Robert Kozakiewicz Gdańsk, 16.05.2015

Dawid Piasecki

Eksploracja Danych

Zrozumienie i przygotowanie danych

1. Ocena jakości danych.

1.1. Brakujące dane

Do zidentyfikowania brakujących danych użyliśmy kontrolki 'Interactive table' wraz z sortowaniem po poszczególnych kolumnach. W pierwszym źródle danych dotyczącym wystawionych mandatów odkryliśmy brakujące wartości w większości kolumn (zazwyczaj występowały one jako znak zapytania), natomiast w drugim źródle danych dotyczącym wypadków drogowych nie zaobserwowaliśmy brakujących wartości.

1.2. Obserwacje odległe

Do wyszukania wartości oddalonych użyliśmy wykresu pudełkowego. W pierwszym źródle danych nie znaleźliśmy punktów ekstremalnie oddalonych, natomiast w drugim źródle danych znaleźliśmy kilka oddalonych wartości w kolumnach oznaczających liczbę poszkodowanych oraz liczbę pojazdów uczestniczących w wypadku.

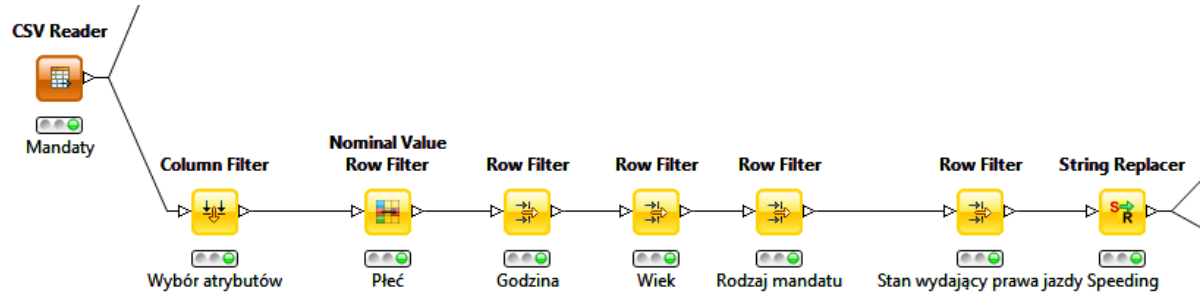
2. Wstępna analiza danych.

Do wyszukiwania zależności między danymi użyliśmy kontrolki 'Linear correlation'. W obu źródłach danych nie znaleźliśmy znaczącej korelacji pomiędzy atrybutami.

3. Wstępne przygotowanie danych.

Pierwsze źródło danych (mandaty):

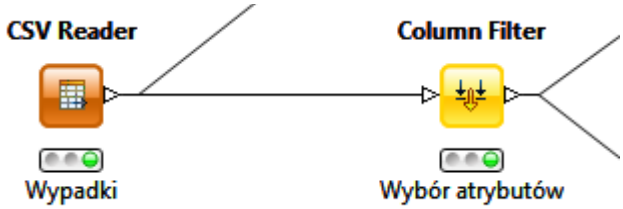
Najpierw dokonaliśmy usunięcia atrybutów, które nie były dla nas użyteczne: kod złamanego przepisu ruchu drogowego, rok wystawienia mandatu, sąd wystawiający mandat, źródło danych o mandacie. Następnie obsłużyliśmy wartości brakujące w kolumnach, w których zaobserwowaliśmy brakujące dane. Jako że wiersze z brakującymi wartościami stanowiły niewielki ułamek całkowitej liczby wierszy, zdecydowaliśmy się na ich usunięcie. Następnie w celu ujednolicenia wartości atrybutu oznaczającego rodzaj mandatu, scaliliśmy wszystkie atrybuty związane z przekraczaniem dozwolonej prędkości do jednego ('SPEEDING').



Obsługa pierwszego źródła danych w programie Knime

Drugie źródło danych (wypadki):

W drugim źródle danych dokonaliśmy jedynie selekcji atrybutów - usunęliśmy kolumny z numerem identyfikacyjnym wypadