węzeł Rozszerzenie Wynik — karta Wynik

Nazwa wyniku. Określa nazwę wyniku uzyskanego po wykonaniu węzła. Gdy wybrana jest opcja **Automatycznie**, wyniki automatycznie otrzymują nazwę „R Output” albo „Python Output”, w zależności od typu skryptu. Opcjonalnie można wybrać opcję **Użytkownika**, aby określić inną nazwę.

Wynik na ekran. Wybierz tę opcję, aby wygenerować i wyświetlić wynik w nowym oknie. Wynik jest także dodawany do Menedżera wyników.

Wynik do pliku. Ta opcja umożliwia zapisanie wyniku w pliku. Powoduje to włączenie przycisków opcji **Wykres wynikowy** i **Plik wynikowy**.

Wykres wynikowy. Opcja aktywna tylko wtedy, gdy wybrano opcję **Wynik do pliku**. Wybierz tę opcję, aby zapisać w pliku wszelkie wykresy będące wynikiem wykonania węzła Rozszerzenie Wynik. W polu **Nazwa pliku** określ nazwę pliku, w którym mają być zapisywane wygenerowane wyniki. Kliknij przycisk z wielokropkiem (**...**), aby wybrać konkretny plik i miejsce. Określ **Typ pliku** na liście rozwijanej Typ pliku. Dostępne są następujące typy plików:

- Obiekt wynikowy (.cou)

- HTML (.html)

Wynik tekstowy. Opcja aktywna tylko wtedy, gdy wybrano opcję **Wynik do pliku**. Wybierz tę opcję, aby zapisać w pliku wszelkie wyniki tekstowe wykonania węzła Rozszerzenie Wynik. W polu **Nazwa pliku** określ nazwę pliku, w którym mają być zapisywane wygenerowane wyniki. Kliknij przycisk z wielokropkiem (**...**), aby określić konkretny plik i miejsce. Określ **Typ pliku** na liście rozwijanej Typ pliku. Dostępne są następujące typy plików:

- HTML (.html)
- Obiekt wynikowy (.cou)
- Dokument tekstowy (.txt)

Przeglądarka wyników rozszerzeń

Jeśli wybrano opcję **Wynik na ekran** na karcie **Wynik** okna dialogowego Rozszerzenie Wynik, wyniki będą wyświetlane na ekranie w oknie przeglądarki wyników. Wynik jest także dodawany do Menedżera wyników. Okno przeglądarki wyników zawiera własny zestaw menu, które umożliwiają drukowanie lub zapisywanie wyniku lub eksportowanie go do innego formatu. Menu **Edycja** zawiera tylko opcję **Kopiuj**. Przeglądarka wyników węzła Rozszerzenie Wynik zawiera dwie karty:

- Na karcie **Wynik tekstowy** wyświetlane są wyniki tekstowe
- Na karcie **Wynik graficzny** wyświetlane są wykresy

Jeśli na karcie **Wynik** okna dialogowego węzła Rozszerzenie Wynik wybrano opcję **Wynik do pliku**, a nie **Wynik na ekran**, to po pomyślnym wykonaniu węzła Rozszerzenie Wynik okno przeglądarki wyników nie będzie wyświetlane.

Przeglądarka wyników rozszerzenia — karta Wynik z konsoli

Na karcie **Wyniki tekstowy** wyświetlane są wszelkie wyniki tekstowe generowane wykonywania skryptu R lub Python for Spark na karcie **Polecenia** węzła Rozszerzenie Wynik.

Uwaga: Komunikaty o błędach lub ostrzeżenia języka R lub Python for Spark generowane podczas wynikowania skryptu są zawsze wyświetlane na karcie **Wynik z konsoli** węzła Rozszerzenie Wynik.

Przeglądarka wyników rozszerzenia — karta Wynik graficzny

Na karcie **Wyniki graficzny** wyświetlane są wszelkie wykresy generowane wykonywania skryptu R lub Python for Spark na karcie **Polecenia** węzła Rozszerzenie Wynik. Na przykład, jeśli skrypt R zawiera wywołanie funkcji plot języka R, to na tej karcie wyświetlony zostanie wynikowy wykres.

Węzły KDE

Jądrowy estymator gęstości — Kernel Density Estimation (KDE)© — używa algorytmów Ball Tree lub KD Tree do efektywnej obsługi zapytań i działa na pograniczu między uczeniem nienadzorowanym, generowaniem cech (feature engineering) i modelowaniem danych. Do najpopularniejszych i najbardziej użytecznych technik estymacji gęstości należą metody oparte na analizie sąsiedztwa, takie jak KDE. Algorytm KDE może być realizowany w dowolnej liczbie wymiarów, jednak w praktyce duża liczba wymiarów powoduje pogorszenie wydajności. Węzły Modelowanie KDE i Symulacja KDE w produkcie SPSS Modeler eksponują podstawowe funkcje i często używane parametry biblioteki KDE. Węzły są zaimplementowane w języku Python. [1]

Aby użyć węzła KDE, należy skonfigurować poprzedzający węzeł Typ. Węzeł KDE odczyta wprowadzane wartości z węzła Typ (lub karty Typy poprzedzającego węzła źródłowego).

Węzeł **Modelowanie KDE** jest dostępny na kartach Modelowanie i Python w programie SPSS Modeler. Węzeł KDE Modeling generuje model użytkowy, a wartości oceniane przez model użytkowy są gęstościami jądra z danych wejściowych.

Węzeł **Symulacja KDE** jest dostępny na karcie wyników i karcie Python. Węzeł Symulacja KDE generuje węzeł źródłowy Gen. KDE, który może utworzyć rekordy o tym samym rozkładzie, co dane wejściowe. Węzeł Gen. KDE zawiera kartę Ustawienia, na której można określić liczbę utworzonych rekordów (domyślnie 1) i wygenerować wartość startową generatora liczb losowych.

Więcej informacji na temat algorytmów KDE, wraz z przykładami, znajduje się w dokumentacji algorytmów KDE dostępnej pod adresem http://scikit-learn.org/stable/modules/density.html#kernel-density-estimation. [1]

[1] "User Guide." *Kernel Density Estimation*. WWW. © 2007-2018, scikit-learn developers.

Węzeł Modelowanie KDE węzeł Symulacja KDE — Zmienne

Na karcie Zmienne określone są zmienne, które będą używane w analizie.

Użyj wstępnie zdefiniowanych ról. Ta opcja używa ustawień wejściowych z wcześniejszego węzła Typ (lub z karty Typy poprzedzającego węzła źródłowego).

Użyj niestandardowych przypisań. Wybierz tę opcję, aby ręcznie przypisać dane wejściowe.

Pola. Za pomocą przycisków strzałek przypisz elementy z listy ręcznie na listę danych wejściowych po prawej stronie ekranu. Ikony wskazują prawidłowe poziomy pomiaru dla każdej zmiennej. Aby wybrać wszystkie zmienne z listy, należy kliknąć przycisk **Wszystkie** lub kliknąć przycisk poziomu pojedynczego pomiaru, aby wybrać wszystkie zmienne dla tego poziomu pomiaru.

Zmienne wejściowe. Wybierz jedną lub więcej zmiennych jako dane wejściowe dla grupowania. Węzeł KDE może działać tylko na zmiennych ciągłych.

Węzły KDE — Opcje budowania

Karta Opcje budowania umożliwia określenie opcji budowania dla węzłów KDE, w tym **opcji podstawowych** dotyczących parametrów gęstości jądra oraz etykiet skupień, a także **opcji zaawansowanych** takich jak tolerancja, wielkość liścia i stosowanie metody „najpierw szerokość”. Więcej informacji na temat tych opcji można znaleźć w następujących źródłach internetowych:

- Skorowidz parametrów węzła jądrowej estymacji gęstości w interfejsie API środowiska Python[1]
- Podręcznik użytkownika jądrowej estymacji gęstości[2]

Podstawowe

Przepustowość. Określ przepustowość jądra.

Jądro. Wybierz jądro (algorytm domyślny), które ma być używane. Jądra dostępne dla węzła Modelowanie KDE to: **Gaussian**, **Tophat**, **Epanechnikov**, **Wykładniczy**, **Liniowy** i **Cosinus**. Jądra dostępne dla węzła Symulacja KDE to: **Gaussian** i **Tophat**. Szczegółowe informacje o dostępnych jądrach zawiera Podręcznik użytkownika jądrowej estymacji gęstości.[2]

Algorytm. Jako algorytm drzewa wybierz **Automatyczny**, **Ball Tree** lub **Drzewo KD**. Aby uzyskać więcej informacji — patrz Ball Tree[3] i KD Tree.[4]

Metric. Wybierz metrykę odległości. Dostępne są metryki: **Euclidean**, **Braycurtis**, **Chebyshev**, **Canberra**, **Cityblock**, **Dice**, **Hamming**, **Infinity**, **Jaccard**, **L1**, **L2**, **Matching**, **Manhattan**, **P**, **Rogerstanimoto**, **Russellrao**, **Sokalmichener**, **Sokalsneath**, **Kulsinski** i **Minkowski**. W przypadku wybrania miary **Minkowski** należy ustawić żądaną wartość w polu **Wartość P**.

To, które metryki dostępne będą w tym menu rozwijanym, zależy od wybranego algorytmu. Należy także zwrócić uwagę, że normalizacja wynikowej gęstości jest prawidłowa tylko dla metryki Euclidean.

Zaawansowane

Tolerancja bezwzględna. Określ żądaną tolerancję bezwzględną wyniku. Większa tolerancja z reguły przyspiesza wykonanie algorytmu. Wartością domyślną jest **0,0**.

Tolerancja względna. Określ żądaną tolerancję względną wyniku. Większa tolerancja z reguły przyspiesza wykonanie algorytmu. Wartością domyślną jest **1E-8**.

Wielkość liścia. Określ wielkość liścia podstawowego drzewa. Wartością domyślną jest **40**. Zmiana wielkości liścia może istotnie wpłynąć na wydajność oraz na zapotrzebowanie na pamięć. Aby uzyskać więcej informacji o algorytmach Ball Tree i KD Tree — patrz Ball Tree[3] i KD Tree.[4]

Najpierw szerokość. Wybierz opcję **Prawda**, jeśli ma być stosowana metoda „najpierw szerokość", a **Fałsz**, jeśli ma być stosowana metoda „najpierw głębokość".

W poniższej tabeli przedstawiono relację między ustawieniami w oknach dialogowych węzłów KDE w programie SPSS Modeler a parametrami biblioteki KDE w środowisku Spark.

Tabela 47. Właściwości węzła odwzorowane na parametry biblioteki Python

Ustawienie w programie SPSS Modeler	Nazwa w skryptach (nazwa właściwości)	Parametr KDE
Predyktory	dane wejściowe	
Przepustowość	bandwidth	bandwidth
Jądro	kernel	kernel
Algorytm	algorithm	algorithm
Metryka	metric	metric
Wartość P	pValue	pValue
Tolerancja bezwzględna	atol	atol
Tolerancja względna	rtol	Rtol
Wielkość liścia	leafSize	leafSize
Najpierw szerokość	breadthFirst	breadthFirst

[1] "API Reference." *sklearn.neighbors.KernelDensity*. WWW. © 2007-2018, scikit-learn developers.

[2] "User Guide." *Kernel Density Estimation*. WWW. © 2007-2018, scikit-learn developers.

[3] "Ball Tree." *Five balltree construction algorithms*. © 1989, Omohundro, S.M., International Computer Science Institute Technical Report.

[4] "K-D Tree." *Multidimensional binary search trees used for associative searching*. © 1975, Bentley, J.L., Communications of the ACM.

Węzeł Modelowanie KDE i węzeł Symulacja KDE — Opcje modelu

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

IBM SPSS Statistics — aplikacje pomocnicze

Jeśli kompatybilna wersja programu IBM SPSS Statistics została zainstalowana na komputerze i użytkownik ma odpowiednią licencję, można skonfigurować program IBM SPSS Modeler, tak aby przetwarzał dane za pomocą funkcji IBM SPSS Statistics przy użyciu węzłów Przekształcenia Statistics, Model Statistics, Wynik Statistics lub Eksport Statistics.

Informacje na temat kompatybilności z bieżącą wersją programu IBM SPSS Modeler można znaleźć na stronie pomocy technicznej pod adresem *http://www.ibm.com/support*.

Aby skonfigurować program IBM SPSS Modeler, tak aby możliwa była praca z produktem IBM SPSS Statistics i innymi aplikacjami, należy wybrać:

Narzędzia > Opcje > Aplikacje pomocnicze

IBM SPSS Statistics — interaktywnie. Wpisz pełną ścieżkę i nazwę komendy (na przykład, C:\Program Files\IBM\SPSS\Statistics\<nn>\stats.exe), która zostanie użyta podczas uruchamiania produktu IBM SPSS Statistics bezpośrednio z pliku danych utworzonego przez węzeł eksportu Plik Statistics. Więcej informacji można znaleźć w “Węzeł eksportu Statistics” na stronie 363.

Połączenie. Jeśli serwer IBM SPSS Statistics znajduje się na tym samym hoście co IBM SPSS Modeler Server, można aktywować połączenie między dwoma aplikacjami, co spowoduje zwiększenie efektywności, dzięki pozostawieniu danych na serwerze podczas analizy. Należy wybrać opcję **Serwer**, aby aktywować poniższą opcję **Port**. Ustawienie domyślne to **Lokalne**.

Port. Należy określić port serwera dla serwera IBM SPSS Statistics.

IBM SPSS Statistics Location Utility. Aby program IBM SPSS Modeler mógł korzystać z węzłów Przekształcenia Statistics, Model Statistics i Wynik Statistics, na komputerze, na którym uruchamiany jest strumień, musi być zainstalowana kopia programu IBM SPSS Statistics i użytkownik musi mieć odpowiednią licencję.

- Jeśli program IBM SPSS Modeler działa w trybie lokalnym (autonomicznym), na komputerze lokalnym musi być zainstalowana licencjonowana kopia produktu IBM SPSS Statistics. Należy kliknąć ten przycisk, aby określić miejsce lokalnej instalacji programu IBM SPSS Statistics na potrzeby licencji.
- Ponadto w przypadku działania w trybie rozproszonym dla serwera zdalnego IBM SPSS Modeler Server konieczne jest również uruchomienie programu narzędziowego na hoście IBM SPSS Modeler Server, aby utworzyć plik *statistics.ini*, który będzie wskazywał programowi IBM SPSS Statistics ścieżkę instalacji produktu IBM SPSS Modeler Server. W tym celu należy bezpośrednio z wiersza komend zmienić katalog *bin* programu IBM SPSS Modeler Server i, w systemie Windows, uruchomić komendę:

  ```
  statisticsutility -location=<IBM SPSS Statistics_ścieżka_instalacji>/
  ```

  Alternatywnie w systemie UNIX należy uruchomić komendę:

  ```
  ./statisticsutility -location=<IBM SPSS Statistics_ścieżka_instalacji>/bin
  ```

Jeśli użytkownik nie ma licencjonowanej kopii programu IBM SPSS Statistics na lokalnym komputerze, nadal można uruchomić węzeł Plik Statistics za pośrednictwem serwera IBM SPSS Statistics, ale próba uruchomienia innych węzłów IBM SPSS Statistics spowoduje wyświetlenie komunikatu o błędzie.

Komentarze

W razie problemów z uruchomieniem węzłów procedur programu IBM SPSS Statistics należy zapoznać się z następującymi wskazówkami:

- Jeśli nazwy zmiennych używane w programie IBM SPSS Modeler składają się z więcej niż ośmiu znaków (dla wersji wcześniejszych niż IBM SPSS Statistics 12.0), więcej niż 64 znaków (wersja IBM SPSS Statistics 12.0 i kolejne) lub zawierają niepoprawne znaki, konieczna jest zmiana nazwy lub skrócenie nazw, zanim będą mogły zostać wczytane do programu IBM SPSS Statistics. Więcej informacji można znaleźć w “Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics” na stronie 364.

- Jeśli program IBM SPSS Statistics był zainstalowany później niż program IBM SPSS Modeler, konieczne może być określenie lokalizacji IBM SPSS Statistics, w sposób omówiony wcześniej.

Rozdział 7. Węzły eksportu

Przegląd węzłów eksportu

Węzły eksportu udostępniają mechanizm do eksportowania danych w różnych formatach, tak aby były dostępne za pomocą innych narzędzi.

Dostępne są następujące węzły eksportu:

Węzeł eksportu do bazy danych zapisuje dane w relacyjnym źródle danych zgodnym z ODBC. Aby możliwy był zapis w źródle danych ODBC, źródło danych musi istnieć, a użytkownik musi mieć prawo do jego obsługi.

Dane wyjściowe węzła eksportu Plik płaski zapisywane są w pliku tekstowym z danymi rozgraniczonymi. Jest to przydatne podczas eksportowania danych, które mogą być odczytywane przez inne oprogramowanie do przeprowadzania analizy lub obsługujące arkusz kalkulacyjny.

Dane wynikowe węzła eksportu Plik Statistics zapisywane są w formacie IBM SPSS Statistics: *.sav* lub *.zsav*. Pliki .sav lub .zsav mogą być odczytywane przez produkty IBM SPSS Statistics Base i inne. Jest to również format używany przez pliki pamięci podręcznej w programie IBM SPSS Modeler.

Dane wyjściowe węzła eksportu Data Collection są w formacie używanym przez oprogramowanie do badań rynku Data Collection. Aby możliwe było korzystanie z tego węzła, konieczne jest zainstalowanie programu Data Data Collection Data Library.

Węzeł eksportu IBM Cognos eksportuje dane w formacie możliwym do odczytania przez bazy danych Cognos.

Węzeł eksportu IBM Cognos TM1 eksportuje dane w formacie możliwym do odczytania przez bazy danych Cognos TM1.

Dane wyjściowe węzła eksportu Plik SAS są zapisywane w formacie SAS, aby mogły zostać odczytane w systemie SAS lub za pomocą pakietu oprogramowania kompatybilnego z systemem SAS. Dostępne są trzy pliki w formacie SAS: SAS dla Windows/OS2, SAS dla UNIX lub SAS wersja 7/8.

Węzeł eksportu Plik Excel powoduje zapisanie danych wyjściowych w pliku Microsoft Excel w formacie .xlsx. Opcjonalnie można zdecydować, aby program Excel był uruchamiany automatycznie i otwierał wyeksportowany plik po wykonaniu węzła.

Węzeł eksportu XML tworzy wyniki dla danych w pliku w formacie XML. Opcjonalnie można utworzyć węzeł źródłowy XML, aby wczytać wyeksportowane dane z powrotem do strumienia.

Węzeł eksportu JSON generuje dane wyjściowe w formacie JSON. Więcej informacji można znaleźć w “Węzeł Eksport JSON” na stronie 374.

Węzeł eksportu do bazy danych

Węzły Baza danych umożliwiają zapis danych w relacyjnych źródłach danych zgodnych z ODBC, co zostało objaśnione w opisie węzła źródłowego bazy danych. Więcej informacji można znaleźć w temacie “Węzeł źródłowy bazy danych” na stronie 17.

Poniżej przedstawiono ogólne kroki pozwalające na zapisanie danych w bazie danych:

1. Zainstaluj sterownik ODBC i skonfiguruj źródło danych dla bazy danych, jaka będzie używana.
2. Na karcie eksportu węzła bazy danych określ źródło danych i tabelę, jaka ma zostać użyta do zapisu. Można utworzyć nową tabelę lub wstawić dane do istniejącej.
3. W razie potrzeby określ opcje dodatkowe.

Te kroki zostały opisane bardziej szczegółowo w kilku kolejnych tematach.

Węzeł Baza danych — karta eksportu

Uwaga: Niektóre bazy danych, do których można eksportować, mogą nie obsługiwać nazw kolumn w tabelach, jeśli ich długość przekracza 30 znaków. Jeśli wyświetlany jest komunikat o błędzie informujący, że tabela zawiera niepoprawną nazwę kolumny, należy skrócić nazwę, tak aby liczba znaków nie przekraczała 30.

Źródło danych. Wyświetla wybrane źródło danych. Należy wpisać nazwę lub wybrać ją z listy rozwijanej. Jeśli na liście nie ma żądanej bazy danych, należy wybrać opcję **Dodaj nowe połączenie z bazą danych** i zlokalizować bazę danych w oknie dialogowym Połączenia z bazą danych. Więcej informacji można znaleźć w “Dodawanie połączenia z bazą danych” na stronie 19.

Nazwa tabeli. Należy wpisać nazwę tabeli, do której dane mają zostać wysłane. Po wybraniu opcji **Wstaw do tabeli** można wybrać tabelę istniejącą w bazie danych, klikając przycisk **Wybierz**.

Utwórz tabelę. Tę opcję należy wybrać, aby utworzyć nową tabelę bazy danych lub zastąpić istniejącą.

Wstaw do tabeli. Tę opcję należy wybrać, aby wstawić dane jako nowe wiersze w istniejącej tabeli bazy danych.

Połącz w tabeli. (O ile dostępna) Tę opcję należy wybrać, aby zaktualizować wybrane kolumny bazy danych przez wartości z odpowiednich zmiennych danych źródłowych. Po wybraniu tej opcji dostępny jest przycisk **Łączenie**, który wyświetla okno dialogowe umożliwiające mapowanie zmiennych danych źródłowych na kolumny bazy danych.

Porzuć istniejącą tabelę. Tę opcję należy wybrać, aby podczas tworzenia nowej tabeli usunąć istniejące tabele o takiej samej nazwie.

Usuń istniejące wiersze. Tę opcję należy wybrać, aby podczas wstawiania do tabeli usunąć z niej wiersze istniejące przed wyeksportowaniem.

Uwaga: Jeśli zaznaczone zostaną dwie z powyższych opcji, podczas wykonywania węzła wyświetlony zostanie komunikat **Nadpisz ostrzeżenie**. Aby pominąć ostrzeżenia, należy usunąć zaznaczenie opcji **Ostrzegaj przed nadpisywaniem tabeli w bazie danych** na karcie Powiadomienia w oknie dialogowym Opcje użytkownika.

Domyślny rozmiar tekstu. Pola oznaczone jako bez typu we wcześniejszym węźle Typy są zapisywane w bazie danych jako zmienne łańcuchowe. Należy określić wielkość łańcuchów, jakie będą stosowane dla zmiennych bez typu.

Kliknięcie opcji **Schemat** umożliwia otwarcie okna dialogowego, w którym można ustawić różne opcje eksportu (w przypadku baz danych, które obsługują tę funkcję), ustawić typy danych SQL dla zmiennych oraz określić klucz główny dla indeksowania bazy danych. Więcej informacji można znaleźć w “Opcje schematu eksportu do bazy danych” na stronie 350.

Aby określić opcje indeksowania wyeksportowanej tabeli w celu zwiększenia wydajności, należy kliknąć opcję **Indeksy**. Więcej informacji można znaleźć w “Opcje indeksu eksportu do bazy danych” na stronie 352.

Po kliknięciu opcji **Zaawansowane** można określić opcje ładowania dużych zbiorów danych i zatwierdzania bazy danych. Więcej informacji można znaleźć w “Zaawansowane opcje eksportu do bazy danych” na stronie 354.

Ujmij w cudzysłów nazwy tabeli i kolumny. Opcje wyboru używane podczas wysyłania instrukcji CREATE TABLE do bazy danych. Tabele lub kolumny zawierające spacje lub znaki niestandardowe muszą być ujęte w cudzysłów.

- **W razie potrzeby.** Tę opcję należy zaznaczyć, aby program IBM SPSS Modeler automatycznie ustalał, kiedy ujęcie w cudzysłów jest konieczne w pojedynczym przypadku.
- **Zawsze.** Po wybraniu tej opcji nazwy tabeli i kolumn zawsze będą ujmowane w cudzysłów.
- **Nigdy.** Ta opcja powoduje wyłączenie funkcji stosowania cudzysłowów.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby wygenerować węzeł źródłowy bazy danych dla danych eksportowanych do określonego źródła i tabeli danych. W czasie wykonywania ten węzeł jest dodawany do obszaru roboczego strumienia.

Opcje łączenia przy eksporcie bazy danych

To okno dialogowe umożliwia mapowanie zmiennych ze źródła danych na kolumny w docelowej tabeli bazy danych. Jeśli zmienna danych źródłowych jest mapowana na kolumnę bazy danych, po uruchomieniu strumienia wartość w kolumnie jest zastępowana wartością danych źródłowych. Niezmapowane zmienne źródłowe pozostają niezmienione w bazie danych.

Mapuj zmienne. W tym miejscu można określić mapowanie pomiędzy zmiennymi danych źródłowych a kolumnami bazy danych. Zmienne danych źródłowych z takimi samymi nazwami jak kolumny w bazie danych są mapowane automatycznie.

- **Mapuj.** Mapuje zmienną danych źródłowych wybraną z listy zmiennych po lewej stronie przycisku na kolumnę bazy danych wybraną z listy po prawej stronie. Można jednocześnie zmapować więcej niż jedną zmienną, ale liczba pozycji wybranych z obu list musi być taka sama.
- **Usuń mapowanie.** Usuwa mapowanie dla co najmniej jednej wybranej kolumny bazy danych. Ten przycisk jest aktywowany po wybraniu zmiennej lub kolumny bazy danych w tabeli po prawej stronie okna dialogowego.
- **Dodaj.** Dodaje co najmniej jedną zmienną danych źródłowych wybraną z listy zmiennych po lewej stronie przycisku na listę po prawej stronie gotową do mapowania. Ten przycisk jest aktywowany, jeśli wybrano zmienną z listy po lewej stronie, a żadna zmienna o takiej nazwie nie istnieje na liście po prawej stronie. Kliknięcie tego przycisku spowoduje zmapowanie wybranej zmiennej na nową kolumnę bazy danych z taką samą nazwą. Słowo **<NEW>** (Nowa) jest wyświetlane po nazwie kolumny bazy danych, wskazując, że jest to nowa zmienna.

Połącz wiersze. Do połączenia rekordów z taką samą wartością w zmiennej kluczowej używana jest zmienna kluczowa, taka jak *Transaction ID* (Id. transakcji). Jest to odpowiednik złączenia równościowego bazy danych. Wartości kluczy muszą być wartościami kluczy głównych; oznacza to, że muszą być unikalne i nie mogą zawierać wartości null.

- **Dostępne klucze.** Lista wszystkich zmiennych w źródłach danych wejściowych. Należy wybrać co najmniej jedną zmienną z listy i za pomocą przycisku strzałki i dodać je jako zmienne kluczowe w celu połączenia rekordów. Każda zmienna mapy z odpowiadającą jej zmapowaną kolumną bazy danych jest dostępna jako kluczowa, z wyjątkiem zmiennych dodanych jako nowe kolumny bazy danych (z słowem **<NEW>** (nowa) po nazwie), które są niedostępne.
- **Użyte klucze.** Wyświetla listę wszystkich zmiennych użytych do połączenia rekordów z wszystkich źródeł danych wejściowych na podstawie wartości zmiennych kluczowych. Aby usunąć klucz z listy, należy go wybrać i za pomocą przycisku strzałki przenieść z powrotem na listę Dostępne klucze. Jeśli wybranych zostanie kilka zmiennych kluczowych, aktywowana jest poniższa opcja.
- **Uwzględnij tylko rekordy istniejące w bazie danych.** Umożliwia przeprowadzenie częściowego złączenia; jeśli rekord znajduje się w bazie danych i w strumieniu, zmapowane zmienne zostaną zaktualizowane.
- **Dodaj rekordy do bazy danych.** Umożliwia wykonanie złączenia zewnętrznego; wszystkie rekordy w strumieniu zostaną połączone (o ile taki sam rekord istnieje w bazie danych) lub dodane (jeśli rekord nie istnieje jeszcze w bazie danych).

Aby zmapować zmienną danych źródłowych na nową kolumnę bazy danych

1. Kliknij nazwę zmiennej źródłowej na liście po lewej stronie, pod opcją **Mapuj zmienne**.
2. Kliknij przycisk **Dodaj**, aby wykonać mapowanie.

Aby zmapować zmienną danych źródłowych na istniejącą kolumnę bazy danych

1. Kliknij nazwę zmiennej źródłowej na liście po lewej stronie, pod opcją **Mapuj zmienne**.
2. Kliknij nazwę kolumny pod obszarem **Kolumna bazy danych** po prawej stronie.
3. Kliknij przycisk **Mapuj**, aby wykonać mapowanie.

Aby usunąć mapowanie

1. Na liście po prawej stronie, pod obszarem Zmienna, kliknij nazwę zmiennej, dla której zamierzasz usunąć mapowanie.
2. Kliknij przycisk **Usuń mapowanie**.

Aby usunąć zaznaczenie zmiennej na wszystkich listach

Naciśnij i przytrzymaj klawisz CTRL i kliknij nazwę zmiennej.

Opcje schematu eksportu do bazy danych

W oknie dialogowym Schemat eksportu do bazy danych można ustawić opcje eksportu (dla baz danych, które obsługują te opcje), ustawić typy danych SQL dla zmiennych, określić, które zmienne są kluczami głównymi, i dostosować instrukcję CREATE TABLE (Utwórz tabelę) generowaną podczas eksportu.

Okno dialogowe składa się z kilku części:

- Sekcja w górnej części (o ile jest wyświetlana) zawiera opcje dla eksportu do bazy danych, która te opcje obsługuje. Ta sekcja nie jest wyświetlana w przypadku braku połączenia z odpowiednią bazą danych.
- Pole tekstowe na środku wyświetla szablon użyty do wygenerowania polecenia CREATE TABLE (Utwórz tabelę), które domyślnie ma następujący format:

  CREATE TABLE <table-name> <(table columns)>
- Tabela w dolnej części umożliwia określenie typu danych SQL dla każdej zmiennej oraz wskazanie zmiennych będących kluczami głównymi (w sposób omówiony poniżej). To okno dialogowe automatycznie generuje wartości parametrów <table-name> (nazwa-tabeli) i <(table columns)> (kolumny tabeli) na podstawie specyfikacji w tabeli.

Ustawianie opcji eksportu do bazy danych

Jeśli ta sekcja jest wyświetlana, można określić wiele ustawień eksportu do bazy danych. Poniżej przedstawiono typy baz danych obsługujących tę funkcję.

- Wydania produktu SQL Server Enterprise i Developer. Więcej informacji można znaleźć w temacie "Opcje dla bazy danych SQL Server".
- Wydania produktu Oracle Enterprise lub Personal. Więcej informacji można znaleźć w temacie "Opcje dla bazy danych Oracle" na stronie 352.

Dostosowywanie instrukcji CREATE TABLE (Utwórz tabelę)

Za pośrednictwem części pola tekstowego w tym oknie dialogowym do instrukcji CREATE TABLE (Utwórz tabelę) można dodać dodatkowe opcje specyficzne dla bazy danych.

1. Zaznacz pole wyboru **Dostosuj komendę CREATE TABLE (utwórz tabelę)**, aby aktywować okno tekstowe.
2. Dodaj do instrukcji opcje specyficzne dla bazy danych. Pamiętaj, aby zachować parametry tekstowe <table-name> (nazwa tabeli) i (<table-columns>) (kolumny tabeli), ponieważ będą one zastępowane przez rzeczywiste definicje nazwy tabeli i kolumny przez program IBM SPSS Modeler.

Ustawianie typów danych SQL

Domyślnie program IBM SPSS Modeler umożliwia serwerowi bazy danych automatyczne przypisanie typów danych SQL. Aby zastąpić automatyczny typ dla zmiennej, należy znaleźć wiersz odpowiadający tej zmiennej i wybrać odpowiedni typ z listy rozwijanej w kolumnie *Type* (Typ) tabeli schematu. Aby wybrać więcej niż jeden wiersz, można użyć metody Shift+kliknięcie.

W przypadku typów, które obejmują argumenty długości, precyzji lub skali (BINARY, VARBINARY, CHAR, VARCHAR, NUMERIC i NUMBER), to użytkownik powinien określić długość; nie należy zezwalać na automatyczne przypisanie długości przez serwer bazy danych. Przykładowo, określenie wartości wrażliwej, takiej jak VARCHAR(25), dla długości spowoduje zastąpienie typu składowania w programie IBM SPSS Modeler, o ile taki był zamysł użytkownika. Aby zastąpić przypisanie automatyczne, należy wybrać opcję **Określ** z listy rozwijanej typu i zastąpić definicję typu odpowiednią instrukcją definicji typu SQL.

Aby to zrobić w najprostszy sposób, należy najpierw wybrać typ najbardziej zbliżony do żądanej definicji typu, a następnie wybrać opcję **Określ**, aby przeprowadzić edycję tej definicji. Przykładowo, aby ustawić typ danych SQL na VARCHAR(25), najpierw należy z listy rozwijanej typu wybrać typ **VARCHAR(length)**, a następnie wybrać opcję **Określ** i zastąpić tekst length (długość) wartością 25.

Klucze główne

Jeśli w co najmniej jednej kolumnie w wyeksportowanej tabeli muszą znajdować się unikalne wartości lub kombinację wartości dla każdego wiersza, można to określić, zaznaczając pole wyboru **Klucz główny** dla każdej zmiennej, dla której ma to zastosowanie. W większości baz danych nie ma możliwości modyfikowania tabel w sposób unieważniający ograniczenie klucza głównego i automatycznie nad kluczem głównym tworzony jest indeks, ułatwiający wymuszenie tego ograniczenia. (Opcjonalnie można utworzyć indeksy dla innych zmiennych w oknie dialogowym Indeksy. Więcej informacji można znaleźć w temacie "Opcje indeksu eksportu do bazy danych" na stronie 352.)

Opcje dla bazy danych SQL Server

Użyj kompresji. Po zaznaczeniu tej opcji tworzone są tabele do eksportu z zastosowaniem kompresji.

Kompresja dla. Należy wybrać poziom kompresji.

- **Wiersz.** Umożliwia kompresję na poziomie wiersza (np. odpowiednik zapisu CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = ROW); w języku SQL).
- **Strona.** Umożliwia kompresję na poziomie strony (np. CREATE TABLE MYTABLE(...) WITH (DATA_COMPRESSION = PAGE); in SQL).

Opcje dla bazy danych Oracle

Ustawienia dla bazy danych Oracle — opcja podstawowa

Użyj kompresji. Po zaznaczeniu tej opcji tworzone są tabele do eksportu z zastosowaniem kompresji.

Kompresja dla. Należy wybrać poziom kompresji.

- **Domyślny.** Umożliwia zastosowanie domyślnej kompresji (np. zapis CREATE TABLE MYTABLE(...) COMPRESS; w języku SQL). W takim przypadku uzyskany zostanie efekt taki sam, jak po zastosowaniu opcji **Podstawowe**.
- **Podstawowe.** Umożliwia zastosowanie podstawowej kompresji (np. zapis CREATE TABLE MYTABLE(...) COMPRESS BASIC; w języku SQL).

Ustawienia dla bazy danych Oracle — opcja zaawansowana

Użyj kompresji. Po zaznaczeniu tej opcji tworzone są tabele do eksportu z zastosowaniem kompresji.

Kompresja dla. Należy wybrać poziom kompresji.

- **Domyślny.** Umożliwia zastosowanie domyślnej kompresji (np. zapis CREATE TABLE MYTABLE(...) COMPRESS; w języku SQL). W takim przypadku uzyskany zostanie efekt taki sam, jak po zastosowaniu opcji **Podstawowe**.
- **Podstawowe.** Umożliwia zastosowanie podstawowej kompresji (np. zapis CREATE TABLE MYTABLE(...) COMPRESS BASIC; w języku SQL).
- **OLTP.** Umożliwia zastosowanie kompresji OLTP (np. zapis CREATE TABLE MYTABLE(...)COMPRESS FOR OLTP; w języku SQL).
- **Zapytanie Niska/Wysoka.** (Tylko serwery Exadata) Umożliwia zastosowanie hybrydowej kompresji kolumnowej dla zapytania (np. zapis CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY LOW; lub CREATE TABLE MYTABLE(...)COMPRESS FOR QUERY HIGH; w języku SQL). Kompresja zapytania jest przydatna w środowiskach zajmujących się magazynowaniem danych; WYSOKA oznacza wyższy współczynnik kompresji niż NISKA.
- **Archiwizacja Niska/Wysoka.** (Tylko serwery Exadata) Umożliwia zastosowanie hybrydowej kompresji kolumnowej dla archiwum (np. zapis CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE LOW; lub CREATE TABLE MYTABLE(...)COMPRESS FOR ARCHIVE HIGH; w języku SQL). Kompresja dla archiwum jest przydatna do kompresji danych, które będą składowane przez długi okres; WYSOKA oznacza wyższy współczynnik kompresji niż NISKA.

Opcje indeksu eksportu do bazy danych

Okno dialogowe Indeksy umożliwia tworzenie indeksów w tabelach baz danych wyeksportowanych z programu IBM SPSS Modeler. Można określić zbiory zmiennych, jakie mają zostać uwzględnione i w razie potrzeby dostosować komendę CREATE INDEX (Utwórz indeks).

Okno dialogowe składa się z dwóch części:

- Pole tekstowe w górnej części wyświetla szablon, jaki może zostać użyty do wygenerowania jednej lub kilku komend CREATE INDEX, które domyślnie mają następujący format:

  CREATE INDEX <index-name> ON <table-name>

- Tabela w dolnej części okna dialogowego umożliwia dodawanie specyfikacji dla każdego indeksu, jaki ma zostać utworzony. Dla każdego indeksu należy określić nazwę indeksu oraz zmienne lub kolumny, jakie mają zostać uwzględnione. Okno dialogowe automatycznie generuje odpowiednio wartości parametrów <index-name> (nazwa indeksu) i <table-name> (nazwa tabeli).

  Przykładowo, wygenerowany kod SQL dla pojedynczego indeksu zmiennych *empid* (id. pracownika) i *deptid* (id. oddziały) może wyglądać następująco:

  CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID)

Aby utworzyć kilka indeksów, można dodać kilka wierszy. Dla każdego wiersza generowana jest osobna komenda CREATE INDEX.

Dostosowywanie komendy CREATE INDEX (Utwórz indeks)

Opcjonalnie można dostosować komendę CREATE INDEX dla wszystkich indeksów lub tylko dla wybranego. Zapewnia to elastyczność podczas dostosowywania konkretnych wymagań lub opcji bazy danych oraz podczas stosowania zmian do wszystkich indeksów lub tylko do wybranych, odpowiednio do potrzeb.

- Należy wybrać opcję **Dostosuj komendę CREATE INDEX** w górnej części okna dialogowego, aby zmodyfikować szablon używany dla wszystkich dodanych później indeksów. Należy pamiętać, że zmiany nie zostaną automatycznie zastosowane do indeksów, które zostały już dodane do tabeli.
- Należy wybrać co najmniej jeden wiersz w tabeli, a następnie kliknąć przycisk **Aktualizuj wybrane indeksy** w górnej części okna dialogowego, aby zastosować bieżące zmiany do wszystkich wybranych wierszy.
- Zaznaczenie pola wyboru **Dostosuj** w każdym wierszu umożliwia zmodyfikowanie szablonu komendy tylko dla danego indeksu.

Należy pamiętać, że wartości parametrów <index-name> (nazwa indeksu) i <table-name> (nazwa tabeli) są generowane automatycznie za pośrednictwem okna dialogowego w oparciu o specyfikacje tabeli i nie można ich bezpośrednio edytować.

Słowo kluczowe BITMAP. W przypadku użycia bazy danych Oracle można dostosować szablon, aby utworzyć indeks bitmapowy, a nie standardowy indeks w następujący sposób:

```
CREATE BITMAP INDEX <index-name> ON <table-name>
```

Indeksy bitmapowe mogą być przydatne podczas indeksowania kolumn z niewielką liczbą odmiennych wartości. Końcowa instrukcja SQL może wyglądać następująco:

```
CREATE BITMAP INDEX MYTABLE_IDX1 ON MYTABLE(COLOR)
```

Słowo kluczowe UNIQUE. Większość baz danych obsługuje słowo kluczowe UNIQUE w komendzie CREATE INDEX. Wymusza ono ograniczenie dotyczące unikalności, podobne do ograniczenia dla klucza głównego w tabeli podstawowej.

```
CREATE UNIQUE INDEX <index-name> ON <table-name>
```

Należy pamiętać, że w przypadku zmiennych rzeczywiście wyznaczonych jako klucze główne, ta specyfikacja nie jest konieczna. Większość baz danych automatycznie tworzy indeks dla wszystkich zmiennych określony jako zmienne klucza głównego za pośrednictwem komendy CREATE TABLE, dlatego jawne tworzenie indeksów dla tych zmiennych nie jest konieczne. Więcej informacji można znaleźć w temacie “Opcje schematu eksportu do bazy danych” na stronie 350.

Słowo kluczowe FILLFACTOR. Niektóre parametry fizyczne dla indeksu mogą zostać dostosowane. Przykładowo, baza danych SQL Server umożliwia użytkownikowi wyśrodkowanie wielkości indeksu (po początkowym utworzeniu) w odniesieniu do kosztów utrzymania na podstawie przyszłych zmian wprowadzonych w tabeli.

```
CREATE INDEX MYTABLE_IDX1 ON MYTABLE(EMPID,DEPTID) WITH FILLFACTOR=20
```

Pozostałe komentarze

- Jeśli istnieje już indeks z określoną nazwą, utworzenie indeksu nie powiedzie się. Wszelkie błędy będą początkowo traktowane jako ostrzeżenia, umożliwiając utworzenie kolejnych indeksów, a następnie po zakończeniu próby utworzenia wszystkich indeksów będą zgłaszane jako błędy w dzienniku komunikatów.
- Aby zapewnić jak najlepszą wydajność, indeksy powinny być tworzone po załadowaniu danych do tabeli. Indeksy muszą zawierać co najmniej jedną kolumnę.

- Przed wykonaniem węzła można wyświetlić podgląd wygenerowanego kodu SQL w dzienniku komunikatów.
- W przypadku tabel tymczasowych zapisanych w bazie danych (to znaczy przy włączonym buforowaniu węzła) opcje do określania kluczy głównych i indeksów są niedostępne. System może jednak utworzyć indeksy w tabeli tymczasowej, w zależności od danych użytych w kolejnych węzłach. Przykładowo, jeśli zbuforowane dane są następnie połączone przez kolumnę *DEPT*, sensowne będzie poindeksowanie zbuforowanych tabel w tej kolumnie.

Optymalizacja indeksów i zapytań

W niektórych systemach zarządzania bazami danych po utworzeniu, załadowaniu i poindeksowaniu bazy danych wymagany jest dodatkowy krok, zanim optymalizator będzie mógł wykorzystać indeksy do przyspieszenia wykonywania zapytania w nowej tabeli. W przypadku bazy danych Oracle optymalizator zapytania opartego na kosztach wymaga przeanalizowania tabeli, zanim jej indeksy będą mogły zostać użyte podczas optymalizacji zapytania. Wewnętrzny plik właściwości ODBC dla bazy danych Oracle (niewidoczny dla użytkownika) zawiera następującą opcję wykonania tej czynności:

```
# Defines SQL to be executed after a table and any associated indexes
# have been created and populated
table_analysis_sql, 'ANALYZE TABLE <table-name> COMPUTE STATISTICS'
```

Ten krok jest wykonywany podczas każdego tworzenia bazy danych Oracle (niezależnie od tego, czy definiowane są klucze główne czy indeksy). W razie konieczności plik właściwości ODBC dla dodatkowych baz danych może być dostosowany w podobny sposób — w celu uzyskania pomocy należy skontaktować się z działem wsparcia technicznego.

Zaawansowane opcje eksportu do bazy danych

Po kliknięciu przycisku Zaawansowane w oknie dialogowym węzła eksportu do bazy danych wyświetlane jest nowe okno dialogowe, w którym można określić techniczne szczegóły dotyczące eksportowania wyników do bazy danych.

Zastosuj wprowadzanie wsadowe. Tę opcję należy wybrać, aby wyłączyć zatwierdzanie w bazie danych wiersz po wierszu.

Liczba rekordów we wsadzie. Określa liczbę rekordów do wysłania do bazy danych przed wprowadzeniem do pamięci. Zmniejszenie tej liczby zapewni większą integralność danych kosztem niższej szybkości transferu. W celu zapewnienia optymalnej wydajności dla bazy danych tę liczbę można dostosować.

Ładowanie dużych zbiorów danych. Określa metodę wsadowego ładowania danych do bazy danych bezpośrednio z programu IBM SPSS Modeler. Aby wybrać, które opcje ładowania wsadowego są odpowiednie dla konkretnego scenariusza, konieczne będzie przeprowadzenie prób.

- **Przez sterowniki ODBC.** Tę opcję należy wybrać, aby do wykonania wstawiania do wielu wierszy użyć interfejsu API ODBC, co zapewni większą wydajność niż w przypadku normalnego eksportu do bazy danych. W opcjach poniżej można wybrać powiązanie wierszowe lub kolumnowe.
- **Przez zewnętrzny program ładujący.** Tę opcję należy wybrać, aby użyć niestandardowego programu ładującego, specyficznego dla danej bazy danych. Zaznaczenie tej opcji aktywuje różne opcje poniżej. Jeśli w systemie AIX wystąpią problemy z kodowaniem danych w standardzie UTF-8, konieczne może być dodanie wpisu locale, en_US.UDF-8 do pliku options.cfg.

Zaawansowane opcje ODBC. Te opcje są dostępne tylko po zaznaczeniu opcji **Przez sterowniki ODBC**. Należy pamiętać, że ta funkcja może nie być obsługiwana przez wszystkie sterowniki ODBC.

- **Wierszami.** Powiązanie wierszowe pozwala użyć wywołania SQLBulkOperations do załadowania danych do bazy danych. Powiązanie wierszowe pozwala zwiększyć szybkość pracy w porównaniu do wstawiania parametryzowanego, w którym dane są wstawiane oddzielnie do poszczególnych rekordów.
- **Kolumnami.** Tę opcję należy wybrać, aby do załadowania danych do bazy danych użyć powiązania kolumnowego. Powiązanie kolumnowe zwiększa wydajność poprzez powiązanie każdej kolumny bazy danych (w sparametryzowanej instrukcji INSERT) z tablicą składającą się z *N* wartości. Jednokrotne wykonanie instrukcji INSERT powoduje wstawienie do bazy danych *N* wierszy. Ta metoda może znacząco zwiększyć wydajność.

Opcje zewnętrznego programu ładującego. Jeśli wybrana jest opcja **Przez zewnętrzny program ładujący**, wyświetlane są różne opcje eksportowania zbioru danych do pliku oraz umożliwiające określanie i wykonywanie działań niestandardowego programu ładującego w celu załadowania danych z tego pliku do bazy danych. Program IBM SPSS Modeler może współpracować z zewnętrznymi programami ładującymi wielu popularnych systemów baz danych. W oprogramowaniu zawarto kilka skryptów, które są dostępne w dokumentacji technicznej w podkatalogu scripts (skrypty). Należy pamiętać, że aby użyć tej funkcji konieczne jest zainstalowanie programu Python 2.7 na tym samym komputerze co program IBM SPSS Modeler lub IBM SPSS Modeler Server, a w pliku options.cfg należy ustawić parametr python_exe_path. Więcej informacji można znaleźć w temacie "Programowanie ładowania wsadowego".

- **Stosuj separator.** Określa, jaki znak separatora powinien być użyty w eksportowanym pliku. Należy wybrać opcję **Tabulator**, aby jako separatora użyć znaku tabulacji lub opcję **Spacja**, aby użyć spacji. Wybranie opcji **Inne** pozwala określić inny znak, na przykład przecinek (,).
- **Określ plik danych.** Tę opcję należy wybrać, aby wprowadzić ścieżkę, jaka będzie użyta podczas zapisywania pliku danych podczas ładowania wsadowego. Domyślnie w katalogu tymczasowym na serwerze tworzony jest plik tymczasowy.
- **Określ program ładujący.** Tę opcję należy wybrać, aby określić program do ładowania wsadowego. Domyślnie oprogramowanie wyszukuje w podkatalogu scripts (skrypty) instalacji programu IBM SPSS Modeler skrypt Python, aby wykonać go dla określonej bazy danych. W oprogramowaniu zawarto kilka skryptów, które są dostępne w dokumentacji technicznej w podkatalogu scripts (skrypty).
- **Generuj dziennik.** Ta opcja umożliwia wygenerowanie pliku dziennika w określonym katalogu. Plik dziennika zawiera informacje o błędach i jest przydatny, jeśli ładowanie wsadowe zakończy się niepowodzeniem.
- **Sprawdź rozmiar tabeli.** Wybranie tej opcji pozwala sprawdzić tabelę, aby upewnić się, że zwiększenie wielkości tabeli odpowiada liczbie wierszy wyeksportowanych z programu IBM SPSS Modeler.
- **Dodatkowe opcje programu ładującego.** Określa dodatkowe argumenty dla programu ładującego. W przypadku argumentów zawierających spacje należy użyć znaków podwójnego cudzysłowu.

Aby uwzględnić podwójne cudzysłowy w opcjonalnych argumentach, należy poprzedzić je lewym ukośnikiem. Na przykład, opcja określona jako -comment "This is a \"comment\"" zawiera flagę -comment (komentarz) i sam komentarz renderowany jako This is a "comment" (To jest "komentarz").

Pojedynczy lewy ukośnik może być poprzedzony kolejnym lewym ukośnikiem. Na przykład opcja określona jako -specialdir "C:\\Test Scripts\\" zawiera flagę -specialdir i katalog directory renderowany jako C:\Test Scripts\.

Programowanie ładowania wsadowego

Węzeł eksportu Baza danych obejmuje opcje ładowania wsadowego dostępne w oknie dialogowym Opcje zaawansowane. Programy ładowania wsadowego mogą być używane do ładowania danych z pliku tekstowego do bazy danych.

Opcja **Ładowanie dużych zbiorów danych - Przez zewnętrzny program ładujący** konfiguruje program IBM SPSS Modeler, tak aby wykonane zostały trzy zadania:

- Utworzenie wymaganych tabel baz danych.
- Wyeksportowanie danych do pliku tekstowego.
- Wywołanie programu ładowania wsadowego, aby załadował dane z pliku do tabeli bazy danych.

Zwykle program ładowania wsadowego nie jest narzędziem do ładowania bazy danych (na przykład narzędzie sqlldr Oracle), ale małym skryptem lub programem, który formuje poprawne argumenty, tworzy pliki dodatkowe specyficzne dla bazy danych (takie jak plik kontrolny), a następnie wywołuje narzędzie do ładowania bazy danych. Informacje w kolejnych sekcjach ułatwią edycję istniejącego programu ładowania wsadowego.

Alternatywnie można napisać własny program do ładowania wsadowego. Więcej informacji można znaleźć w temacie "Opracowywanie programów ładowania wsadowego" na stronie 359. Należy pamiętać, że nie jest to objęte standardową umową dotyczącą wsparcia technicznego i w celu uzyskania wsparcia należy skontaktować się z przedstawicielem serwisu usług IBM.

Skrypty ładowania wsadowego

Program IBM SPSS Modeler jest dostarczany wraz z programami ładowania wsadowego przeznaczonymi dla różnych baz danych, które są implementowane za pośrednictwem skryptów Python. Po uruchomieniu strumienia zawierającego węzeł eksportu Baza danych z wybraną opcją **Przez zewnętrzny program ładujący** program IBM SPSS Modeler tworzy tabelę bazy danych (o ile jest wymagana) za pośrednictwem skryptów ODBC, eksportuje dane do pliku tymczasowego na hoście, na którym działa program IBM SPSS Modeler Server, a następnie wywołuje skrypt ładowania wsadowego. Ten skrypt z kolei powoduje wykonanie programów narzędziowych udostępnionych przez dostawcę DBMS, aby pobrać dane z plików tymczasowych do bazy danych.

Uwaga: Instalacja programu IBM SPSS Modeler nie obejmuje instalacji interpretera Pythona, dlatego konieczna jest osobna instalacja programu Python. Więcej informacji można znaleźć w temacie “Zaawansowane opcje eksportu do bazy danych” na stronie 354.

Skrypty są udostępnione (w folderze *\scripts* w katalogu instalacyjnym produktu IBM SPSS Modeler) dla baz danych wymienionych w poniższej tabeli.

Tabela 48. Udostępnione skrypty programu ładowania wsadowego

Baza danych	Nazwa skryptu	Dalsze informacje
IBM Db2	db2_loader.py	Więcej informacji można znaleźć w temacie “Ładowanie wsadowe danych do baz danych IBM Db2”.
IBM Netezza	netezza_loader.py	Więcej informacji można znaleźć w temacie “Ładowanie wsadowe danych do baz danych IBM Netezza” na stronie 357.
Oracle	oracle_loader.py	Więcej informacji można znaleźć w temacie “Ładowanie wsadowe danych do baz danych Oracle” na stronie 357.
SQL Server	mssql_loader.py	Więcej informacji można znaleźć w temacie “Ładowanie wsadowe danych do bazy danych SQL Server” na stronie 358.
Teradata	teradata_loader.py	Więcej informacji można znaleźć w temacie “Ładowanie wsadowe danych do baz danych Teradata” na stronie 359.

Ładowanie wsadowe danych do baz danych IBM Db2

Omówione poniżej punkty mogą ułatwić skonfigurowanie ładowania wsadowego z programu IBM SPSS Modeler do bazy danych IBM Db2 przy użyciu opcji Przez zewnętrzny program ładujący dostępnej w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane.

Należy upewnić się, czy zainstalowano program narzędziowy do przetwarzania wiersza komend Db2 (CLP).

Skrypt db2_loader.py wywołuje komendę Db2 LOAD. Należy upewnić się, czy program do przetwarzania wiersza komend (db2 w systemie UNIX, db2cmd w systemie Windows) jest zainstalowany na serwerze, na którym wykonywany będzie skrypt db2_loader.py (zwykle jest to host, na którym działa program IBM SPSS Modeler Server).

Należy sprawdzić, czy nazwa aliasu lokalnej bazy danych jest taka sama, jak rzeczywista nazwa bazy danych.

Alias lokalnej bazy danych Db2 to nazwa używana przez oprogramowanie kliencie DB2 do odnoszenia się do bazy danych znajdującej się w lokalnej lub zdalnej instancji Db2. Jeśli alias lokalnej bazy danych różni się od nazwy zdalnej bazy danych, wprowadzić dodatkową opcję programu ładującego:

```
-alias <local_database_alias>
```

Przykładowo, nazwa zdalnej bazy danych to STARS na hoście GALAXY, ale alias lokalnej bazy danych Db2 na hoście, na którym działa program IBM SPSS Modeler Server, to STARS_GALAXY. Należy użyć dodatkowej opcji programu ładującego

```
-alias STARS_GALAXY
```

Kodowanie danych z użyciem znaków innych niż ASCII

Jeśli ładowanie wsadowe obejmuje dane, które nie są zapisane w formacie ASCII, należy upewnić się, czy zmienna strony kodowej w sekcji konfiguracji skryptu db2_loader.py jest poprawnie skonfigurowana w systemie.

Puste łańcuchy znaków

Puste łańcuchy znaków są eksportowane do bazy danych jako wartości NULL.

Ładowanie wsadowe danych do baz danych IBM Netezza

Omówione poniżej punkty mogą ułatwić skonfigurowanie ładowania wsadowego z programu IBM SPSS Modeler do bazy danych IBM Netezza przy użyciu opcji Przez zewnętrzny program ładujący dostępnej w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane.

Należy upewnić się, czy zainstalowano program narzędziowy nzload Netezza.

Skrypt *netezza_loader.py* wywołuje program narzędziowy *nzload* Netezza. Należy upewnić się, czy program *nzload* jest zainstalowany i poprawnie skonfigurowany na serwerze, na którym wykonany zostanie skrypt *netezza_loader.py*.

Eksportowanie danych w formacie innym niż ASCII

Jeśli eksport obejmuje dane, które nie są zapisane w formacie ASCII, konieczne może być dodanie zapisu -encoding UTF8 w polu **Dodatkowe opcje programu ładującego** w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane. Dzięki temu dane w formacie innym niż ASCII powinny zostać poprawnie pobrane.

Dane w formacie data, czas i znacznik czasu

We właściwościach strumienia należy ustawić format daty na **DD-MM-RRRR**, a format godziny na **GG:MM:SS**.

Puste łańcuchy znaków

Puste łańcuchy znaków są eksportowane do bazy danych jako wartości NULL.

Inna kolejność kolumn w strumieniu i tabeli docelowej podczas wstawiania danych do istniejącej tabeli

Jeśli kolejność kolumn w strumieniu różni się od kolejności w tabeli docelowej, wartości danych zostaną wstawione do niewłaściwych kolumn. Należy użyć węzła Reorganizacja, aby upewnić się, że kolejność kolumn w strumieniu jest zgodna z kolejnością w tabeli docelowej. Więcej informacji można znaleźć w temacie “Węzeł Reorganizacja” na stronie 183.

Śledzenie postępów programu nzload

Jeśli program IBM SPSS Modeler działa w trybie lokalnym, należy dodać wpis -sts w polu **Dodatkowe opcje programu ładującego** w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane, aby w oknie komend otwartym za pomocą programu *nzload* co 10000 wierszy wyświetlane były komunikaty o statusie.

Ładowanie wsadowe danych do baz danych Oracle

Omówione poniżej punkty mogą ułatwić skonfigurowanie ładowania wsadowego z programu IBM SPSS Modeler do bazy danych Oracle przy użyciu opcji Przez zewnętrzny program ładujący dostępnej w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane.

Należy upewnić się, czy zainstalowano program narzędziowy sqlldr Oracle.

Skrypt *oracle_loader.py* wywołuje program narzędziowy *sqlldr* Oracle. Należy pamiętać, że program *sqlldr* nie jest automatycznie uwzględniany na kliencie Oracle. Należy upewnić się, czy program *sqlldr* jest zainstalowany na serwerze, na którym ma zostać wykonany skrypt *oracle_loader.py*.

Należy określić identyfikator SID bazy danych lub nazwę usługi

Jeśli dane są eksportowane na nielokalny serwer Oracle lub jeśli na lokalnym serwerze Oracle znajduje się wiele baz danych, konieczne będzie określenie w polu **Dodatkowe opcje programu ładującego** w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane następującego wpisu, umożliwiającego przekazanie identyfikatora SID lub nazwy usługi:

```
-database <SID>
```

Edytowanie sekcji konfiguracji w skrypcie oracle_loader.py

W systemach UNIX (i opcjonalnie w systemach Windows) należy przeprowadzić edycję sekcji konfiguracji na początku skryptu *oracle_loader.py*. Tutaj można określić wartości dla zmiennych środowiskowych ORACLE_SID, NLS_LANG, TNS_ADMIN i ORACLE_HOME (o ile mają zastosowanie), a także pełną ścieżkę programu narzędziowego *sqlldr*.

Dane w formacie data, czas i znacznik czasu

We właściwościach strumienia należy zwykle ustawić format daty na **RRRR-MM-DD**, a format godziny na **GG:MM:SS**.

Jeśli konieczne jest użycie innych formatów daty i godziny, należy zapoznać się z dokumentacją Oracle i dokonać edycji pliku skryptu *oracle_loader.py*.

Kodowanie danych z użyciem znaków innych niż ASCII

Jeśli ładowanie wsadowe dotyczy danych, które nie są zapisane w formacie ASCII, należy upewnić się, czy zmienna środowiskowa NLS_LANG została poprawnie skonfigurowana w systemie. Jest ona odczytywana przez program narzędziowy do ładowania *sqlldr* Oracle. Przykładowo, poprawna wartość dla zmiennej NLS_LANG dla Shift-JIS w systemach Windows to Japanese_Japan.JA16SJIS. Więcej informacji na temat zmiennej NLS_LANG można uzyskać w dokumentacji Oracle.

Puste łańcuchy znaków

Puste łańcuchy znaków są eksportowane do bazy danych jako wartości NULL.

Ładowanie wsadowe danych do bazy danych SQL Server

Omówione poniżej punkty mogą ułatwić skonfigurowanie ładowania wsadowego z programu IBM SPSS Modeler do bazy danych SQL Server przy użyciu opcji Przez zewnętrzny program ładujący dostępnej w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane.

Należy upewnić się, czy program narzędziowy bcp.exe bazy SQL Server został zainstalowany
: Skrypt *mssql_loader.py* wywołuje program narzędziowy *bcp.exe* SQL Server. Należy upewnić się, czy program narzędziowy *bcp.exe* jest zainstalowany na serwerze, na którym zostanie wykonany skrypt *mssql_loader.py*.

Użycie spacji jako separatora nie działa
: Należy unikać użycia spacji jako separatora w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane.

Zalecane jest użycie opcji sprawdzania rozmiaru tabeli
: Zaleca się włączenie opcji **Sprawdź rozmiar tabeli** w oknie dialogowym Eksport do bazy danych: Opcje

zaawansowane. Nieprawidłowości w procesie ładowania wsadowego nie zawsze są wykrywane; aktywowanie tej opcji umożliwi przeprowadzenie dodatkowego sprawdzenia, czy załadowana została poprawna liczba wierszy.

Puste łańcuchy znaków
Puste łańcuchy znaków są eksportowane do bazy danych jako wartości NULL.

Należy określić w pełni kwalifikowaną nazwę instancji serwera SQL
W niektórych sytuacjach program SPSS Modeler nie może uzyskać dostępu do bazy danych SQL z powodu braku pełnej nazwy hosta; wyświetlany jest wówczas następujący błąd:

```
Napotkano błąd w działaniu zewnętrznego programu ładowania wsadowego. Plik dziennika może zawierać dalsze szczeg
```

Aby usunąć ten błąd, należy dodać w polu **Dodatkowe opcje programu ładującego** następujący łańcuch, z uwzględnieniem podwójnych cudzysłowów:

```
"-S mhreboot.spss.com\SQLEXPRESS"
```

Ładowanie wsadowe danych do baz danych Teradata

Omówione poniżej punkty mogą ułatwić skonfigurowanie ładowania wsadowego z programu IBM SPSS Modeler do bazy danych Teradata przy użyciu opcji Przez zewnętrzny program ładujący dostępnej w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane.

Należy upewnić się, czy zainstalowano program narzędziowy fastload Teradata.

Skrypt *teradata_loader.py* wywołuje program narzędziowy *fastload* Teradata. Należy upewnić się, czy program *fastload* jest zainstalowany i poprawnie skonfigurowany na serwerze, na którym uruchomiony zostanie skrypt *teradata_loader.py*.

Ładowanie wsadowe danych może zostać wykonane tylko do pustych tabel

Puste tabele mogą być użyte jako tabele docelowe dla ładowania wsadowego. Jeśli tabela docelowa będzie zawierać jakieś dane przed wykonaniem ładowania wsadowego, operacja nie powiedzie się.

Dane w formacie data, czas i znacznik czasu

We właściwościach strumienia należy ustawić format daty na **RRRR-MM-DD**, a format godziny na **GG:MM:SS**.

Puste łańcuchy znaków

Puste łańcuchy znaków są eksportowane do bazy danych jako wartości NULL.

Identyfikator procesu Teradata (tdpid)

Domyślnie program *fastload* eksportuje dane do systemu Teradata z identyfikatorem tdpid=dbc. Zwykle w pliku HOSTS znajduje się wpis, który umożliwia powiązanie dbccop1 z adresem IP serwera Teradata. Aby użyć innego serwera, należy w polu **Dodatkowe opcje programu ładującego** w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane dodać następujący wpis, aby przekazać identyfikator tdpid tego serwera:

```
-tdpid <id>
```

Spacje w nazwach tabeli i kolumny

Jeśli nazwy tabeli lub kolumny zawierają spacje, operacja ładowania wsadowego zakończy się niepowodzeniem. O ile to możliwe, należy zmienić nazwę tabeli lub kolumny, tak aby nie zawierała spacji.

Opracowywanie programów ładowania wsadowego

W tym temacie omówiono sposób opracowania programu ładowania wsadowego, jaki może zostać uruchomiony za pośrednictwem programu IBM SPSS Modeler w celu załadowania danych z pliku tekstowego do bazy danych. Należy

pamiętać, że nie jest to objęte standardową umową dotyczącą wsparcia technicznego i w celu uzyskania wsparcia należy skontaktować się z przedstawicielem serwisu usług IBM.

Użycie języka Python do budowania programów ładowania wsadowego

Domyślnie IBM SPSS Modeler wyszukuje program ładowania wsadowego na podstawie typu bazy danych. Patrz Tabela 48 na stronie 356.

Skrypt *test_loader.py* może być pomocny w opracowaniu programów ładowania wsadowego. Więcej informacji można znaleźć w temacie “Testowanie programów ładowania wsadowego” na stronie 362.

Obiekty przekazywane do programu ładowania wsadowego

IBM SPSS Modeler zapisuje dwa pliki, które są przekazywane do programu ładowania wsadowego.

- **Plik danych.** Ten plik zawiera dane w formacie tekstowym przeznaczone do załadowania.
- **Plik schematu.** Jest to plik XML, który opisuje nazwy i typy kolumn oraz udostępnia informacje na temat sposobu formatowania danych (np. jaki znak został użyty jako separator zmiennych).

Ponadto, program IBM SPSS Modeler przekazuje inne informacje, takie jak nazwa tabeli, nazwa użytkownika i hasło jako argumenty w czasie wywoływania programu ładowania wsadowego.

Uwaga: Aby zasygnalizować programowi IBM SPSS Modeler pomyślne ukończenie, program ładowania wsadowego powinien usunąć plik schematu.

Argumenty przekazywane do programu ładowania wsadowego

Argumenty przekazywane do programu są wyświetlane w następującej tabeli.

Tabela 49. Argumenty przekazane do programu ładowania wsadowego

Argument	Opis
schemafile	Ścieżka pliku schematu.
data file	Ścieżka pliku danych.
servername	Nazwa serwera DBMS; może być pusta.
databasename	Nazwa bazy danych na serwerze DBMS; może być pusta.
username	Nazwa użytkownika do zalogowania się w bazie danych.
hasło	Hasło do zalogowania się w bazie danych.
tablename	Nazwa tabeli do załadowania.
ownername	Nazwa właściciela tabeli (znana również jako nazwa schematu).
logfilename	Nazwa pliku dziennika (pusta, jeśli dziennik nie został utworzony).
rowcount	Liczba wierszy w zbiorze danych.

Opcje określone w polu **Dodatkowe opcje programu ładującego** w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane są przekazywane do programu ładowania wsadowego po załadowaniu standardowych argumentów.

Format pliku danych

Dane są zapisywane w pliku danych w formacie tekstowym, w którym poszczególne zmienne mogą zostać rozdzielone znakiem separatora, jaki jest określany w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane. Poniżej podano przykład, jak może wyglądać plik danych rozdzielonych tabulatorami.

```
48  F  HIGH    NORMAL  0.692623  0.055369  drugA
15  M  NORMAL  HIGH    0.678247  0.040851  drugY
37  M  HIGH    NORMAL  0.538192  0.069780  drugA
35  F  HIGH    HIGH    0.635680  0.068481  drugA
```

Plik jest napisany zgodnie z lokalnym kodowaniem używanym przez program IBM SPSS Modeler Server (lub program IBM SPSS Modeler w przypadku braku połączenia z programem IBM SPSS Modeler Server). Niektóre opcje formatowania mogą być kontrolowane za pośrednictwem ustawień strumienia IBM SPSS Modeler.

Format pliku schematu

Plik schematu jest zapisany w formacie XML, który opisuje plik danych. Poniżej przedstawiono przykład towarzyszący poprzedniemu plikowi danych.

```
<?xml version="1.0" encoding="UTF-8"?>
<DBSCHEMA version="1.0">
 <table delimiter="\t" commit_every="10000" date_format="YYYY-MM-DD" time_format="HH:MM:SS"
append_existing="false" delete_datafile="false">
  <column name="Age" encoded_name="416765" type="integer"/>
  <column name="Sex" encoded_name="536578" type="char" size="1"/>
  <column name="BP" encoded_name="4250" type="char" size="6"/>
  <column name="Cholesterol" encoded_name="43686F6C65737465726F6C" type="char" size="6"/>
  <column name="Na" encoded_name="4E61" type="real"/>
  <column name="K" encoded_name="4B" type="real"/>
  <column name="Drug" encoded_name="44727567" type="char" size="5"/>
 </table>
</DBSCHEMA>
```

W poniższych dwóch tabelach przedstawiono atrybuty elementów <table> (tabela) i <column> (kolumna) z pliku schematu.

Tabela 50. Atrybuty elementu <table> (tabela)

Atrybut	Opis
delimiter	Znak separatora zmiennych (znak tabulacji jest reprezentowany jako \t).
commit_every	Przedział liczby regionów we wsadzie (jak w oknie dialogowym Eksport do bazy danych: Opcje zaawansowane).
date_format	Format użyty do reprezentowania dat.
time_format	Format użyty do reprezentowania godzin.
append_existing	Wartość true (prawda), jeśli tabela do załadowania zawiera już dane, w przeciwnym razie wartość false (fałsz).
delete_datafile	Wartość true (prawda), jeśli program ładowania wsadowego powinien usunąć plik danych po zakończeniu ładowania.

Tabela 51. Atrybuty elementu <column> (kolumna)

Atrybut	Opis
name	Nazwa kolumny.
encoded_name	Nazwa kolumny przekształcona zgodnie z tym samym kodowaniem, co plik danych; wynik w postaci serii dwucyfrowych cyfr szesnastkowych.
typ	Typ danych w kolumnie: integer, real, char, time, date lub datetime.
rozmiar	W przypadku typu danych char jest to maksymalna szerokość kolumny określana przez liczbę znaków.

Testowanie programów ładowania wsadowego

Ładowanie wsadowe można przetestować za pośrednictwem skryptu *test_loader.py* znajdującego się w folderze *\scripts* w katalogu instalacyjnym produktu IBM SPSS Modeler. Jest to przydatne w przypadku próby opracowania, debugowania lub rozwiązywania problemów związanych z programami ładowania wsadowego lub skryptami w celu użycia ich z programem IBM SPSS Modeler.

Aby użyć opcji testowania skryptu, należy wykonać następujące czynności.

1. Uruchom skrypt *test_loader.py*, aby skopiować pliki schematu i danych do plików *schema.xml* i *data.txt* i utworzyć plik wsadowy systemu Windows (*test.bat*).
2. Przeprowadź edycję pliku *test.bat*, aby wybrać program ładowania wsadowego lub skrypt do przetestowania.
3. Uruchom plik *test.bat* z wiersza komend, aby przetestować wybrany program ładowania wsadowego lub skrypt.

Uwaga: Uruchomienie pliku *test.bat* nie powoduje rzeczywistego załadowania danych do bazy danych.

Węzeł eksportu do pliku płaskiego

Węzeł eksportu do pliku płaskiego umożliwia zapisanie danych w rozdzielonym pliku tekstowym. Jest to przydatne podczas eksportowania danych, które mogą być odczytywane przez inne oprogramowanie do przeprowadzania analizy lub obsługujące arkusze kalkulacyjne.

Jeśli dane zawierają informacje geoprzestrzenne, można je wyeksportować jako plik płaski i w przypadku wygenerowania węzła źródłowego pliku zmiennych przeznaczonego do użycia w tym samym strumieniu wszystkie metadane składowania, pomiarów i geoprzestrzenne zostaną zachowane w nowym węźle źródłowym. Jeśli jednak dane mają zostać wyeksportowane, a następnie zaimportowane do innego strumienia, należy wykonać kilka dodatkowych kroków, które pozwolą ustawić metadane geoprzestrzenne w nowym węźle źródłowym. Więcej informacji można znaleźć w temacie “Węzeł Plik zmiennych” na stronie 26.

Uwaga: Nie można zapisywać plików w starym formacje pamięci podręcznej, ponieważ program IBM SPSS Modeler nie korzysta już z tego formatu dla plików pamięci podręcznej. Pliki pamięci podręcznej IBM SPSS Modeler są teraz zapisywane w formacie IBM SPSS Statistics *.sav*, do zapisu którego można użyć węzła eksportu Statistics. Więcej informacji można znaleźć w temacie “Węzeł eksportu Statistics” na stronie 363.

Karta eksportu do pliku płaskiego

Eksportuj plik. Określa nazwę pliku. Należy wprowadzić nazwę pliku lub kliknąć przycisk selektora plików, aby wybrać lokalizację pliku.

Tryb zapisu. Jeśli wybrana zostanie opcja **Nadpisz**, wszystkie dane istniejące w określonym pliku zostaną nadpisane. Jeśli wybrana zostanie opcja **Dopisz**, wynik zostanie dodany na końcu istniejącego pliku, a dane, jakie zawiera, zostaną zachowane.

- **Uwzględnij nazwy zmiennych.** Jeśli ta opcja zostanie wybrana, nazwy zmiennych zostaną zapisane w pierwszym wierszu pliku wynikowego. Ta opcja jest dostępna tylko w trybie **Nadpisz**.

Nowy wiersz po każdym rekordzie. Jeśli ta opcja zostanie wybrana, każdy rekord zostanie zapisany w nowym wierszu pliku wynikowego.

Separator zmiennych. Określa znak, jaki będzie wstawiony między wartościami zmiennych w wygenerowanym pliku tekstowym. Możliwe opcje to: **Przecinek**, **Tabulator**, **Spacja** i **Inne**. Wybranie opcji **Inne** umożliwia wstawienie wybranego znaku separatora w polu tekstowym.

Symbol cudzysłowu. Określa typ cudzysłowów używanych dla wartości zmiennych symbolicznych. Możliwe opcje to: **Brak** wartości nie są ujmowane w cudzysłów), **Apostrof (')**, **Cudzysłów (")** i **Inne**. Wybranie opcji **Inne** umożliwia wstawienie wybranego znaku cudzysłowu w polu tekstowym.

Kodowanie. Określa typ używanej metody kodowania tekstu. Można wybrać domyślne ustawienie systemowe, domyślne ustawienie strumienia lub UTF-8.

- Domyślne ustawienia systemowe są określone w Panelu sterowania systemu Windows (lub w przypadku trybu rozproszonego — na serwerze).
- Ustawienie domyślne strumienia jest określane w oknie dialogowym Właściwości strumienia.

Separator dziesiętny. Określa, w jaki sposób wartości dziesiętne będą reprezentowane w danych.

- **Jak dla strumienia.** Użyty zostanie separator dziesiętny zdefiniowany przez domyślne ustawienie dla bieżącego strumienia. Zwykle jest to separator dziesiętny zdefiniowany przez ustawienia regionalne komputera.
- **Kropka (.).** Jako separator dziesiętny użyty będzie znak kropki.
- **Przecinek (,).** Jako separator dziesiętny użyty będzie znak przecinka.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby automatycznie wygenerować węzeł źródłowy pliku zmiennych, który odczyta wyeksportowany plik danych. Więcej informacji można znaleźć w temacie “Węzeł Plik zmiennych” na stronie 26.

Węzeł eksportu Statistics

Węzeł eksportu Statistics umożliwia eksportowanie danych w formacie *.sav* programu IBM SPSS Statistics. Pliki *.sav* programu IBM SPSS Statistics mogą być odczytywane przez produkt IBM SPSS Statistics i inne moduły. Jest to również format używany przez pliki pamięci podręcznej w programie IBM SPSS Modeler.

Mapowanie nazw zmiennych IBM SPSS Modeler na nazwy zmiennych IBM SPSS Statistics może czasami powodować błędy, ponieważ nazwy zmiennych IBM SPSS Statistics są ograniczone do 64 znaków i nie mogą zawierać niektórych znaków, takich jak spacje, znaki dolara ($) i myślniki (–). Istnieją dwa sposoby na skorygowanie tych ograniczeń:

- Można zmienić nazwy zmiennych, tak aby były zgodne z wymaganiami nazw zmiennych w programie IBM SPSS Statistics, klikając w tym celu zakładkę Filtr. Więcej informacji można znaleźć w temacie “Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics” na stronie 364.
- Wybrać opcję wyeksportowania nazw zmiennych wraz z etykietami z programu IBM SPSS Modeler.

Uwaga: IBM SPSS Modeler zapisuje pliki *.sav* w formacie Unicode UTF-8. IBM SPSS Statistics obsługuje tylko pliki w formacie Unicode UTF-8 począwszy od wersji 16.0. Aby uniknąć możliwości uszkodzenia danych, pliki *.sav* zapisane zgodnie z kodowaniem Unicode nie powinny być używane w programie IBM SPSS Statistics w wersjach wcześniejszych niż 16.0. Więcej informacji zawiera pomoc programu IBM SPSS Statistics.

Zestawy wielokrotnych odpowiedzi. Wszystkie zestawy wielokrotnych odpowiedzi zdefiniowane w strumieniu zostaną automatycznie zachowane po wyeksportowaniu pliku. Zestawy wielokrotnych odpowiedzi można wyświetlać i edytować w dowolnym węźle po wybraniu zakładki Filtr. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi” na stronie 153.

Węzeł eksportu Statistics — karta eksportu

Eksportuj plik Określa nazwę pliku. Należy wprowadzić nazwę pliku lub kliknąć przycisk selektora plików, aby wybrać lokalizację pliku.

Typ pliku Tę opcję należy wybrać, jeśli plik ma zostać zapisany w normalnym pliku *.sav* lub w formacie skompresowanym *.zsav*.

Zaszyfruj plik hasłem Aby zabezpieczyć plik hasłem, należy zaznaczyć to pole wyboru; zostanie wyświetlony monit o wprowadzenie i potwierdzenie **hasła** w osobnym oknie dialogowym.

Uwaga: Pliki zabezpieczone hasłem można otwierać tylko w programie SPSS Modeler w wersji 16 lub wyższej lub w programie SPSS Statistics w wersji 21 lub wyższej.

Eksportuj nazwy zmiennych Określa metodę obsługi nazw zmiennych i etykiet podczas eksportowania z programu SPSS Modeler do pliku SPSS Statistics *.sav* lub *.zsav*.

- **Nazwy i etykiety zmiennych** Tę opcję należy wybrać, aby wyeksportować nazwy zmiennych i etykiety zmiennych SPSS Modeler. Nazwy są eksportowane jako nazwy zmiennych SPSS Statistics, a etykiety jako etykiety zmiennych SPSS Statistics.
- **Nazwy jako etykiety zmiennych** Ta opcja umożliwia użycie nazw zmiennych SPSS Modeler jako etykiety zmiennych w systemie SPSS Statistics. W programie SPSS Modeler dozwolone jest użycie w nazwach zmiennych znaków, które w systemie SPSS Statistics są niepoprawne. Aby uniknąć utworzenia niepoprawnych nazw w programie SPSS Statistics, należy wybrać opcję **Nazwy jako etykiety zmiennych** lub użyć karty Filtrowanie, aby skorygować nazwy zmiennych.

Uruchom aplikację Jeśli program SPSS Statistics jest zainstalowany na komputerze, można wybrać tę opcję, aby wywołać aplikację bezpośrednio z zapisanego pliku danych. Opcje uruchamiania aplikacji muszą być określone w oknie dialogowym Aplikacje pomocnicze. Więcej informacji można znaleźć w temacie “IBM SPSS Statistics — aplikacje pomocnicze” na stronie 345. Aby utworzyć plik *.sav* lub *.zsav* programu SPSS Statistics bez otwierania programu zewnętrznego, należy usunąć zaznaczenie tej opcji.

Uwaga: Podczas uruchamiania SPSS Modeler i SPSS Statistics w trybie serwera (rozproszonym) zapisywanie danych i uruchamianie sesji SPSS Statistics nie powoduje automatycznego otwarcia klienta SPSS Statistics i wyświetlenia zbioru danych w aktywnym zbiorze danych. Aby tego uniknąć, należy ręcznie otworzyć plik danych w kliencie SPSS Statistics po jego uruchomieniu.

Wygeneruj węzeł importu dla tych danych Tę opcję należy wybrać, aby automatycznie wygenerować węzeł źródłowy Plik Statistics, który odczyta wyeksportowany plik danych. Więcej informacji można znaleźć w temacie “Węzeł Plik Statistics” na stronie 31.

Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics

Przed wyeksportowaniem lub wdrożeniem danych z programu IBM SPSS Modeler do aplikacji zewnętrznych, takich jak IBM SPSS Statistics, konieczna może być zmiana nazwy lub skorygowanie nazw zmiennych. Okna dialogowe Przekształcenia Statistics, Wynik Statistics i Plik Statistics zawierają zakładkę Filtr, która upraszcza ten proces.

Główny opis funkcji karty Filtr został zamieszczony w innym miejscu. Więcej informacji można znaleźć w temacie “Ustawianie opcji filtrowania” na stronie 152. Ten temat zawiera wskazówki dotyczące odczytu danych w programie IBM SPSS Statistics.

Aby skorygować nazwy, tak aby były zgodne z konwencjami nadawania nazw IBM SPSS Statistics:

1. Na karcie Filtr kliknij pasek narzędzi Menu opcji filtrowania (pierwszy na pasku narzędzi).
2. Wybierz opcję Zmień nazwy dla IBM SPSS Statistics.
3. W oknie dialogowym Zmień nazwy dla IBM SPSS Statistics można zastąpić niepoprawne znaki w nazwach plików, wybierając **Krzyżyk (#)** lub **Podkreślenie (_)**.

Zmień nazwę zestawów wielokrotnych odpowiedzi. Tę opcję należy wybrać, aby skorygować nazwy zestawów wielokrotnych odpowiedzi, które można zaimportować do IBM SPSS Modeler za pośrednictwem węzła źródłowego Plik Statistics. Zestawy służą do rejestrowania danych, które mogą mieć więcej niż jedną wartość dla każdej obserwacji, np. odpowiedzi w ankiecie.

Węzeł eksportu Data Collection

Węzeł eksportu Data Collection zapisuje dane w formacie używanym przez oprogramowanie do badań rynku Data Collection, na podstawie modelu danych Data Collection. Ten format odróżnia obserwacje — rzeczywiste odpowiedzi na pytania zgromadzone w czasie ankiety — od metadanych, które opisują sposób gromadzenia i rozmieszczania obserwacji. Metadane składają się z informacji, takich jak teksty pytań, nazwy i opisy zmiennych, zestawy wielokrotnych odpowiedzi, tłumaczenia różnych tekstów oraz definicje struktury obserwacji. Więcej informacji można znaleźć w temacie "Węzeł Data Collection" na stronie 33.

Plik metadanych. Określa nazwę pliku definicji kwestionariusza (*.mdd*), w którym zapisane zostaną wyeksportowane metadane. Domyślny kwestionariusz jest tworzony na podstawie informacji o typie zmiennej. Przykładowo zmienna nominalna (zbiór) może być reprezentowana jako pojedyncze pytanie z opisem zmiennej użytym jako tekst pytania i osobnym polem wyboru dla każdej zdefiniowanej wartości.

Łącz metadane. Określa, czy metadane zastąpią istniejące wersje, czy zostaną połączone z istniejącymi metadanymi. Jeśli wybrana zostanie opcja łączenia, po każdym uruchomieniu strumienia zostanie utworzona nowa wersja. Dzięki temu możliwe jest śledzenie wersji kwestionariusza, kiedy ulegnie on zmianie. Każda wersja może być traktowana jako obraz stanu metadanych użytych do zgromadzenia konkretnego zbioru obserwacji.

Włącz zmienne systemowe. Określa, czy zmienne systemowe będą dołączane do pliku eksportu .mdd. Dotyczy to zmiennych, takich jak *Respondent.Serial*, *Respondent.Origin* i *DataCollection.StartTime*.

Ustawienia danych z obserwacji. Określa plik danych IBM SPSS Statistics (*.sav*), do którego obserwacje są eksportowane. Należy pamiętać, że wszystkie ograniczenia dotyczące nazw zmiennych i wartości mają tutaj zastosowanie, dlatego konieczne może być na przykład przełączenie karty filtrowania i użycie opcji „Zmień nazwy dla IBM SPSS Statistics" w menu opcji filtrowania, aby poprawić niepoprawne znaki w nazwach zmiennych.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby automatycznie wygenerować węzeł źródłowy Data Collection, który odczyta wyeksportowany plik danych.

Zestawy wielokrotnych odpowiedzi. Wszystkie zestawy wielokrotnych odpowiedzi zdefiniowane w strumieniu zostaną automatycznie zachowane po wyeksportowaniu pliku. Zestawy wielokrotnych odpowiedzi można wyświetlać i edytować w dowolnym węźle po wybraniu zakładki Filtr. Więcej informacji można znaleźć w temacie "Edytowanie zestawów wielokrotnych odpowiedzi" na stronie 153.

Węzeł eksportu Analytic Server

Węzeł eksportu Analytic Server umożliwia zapisywanie danych z analizy w istniejącym źródle danych Analytic Server. Mogą to być na przykład pliki testowe w systemie HDFS (Hadoop Distributed File System) lub bazie danych.

Zwykle strumień z węzłem eksportu Analytic Server rozpoczyna się również od węzłów źródłowych Analytic Server i jest przesyłany do programu Analytic Server i wykonywany w systemie HDFS. Alternatywnie, strumień z „lokalnymi" źródłami danych może kończyć się węzłem eksportu Analytic Server, aby pobrać relatywnie niewielkie zbiory danych (nie więcej niż 100 000 rekordów) do użycia z programem Analytic Server.

Aby użyć własnego połączenia Analytic Server zamiast połączenia domyślnego zdefiniowanego przez administratora, należy usunąć zaznaczenie pola **Użyj domyślnego serwera analitycznego** i wybrać własne połączenie. Szczegółowe informacje na temat konfigurowania kilku połączeń Analytic Server zawiera temat Łączenie się z serwerem Analytic Server.

Źródło danych. Należy wybrać źródło danych zawierające dane, jakie mają zostać użyte. Źródło danych zawiera zmienne i metadane powiązane ze źródłem. Aby wyświetlić listę dostępnych źródeł danych, należy kliknąć przycisk **Wybierz**. Więcej informacji można znaleźć w temacie "Wybieranie źródła danych" na stronie 13.

Jeśli konieczne jest utworzenie nowego źródła danych lub przeprowadzenie edycji istniejącego, należy kliknąć opcję **Uruchom edytora źródła danych...**.

Dominanta. Można wybrać opcję **Dopisz**, aby dodać do istniejącego źródła danych lub **Nadpisz**, aby zastąpić zawartość źródła danych.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby wygenerować węzeł źródłowy dla danych eksportowanych do określonego źródła danych. Ten węzeł jest dodawany do obszaru roboczego strumienia.

Należy zauważyć, że użycie kilku połączeń Analytic Server może być przydatne podczas sterowania przepływem danych. Przykładowo, w przypadku użycia węzła źródłowego i węzła eksportu Analytic Server użytkownik może chcieć użyć różnych połączeń Analytic Server w różnych gałęziach strumienia, tak aby po uruchomieniu poszczególnych gałęzi korzystały one z własnego serwera Analytic Server i aby żadne dane nie były przekazywane do serwera IBM SPSS Modeler Server. Należy pamiętać, że jeśli gałąź zawiera więcej niż jedno połączenie Analytic Server, dane będą pobierane z serwerów Analytic Server na serwer IBM SPSS Modeler Server. Więcej informacji, w tym informacje o ograniczeniach, można znaleźć w sekcji Właściwości strumienia Analytic Server.

Węzeł eksportu IBM Cognos

Węzeł eksportu IBM Cognos umożliwia eksportowanie danych ze strumienia IBM SPSS Modeler do programu Cognos Analytics, w formacie UTF-8. Dzięki temu Cognos może wykorzystać dane przekształcone lub ocenione z programu IBM SPSS Modeler. Przykładowo można użyć programu Cognos Report Studio, aby utworzyć raport na podstawie wyeksportowanych danych, z uwzględnieniem wartości predykcji i ufności. Raport może być następnie zapisany na serwerze Cognos i rozdzielony do użytkowników Cognos.

Uwaga: Można wyeksportować tylko dane relacyjne; nie można wyeksportować danych OLAP.

Aby wyeksportować dane do programu Cognos, należy określić:

- Połączenie Cognos — połączenie z serwerem Cognos Analytics (obsługiwana jest wersja 11 lub nowsza)
- Połączenie ODBC — połączenie z serwerem danych Cognos, z jakiego korzysta serwer Cognos

W przypadku korzystania z połączenia Cognos określane jest źródło danych Cognos, jakie będzie używane. To źródło danych musi korzystać z tych samych danych logowania, jak w przypadku źródła danych ODBC.

Rzeczywiste dane strumienia są eksportowane na serwer danych, a pakiet metadanych jest umieszczany na serwerze Cognos.

Jak w przypadku innych węzłów eksportu, można również użyć karty Publikuj z okna dialogowego węzła, aby opublikować strumień w celu opracowania za pośrednictwem programu IBM SPSS Modeler Solution Publisher.

Uwaga: Węzeł źródłowy Cognos obsługuje jedynie pakiety CQM Cognos. Pakiety DQM nie są obsługiwane.

Połączenie Cognos

Tutaj określane jest połączenie z serwerem Cognos Analytics (obsługiwana jest wersja 11 lub nowsza), jakie ma być używane do eksportu. Procedura obejmuje eksportowanie metadanych do nowego pakietu na serwerze Cognos, podczas gdy dane strumienia są eksportowane na serwer danych Cognos.

Połączenie. Kliknięcie przycisku **Edytuj** umożliwia wyświetlenie okna dialogowego, w którym można zdefiniować adres URL i inne szczegóły dotyczące serwera Cognos, na który dane mają zostać wyeksportowane. Jeśli użytkownik jest już zalogowany na serwerze Cognos za pośrednictwem programu IBM SPSS Modeler, może również edytować szczegóły dotyczące bieżącego połączenia. Więcej informacji można znaleźć w “Połączenia Cognos” na stronie 40.

Źródło danych. Nazwa źródła danych Cognos (zwykle bazy danych), do którego są eksportowane dane. Na liście rozwijanej wyświetlane są wszystkie źródła danych Cognos, do których można uzyskać dostęp za pośrednictwem bieżącego połączenia. Kliknięcie przycisku **Odśwież** umożliwia zaktualizowanie listy.

Folder. Ścieżka i nazwa folderu na serwerze Cognos, na którym tworzony jest pakiet eksportu.

Nazwa pakietu. Nazwa pakietu w określonym folderze, który będzie zawierał wyeksportowane metadane. Musi to być pojedynczy pakiet z pojedynczym tematem zapytania; nie można eksportować do istniejącego pakietu.

Dominanta. Określa, w jaki sposób eksport ma zostać przeprowadzony:

- **Teraz publikuj pakiet.** (ustawienie domyślne) Wykonuje operację eksportu od razu po kliknięciu przycisku **Uruchom**.
- **Eksportuj skrypt działania.** Tworzy skrypt XML, jaki można uruchomić później (na przykład za pomocą programu Framework Manager), aby przeprowadzić eksport. Należy wpisać ścieżkę i nazwę pliku skryptu w polu **Plik** lub użyć przycisku **Edytuj**, aby określić nazwę i lokalizację pliku skryptu.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby wygenerować węzeł źródłowy dla danych eksportowanych do określonego źródła i tabeli danych. Po kliknięciu przycisku **Uruchom** ten węzeł jest dodawany do obszaru roboczego strumienia.

Połączenie ODBC

Tutaj można określić połączenie z serwerem danych Cognos (czyli z bazą danych), do którego wyeksportowane zostaną dane strumienia.

Uwaga: Należy upewnić się, że dane źródłowe określone tutaj wskazują te same dane określone w panelu **Połączenia Cognos**. Należy również upewnić się, że źródło danych połączenia Cognos korzysta z tych samych danych logowania jak źródło danych ODBC.

Źródło danych. Wyświetla wybrane źródło danych. Należy wpisać nazwę lub wybrać ją z listy rozwijanej. Jeśli na liście nie ma żądanej bazy danych, należy wybrać opcję **Dodaj nowe połączenie z bazą danych** i zlokalizować bazę danych w oknie dialogowym Połączenia z bazą danych. Więcej informacji można znaleźć w “Dodawanie połączenia z bazą danych” na stronie 19.

Nazwa tabeli. Należy wpisać nazwę tabeli, do której dane mają zostać wysłane. Po wybraniu opcji **Wstaw do tabeli** można wybrać tabelę istniejącą w bazie danych, klikając przycisk **Wybierz**.

Utwórz tabelę. Tę opcję należy wybrać, aby utworzyć nową tabelę bazy danych lub zastąpić istniejącą.

Wstaw do tabeli. Tę opcję należy wybrać, aby wstawić dane jako nowe wiersze w istniejącej tabeli bazy danych.

Połącz w tabeli. (O ile dostępna) Tę opcję należy wybrać, aby zaktualizować wybrane kolumny bazy danych przez wartości z odpowiednich zmiennych danych źródłowych. Po wybraniu tej opcji dostępny jest przycisk **Łączenie**, który wyświetla okno dialogowe umożliwiające mapowanie zmiennych danych źródłowych na kolumny bazy danych.

Porzuć istniejącą tabelę. Tę opcję należy wybrać, aby podczas tworzenia nowej tabeli usunąć istniejące tabele o takiej samej nazwie.

Usuń istniejące wiersze. Tę opcję należy wybrać, aby podczas wstawiania do tabeli usunąć z niej wiersze istniejące przed wyeksportowaniem.

Uwaga: Jeśli zaznaczone zostaną dwie z powyższych opcji, podczas wykonywania węzła wyświetlony zostanie komunikat **Nadpisz ostrzeżenie**. Aby pominąć ostrzeżenia, należy usunąć zaznaczenie opcji **Ostrzegaj przed nadpisywaniem tabeli w bazie danych** na karcie Powiadomienia w oknie dialogowym Opcje użytkownika.

Domyślny rozmiar tekstu. Pola oznaczone jako bez typu we wcześniejszym węźle Typy są zapisywane w bazie danych jako zmienne łańcuchowe. Należy określić wielkość łańcuchów, jakie będą stosowane dla zmiennych bez typu.

Kliknięcie opcji **Schemat** umożliwia otwarcie okna dialogowego, w którym można ustawić różne opcje eksportu (w przypadku baz danych, które obsługują tę funkcję), ustawić typy danych SQL dla zmiennych oraz określić klucz główny dla indeksowania bazy danych. Więcej informacji można znaleźć w “Opcje schematu eksportu do bazy danych” na stronie 350.

Aby określić opcje indeksowania wyeksportowanej tabeli w celu zwiększenia wydajności, należy kliknąć opcję **Indeksy**. Więcej informacji można znaleźć w “Opcje indeksu eksportu do bazy danych” na stronie 352.

Po kliknięciu opcji **Zaawansowane** można określić opcje ładowania dużych zbiorów danych i zatwierdzania bazy danych. Więcej informacji można znaleźć w “Zaawansowane opcje eksportu do bazy danych” na stronie 354.

Ujmij w cudzysłów nazwy tabeli i kolumny. Opcje wyboru używane podczas wysyłania instrukcji CREATE TABLE do bazy danych. Tabele lub kolumny zawierające spacje lub znaki niestandardowe muszą być ujęte w cudzysłów.

- **W razie potrzeby.** Tę opcję należy zaznaczyć, aby program IBM SPSS Modeler automatycznie ustalał, kiedy ujęcie w cudzysłów jest konieczne w pojedynczym przypadku.
- **Zawsze.** Po wybraniu tej opcji nazwy tabeli i kolumn zawsze będą ujmowane w cudzysłów.
- **Nigdy.** Ta opcja powoduje wyłączenie funkcji stosowania cudzysłowów.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby wygenerować węzeł źródłowy dla danych eksportowanych do określonego źródła i tabeli danych. Po kliknięciu przycisku **Uruchom** ten węzeł jest dodawany do obszaru roboczego strumienia.

Węzeł eksportu IBM Cognos TM1

Węzeł eksportu IBM Cognos umożliwia eksportowanie danych ze strumienia SPSS Modeler do programu Cognos TM1. Dzięki temu Cognos Analytics może wykorzystać dane przekształcone lub ocenione z programu SPSS Modeler.

Uwaga: Można wyeksportować tylko miary bez kontekstu z danymi wymiaru; alternatywnie, można dodać nowe elementy do kostki.

Aby wyeksportować dane do programu Cognos Analytics (obsługiwana jest wersja 11 lub nowsza), należy określić:

- Połączenie z serwerem Cognos TM1.
- Kostkę, do której dane zostaną wyeksportowane.
- Mapowanie z nazw danych SPSS na równoważne wymiary i miary TM1.

Uwaga: Użytkownik TM1 musi mieć następujące uprawnienia: prawo do zapisu kostek, prawdo do odczytu wymiarów oraz prawo do zapisu elementów wymiaru. Ponadto do zaimportowania i wyeksportowania danych Cognos TM1 za pośrednictwem programu SPSS Modeler wymagana jest instalacja IBM Cognos TM1 10.2, pakiet poprawek 3. Istniejące strumienie, które utworzono na podstawie poprzednich wersji, będą nadal działać.

Dla tego węzła nie są wymagane dane uwierzytelniające administratora. Jeśli jednak nadal używana jest starsza wersja niż węzeł TM1 17.1, dane uwierzytelniające administratora są nadal wymagane.

SPSS Modeler umożliwia współpracę z Cognos TM1 wyłącznie w trybach IntegratedSecurityMode 1, 4 i 5.

Jak w przypadku innych węzłów eksportu, można również użyć karty Publikuj z okna dialogowego węzła, aby opublikować strumień w celu opracowania za pośrednictwem programu IBM SPSS Modeler Solution Publisher.

Uwaga: Przed użyciem węzła źródłowego lub eksportu TM1 w programie SPSS Modeler należy zweryfikować niektóre ustawienia w pliku tm1s.cfg; jest to plik konfiguracji serwera TM1 znajdujący się w katalogu głównym serwera TM1.

- HTTPPortNumber — należy ustawić poprawny numer portu; zwykle jest to 1-65535. Należy pamiętać, że nie jest to numer portu, który wcześniej podawany był dla połączenia w węźle; jest to używany przez TM1 port wewnętrzny, który został domyślnie wyłączony. W razie konieczności należy skontaktować się z administratorem TM1, aby potwierdzić poprawność ustawienia dla tego portu.
- UseSSL — jeśli ta opcja zostanie ustawiona na *True* (Prawda), jako protokół transportu użyty zostanie protokół HTTPS. W tym przypadku należy zaimportować certyfikat TM1 do środowiska JRE serwera SPSS Modeler Server.

Nawiązywanie połączenia z kostką IBM Cognos TM1 w celu wyeksportowania danych

Pierwszym krokiem podczas eksportowania danych do bazy danych IBM Cognos TM1 jest wybranie odpowiedniego hosta administratora TM1 oraz powiązanego serwera i kostki na karcie **Połączenie** w oknie dialogowym IBM Cognos TM1.

Uwaga: Podczas eksportowania danych do TM1 odrzucane będą tylko rzeczywiste wartości "null". Wartości zerowe (0) zostaną wyeksportowane jako poprawne wartości. Należy również pamiętać, że tylko zmienne z typem składowania *łańcuch* mogą być mapowane na wymiary na karcie mapowania. Przed wyeksportowaniem do TM1 należy użyć klienta IBM SPSS Modeler, aby przekształcić dane, które nie są łańcuchami, na łańcuchy.

Host administracyjny Należy wpisać adres URL hosta administracyjnego, z zainstalowanym serwerem TM1, z którym ma zostać nawiązane połączenie. Host administracyjny jest zdefiniowany jako pojedynczy adres URL dla wszystkich serwerów TM1. Za pomocą tego adresu URL można wykryć wszystkie zainstalowane i uruchomione w danym środowisku serwery IBM Cognos TM1 oraz uzyskać do nich dostęp.

TM1 Server Po nawiązaniu połączenia z hostem administracyjnym należy wybrać serwer, który zawiera dane do zaimportowania, i kliknąć przycisk **Login**. Jeśli wcześniej nie zostało nawiązane połączenie z tym serwerem, zostanie wyświetlony monit o wprowadzenie danych w polach **Nawa użytkownika** i **Hasło**; alternatywnie, można wyszukać wcześniej wprowadzone dane logowania, zapisane jako **Zapisane dane uwierzytelniające**.

Wybierz kostkę TM1 do wyeksportowania Wyświetla nazwy kostek na serwerze TM1, do których można wyeksportować dane.

Aby wybrać dane do wyeksportowania, należy wybrać kostkę i kliknąć strzałkę w prawo, aby przenieść kostkę do pola **Eksportuj do kostki**. Po wybraniu kostki można użyć karty Mapowanie, aby zmapować wymiary i miary TM1 na odpowiednie zmienne SPSS lub wartości stałe (operacja *Wybierz*).

Mapowanie danych IBM Cognos TM1 do eksportu

Po wybraniu hosta administracyjnego TM1 oraz powiązanego serwera i kostki TM1 karta Mapowanie w oknie dialogowym eksportu IBM Cognos TM1 umożliwia zmapowanie wymiarów i miar TM1 na zmienne SPSS lub ustawienie wymiarów TM1 jako wartości stałych.

Uwaga: Na wymiary można mapować tylko zmienne z typem składowania *łańcuch*. Przed wyeksportowaniem do TM1 należy użyć klienta IBM SPSS Modeler, aby przekształcić dane, które nie są łańcuchami, na łańcuchy.

Zmienne Wyświetla nazwy zmiennych danych z pliku danych SPSS, które są dostępne do wyeksportowania.

Wymiary TM1 Wyświetla kostkę TM1 wybraną na karcie Połączenie, wraz z jej wymiarami, wymiarem miary oraz elementami wybranego wymiaru miary. Aby zmapować zmienną danych SPSS, należy wybrać wymiar lub miarę TM1.

Na karcie Mapowanie dostępne są następujące opcje.

Wybierz wymiar miary Z listy wymiarów dla wybranej kostki należy wybrać ten, który będzie wymiarem miary.

Po wybraniu wymiaru, z wyjątkiem wymiaru miary, i kliknięciu przycisku **Wybierz** wyświetlane jest okno dialogowe z elementami-liśćmi wybranego wymiaru. Wybrać można tylko elementy-liście. Wybrane elementy są oznaczane literą **S**.

Mapuj Umożliwia zmapowanie wybranych zmiennych danych SPSS na wybrany wymiar lub miarę TM1 (stały wymiar lub konkretna miara lub element z wymiaru miary). Zmapowane zmienne są oznaczane literą **M**.

Usuń mapowanie Usuwa mapowanie wybranej zmiennej danych SPSS z wybranego wymiaru lub wybranej miary TM1. Należy pamiętać, że mapowania można usuwać tylko pojedynczo. Zmienne danych SPSS, dla których usunięto mapowanie, są przenoszone z powrotem do kolumny po lewej stronie.

Utwórz nowe Tworzy nową miarę w wymiarze miary TM1. Wyświetlane jest okno dialogowe, w którym można wprowadzić nową wartość **Nazwa miary TM1**. Ta opcja jest dostępna tylko dla wymiarów miary; nie jest dostępna dla stałych wymiarów.

Więcej informacji na temat TM1 zawiera dokumentacja produktu IBM Cognos TM1 pod adresem http://www-01.ibm.com/support/knowledgecenter/SS9RXT_10.2.2/com.ibm.swg.ba.cognos.ctm1.doc/welcome.html.

Węzeł eksportu SAS

Ta funkcja jest dostępna w programach SPSS Modeler Professional i SPSS Modeler Premium.

Węzeł eksportu SAS umożliwia zapisywanie danych w formacie SAS, aby mogły zostać odczytane w systemie SAS lub za pomocą pakietu oprogramowania kompatybilnego z systemem SAS. Dane można wyeksportować w trzech formatach plików SAS: SAS dla systemu Windows/OS2, SAS dla systemu UNIX lub SAS.

Węzeł eksportu SAS — karta eksportu

Eksportuj plik. Należy określić nazwę pliku. Należy wprowadzić nazwę pliku lub kliknąć przycisk selektora plików, aby wybrać lokalizację pliku.

Eksportuj. Należy określić format pliku. Opcje do wyboru to: **SAS dla systemu Windows/OS2**, **SAS dla systemu UNIX** lub **SAS wersja 7/8/9**.

Eksportuj nazwy zmiennych. Należy wybrać opcje dla wyeksportowania nazw i etykiet zmiennych z programu IBM SPSS Modeler, jakie będą używane w systemie SAS.

- **Nazwy i etykiety zmiennych.** Tę opcję należy wybrać, aby wyeksportować nazwy zmiennych i etykiety zmiennych IBM SPSS Modeler. Nazwy są eksportowane jako nazwy zmiennych SAS, a etykiety jako etykiety zmiennych SAS.
- **Nazwy jako etykiety zmiennych.** Ta opcja umożliwia użycie nazw zmiennych IBM SPSS Modeler jako etykiety zmiennych w systemie SAS. W programie IBM SPSS Modeler dozwolone jest użycie w nazwach zmiennych znaków, które w systemie SAS są niepoprawne. Aby uniknąć utworzenia niepoprawnych nazw SAS, należy wybrać opcję **Nazwy i etykiety zmiennych**.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby automatycznie wygenerować węzeł źródłowy SAS, który odczyta wyeksportowany plik danych. Więcej informacji można znaleźć w temacie “Węzeł źródłowy SAS” na stronie 43.

Uwaga: Maksymalna dozwolona długość łańcucha wynowi 255 bajtów. Jeśli długość łańcucha przekracza 255 bajtów, to łańcuch zostanie obcięty podczas eksportowania.

Węzeł eksportu programu Excel

Węzeł eksportu programu Excel powoduje zapisanie danych wyjściowych w pliku Microsoft Excel w formacie .xlsx. Opcjonalnie można zdecydować, aby program Excel był uruchamiany automatycznie i otwierał wyeksportowany plik po wykonaniu węzła.

Węzeł programu Excel — karta eksportu

Nazwa pliku. Należy wprowadzić nazwę pliku lub kliknąć przycisk selektora plików, aby wybrać lokalizację pliku. Domyślna nazwa pliku to *excelexp.xlsx*.

Typ pliku. Obsługiwany jest plik programu Excel w formacie .xlsx.

Utwórz nowy plik. Tworzy nowy plik programu Excel.

Wstaw do istniejącego pliku. Zawartość jest zastępowana, począwszy od komórki wskazanej w polu **Rozpocznij w komórce**. Pozostałe komórki arkusza zachowują swoją pierwotną zawartość.

Uwzględnij nazwy zmiennych. Określa, czy nazwy zmiennych powinny być uwzględniane w pierwszym wierszu arkusza.

Rozpocznij w komórce. Lokalizacja komórki użyta dla pierwszego rekordu eksportu (lub nazwa pierwszej zmiennej, jeśli wybrano opcję **Uwzględnij nazwy zmiennych**). Dane są wstawiane po prawej stronie i w dół względem komórki początkowej.

Wybierz arkusz. Określa arkusz, do jakiego dane mają zostać wyeksportowane. Arkusz można zidentyfikować na podstawie indeksu lub nazwy:

- **Według indeksu.** Jeśli tworzony jest nowy plik, należy określić liczbę do 0 do 9, aby zidentyfikować arkusz, do którego dane mają zostać wyeksportowane, rozpoczynając od 0 dla pierwszego arkusza, 1 dla drugiego itd. Jeśli arkusz istnieje już w danym miejscu, można użyć wartości 10 lub wyższej.
- **Według nazwy.** Jeśli tworzony jest nowy plik, należy określić nazwę używaną dla arkusza. Jeśli dane wstawiane są do istniejącego pliku, zostaną wstawione do tego arkusza, o ile istnieje; w przeciwnym razie tworzony jest nowy arkusz o tej nazwie.

Uruchom program Excel. Określa, czy program Excel będzie automatycznie uruchamiany dla wyeksportowanego pliku po wykonaniu węzła. Należy pamiętać, że w przypadku działania w trybie rozproszonym dla programu IBM SPSS Modeler Server wynik jest zapisywany w systemie plików serwera, a program Excel jest uruchamiany na kliencie z kopią wyeksportowanego pliku.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby automatycznie wygenerować węzeł źródłowy programu Excel, który odczyta wyeksportowany plik danych. Więcej informacji można znaleźć w temacie “Węzeł źródłowy programu Excel” na stronie 44.

Węzeł Rozszerzenie Eksport

Węzeł Rozszerzenie Eksport umożliwia wykonywanie skryptów R lub Python for Spark służących do eksportowania danych.

Węzeł Rozszerzenie Eksport — karta Polecenia

Wybierz język poleceń: **R** albo **Python for Spark**. Więcej informacji można znaleźć w następujących sekcjach. Gdy polecenia będą gotowe, można kliknąć przycisk **Wykonaj**, aby wykonać węzeł Rozszerzenie Eksport.

Polecenia R

Polecenia R. Do tego pola można wpisać lub wkleić własny skrypt R służący do analizy danych.

Konwertuj zmienne typu flaga. Określa sposób traktowania zmiennych typu flaga. Dostępne są dwie opcje: **Łańcuchy na czynnik, liczby całkowite i rzeczywiste na liczby typu double** oraz **Wartości logiczne (Prawda, Fałsz)**. W przypadku wybrania opcji **Wartości logiczne (Prawda, Fałsz)** pierwotne wartości zmiennych typu flaga zostaną utracone. Na przykład, jeśli zmienna ma wartości Mężczyzna i Kobieta, to zostaną zamienione na Prawda i Fałsz.

Konwertuj brakujące wartości na wartość niedostępności danych (NA) pakietu R. Gdy ta opcja jest wybrana, wszelkie brakujące wartości są przekształcane w wartość NA w języku R. W języku R wartość NA oznacza brakujące wartości. Niektóre funkcje R przyjmują argument sterujący zachowaniem funkcji w przypadku, gdy dane zawierają wartość NA. Na przykład funkcja może oferować opcję automatycznego wykluczania rekordów zawierających wartość NA. Jeśli ta opcja nie będzie wybrana, wszelkie brakujące wartości będą przekazywane do skryptu R bez zmian, co może powodować błędy podczas jego wykonywania.

Konwertuj zmienne daty/czasu na klasy pakietu R ze specjalną kontrolą stref czasowych. Gdy ta opcja jest wybrana, zmienna typu data lub data/czas są przekształcane w obiekty date/time języka R. Należy wybrać jedną z następujących opcji:

- **R POSIXct.** Zmienne typu data lub data/czas są przekształcane w obiekty POSIXct języka R.

- **R POSIXlt (lista).** Zmienne typu data lub data/czas są przekształcane w obiekty POSIXlt języka R.

Uwaga: Formaty POSIX są opcjami zaawansowanymi. Opcji tych należy używać tylko wtedy, gdy w skrypcie R nakazano traktowanie zmiennych daty/czasu w sposób wymagający zastosowania tych formatów. Formaty POSIX nie mają zastosowania względem zmiennych z formatami czasu.

Polecenia Python

Polecenia Python. Do tego pola można wpisać lub wkleić własny skrypt Python służący do analizy danych. Aby uzyskać więcej informacji na temat języka Python for Spark, patrz Python for Spark i Pisanie skryptów w języku Python for Spark.

Węzeł Rozszerzenie Eksport — karta Wynik z konsoli

Karta **Wynik z konsoli** zawiera wszelkie wyniki odbierane podczas wykonywania skryptu w języku R lub Python for Spark na karcie Polecenia (na przykład, jeśli używany jest skrypt R, to na tej karcie wyświetlane są wyniki odbierane z konsoli R podczas wykonywania skryptu z pola **Polecenia R** na karcie **Polecenie**). Wyniki te mogą zawierać komunikaty o błędach lub ostrzeżenia generowane podczas wykonywania skryptu w języku R lub Python. Wyniki można wykorzystać przede wszystkim do debugowania skryptu. Karta **Wynik z konsoli** zawiera także skrypt z pola **Polecenia R** lub **Polecenia Python**.

Po każdym wykonaniu skryptu Rozszerzenie Eksport zawartość karty **Wynik z konsoli** jest nadpisywana wynikami z konsoli R lub środowiska Python for Spark. Wyników nie można edytować.

Węzeł eksportu XML

Węzeł eksportu XML umożliwia utworzenie danych wynikowych w formacie XML, z zastosowaniem kodowania UTF-8. Opcjonalnie można utworzyć węzeł źródłowy XML, aby wczytać wyeksportowane dane z powrotem do strumienia.

Plik eksportu XML. Pełna ścieżka i nazwa pliku XML, do którego dane mają zostać wyeksportowane.

Użyj schematu XML. To pole wyboru należy zaznaczyć, aby użyć schematu lub pliku DTD w celu kontrolowania struktury wyeksportowanych danych. Spowoduje to aktywowanie przycisku **Mapuj** opisanego poniżej.

Jeśli nie zostanie użyty schemat ani plik DTD, dla wyeksportowanych danych stosowana jest następująca struktura domyślna:

```
<records>
     <record>
          <fieldname1>value</fieldname1>
          <fieldname2>value</fieldname2>
      :
          <fieldnameN>value</fieldnameN>
     </record>
     <record>
      :
      :
     </record>
      :
      :
</records>
```

Spacje w nazwie zmiennej są zastępowane podkreśleniami; przykładowo, „Moja zmienna” jest zapisywana jako <Moja_zmienna>.

Mapuj. Jeśli wybrano użycie schematu XML, ten przycisk otwiera okno dialogowe, w którym można określić, która część struktury XML powinna być użyta w celu rozpoczęcia każdego nowego rekordu. Więcej informacji można znaleźć w temacie “Mapowanie XML — opcje rekordów”.

Zmapowane zmienne. Wskazuje liczbę zmiennych, które zostały zmapowane.

Wygeneruj węzeł importu dla tych danych. Tę opcję należy wybrać, aby automatycznie wygenerować węzeł źródłowy XML, który wczyta wyeksportowany plik danych z powrotem do strumienia. Więcej informacji można znaleźć w temacie “Węzeł źródłowy XML” na stronie 45.

Zapisywanie danych XML

Po określeniu elementu XML wartość zmiennej jest umieszczana w znaczniku elementu:

```
<element>value</element>
```

W przypadku mapowania atrybutu wartość zmiennej jest wstawiana jako wartość atrybutu:

```
<element attribute="value">
```

Jeśli zmienna jest mapowana na element powyżej elementu <records>, zmienna jest zapisywana tylko raz i będzie stała dla wszystkich rekordów. Wartość dla tego elementu będzie pochodziła z pierwszego rekordu.

Jeśli zapisana ma zostać wartość null, wprowadzana jest pusta zawartość. Dla elementów wygląda to następująco:

```
<element></element>
```

W przypadku atrybutów jest to zapis:

```
<element attribute="">
```

Mapowanie XML — opcje rekordów

Karta Rekordy umożliwia określenie, która część struktury XML będzie użyta do rozpoczęcia każdego nowego rekordu. Aby możliwe było poprawne zmapowanie na schemat, należy określić separator rekordów.

Struktura XML. Hierarchiczne drzewo przedstawiające strukturę schematu XML określoną na wcześniejszym ekranie.

Rekordy (XPath). Aby ustawić separator rekordów, należy wybrać element w strukturze XML i kliknąć przycisk strzałki w prawo. Za każdym razem, kiedy ten element zostanie napotkany w danych źródłowych, w pliku wynikowym utworzony zostanie nowy rekord.

Uwaga: Jeśli w strukturze XML wybrany zostanie element główny, zapisany może zostać tylko jeden rekord, a pozostałe rekordy będą pominięte.

Mapowanie XML — opcje zmiennych

Karta Zmienne jest używana do mapowania zmiennych w zbiorze danych na elementy lub atrybuty w strukturze XML, o ile używany jest plik schematu.

Nazwy zmiennych zgodne z nazwami elementów lub atrybutów są mapowane automatycznie, o ile nazwa elementu lub atrybutu jest unikalna. Dlatego jeśli istnieje element i atrybut o nazwie field1 (zmienna1), mapowanie automatyczne nie zostanie przeprowadzone. Jeśli w strukturze znajduje się tylko jedna pozycja o nazwie field1, zmienna z tą nazwą w strumieniu zostanie zmapowana automatycznie.

Pola. Lista zmiennych w modelu. Należy wybrać co najmniej jedną zmienną jako część źródła mapowania. Można użyć przycisków u dołu listy, aby wybrać wszystkie zmienne lub wszystkie zmienne z określonym poziomem pomiaru.

Struktura XML. Należy wybrać element w strukturze XML jako element docelowy mapowania. Aby utworzyć mapowanie, należy kliknąć przycisk Mapuj. Zostanie wyświetlone mapowanie. Poniżej listy wyświetlana jest liczba zmiennych zmapowanych w ten sposób.

Aby usunąć mapowanie, należy wybrać pozycję z listy struktury XML i kliknąć przycisk **Usuń mapowanie**.

Pokaż atrybuty. Umożliwia wyświetlanie lub ukrywanie atrybutów, o ile są dostępne, elementów XML w strukturze XML.

Podgląd mapowania XML

Na karcie Podgląd należy kliknąć przycisk **Aktualizuj**, aby wyświetlić podgląd pliku XML, jaki został zapisany.

Jeśli mapowanie jest niepoprawne, należy wrócić do karty Rekordy lub zmienne, aby poprawić błędy, i kliknąć ponownie przycisk **Aktualizuj**, aby wyświetlić wynik.

Węzeł Eksport JSON

Węzeł eksportu JSON umożliwia utworzenie danych wynikowych w formacie JSON, z zastosowaniem kodowania UTF-8. Opcjonalnie można utworzyć węzeł źródłowy JSON, aby wczytać wyeksportowane dane z powrotem do strumienia.

Gdy program SPSS Modeler zapisuje dane w pliku JSON, wykonuje następujące przekształcenia.

Tabela 52. Przekształcenia podczas eksportu danych JSON

Typ składowania danych w programie SPSS Modeler	Wartość JSON
Łańcuch	string
Liczba całkowita	number(int)
Liczba rzeczywista	number(real)
Data	string
Czas	string
Znacznik czasu	string
Lista	Nieobsługiwane Pola list zostaną wykluczone z eksportu.
Brakujące wartości	null

Plik eksportu JSON. Pełna ścieżka i nazwa pliku JSON, do którego zostaną wyeksportowane dane.

Format łańcucha JSON. Określ format łańcucha JSON. Wybierz opcję **Rekordy**, jeśli chcesz, aby węzeł eksportu JSON generował zbiór par nazwa-wartość. Lub wybierz opcję **Wartości**, jeśli wymagane jest tylko wyeksportowanie wartości (bez nazw).

Format łańcucha JSON. Określ format łańcucha JSON. Opcja Rekordy spowoduje wyeksportowanie zbioru par nazwa-wartość. Wybierz opcję Wartości, jeśli wymagane jest tylko wyeksportowanie wartości (bez nazw).

Wygeneruj węzeł importu dla tych danych. Wybierz tę opcję, aby automatycznie wygenerować węzeł źródłowy JSON, który wczyta wyeksportowany plik danych z powrotem do strumienia. Aby uzyskać więcej informacji, patrz “Węzeł źródłowy JSON” na stronie 67.

Wspólne karty węzła eksportu

Poniżej przedstawiono opcje, jakie można określić dla wszystkich węzłów eksportu po kliknięciu odpowiedniej zakładki:

- **Karta Publikuj.** Umożliwia publikowanie wyników w strumieniu.

- **Zakładka Adnotacje.** Ta karta jest używana dla wszystkich węzłów i oferuje opcje zmiany nazwy węzłów, obsługi podpowiedzi użytkownika i zapisywania długich powiadomień.

Publikowanie strumieni

Publikowanie strumieni odbywa się bezpośrednio z IBM SPSS Modeler przy użyciu dowolnego standardowego węzła eksportu: Baza danych, Plik płaski, Eksport Statistics, Eksportowanie przez rozszerzenie, Eksport Data Collection, Eksport SAS, Excel i Eksport XML. Typ węzła eksportu określa format wyników zapisywanych po każdym opublikowaniu strumienia za pośrednictwem programu IBM SPSS Modeler Solution Publisher Runtime lub aplikacji zewnętrznej. Przykładowo, aby po każdym uruchomieniu opublikowanego strumienia zapisywać wyniki w bazie danych, należy użyć węzła eksportu Baza danych.

Publikowanie strumienia

1. Otwórz lub zbuduj strumień w zwykły sposób i na końcu dołącz węzeł eksportu.
2. Na karcie Publikuj w węźle eksportu określ trzon nazwy dla opublikowanych plików (czyli nazwę pliku, do której dodane zostanie rozszerzenie .pim, .par i .xml).
3. Kliknij **Publikuj**, aby opublikować strumień lub wybierz opcję **Opublikuj strumień**, aby opublikować strumień automatycznie po każdym wykonaniu węzła.

Publikowana nazwa. Określ trzon nazwy dla publikowanych plików obrazów i parametrów.

- **Plik obrazu** (*.pim) udostępnia wszystkie informacje potrzebne w środowisku wykonawczym do wykonania opublikowanego strumienia dokładnie w taki sposób, jak podczas eksportu. Jeśli nie ma potrzeby wprowadzania zmian w ustawieniach strumienia (np. źródła danych wejściowych lub pliku danych wynikowych), można jedynie wdrożyć plik obrazu.
- **Plik parametru** (*.par) zawiera możliwe do skonfigurowania informacje na temat źródeł danych, plików wynikowych i opcji wykonywania. Aby możliwe było kontrolowanie informacji wejściowych i wynikowych w strumieniu bez konieczności ponownego publikowania strumienia, potrzebny będzie plik parametru oraz plik obrazu.
- **Plik metadanych** (*.xml) opisuje dane wejściowe i dane wynikowe obrazu oraz ich modele danych. Jest przeznaczony do użycia w aplikacjach, w których osadzono bibliotekę środowiska wykonawczego i które wymagają znajomości struktury danych wejściowych i wynikowych.

  Uwaga: Ten plik jest tworzony tylko po zaznaczeniu opcji **Generuj metadane**.

Publikuj parametry. W razie potrzeby parametry strumienia można dołączyć do pliku *.par. Wartości tych parametrów strumienia można zmieniać po wykonaniu obrazu poprzez edycję pliku *.par lub za pośrednictwem interfejsu API środowiska wykonawczego.

Ta opcja aktywuje przycisk **Parametry**. Po kliknięciu tego przycisku wyświetlane jest okno dialogowe Publikuj parametry.

Należy wybrać parametry, jakie mają zostać dołączone do publikowanego obrazu, zaznaczając odpowiednią opcję w kolumnie **Publikuj**.

Przy wykonaniu strumienia. Określa, czy strumień jest publikowany automatycznie po wykonaniu węzła.

- **Eksportuj dane.** Umożliwia wykonanie węzła eksportu w standardowy sposób, bez publikowania strumienia. (Zwykle węzeł jest wykonywany w IBM SPSS Modeler w taki sam sposób, w jaki odbywałoby się to w razie braku dostępu do komponentu IBM SPSS Modeler Solution Publisher). Po naciśnięciu tego przycisku strumień nie zostanie opublikowany do czasu wydania takiego polecenia w sposób jawny poprzez kliknięcie przycisku **Publikuj** w oknie dialogowym węzła eksportu. Alternatywnie można opublikować bieżący strumień za pośrednictwem narzędzia publikowania na pasku narzędzi lub za pośrednictwem skryptu.
- **Opublikuj strumień.** Umożliwia opublikowanie strumienia z wdrożenia przy użyciu narzędzia IBM SPSS Modeler Solution Publisher. Tę opcję należy zaznaczyć, jeśli strumień ma być publikowany automatycznie po każdym jego wykonaniu.

Uwaga:

- Jeśli opublikowany strumień ma zostać uruchomiony z użyciem nowych lub zaktualizowanych danych, ważne jest, aby pamiętać, że kolejność zmiennych w pliku wejściowym musi być taka sama, jak kolejność zmiennych w pliku wejściowym węzła źródłowego określonego w opublikowanym strumieniu.
- Podczas publikowania do aplikacji zewnętrznych należy rozważyć filtrowanie zmiennych zewnętrznych lub zmianę nazwy zmiennych, tak aby były zgodne z wymogami dotyczącymi danych wejściowych. Obie te funkcje są dostępne po wprowadzeniu węzła filtrowania przed węzłem eksportu.

Rozdział 8. Węzły programu IBM SPSS Statistics

Węzły programu IBM SPSS Statistics — Przegląd

Aby uzupełnić program IBM SPSS Modeler i jego możliwości eksploracji danych, IBM SPSS Statistics udostępnia możliwość przeprowadzenia dodatkowych analiz statystycznych oraz funkcje zarządzania danymi.

Jeśli użytkownik dysponuje zainstalowaną, kompatybilną, kopią programu IBM SPSS Statistics z odpowiednią licencją, może się z tym programem połączyć za pośrednictwem programu IBM SPSS Modeler i przeprowadzić kompleksowe, wieloetapowe manipulacje dla danych oraz analizy, które w przeciwnym razie nie są obsługiwane w programie IBM SPSS Modeler. Dla użytkowników zaawansowanych dostępna jest również opcja dalszego modyfikowania analiz za pośrednictwem składni komend. Informacje na temat kompatybilności wersji zawiera publikacja Uwagi do wydania.

Węzły programu IBM SPSS Statistics (o ile są dostępne) są wyświetlane w wyznaczonej części palety węzłów.

Uwaga: Zalecamy określenie danych w węźle Typy przed użyciem węzłów Przekształcenia, Model lub Wynik IBM SPSS Statistics. Jest to również wymagane w przypadku użycia komend składni AUTORECODE.

Paleta IBM SPSS Statistics zawiera następujące węzły:

Węzeł Plik Statistics odczytuje dane z pliku w formacie *.sav* lub *.zsav* używanym przez program IBM SPSS Statistics, jak również pliki pamięci podręcznej zapisane w programie IBM SPSS Modeler, które również używają tego samego formatu.

Węzeł Przekształceń Statistics uruchamia wybór komend składni IBM SPSS Statistics dla źródeł danych w programie IBM SPSS Modeler. Ten węzeł wymaga licencjonowanej kopii programu IBM SPSS Statistics.

Węzeł Model Statistics umożliwia analizowanie danych i pracę z nimi poprzez uruchomienie procedur IBM SPSS Statistics tworzących PMML. Ten węzeł wymaga licencjonowanej kopii programu IBM SPSS Statistics.

Węzeł Wynik Statistics umożliwia wywołanie procedury IBM SPSS Statistics w celu przeprowadzenia analizy danych IBM SPSS Modeler. Dostępne są różnorodne procedury analityczne programu IBM SPSS Statistics. Ten węzeł wymaga licencjonowanej kopii programu IBM SPSS Statistics.

Dane wynikowe węzła eksportu Plik Statistics zapisywane są w formacie IBM SPSS Statistics: *.sav* lub *.zsav*. Pliki .sav lub .zsav mogą być odczytywane przez produkty IBM SPSS Statistics Base i inne. Jest to również format używany przez pliki pamięci podręcznej w programie IBM SPSS Modeler.

Uwaga: Jeśli kopia programu SPSS Statistics jest objęta licencją dla jednego użytkownika i uruchomiony zostanie strumień obejmujący dwie lub więcej gałęzi, z których każda będzie zawierała węzeł SPSS Statistics, może wystąpić błąd związany z licencją. Dzieje się tak, kiedy sesja programu SPSS Statistics dla jednej gałęzi nie zakończy się przed podjęciem próby uruchomienia sesji dla drugiej gałęzi. O ile to możliwe, należy ponownie zaprojektować strumień, tak aby gałęzie zawierające węzeł SPSS Statistics nie były wykonywane równolegle.

Węzeł Plik Statistics

Węzeł Plik Statistics umożliwia odczyt danych bezpośrednio z zapisanego pliku IBM SPSS Statistics (*.sav* lub *.zsav*). Ten format zastępuje teraz plik pamięci podręcznej z wcześniejszych wersji programu IBM SPSS Modeler. Aby zaimportować zapisany plik pamięci podręcznej, należy użyć węzła Plik IBM SPSS Statistics.

Importuj Plik. Należy określić nazwę pliku. Można wpisać nazwę pliku lub kliknąć przycisk wielokropka (**...**), aby wybrać plik. Po wybraniu pliku zostanie wyświetlona jego ścieżka.

Plik jest zaszyfrowany hasłem. Należy zaznaczyć to pole, jeśli wiadomo, że plik jest zabezpieczony hasłem; po wyświetleniu monitu należy wprowadzić **Hasło**. Jeśli plik jest zabezpieczony hasłem i nie zostanie ono wprowadzone, przy próbie przejścia do innej karty, odświeżenia danych, wyświetlenia podglądu zawartości węzła lub próbie wykonania strumienia zawierającego węzeł zostanie wyświetlone ostrzeżenie.

Uwaga: Pliki zabezpieczone hasłem można otwierać tylko w programie IBM SPSS Modeler w wersji 16 lub wyższej.

Nazwy zmiennych. Należy wybrać metodę obsługi nazw zmiennych i etykiet podczas importowania z pliku *.sav* lub *.zsav* programu IBM SPSS Statistics. Wybrane tutaj metadane do uwzględnienia będą dostępne podczas całej pracy w programie IBM SPSS Modeler i można je wyeksportować ponownie w celu użycia z narzędziem IBM SPSS Statistics.

- **Odczytaj nazwy i etykiety.** To pole należy zaznaczyć, aby w programie IBM SPSS Modeler odczytywane były nazwy i etykiety zmiennych. Domyślnie ta opcja jest zaznaczona i w węźle typu wyświetlane są nazwy zmiennych. Etykiety mogą być wyświetlane na wykresach, w przeglądarkach modeli oraz w innego typu wynikach, w zależności od opcji określonych w oknie dialogowym właściwości strumienia. Domyślnie opcja wyświetlania etykiet w wynikach jest wyłączona.
- **Odczytaj etykiety jako nazwy.** Tę opcję należy wybrać, aby odczytywać opisowe etykiety zmiennej z pliku *.sav* lub *.zsav* programu IBM SPSS Statistics zamiast krótkich nazw zmiennych i używać tych etykiet jako nazwy zmiennych w programie IBM SPSS Modeler.

Wartości. Należy wybrać metodę obsługi wartości i etykiet podczas importowania z pliku *.sav* lub *.zsav* programu IBM SPSS Statistics. Wybrane tutaj metadane do uwzględnienia będą dostępne podczas całej pracy w programie IBM SPSS Modeler i można je wyeksportować ponownie w celu użycia z narzędziem IBM SPSS Statistics.

- **Odczytaj dane i etykiety.** Tę opcję należy wybrać, aby odczytywać rzeczywiste wartości i wartości etykiet w programie IBM SPSS Modeler. Domyślnie ta opcja jest zaznaczona i wartości są wyświetlane w węźle typu. Etykiety wartości mogą być wyświetlane w konstruktorze wyrażeń, na wykresach, w przeglądarkach modeli oraz w innego typu wynikach, w zależności od opcji określonych w oknie dialogowym właściwości strumienia.
- **Odczytaj etykiety jako dane.** Tę opcję należy wybrać, jeśli zamiast kodów numerycznych lub symbolicznych używanych do reprezentowania wartości używane były etykiety wartości z pliku *.sav* lub *.zsav*. Na przykład, po wybraniu tej opcji dla danych zawierających zmienną płci, których wartości 1 i 2 w rzeczywistości oznaczają odpowiednio *male* (mężczyzna) i *female* (kobieta), nastąpi przekształcenie zmiennej na łańcuch i zaimportowanie zmiennych *male* i *female* jako wartości rzeczywiste.

  Istotne jest, aby przed zaznaczeniem tej opcji w danych programu IBM SPSS Statistics uwzględnić braki danych. Przykładowo, jeśli zmienna numeryczna używa etykiet tylko dla braków danych (0 = *No Answer* (Brak odpowiedzi), –99 = *Unknown* (Nieznane)), wówczas zaznaczenie powyższej opcji spowoduje zaimportowanie tylko etykiet wartości *No Answer* i *Unknown* i przekształcenie zmiennej na łańcuch. W takich przypadkach należy importować same wartości i ustawić braki danych w węźle typu.

Użyj informacji o formacie zmiennej w celu wymuszenia typu składowania. Jeśli to pole nie jest zaznaczone, wartości zmiennych sformatowane w pliku *.sav* jako liczby całkowite (tj. zmienne określone jako F*n*.0 w widoku zmiennych w programie IBM SPSS Statistics) są importowane z użyciem składowania jako liczba całkowita. Wszystkie pozostałe wartości zmiennych, z wyjątkiem łańcuchów, są importowane jako liczby rzeczywiste.

Jeśli to pole jest zaznaczone (ustawienie domyślne), wszystkie wartości zmiennych, z wyjątkiem łańcuchów, są importowane jako liczby rzeczywiste, niezależnie od tego, czy w pliku *.sav* zostały sformatowane jaki liczby całkowite.

Zestawy wielokrotnych odpowiedzi. Wszystkie zestawy wielokrotnych odpowiedzi zdefiniowane w pliku IBM SPSS Statistics zostaną automatycznie zachowane po zaimportowaniu pliku. Zestawy wielokrotnych odpowiedzi można wyświetlać i edytować w dowolnym węźle po wybraniu zakładki Filtr. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi” na stronie 153.

Węzeł Przekształcenia Statistics

Węzeł Przekształcenia Statistics umożliwia przeprowadzenie przekształceń danych za pośrednictwem składni komend IBM SPSS Statistics. Umożliwia on wykonanie wielu przekształceń, które nie są obsługiwane w programie IBM SPSS Modeler oraz zezwala na automatyzację złożonych, wieloetapowych przekształceń, w tym na tworzenie wielu zmiennych z jednego węzła. Węzeł ten jest podobny do węzła Wynik Statistics, z tym że dane są tu zwracane do programu IBM SPSS Modeler w celu przeprowadzenia dalszej analizy, podczas gdy dane węzła wynikowego są zwracane w postaci obiektów wynikowych, takich jak wykresy lub tabele.

Aby użyć tego węzła, na komputerze musi być zainstalowana kompatybilna wersja programu IBM SPSS Statistics z odpowiednią licencją. Więcej informacji można znaleźć w “IBM SPSS Statistics — aplikacje pomocnicze” na stronie 345. Informacje na temat kompatybilności zawiera publikacja Uwagi do wydania.

W razie potrzeby można użyć karty Filtrowanie, aby przeprowadzić filtrowanie lub zmienić nazwy zmiennych, tak aby były zgodne ze standardami nadawania nazw w programie IBM SPSS Statistics. Więcej informacji można znaleźć w “Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics” na stronie 364.

Materiały o składni. Szczegóły na temat procedur specyficznych dla programu IBM SPSS Statistics zawiera publikacja *IBM SPSS Statistics — Materiały o składni komend* dołączona do oprogramowania IBM SPSS Statistics. Aby wyświetlić tę publikację z karty Składnia, należy wybrać opcję **Edytor komend** i kliknąć przycisk **Uruchom pomoc dotyczącą poleceń programu Statistics**.

Uwaga: Nie wszystkie elementy składni IBM SPSS Statistics są obsługiwane w tym węźle. Więcej informacji można znaleźć w temacie “Dozwolona składnia” na stronie 380.

Węzeł Przekształcenia Statistics — karta Składnia

Opcja okna dialogowego programu IBM SPSS Statistics

Jeśli użytkownik nie jest zaznajomiony ze składnią IBM SPSS Statistics, najprostszym sposobem na jej utworzenie w programie IBM SPSS Modeler jest wybranie opcji **Okno dialogowe IBM SPSS Statistics**, wybranie okna dialogowego dla procedury, wypełnienie tego okna i kliknięcie przycisku OK. W ten sposób składnia zostanie zawarta na karcie Składnia w węźle IBM SPSS Statistics, który jest używany w programie IBM SPSS Modeler. Następnie należy uruchomić strumień, aby uzyskać wynik dla procedury.

IBM SPSS Statistics — opcja Edytor komend

Sprawdź. Po wprowadzeniu komend składni w górnej części okna dialogowego należy użyć tego przycisku, aby przeprowadzić walidację wpisów. Jeśli wykryta zostanie niepoprawna składnia, zostanie ona wskazana w dolnej części okna dialogowego.

Aby proces sprawdzania nie trwał zbyt długo, podczas walidacji składni sprawdzana jest reprezentatywna próba danych, co pozwoli na upewnienie się, że zapisy są poprawne bez konieczności sprawdzania całego zbioru danych.

Dozwolona składnia

Jeśli dostępna jest składnia z poprzednich wersji programu IBM SPSS Statistics lub jeśli użytkownik jest zaznajomiony z funkcjami przygotowywania danych w programie IBM SPSS Statistics, można użyć węzła Przekształcenia Statistics, aby uruchomić wiele z istniejących przekształceń. Dla ułatwienia węzeł umożliwia przekształcenie danych w przewidywalny sposób — na przykład, poprzez uruchomienie zapętlonych komend lub poprzez zmianę, dodanie, sortowanie, filtrowanie lub wybieranie danych.

Poniżej przedstawiono przykładowe komendy, jakie można wykonać:

- Obliczanie liczb losowych zgodnie z rozkładem dwumianowym:
  ```
  COMPUTE newvar = RV.BINOM(10000,0.1)
  ```
- Rekodowanie zmiennej na nową zmienną:
  ```
  RECODE Age (Lowest thru 30=1) (30 thru 50=2) (50 thru Highest=3) INTO AgeRecoded
  ```
- Zastępowanie braków danych:

```
RMV Age_1=SMEAN(Age)
```

Poniżej wymieniono składnię IBM SPSS Statistics obsługiwaną przez węzeł Przekształcenia Statistics.

Nazwa komendy

ADD VALUE LABELS

APPLY DICTIONARY

AUTORECODE

BREAK

CD

CLEAR MODEL PROGRAMS

CLEAR TIME PROGRAM

CLEAR TRANSFORMATIONS

COMPUTE

COUNT

CREATE

DATA

DEFINE-!ENDDEFINE

DELETE VARIABLES

DO IF

DO REPEAT

ELSE

ELSE IF

END CASE

END FILE

END IF

END INPUT PROGRAM

END LOOP

END REPEAT

EXECUTE

FILE HANDLE

FILE LABEL

Nazwa komendy

FILE TYPE-END FILE TYPE
FILTER
FORMATS
IF
INCLUDE
INPUT PROGRAM-END INPUT PROGRAM
INSERT
LEAVE
LOOP-END LOOP
MATRIX-END MATRIX
MISSING VALUES
N OF CASES
NUMERIC
PERMISSIONS
PRESERVE
RANK
RECODE
RENAME VARIABLES
RESTORE
RMV
SAMPLE
SELECT IF
SET
SORT CASES
STRING
SUBTITLE
TEMPORARY
TITLE
UPDATE
V2C
VALIDATEDATA
VALUE LABELS
VARIABLE ATTRIBUTE
VARSTOCASES
VECTOR

Węzeł Model Statistics

Węzeł Model Statistics umożliwia analizowanie danych i pracę z nimi poprzez uruchomienie procedur IBM SPSS Statistics tworzących PMML. Utworzony model użytkowy może być wówczas używany w normalny sposób w strumieniach IBM SPSS Modeler do przeprowadzania oceny itd.

Aby użyć tego węzła, na komputerze musi być zainstalowana kompatybilna wersja programu IBM SPSS Statistics z odpowiednią licencją. Więcej informacji można znaleźć w “IBM SPSS Statistics — aplikacje pomocnicze” na stronie 345. Informacje na temat kompatybilności zawiera publikacja Uwagi do wydania.

Dostępność procedur analitycznych IBM SPSS Statistics będzie zależała od typu używanej licencji.

Węzeł Model Statistics — karta Model

Nazwa modelu. Można wygenerować nazwę modelu, która będzie automatycznie oparta na zmiennej przewidywanej lub identyfikacyjnej (albo na typie modelu w przypadkach, gdy żadna taka zmienna nie jest określona), albo określić nazwę niestandardową.

Wybierz okno dialogowe. Kliknij, aby wyświetlić listę dostępnych procedur IBM SPSS Statistics, które można wybrać i uruchomić. Na liście wyświetlone są tylko te procedury, które tworzą PMML i dla których użytkownik ma licencję; lista nie zawiera procedur napisanych przez użytkownika.

1. Kliknij wybraną procedurę; zostanie wyświetlone odpowiednie okno dialogowe IBM SPSS Statistics.
2. W oknie dialogowym IBM SPSS Statistics wprowadź szczegóły dotyczące procedury.
3. Kliknij przycisk **OK**, aby powrócić do węzła Model Statistics; na karcie Model zostanie wyświetlona składnia IBM SPSS Statistics.
4. Aby powrócić do okna dialogowego IBM SPSS Statisticsw dowolnym czasie, na przykład w celu zmodyfikowania zapytania, kliknij przycisk wyświetlania okna dialogowego IBM SPSS Statistics po prawej stronie przycisku wyboru procedury.

Węzeł Model Statistics — podsumowanie modelu użytkowego

Po uruchomieniu węzła Model Statistics wykonywana jest powiązana procedura IBM SPSS Statistics oraz tworzony jest model użytkowy, którego można używać w strumieniach IBM SPSS Modeler do przeprowadzania oceny.

Karta Podsumowanie modelu użytkowego umożliwia wyświetlenie informacji na temat zmiennych, ustawień budowania i procesu estymacji modelu. Wyniki są prezentowane w widoku drzewa, które może być rozwijane i zwijane poprzez kliknięcie konkretnych elementów.

Przycisk **Wyświetl model** wyświetla wyniki w zmodyfikowanej postaci dla przeglądarki wyników IBM SPSS Statistics. Więcej informacji na temat tej przeglądarki zawiera dokumentacja programu IBM SPSS Statistics.

Typowe opcje eksportowania i drukowania są dostępne w menu Plik. Więcej informacji można znaleźć w temacie “Wyświetlanie wyników” na stronie 303.

Węzeł Wynik Statistics

Węzeł Wynik Statistics umożliwia wywołanie procedury IBM SPSS Statistics w celu przeprowadzenia analizy danych IBM SPSS Modeler. Wyniki można wyświetlić w oknie przeglądarki lub zapisać je w formacie pliku wynikowego programu IBM SPSS Statistics. W programie IBM SPSS Modeler dostępne są różnorodne procedury analityczne modułu IBM SPSS Statistics.

Aby użyć tego węzła, na komputerze musi być zainstalowana kompatybilna wersja programu IBM SPSS Statistics z odpowiednią licencją. Więcej informacji można znaleźć w “IBM SPSS Statistics — aplikacje pomocnicze” na stronie 345. Informacje na temat kompatybilności zawiera publikacja Uwagi do wydania.

W razie potrzeby można użyć karty Filtrowanie, aby przeprowadzić filtrowanie lub zmienić nazwy zmiennych, tak aby były zgodne ze standardami nadawania nazw w programie IBM SPSS Statistics. Więcej informacji można znaleźć w “Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics” na stronie 364.

Materiały o składni. Szczegóły na temat procedur specyficznych dla programu IBM SPSS Statistics zawiera publikacja *IBM SPSS Statistics — Materiały o składni komend* dołączona do oprogramowania IBM SPSS Statistics. Aby wyświetlić tę publikację z karty Składnia, należy wybrać opcję **Edytor komend** i kliknąć przycisk **Uruchom pomoc dotyczącą poleceń programu Statistics**.

Węzeł Wynik Statistics — karta Składnia

Ta karta umożliwia tworzenie składni dla procedury programu SPSS Statistics, jaka ma zostać użyta do analizy danych. Składnia składa się z dwóch części: **instrukcja** i powiązane z nią **opcje**. Instrukcja określa analizę lub operacje, jakie mają zostać wykonane oraz zmienne, jakie będą użyte. Opcje określają całą resztę, w tym: które statystyki będą wyświetlane, zmienne pochodne do zapisania itd.

Opcja okna dialogowego programu SPSS Statistics

Jeśli użytkownik nie jest zaznajomiony ze składnią IBM SPSS Statistics, najprostszym sposobem na jej utworzenie w programie IBM SPSS Modeler jest wybranie opcji **Okno dialogowe IBM SPSS Statistics**, wybranie okna dialogowego dla procedury, wypełnienie tego okna i kliknięcie przycisku OK. W ten sposób składnia zostanie zawarta na karcie Składnia w węźle IBM SPSS Statistics, który jest używany w programie IBM SPSS Modeler. Następnie należy uruchomić strumień, aby uzyskać wynik dla procedury.

Opcjonalnie można wygenerować węzeł źródłowy Plik Statistics w celu zaimportowania danych wynikowych. Jest to przydatne, na przykład, jeśli procedura oprócz wyświetlania wyniku zapisuje zmienne, takie jak oceny, w aktywnym zbiorze danych.

Uwaga:

- Podczas generowania wyniku w języku innym niż angielski zaleca się określenie języka w składni.
- Opcja Styl wyniku nie jest obsługiwana w węźle Wynik Statistics.

Aby utworzyć składnię

1. Kliknij przycisk **Wybierz okno dialogowe**.
2. Wybierz jedną z opcji:
   - **Analiza** Wyświetla zawartość menu analizy programu SPSS Statistics; wybierz procedurę, jakiej zamierzasz użyć.
   - **Inne** Jeśli ta opcja jest widoczna, wyświetla okna dialogowe utworzone za pomocą kreatora niestandardowych okien dialogowych w programie SPSS Statistics, jak również pozostałe okna dialogowe SPSS Statistics, które nie są wyświetlane w menu Analiza i dla których użytkownik ma licencję. Jeśli nie ma odpowiednich okien, ta opcja nie jest widoczna.

Uwaga: Okna dialogowe automatycznego przygotowania danych nie są widoczne.

Jeśli dostępne jest niestandardowe okno dialogowe SPSS Statistics, które tworzy nowe zmienne, zmienne te nie mogą być używane w programie SPSS Modeler, ponieważ węzeł Wynik Statistics jest węzłem końcowym.

Opcjonalnie można zaznaczyć pole wyboru **Wygeneruj węzeł importu dla danych wynikowych**, aby utworzyć węzeł źródłowy Plik Statistics, którego można będzie użyć do zaimportowania danych wynikowych do innego strumienia. Węzeł jest umieszczany w obszarze roboczym ekranu, wraz z danymi zawartymi w pliku .sav określonym przez zmienną **Plik** (domyślna lokalizacja to katalog instalacyjny programu SPSS Modeler).

Opcja edytora komend

Aby zapisać składnię utworzoną dla często używanej procedury:

1. Kliknij przycisk **Opcje pliku** (pierwszy na pasku narzędzi).
2. Wybierz z menu polecenie **Zapisz** lub **Zapisz jako**.
3. Zapisz plik jako plik .sps.

Aby użyć wcześniej utworzonych plików składni, zastępując aktualną zawartość, o ile istnieje, edytora komend:

1. Kliknij przycisk Opcje pliku (pierwszy na pasku narzędzi).
2. Wybierz z menu opcję **Otwórz**.
3. Wybierz plik .sps, aby wkleić jego zawartość na karcie Składnia węzła wynikowego.

Aby wstawić wcześniej zapisaną składnię bez zastępowania aktualnej zawartości:

1. Kliknij przycisk Opcje pliku (pierwszy na pasku narzędzi).
2. Wybierz z menu opcję **Wstaw**.
3. Wybierz plik .sps, aby wkleić jego zawartość w węźle wynikowym w miejscu wskazanym przez kursor.

Opcjonalnie można zaznaczyć pole wyboru **Wygeneruj węzeł importu dla danych wynikowych**, aby utworzyć węzeł źródłowy Plik Statistics, którego można będzie użyć do zaimportowania danych wynikowych do innego strumienia. Węzeł jest umieszczany w obszarze roboczym ekranu, wraz z danymi zawartymi w pliku .sav określonym przez zmienną **Plik** (domyślna lokalizacja to katalog instalacyjny programu SPSS Modeler).

Po kliknięciu przycisku **Uruchom** wyniki są wyświetlane w przeglądarce wyników SPSS Statistics. Więcej informacji na temat przeglądarki zawiera dokumentacja programu SPSS Statistics.

Uwaga: Dla poniższych pozycji (i odpowiadających im opcji w oknie dialogowym SPSS Statistics) składnia nie jest obsługiwana. Nie mają one wpływu na wyniki.

- OUTPUT ACTIVATE
- OUTPUT CLOSE
- OUTPUT DISPLAY
- OUTPUT EXPORT
- OUTPUT MODIFY
- OUTPUT NAME
- OUTPUT NEW
- OUTPUT OPEN
- OUTPUT SAVE

Węzeł Wynik Statistics — karta Wynik

Karta Wynik umożliwia określenie formatu i lokalizacji wyniku. Wyniki mogą być wyświetlane na ekranie lub można je wysłać do jednego z dostępnych typów plików.

Nazwa wyniku. Określa nazwę uzyskanego wyniku po wykonaniu węzła. **Automatycznie** wybiera nazwę na podstawie węzła, który spowodował wygenerowanie wyniku. Opcjonalnie można wybrać opcję **Użytkownika**, aby określić inną nazwę.

Wynik na ekran (ustawienie domyślne)**.** Tworzy obiekty wynikowe w celu ich wyświetlenia online. Obiekt wynikowy zostanie wyświetlony na kacie Wyniki w oknie menedżera po wykonaniu węzła wynikowego.

Wynik do pliku. Po uruchomieniu węzła zapisuje wynik w pliku. Jeśli ta opcja zostanie wybrana, należy wprowadzić nazwę pliku w polu **Nazwa pliku** (lub przejść do katalogu i określić nazwę pliku, używając przycisku selektora plików) i wybrać typ pliku.

Typ pliku. Należy wybrać typ pliku, do którego ma zostać wysłany wynik.

- **Dokument HTML (*.html).** Zapisuje wynik w formacie HTML.
- Plik **IBM SPSS Statistics Viewer (*.spv)** Zapisuje wynik w formacie, który może być odczytamy przez przeglądarkę wyników IBM SPSS Statistics.

- Plik **IBM SPSS Statistics Web Reports (*.spw).** Zapisuje wynik w formacie IBM SPSS Statistics Web Reports, który może być opublikowany w repozytorium IBM SPSS Collaboration and Deployment Services i następnie wyświetlany w przeglądarce WWW. Więcej informacji można znaleźć w temacie “Publikowanie w sieci WWW” na stronie 303.

Uwaga: Jeśli wybrana zostanie opcja **Wynik na ekran** dyrektywa OMS IBM SPSS Statistics VIEWER=NO nie będzie działać; również interfejsy API tworzenia skryptów (moduły *Basic* i *Python SpssClient*) nie będą dostępne w programie IBM SPSS Modeler.

Węzeł eksportu Statistics

Węzeł eksportu Statistics umożliwia eksportowanie danych w formacie *.sav* programu IBM SPSS Statistics. Pliki *.sav* programu IBM SPSS Statistics mogą być odczytywane przez produkt IBM SPSS Statistics i inne moduły. Jest to również format używany przez pliki pamięci podręcznej w programie IBM SPSS Modeler.

Mapowanie nazw zmiennych IBM SPSS Modeler na nazwy zmiennych IBM SPSS Statistics może czasami powodować błędy, ponieważ nazwy zmiennych IBM SPSS Statistics są ograniczone do 64 znaków i nie mogą zawierać niektórych znaków, takich jak spacje, znaki dolara ($) i myślniki (–). Istnieją dwa sposoby na skorygowanie tych ograniczeń:

- Można zmienić nazwy zmiennych, tak aby były zgodne z wymaganiami nazw zmiennych w programie IBM SPSS Statistics, klikając w tym celu zakładkę Filtr. Więcej informacji można znaleźć w temacie “Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics” na stronie 364.
- Wybrać opcję wyeksportowania nazw zmiennych wraz z etykietami z programu IBM SPSS Modeler.

Uwaga: IBM SPSS Modeler zapisuje pliki *.sav* w formacie Unicode UTF-8. IBM SPSS Statistics obsługuje tylko pliki w formacie Unicode UTF-8 począwszy od wersji 16.0. Aby uniknąć możliwości uszkodzenia danych, pliki *.sav* zapisane zgodnie z kodowaniem Unicode nie powinny być używane w programie IBM SPSS Statistics w wersjach wcześniejszych niż 16.0. Więcej informacji zawiera pomoc programu IBM SPSS Statistics.

Zestawy wielokrotnych odpowiedzi. Wszystkie zestawy wielokrotnych odpowiedzi zdefiniowane w strumieniu zostaną automatycznie zachowane po wyeksportowaniu pliku. Zestawy wielokrotnych odpowiedzi można wyświetlać i edytować w dowolnym węźle po wybraniu zakładki Filtr. Więcej informacji można znaleźć w temacie “Edytowanie zestawów wielokrotnych odpowiedzi” na stronie 153.

Węzeł eksportu Statistics — karta eksportu

Eksportuj plik Określa nazwę pliku. Należy wprowadzić nazwę pliku lub kliknąć przycisk selektora plików, aby wybrać lokalizację pliku.

Typ pliku Tę opcję należy wybrać, jeśli plik ma zostać zapisany w normalnym pliku *.sav* lub w formacie skompresowanym *.zsav*.

Zaszyfruj plik hasłem Aby zabezpieczyć plik hasłem, należy zaznaczyć to pole wyboru; zostanie wyświetlony monit o wprowadzenie i potwierdzenie **hasła** w osobnym oknie dialogowym.

Uwaga: Pliki zabezpieczone hasłem można otwierać tylko w programie SPSS Modeler w wersji 16 lub wyższej lub w programie SPSS Statistics w wersji 21 lub wyższej.

Eksportuj nazwy zmiennych Określa metodę obsługi nazw zmiennych i etykiet podczas eksportowania z programu SPSS Modeler do pliku SPSS Statistics *.sav* lub *.zsav*.

- **Nazwy i etykiety zmiennych** Tę opcję należy wybrać, aby wyeksportować nazwy zmiennych i etykiety zmiennych SPSS Modeler. Nazwy są eksportowane jako nazwy zmiennych SPSS Statistics, a etykiety jako etykiety zmiennych SPSS Statistics.
- **Nazwy jako etykiety zmiennych** Ta opcja umożliwia użycie nazw zmiennych SPSS Modeler jako etykiety zmiennych w systemie SPSS Statistics. W programie SPSS Modeler dozwolone jest użycie w nazwach zmiennych

znaków, które w systemie SPSS Statistics są niepoprawne. Aby uniknąć utworzenia niepoprawnych nazw w programie SPSS Statistics, należy wybrać opcję **Nazwy jako etykiety zmiennych** lub użyć karty Filtrowanie, aby skorygować nazwy zmiennych.

Uruchom aplikację Jeśli program SPSS Statistics jest zainstalowany na komputerze, można wybrać tę opcję, aby wywołać aplikację bezpośrednio z zapisanego pliku danych. Opcje uruchamiania aplikacji muszą być określone w oknie dialogowym Aplikacje pomocnicze. Więcej informacji można znaleźć w temacie “IBM SPSS Statistics — aplikacje pomocnicze” na stronie 345. Aby utworzyć plik *.sav* lub *.zsav* programu SPSS Statistics bez otwierania programu zewnętrznego, należy usunąć zaznaczenie tej opcji.

Uwaga: Podczas uruchamiania SPSS Modeler i SPSS Statistics w trybie serwera (rozproszonym) zapisywanie danych i uruchamianie sesji SPSS Statistics nie powoduje automatycznego otwarcia klienta SPSS Statistics i wyświetlenia zbioru danych w aktywnym zbiorze danych. Aby tego uniknąć, należy ręcznie otworzyć plik danych w kliencie SPSS Statistics po jego uruchomieniu.

Wygeneruj węzeł importu dla tych danych Tę opcję należy wybrać, aby automatycznie wygenerować węzeł źródłowy Plik Statistics, który odczyta wyeksportowany plik danych. Więcej informacji można znaleźć w temacie “Węzeł Plik Statistics” na stronie 31.

Zmiana nazw lub filtrowanie zmiennych dla programu IBM SPSS Statistics

Przed wyeksportowaniem lub wdrożeniem danych z programu IBM SPSS Modeler do aplikacji zewnętrznych, takich jak IBM SPSS Statistics, konieczna może być zmiana nazwy lub skorygowanie nazw zmiennych. Okna dialogowe Przekształcenia Statistics, Wynik Statistics i Plik Statistics zawierają zakładkę Filtr, która upraszcza ten proces.

Główny opis funkcji karty Filtr został zamieszczony w innym miejscu. Więcej informacji można znaleźć w temacie “Ustawianie opcji filtrowania” na stronie 152. Ten temat zawiera wskazówki dotyczące odczytu danych w programie IBM SPSS Statistics.

Aby skorygować nazwy, tak aby były zgodne z konwencjami nadawania nazw IBM SPSS Statistics:

1. Na karcie Filtr kliknij pasek narzędzi Menu opcji filtrowania (pierwszy na pasku narzędzi).
2. Wybierz opcję Zmień nazwy dla IBM SPSS Statistics.
3. W oknie dialogowym Zmień nazwy dla IBM SPSS Statistics można zastąpić niepoprawne znaki w nazwach plików, wybierając **Krzyżyk (#)** lub **Podkreślenie (_)**.

Zmień nazwę zestawów wielokrotnych odpowiedzi. Tę opcję należy wybrać, aby skorygować nazwy zestawów wielokrotnych odpowiedzi, które można zaimportować do IBM SPSS Modeler za pośrednictwem węzła źródłowego Plik Statistics. Zestawy służą do rejestrowania danych, które mogą mieć więcej niż jedną wartość dla każdej obserwacji, np. odpowiedzi w ankiecie.

Rozdział 9. Superwęzły

Przegląd informacji dotyczących Superwęzłów

Wizualny interfejs programistyczny IBM SPSS Modeler jest łatwy w obsłudze m.in. dzięki temu, że każdy węzeł pełni wyraźnie zdefiniowaną funkcję. W przypadku złożonego przewarzania konieczne może być jednak użycie długiej sekwencji węzłów. Może to jednak czasami spowodować chaos w obszarze roboczym strumienia i utrudnić śledzenie diagramów strumienia. Istnieją dwa sposoby, aby uniknąć zakłóceń spowodowanych przez długi i złożony strumień:

- Można podzielić sekwencję przetwarzania na kilka zależnych od siebie strumieni. Przykładowo, pierwszy strumień tworzy plik danych, który w drugim węźle jest używany jako dane wejściowe. Drugi węzeł tworzy plik, który w trzecim węźle stanowi dane wejściowe itd. Można zarządzać tymi strumieniami, zapisując je w **projekcie**. Projekt zapewnia uporządkowanie wielu strumieni i ich wyników. Plik projektu zawiera jednak tylko odwołanie do obiektów i nadal konieczne jest zarządzanie plikami wielu strumieni.
- Podczas pracy ze złożonymi procesami strumienia uproszczonym rozwiązaniem jest utworzenie **Superwęzła**.

Superwęzły grupują wiele węzłów w jednym węźle, zamykając osobne sekcje strumienia danych. Zapewnia to wiele korzyści dla specjalisty ds. eksploracji danych:

- Strumienie są bardziej przejrzyste i łatwiejsze do zarządzania.
- Węzły można łączyć w Superwęzeł specyficzny dla danej branży.
- Superwęzły mogą być eksportowane do bibliotek, aby mogły zostać ponownie użyte w innych projektach eksploracji danych.

Typy Superwęzłów

Superwęzły są reprezentowane w strumieniu danych przez ikonę gwiazdy. Ikona ma różne odcienie reprezentujące typ Superwęzła oraz kierunek, w którym strumień musi przebiegać do lub z Superwęzła.

Dostępne są trzy typy Superwęzłów:

- Superwęzły źródłowe
- Superwęzły procesowe
- Superwęzły końcowe

Superwęzły źródłowe

Superwęzeł źródłowy zawiera źródło danych, tak jak zwykły węzeł źródłowy i może być użyty wszędzie tam, gdzie zwykły węzeł źródłowy. Lewa strona Superwęzła źródła jest zacieniowana, co wskazuje, że jest on „zamknięty” po lewej stronie i dane muszą przepływać *od* Superwęzła w dół strumienia.

Superwęzły źródłowe mają tylko jeden punkt połączenia po prawej stronie, co oznacza, że dane opuszczają Superwęzeł i wpływają do strumienia.

Superwęzły procesowe

Superwęzły procesowe zawierają tylko węzły procesowe i nie są cieniowane, co oznacza, że dane mogą przepływać *do* i *z* tego Superwęzła.

W Superwęzłach procesowych punkty połączeń znajdują się po lewej i po prawej stronie, co oznacza, że dane są wprowadzane do Superwęzła i wychodzą z niego, aby wpłynąć z powrotem do strumienia. Superwęzły mogą zawierać dodatkowe fragmenty strumienia, a nawet dodatkowe strumienie, jednak oba punkty połączeń muszą przepływać przez pojedynczą ścieżkę łączącą punkty *Ze strumienia* i *Do strumienia*.

Uwaga: Superwęzły procesowe są niekiedy zwane *Superwęzłami operacyjnymi*.

Superwęzły końcowe

Superwęzeł końcowy zawiera jeden lub kilka węzłów końcowych (wykres, tabela itd.) i mogą być używane w taki sam sposób, jak węzeł końcowy. Superwęzeł końcowy jest zacieniowany po prawej stronie, co oznacza, że jest „zamknięty" po prawej stronie i dane mogą jedynie wpływać *do* końcowego Superwęzła.

Superwęzły końcowe mają tylko jeden punkt połączenia po lewej stronie, co oznacza, że dane są wprowadzane do Superwęzła ze strumienia i ich przepływ kończy się wewnątrz Superwęzła.

Superwęzły końcowe mogą również zawierać skrypty, które są używane do wykonywania wielu węzłów końcowych wewnątrz tego Superwęzła. Więcej informacji można znaleźć w temacie "Superwęzły i tworzenie skryptów" na stronie 393.

Tworzenie Superwęzłów

Utworzenie Superwęzła „skraca" strumień danych poprzez zamknięcie (opakowanie) kilku węzłów w jednym. Po utworzeniu lub załadowaniu strumienia do obszaru roboczego można utworzyć Superwęzeł na kilka sposobów.

Wybór wielokrotny

Najprostszym sposobem na utworzenie Superwęzła jest wybranie wszystkich węzłów, jakie mają zostać opakowane (zamknięte w obudowie):

1. Za pomocą myszy wybierz wiele węzłów z obszaru roboczego strumienia. Można również użyć metody Shift+kliknięcie, aby wybrać strumień lub część strumienia.

   Uwaga: Wybierane węzły muszą należeć do strumienia ciągłego lub rozgałęzionego. Nie można wybrać węzłów, które nie są sąsiadujące lub połączone w jakiś sposób.
2. Następnie, wybierając jedną z poniższych metod, opakuj (zamknij) wybrane węzły:
   - Kliknij ikonę Superwęzła (kształt gwiazdy) na pasku narzędzi.
   - Kliknij prawym przyciskiem myszy Superwęzeł i z menu kontekstowego wybierz opcje:
     Utwórz Superwęzeł > Z wyboru
   - Z menu Superwęzła wybierz opcje:
     Utwórz Superwęzeł > Z wyboru

Wszystkie te trzy opcje powodują opakowanie (zamknięcie) węzłów i utworzenie zacieniowanego Superwęzła; cieniowanie odzwierciedla typ Superwęzła — źródłowy, procesowy lub końcowy — w oparciu o jego zawartość.

Pojedynczy wybór

Superwęzeł można również utworzyć, wybierając jeden węzeł i za pomocą opcji menu ustalić, gdzie jest początek i koniec Superwęzła lub opakować (zamknąć) wszystko, co znajduje się poniżej wybranego węzła.

1. Kliknij węzeł, który określi początek opakowania.
2. Z menu Superwęzła wybierz opcje:
   Utwórz Superwęzeł > Z tego miejsca

Superwęzły można również tworzyć w sposób bardziej interaktywny poprzez wybranie początku i końca sekcji strumienia w celu opakowania węzłów:

1. Kliknij pierwszy lub ostatni węzeł, jaki miał zostać uwzględniony w Superwęźle.
2. Z menu Superwęzła wybierz opcje:
   Utwórz Superwęzeł > Wybierz...
3. Alternatywnie można użyć opcji menu kontekstowego, klikając prawym przyciskiem myszy na wybranym węźle.

4. Kursor zmieni się w ikonę Superwęzła, wskazując, że należy wybrać kolejny punkt w strumieniu. Można przejść w górę lub w dół strumienia, aby wybrać drugi koniec Superwęzła, a następnie kliknąć węzeł. Ta czynność zastąpi wszystkie węzły znajdujące się pomiędzy wybranymi punktami ikoną gwiazdy Superwęzła.

Uwaga: Wybierane węzły muszą należeć do strumienia ciągłego lub rozgałęzionego. Nie można wybrać węzłów, które nie są sąsiadujące lub połączone w jakiś sposób.

Zagnieżdżanie Superwęzłów

Superwęzły można zagnieżdżać w innych Superwęzłach. W przypadku zagnieżdżonych Superwęzłów obowiązują te same reguły dotyczące typów Superwęzłów (źródłowy, procesowy i końcowy). Przykładowo, Superwęzeł procesowy z zagnieżdżaniem musi zapewniać ciągły przepływ danych przez wszystkie zagnieżdżone Superwęzły, aby węzeł ten pozostał Superwęzłem procesowym. Jeśli jeden z zagnieżdżonych Superwęzłów będzie końcowy, wówczas dane nie będą przepływać wg hierarchii.

Superwęzły źródłowe i wyników mogą zawierać zagnieżdżone Superwęzły innego typu, ale zastosowanie mają te same podstawowe reguły, jak w przypadku tworzenia Superwęzłów.

Blokowanie Superwęzłów

Po utworzeniu Superwęzła można go zablokować hasłem, aby uniemożliwić wprowadzanie zmian. Przykładowo można to zrobić w przypadku tworzenia strumieni lub części strumieni jaki szablony stałych wartości przeznaczone do użycia przez innych użytkowników w organizacji, którzy mają mniejsze doświadczenie w konfigurowaniu zapytań IBM SPSS Modeler.

Jeśli Superwęzeł jest zablokowany, użytkownicy mogą wprowadzać wartości na karcie Parametry dla wszystkich zdefiniowanych parametrów, a zablokowany Superwęzeł może być wykonywany bez wprowadzania hasła.

Uwaga: Blokowania i usuwania blokady nie można wykonać za pomocą skryptów.

Blokowanie i usuwanie blokady Superwęzła

Uwaga: Utraconych haseł nie można odzyskać.

Blokowania i usuwania blokady Superwęzła można dokonać za pośrednictwem jednej z trzech kart.

1. Kliknij opcję **Blokada węzła**.
2. Wprowadź i potwierdź hasło.
3. Kliknij przycisk **OK**.

Superwęzeł zabezpieczony hasłem jest w obszarze roboczym strumienia oznaczany symbolem małej kłódki znajdującej się u góry po lewej stronie ikony Superwęzła.

Usuwanie blokady Superwęzła

1. Aby trwale usunąć zabezpieczenie hasłem, kliknij opcję **Odblokuj węzeł**. Zostanie wyświetlony monit o podanie hasła.
2. Wprowadź hasło i kliknij przycisk **OK**. Superwęzeł nie będzie już zabezpieczony hasłem, a symbol kłódki przestanie być wyświetlany obok ikony w strumieniu.

W przypadku strumienia zapisanego w programie SPSS Modeler w wersji od 16 do 17.0, która zawiera zablokowany Superwęzeł, a po otwarciu strumienia w innym środowisku, takim jak IBM SPSS Collaboration and Deployment Services lub na komputerze Mac, zainstalowane środowisko JRE SPSS Modeler jest inne, należy go najpierw otworzyć, odblokować i ponownie zapisać w wersji 17.1 lub nowszej w starym środowisku, w którym był ostatnio zapisany.

Czasami podczas usuwania blokady węzła w strumieniu w wersji starszej niż 18 zostaje wyświetlony błąd informujący o nieprawidłowym haśle. Aby rozwiązać ten problem, należy ponownie otworzyć węzeł i usunąć blokadę wyłącznie za pomocą wersji IBM SPSS Modeler (lub nowszej) na tej samej platformie i z tymi samymi ustawieniami lokalnymi jak przy ostatnim zapisie. Następnie należy otworzyć go w wersji 18 lub nowszej. Potem należy zablokować węzeł i ponownie zapisać strumień.

Edytowanie zablokowanego Superwęzła

Podczas próby zdefiniowania parametrów lub wejścia do zablokowanego Superwęzła w celu wyświetlenia jego zawartości zostanie wyświetlony monit o wprowadzenie hasła.

Wprowadź hasło i kliknij przycisk **OK**.

Można teraz edytować definicje parametrów oraz wchodzić do i wychodzić z Superwęzła tak często, jak to konieczne, do czasu zamknięcia strumienia, w którym Superwęzeł się znajduje.

Należy pamiętać, że nie powoduje to usunięcia zabezpieczenia hasłem, a jedynie umożliwia dostęp do Superwęzła w celu wykonania stosownych działań. Więcej informacji można znaleźć w temacie “Blokowanie i usuwanie blokady Superwęzła” na stronie 389.

Edytowanie Superwęzłów

Po utworzeniu Superwęzła można zbadać go dokładniej, wchodząc do jego środka; jeśli Superwęzeł jest zablokowany, zostanie wyświetlony monit o wprowadzenie hasła. Więcej informacji można znaleźć w temacie “Edytowanie zablokowanego Superwęzła”.

Aby wyświetlić zawartość Superwęzła, można użyć ikony powiększania na pasku narzędzi programu IBM SPSS Modeler lub skorzystać z następującej metody:

1. Kliknij Superwęzeł prawym przyciskiem myszy.
2. Z menu kontekstowego wybierz opcję **Powiększ**.

Zawartość wybranego Superwęzła zostanie wyświetlona w nieco innym środowisku IBM SPSS Modeler, w którym łączniki będą wskazywały przepływ danych w strumieniu lub fragmencie strumienia. Na tym poziomie obszaru roboczego strumienia można wykonać kilka zadań:

- Modyfikowanie typu Superwęzła — źródłowy, procesowy lub końcowy.
- Tworzenie parametrów lub edytowanie wartości parametru. Parametry są używane podczas tworzenia skryptów i w wyrażeniach CLEM.
- Określenie opcji buforowania Superwęzła i jego podwęzłów.
- Tworzenie lub modyfikowanie skryptu Superwęzła (tylko Superwęzeł końcowy).

Modyfikowanie typów Superwęzłów

W niektórych przypadkach przydatna jest zmiana typu Superwęzła. Ta opcja jest dostępna tylko po wejściu do Superwęzła i dotyczy tylko Superwęzła na tym poziomie. W poniższej tabeli omówiono trzy typy Superwęzłów.

Tabela 53. Typy Superwęzłów

Typ Superwęzła	Opis
Superwęzeł źródłowy	Jedno połączenie wyjściowe
Superwęzeł procesowy	Dwa połączenia: jedno wejściowe, drugie wyjściowe
Superwęzeł końcowy	Jedno połączenie wejściowe

Aby zmienić typ Superwęzła

1. Wejdź do Superwęzła.
2. Z menu Superwęzła wybierz opcję **Typ Superwęzła**, a następnie wybierz typ.

Adnotacje i zmiana nazwy Superwęzłów

Istnieje możliwość zmiany nazwy Superwęzła, jaka jest wyświetlana w strumieniu oraz zapisania adnotacji używanych w projekcie lub raporcie. Aby uzyskać dostęp do tych właściwości:

- Kliknij Superwęzeł prawym przyciskiem myszy (nie wchodząc do węzła) i wybierz opcję **Zmień nazwę i skomentuj**.
- Alternatywnie, z menu Superwęzła wybierz opcję **Zmień nazwę i skomentuj**. Ta opcja jest dostępna po wejściu do Superwęzła i bez wchodzenia do Superwęzła.

W obu przypadkach otwierane jest okno dialogowe z wybraną kartą Adnotacje. Dostępne tutaj opcje umożliwiają dostosowanie nazwy wyświetlanej w obszarze roboczym strumienia i udostępnienie dokumentacji dotyczącej operacji związanych z Superwęzłem.

Używanie komentarzy z Superwęzłami

Jeśli Superwęzeł jest tworzony na podstawie węzła lub modelu użytkowego opatrzonego komentarzem, podczas dokonywania wyboru należy uwzględnić komentarz, aby był on wyświetlany w Superwęźle. Jeśli przy dokonywaniu wyboru opcja komentarza nie zostanie wybrana, po utworzeniu Superwęzła komentarz pozostanie w strumieniu.

Po rozwinięciu Superwęzła, który zawiera komentarze, zostaną one przywrócone do miejsca, w którym znajdowały się przed utworzeniem Superwęzła.

Po rozwinięciu Superwęzła, który zawiera obiekty opatrzone komentarzem, ale komentarze nie zostały uwzględnione w Superwęźle, obiekty zostaną przywrócone do miejsca, w którym się znajdowały, ale komentarze nie zostaną ponownie dołączone.

parametry Superwęzła

W programie IBM SPSS Modeler istnieje możliwość ustawienia zmiennych zdefiniowanych przez użytkownika, takich jak Minvalue (Wartość minimalna), których wartości mogą być określane podczas tworzenia skryptów lub wyrażeń CLEM. Zmienne te są nazywane **parametrami**. Parametry można ustawić dla strumieni, sesji i Superwęzłów. Wszystkie parametry ustawione dla Superwęzła są dostępne podczas tworzenia wyrażeń CLEM w tym Superwęźle lub w dowolnym węźle zagnieżdżonym. Parametry ustawione dla zagnieżdżonych Superwęzłów nie są dostępne dla ich nadrzędnego Superwęzła.

Tworzenie i ustawianie parametrów dla Superwęzłów odbywa się w dwóch krokach:

1. Zdefiniowanie parametrów dla Superwęzła.
2. Następnie określenie wartości dla każdego parametru Superwęzła.

Parametry te mogą być następnie użyte w wyrażeniach CLEM dla dowolnych opakowanych (zamkniętych) węzłów.

Definiowanie parametrów Superwęzła

Parametry dla Superwęzła można zdefiniować po wejściu do Superwęzła i bez wchodzenia do jego środka. Zdefiniowane parametry mają zastosowanie do wszystkich opakowanych (zamkniętych) węzłów. Aby zdefiniować parametry Superwęzła, najpierw należy przejść do karty Parametry w oknie dialogowym Superwęzła. Aby otworzyć to okno dialogowe, można użyć jednej z następujących metod:

- Dwukrotne kliknięcie Superwęzła w strumieniu.
- Wybranie opcji **Ustaw parametry** w menu Superwęzła.
- Alternatywnie, po wejściu do Superwęzła, wybranie opcji **Ustaw parametry** z menu kontekstowego.

Po otwarciu okna dialogowego karta Parametry jest wyświetlana bez zdefiniowanych parametrów.

Aby zdefiniować nowy parametr

Kliknij przycisk **Definiuj parametry**, aby otworzyć okno dialogowe.

Nazwa. Nazwy parametrów wymieniono tutaj. Nowy parametr można utworzyć, wprowadzając nazwę w tym polu. Na przykład, aby utworzyć parametr dla temperatury minimalnej, można wpisać: minvalue. Nie należy uwzględniać prefiksu $P-, który określa parametr w wyrażeniach CLEM. Nazwa ta jest także używana do wyświetlania w Konstruktorze wyrażeń CLEM.

Długa nazwa. Przedstawia nazwy opisowe każdego utworzonego parametru.

Składowanie. Umożliwia wybór typu składowania z listy. Składowanie wskazuje, w jaki sposób wartości danych są składowane w parametrze. Na przykład w przypadku pracy z wartościami zawierającymi wiodące zera, które użytkownik chce zachować (takie jak 008), należy jako typ składowania wybrać **Łańcuch**. W przeciwnym wypadku zera zostaną odrzucone. Dostępne typy składowania to: łańcuch, liczba całkowita, liczba rzeczywista, czas, data oraz znacznik czasu. W przypadku parametrów typu data należy zwrócić uwagę, że wartości należy określić, korzystając z zapisu standardowego ISO zgodnie z informacją w następnym akapicie.

Wartość. Zawiera listę bieżących wartości dla każdego parametru. Parametr należy dostosować odpowiednio do potrzeb. Należy zwrócić uwagę, że w przypadku parametrów typu data wartości muszą być zapisane w notacji określonej normą ISO, to jest: RRRR-MM-DD. Daty w innych formatach nie będą akceptowane.

Typ (opcja). Jeśli planuje się wdrożenie strumienia do aplikacji zewnętrznej, należy wybrać poziom pomiaru z listy. W przeciwnym wypadku zaleca się pozostawienie kolumny *Typ* bez zmian. W przypadku określenia ograniczeń wartości dla parametru, takich jak dolna i górna granica przedziału liczbowego, należy wybrać z listy opcję **Określ**.

Należy pamiętać, że opcje długiej nazwy, składowania i typu parametru można ustawić tylko za pośrednictwem interfejsu użytkownika. Opcji tych nie można ustawić za pomocą skryptów.

Kliknij strzałki po prawej stronie, aby przenieść wybrany parametr w górę lub w dół listy dostępnych parametrów. Użyj przycisku usuwania (oznaczonego symbolem *X*) w celu usunięcia wybranego parametru.

Ustawianie wartości dla parametrów Superwęzła

Po zdefiniowaniu parametrów dla Superwęzła można określić jego wartości, używając parametrów w wyrażeniu CLEM lub w skrypcie.

Aby określić parametry Superwęzła

1. Dwukrotnie kliknij ikonę Superwęzła, aby otworzyć jego okno dialogowe.
2. Alternatywnie, w menu Superwęzła wybierz opcję **Ustaw parametry**.
3. Kliknij zakładkę **Parametry**. *Uwaga*: Pola w tym oknie dialogowym są polami definiowanymi po kliknięciu przycisku **Definiuj parametry** na tej karcie.
4. Dla każdego utworzonego parametru wprowadź wartość w polu tekstowym. Przykładowo można ustawić wartość *minvalue* (wartość minimalna) na konkretną wartość graniczną. Ten parametr może być następnie użyty w wielu operacjach, takich jak wybór rekordów powyżej lub poniżej tej wartości granicznej, w celu przeprowadzenia dalszej eksploracji.

Użycie parametrów Superwęzła w celu uzyskania dostępu do właściwości węzła

Parametry Superwęzła mogą być również używane do zdefiniowania właściwości węzła (znane również jako **parametry zagnieżdżania**) dla opakowanych (zamkniętych) węzłów. Załóżmy na przykład, że użytkownik chce określić, aby Superwęzeł uczył opakowany węzeł sieci neuronowej przez określony czas z zastosowaniem próby losowej dla dostępnych danych. Używając parametrów można określić wartości dla czasu oraz wartości procentowej próby.

Załóżmy, że przykładowy Superwęzeł zawiera węzeł przykładowy o nazwie *Sample* (Losowanie) oraz węzeł sieci neuronowej o nazwie *Train* (Uczenie). Za pomocą okien dialogowych można określić ustawienie **Próba** dla węzła

próby jako **Losowo % ze wszystkich**, a ustawienie **Zatrzymywanie uczenia** dla węzła sieci neuronowej na **Czas**. Po określeniu tych opcji można uzyskać dostęp do właściwości węzła z parametrami i określić konkretne wartości dla Superwęzła. W oknie dialogowym Superwęzła należy kliknąć opcję **Definiuj parametry** i utworzyć parametry przedstawione w poniższej tabeli.

Tabela 54. Parametry do utworzenia

Parametr	Wartość	Długa nazwa
Train.time	5	Czas uczenia (w minutach)
Sample.random	10	Wartość procentowa próby losowej

Uwaga: W nazwach parametrów, takich jak *Sample.random*, używana jest poprawna składnia, umożliwiająca odwołanie do właściwości węzła, gdzie *Sample* (Losowanie) reprezentuje nazwę węzła, a *random* (losowa) jest właściwością węzła.

Po zdefiniowaniu tych parametrów można w prosty sposób zmodyfikować wartości właściwości węzłów próby i sieci neuronowej bez konieczności ponownego otwierania każdego okna dialogowego. Wystarczy wybrać opcję **Ustaw parametry** z menu Superwęzła, aby uzyskać dostęp do karty Parametry w oknie dialogowym Superwęzła, na której można określić nowe wartości dla właściwości **Losowo % ze wszystkich** i **Czas**. Jest to szczególnie przydatne podczas eksploracji danych przy wielu iteracjach budowania modelu.

Superwęzły i buforowanie

W przypadku Superwęzłów wszystkie węzły, z wyjątkiem węzłów końcowych, można zapisywać w pamięci podręcznej (buforować). Aby kontrolować buforowanie, należy kliknąć węzeł prawym przyciskiem myszy i wybrać jedną z kilku opcji z menu kontekstowego pamięci podręcznej. Ta opcja menu jest dostępna z zewnątrz Superwęzła oraz dla węzłów opakowanych (zamkniętych) w ramach Superwęzła.

Dostępnych jest kilka wytycznych dotyczących pamięci podręcznych Superwęzłów:

- Jeśli dla dowolnego węzła opakowanego w Superwęźle została włączona funkcja buforowania, będzie ona również włączona dla Superwęzła.
- Wyłączenie pamięci podręcznej w Superwęźle powoduje wyłączenie pamięci podręcznej dla *wszystkich* opakowanych (zamkniętych) węzłów.
- Aktywowanie buforowania w Superwęźle w rzeczywistości aktywuje pamięć podręczną w ostatnim możliwym do zbuforowania podwęźle. Innymi słowy, jeśli ostatnim podwęzłem jest węzeł wyboru, pamięć podręczna będzie aktywowana dla tego węzła wyboru. Jeśli ostatnim podwęzłem jest węzeł końcowy (który nie umożliwia buforowania), buforowanie zostanie aktywowane dla wcześniejszego węzła w strumieniu, który obsługuje buforowanie.
- Po ustawieniu pamięci podręcznych dla podwęzłów Superwęzła wszelkie działania powyżej zbuforowanego węzła, takie jak dodawanie lub edytowanie węzłów, będą opróżniały pamięci podręczne.

Superwęzły i tworzenie skryptów

Korzystając z języka skryptów SPSS Modeler można napisać proste programy, które umożliwiają manipulowanie zawartością i wykonywanie końcowego Superwęzła. Na przykład, użytkownik chce określić kolejność wykonywania złożonego strumienia. Jeśli Superwęzeł zawiera węzeł Globalne, który wymaga wykonania przed węzłem wykresu, można utworzyć skrypt, który spowoduje, że węzeł Globalne zostanie wykonany jako pierwszy. Wartości obliczone przez ten węzeł, np. średnia lub odchylenie standardowe, mogą być następnie użyte podczas wykonywania węzła wykresu.

Karta Skrypt w oknie dialogowym Superwęzła jest dostępna tylko dla końcowych Superwęzłów.

Aby otworzyć okno dialogowe tworzenia skryptów dla końcowego Superwęzła:

- Należy kliknąć obszar roboczy Superwęzła prawym przyciskiem myszy i wybrać opcję **Skrypt Superwęzła**.

- Alternatywnie, zarówno po wejściu do węzła, jak i dla trybu bez wchodzenia do węzła, można wybrać opcję **Skrypt Superwęzła** z menu Superwęzła.

Uwaga: Skrypty Superwęzła są wykonywane tylko ze strumieniem i Superwęzłem po wybraniu opcji **Wykonaj ten skrypt** w oknie dialogowym.

Konkretne opcje dotyczące skryptów i ich użycia w programie SPSS Modeler zostały omówione w *podręczniku tworzenia skryptów i automatyzacji*, który jest dostępny jako plik PDF w materiałach do pobrania dla produktu.

Zapisywanie i ładowanie Superwęzłów

Jedną z zalet Superwęzłów jest możliwość zapisywania ich i ponownego użycia w innych strumieniach. Podczas zapisywania i ładowania Superwęzłów należy pamiętać, że używają one rozszerzenia .slb.

Aby zapisać Superwęzeł

1. Wejdź do Superwęzła.
2. Z menu Superwęzła wybierz opcję **Zapisz Superwęzeł**.
3. W oknie dialogowym określ nazwę pliku i katalog.
4. Wybierz, czy zapisany Superwęzeł ma zostać dodany do bieżącego projektu.
5. Kliknij przycisk **Zapisz**.

Aby załadować Superwęzeł

1. Z menu Wstaw w oknie IBM SPSS Modeler wybierz opcje **Superwęzeł**.
2. Wybierz plik Superwęzła (.slb) z bieżącego katalogu lub przejdź do innego.
3. Kliknij przycisk **Wczytaj**.

Uwaga: Zaimportowane Superwęzły zawierają domyślne wartości dla wszystkich parametrów. Aby zmienić te parametry, kliknij dwukrotnie Superwęzeł w obszarze roboczym strumienia.

Uwagi

Niniejsza publikacja została przygotowana z myślą o produktach i usługach oferowanych w Stanach Zjednoczonych. Materiał ten jest również dostępny w IBM w innych językach. Jednakże w celu uzyskania dostępu do takiego materiału istnieje konieczność posiadania egzemplarza produktu w takim języku.

Produktów, usług lub opcji opisywanych w tym dokumencie IBM nie musi oferować we wszystkich krajach. Informacje o produktach i usługach dostępnych w danym kraju można uzyskać od lokalnego przedstawiciela IBM. Odwołanie do produktu, programu lub usługi IBM nie oznacza, że można użyć wyłącznie tego produktu, programu lub usługi IBM. Zamiast nich można zastosować ich odpowiednik funkcjonalny pod warunkiem, że nie narusza to praw własności intelektualnej IBM. Jednakże cała odpowiedzialność za ocenę przydatności i sprawdzenie działania produktu, programu lub usługi pochodzących od producenta innego niż IBM spoczywa na użytkowniku.

IBM może posiadać patenty lub złożone wnioski patentowe na towary i usługi, o których mowa w niniejszej publikacji. Przedstawienie niniejszej publikacji nie daje żadnych uprawnień licencyjnych do tychże patentów. Pisemne zapytania w sprawie licencji można przesyłać na adres:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Zapytania dotyczące zestawów znaków dwubajtowych (DBCS) należy kierować do lokalnych działów własności intelektualnej IBM (IBM Intellectual Property Department) lub wysłać je na piśmie na adres:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan, Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokio 103-8510, Japonia

INTERNATIONAL BUSINESS MACHINES CORPORATION DOSTARCZA TĘ PUBLIKACJĘ W STANIE, W JAKIM SIĘ ZNAJDUJE ("AS IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH LUB DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU ORAZ GWARANCJI, ŻE PUBLIKACJA TA NIE NARUSZA PRAW OSÓB TRZECICH. Ustawodawstwa niektórych krajów nie dopuszczają zastrzeżeń dotyczących gwarancji wyraźnych lub domniemanych w odniesieniu do pewnych transakcji; w takiej sytuacji powyższe zdanie nie ma zastosowania.

Informacje zawarte w niniejszej publikacji mogą zawierać nieścisłości techniczne lub błędy drukarskie. Informacje te są okresowo aktualizowane, a zmiany te zostaną uwzględnione w kolejnych wydaniach tej publikacji. IBM zastrzega sobie prawo do wprowadzania ulepszeń i/lub zmian w produktach i/lub programach opisanych w tej publikacji w dowolnym czasie, bez wcześniejszego powiadomienia.

Wszelkie wzmianki w tej publikacji na temat stron internetowych innych podmiotów zostały wprowadzone wyłącznie dla wygody użytkownika i w żadnym wypadku nie stanowią zachęty do ich odwiedzania. Materiały dostępne na tych stronach nie są częścią materiałów opracowanych dla tego produktu IBM, a użytkownik korzysta z nich na własną odpowiedzialność.

IBM ma prawo do używania i rozpowszechniania informacji przysłanych przez użytkownika w dowolny sposób, jaki uzna za właściwy, bez żadnych zobowiązań wobec ich autora.

Licencjobiorcy tego programu, którzy chcieliby uzyskać informacje na temat programu w celu: (i) wdrożenia wymiany informacji między niezależnie utworzonymi programami i innymi programami (łącznie z tym opisywanym) oraz (ii) wspólnego wykorzystywania wymienianych informacji, powinni skontaktować się z:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
U.S.A.

Informacje takie mogą być udostępnione, o ile spełnione zostaną odpowiednie warunki, w tym, w niektórych przypadkach, zostanie uiszczona stosowna opłata.

Licencjonowany program opisany w niniejszej publikacji oraz wszystkie inne licencjonowane materiały dostępne dla tego programu są dostarczane przez IBM na warunkach określonych w Umowie IBM z Klientem, Międzynarodowej Umowie Licencyjnej IBM na Program lub w innych podobnych umowach zawartych między IBM i użytkownikami.

Dane dotyczące wydajności i cytowane przykłady zostały przedstawione jedynie w celu zobrazowania sytuacji. Faktyczne wyniki dotyczące wydajności mogą się różnić w zależności do konkretnych warunków konfiguracyjnych i operacyjnych.

Informacje dotyczące produktów innych podmiotów niż IBM zostały uzyskane od dostawców tych produktów, z ich publicznych ogłoszeń lub innych dostępnych publicznie źródeł. IBM nie testował tych produktów i nie może potwierdzić dokładności pomiarów wydajności, kompatybilności ani żadnych innych danych związanych z tymi produktami. Pytania dotyczące możliwości produktów innych podmiotów należy kierować do dostawców tych produktów.

Wszelkie stwierdzenia dotyczące przyszłych kierunków rozwoju i zamierzeń IBM mogą zostać zmienione lub wycofane bez powiadomienia.

Publikacja ta zawiera przykładowe dane i raporty używane w codziennej pracy. W celu kompleksowego ich zilustrowania, podane przykłady zawierają nazwiska osób prywatnych, nazwy przedsiębiorstw oraz nazwy produktów. Wszystkie te nazwy/nazwiska są fikcyjne i jakiekolwiek podobieństwo do istniejących nazw/nazwisk jest całkowicie przypadkowe.

Znaki towarowe

IBM, logo IBM i ibm.com są znakami towarowymi lub zastrzeżonymi znakami towarowymi International Business Machines Corp. zarejestrowanymi w wielu systemach prawnych na całym świecie. Pozostałe nazwy produktów i usług mogą być znakami towarowymi IBM lub innych przedsiębiorstw. Aktualna lista znaków towarowych IBM dostępna jest w serwisie WWW IBM, w sekcji "Copyright and trademark information" Informacje o prawach autorskich i znakach towarowych), pod adresem www.ibm.com/legal/copytrade.shtml.

Adobe, logo Adobe, PostScript oraz logo PostScript są znakami towarowymi lub zastrzeżonymi znakami towarowymi Adobe Systems Incorporated w Stanach Zjednoczonych i/lub w innych krajach.

Intel, logo Intel, Intel Inside, logo Intel Inside, Intel Centrino, logo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium i Pentium są znakami towarowymi lub zastrzeżonymi znakami towarowymi Intel Corporation lub przedsiębiorstw podporządkowanych Intel Corporation w Stanach Zjednoczonych i w innych krajach.

Linux jest zastrzeżonym znakiem towarowym Linusa Torvaldsa w Stanach Zjednoczonych i/lub w innych krajach.

Microsoft, Windows, Windows NT oraz logo Windows są znakami towarowymi Microsoft Corporation w Stanach Zjednoczonych i/lub w innych krajach.

UNIX jest zastrzeżonym znakiem towarowym The Open Group w Stanach Zjednoczonych i/lub w innych krajach.

Java oraz wszystkie znaki towarowe i logo dotyczące języka Java są znakami towarowymi lub zastrzeżonymi znakami towarowymi Oracle i/lub przedsiębiorstw afiliowanych.

Warunki dotyczące dokumentacji produktu

Zezwolenie na korzystanie z tych publikacji jest przyznawane na poniższych warunkach.

Zakres stosowania

Niniejsze warunki stanowią uzupełnienie warunków używania serwisu WWW IBM.

Użytek osobisty

Użytkownik ma prawo kopiować te publikacje do własnego, niekomercyjnego użytku pod warunkiem zachowania wszelkich uwag dotyczących praw własności. Użytkownik nie ma prawa dystrybuować ani wyświetlać tych publikacji czy ich części, ani też wykonywać na ich podstawie prac pochodnych bez wyraźnej zgody IBM.

Użytek służbowy

Użytkownik ma prawo kopiować te publikacje, dystrybuować je i wyświetlać wyłącznie w ramach przedsiębiorstwa Użytkownika pod warunkiem zachowania wszelkich uwag dotyczących praw własności. Użytkownik nie ma prawa wykonywać na podstawie tych publikacji ani ich fragmentów prac pochodnych, kopiować ich, dystrybuować ani wyświetlać poza przedsiębiorstwem Użytkownika bez wyraźnej zgody IBM.

Prawa

Z wyjątkiem zezwoleń wyraźnie udzielonych w niniejszym dokumencie, nie udziela się jakichkolwiek innych zezwoleń, licencji ani praw, wyraźnych czy domniemanych, odnoszących się do tych publikacji czy jakichkolwiek informacji, danych, oprogramowania lub innej własności intelektualnej, o których mowa w niniejszym dokumencie.

IBM zastrzega sobie prawo do anulowania zezwolenia przyznanego w niniejszym dokumencie w każdej sytuacji, gdy, według uznania IBM, korzystanie z tych publikacji jest szkodliwe dla IBM lub jeśli IBM uzna, że warunki niniejszego dokumentu nie są przestrzegane.

Użytkownik ma prawo pobierać, eksportować lub reeksportować niniejsze informacje pod warunkiem zachowania bezwzględnej i pełnej zgodności z obowiązującym prawem i przepisami, w tym ze wszelkimi prawami i przepisami eksportowymi Stanów Zjednoczonych.

IBM NIE UDZIELA JAKICHKOLWIEK GWARANCJI, W TYM TAKŻE RĘKOJMI, DOTYCZĄCYCH TREŚCI TYCH PUBLIKACJI. PUBLIKACJE TE SĄ DOSTARCZANE W STANIE, W JAKIM SIĘ ZNAJDUJĄ ("AS-IS") BEZ UDZIELANIA JAKICHKOLWIEK GWARANCJI (RĘKOJMIĘ RÓWNIEŻ WYŁĄCZA SIĘ), WYRAŹNYCH CZY DOMNIEMANYCH, A W SZCZEGÓLNOŚCI DOMNIEMANYCH GWARANCJI PRZYDATNOŚCI HANDLOWEJ, PRZYDATNOŚCI DO OKREŚLONEGO CELU CZY NIENARUSZANIA PRAW OSÓB TRZECICH.

Glosariusz

B

*Błąd standardowy***.** Miara tego, jak bardzo wartość statystyki testowej (sprawdzianu testu) zmienia się pomiędzy próbami. Jest to odchylenie standardowe rozkładu wartości danej statystyki dla poszczególnych prób. Na przykład błąd standardowy średniej to odchylenie standardowe średnich z prób.

*Błąd standardowy kurtozy***.** Iloraz kurtozy i jej błędu standardowego; może być wykorzystywany jako test normalności (tzn. można odrzucić założenie normalności, jeśli wartość ilorazu jest mniejsza od -2 lub większa od +2). Wysoka dodatnia wartość dla kurtozy wskazuje na to, iż krańce rozkładu są dłuższe niż te dla rozkładu normalnego; ujemna wartość dla kurtozy wskazuje na krótsze krańce (podobnie jak w rozkładach prostokątnych).

*Błąd standardowy skośności***.** Iloraz skośności i jej błędu standardowego; może być wykorzystywany jako test normalności (tzn. można odrzucić założenie o normalności, jeśli wartość ilorazu jest mniejsza od -2 lub większa od +2). Wysoka dodatnia wartość dla skośności wskazuje na długi prawy kraniec; skrajnie ujemna wartość wskazuje na długi lewy kraniec.

*Błąd standardowy średniej***.** Miara tego, jak bardzo wartość średniej może się zmieniać dla różnych prób losowanych z tego samego rozkładu. Może być wykorzystywana do pobieżnego porównania rzeczywistej wartości średniej z wartością hipotetyczną (tj. można sądzić, że te dwie wartości są różne, jeśli iloraz różnicy i błędu standardowego jest mniejszy od -2 lub większy od +2).

D

*Dominanta***.** Wartość występująca najczęściej. Jeśli więcej niż jedna wartość występuje z taką samą, największą częstością, każda z nich jest dominantą (wartością modalną).

K

*Kowariancja***.** Nieustandaryzowana miara powiązania dwóch zmiennych, równa sumie iloczynów wektorowych odchyleń wartości tych zmiennych od ich średnich podzielonej przez N-1.

*Kurtoza***.** Miara ilości skrajnych wartości odstających. W przypadku rozkładu normalnego wartość statystyki kurtozy wynosi zero. Kurtoza dodatnia oznacza, że w danych jest więcej skrajnych wartości odstających niż w rozkładzie normalnym. Kurtoza ujemna oznacza, że w danych jest mniej skrajnych wartości odstających niż w rozkładzie normalnym.

M

*Maksimum***.** Największa wartość zmiennej numerycznej.

*Mediana***.** Jest to 50. percentyl, czyli taka wartość, że połowa obserwacji ma wartości mniejsze, a druga połowa ma wartości większe od niej. W sytuacji parzystej liczby obserwacji mediana jest średnią dwóch środkowych obserwacji w próbie posortowanej rosnąco lub malejąco. W przeciwieństwie do średniej, na którą wpływ może mieć nawet kilka ekstremalnie dużych lub małych wartości, mediana jest miarą tendencji centralnej niewrażliwą na wartości odstające.

*Minimum***.** Najmniejsza wartość zmiennej numerycznej.

O

*Odchylenie standardowe***.** Miara rozproszenia wokół wartości średniej, równa pierwiastkowi z wariancji. Odchylenie standardowe mierzy się w tych samych jednostkach co pierwotną wartość.

*Odchylenie standardowe***.** Miara rozproszenia wokół średniej. W przypadku rozkładu normalnego, 68% obserwacji znajduje się w obszarze oddalonym o jedno odchylenie standardowe od średniej, zaś 95% — w przedziale oddalonym o dwa odchylenia standardowe. Na przykład, jeśli średnia wieku osób wynosi 45 lat, a odchylenie standardowe wynosi 10, wówczas 95% rozważanych osób znajduje się w przedziale wiekowym między 25 a 65 lat.

P

Przedział. Różnica między największą a najmniejszą wartością zmiennej numerycznej; maksimum minus minimum.

S

Skośność. Miara asymetrii rozkładu. Rozkład normalny jest symetryczny, a jego wartość skośności wynosi 0. Rozkład o dużej skośności dodatniej ma długi kraniec z prawej strony. Gdy zaś współczynnik skośności jest ujemny, rozkład ma długi kraniec z lewej strony. Jako wytyczna, wartość skośności przekraczająca dwukrotnie swój błąd standardowy na ogół oznacza odstępstwo od symetrii rozkładu.

Suma. Suma wartości wszystkich obserwacji nieposiadających braków danych.

Swoista. Oszacowuje wielkość wszystkich efektów równocześnie, korygując każdy z nich ze względu na wszystkie inne efekty każdego typu.

Ś

Średnia. Miara tendencji centralnej. Średnia arytmetyczna; suma podzielona przez liczbę obserwacji.

W

Wariancja. Miara rozproszenia wokół średniej, równa sumie podniesionych do kwadratu odchyleń od średniej, podzielonej przez liczbę obserwacji minus jeden. Wariancja jest mierzona w jednostkach będących kwadratami jednostek miary dla zmiennej, do której wariancja się odnosi.

Ważne. Poprawne obserwacje nieposiadające systemowych ani zdefiniowanych przez użytkownika braków danych.

Indeks

Znaki specjalne

A

B

C

D

E

F

G

H

I

J

K

L

Ł

M

N

O

P

R

S

Ś

Z

Ź

Drukowane w USA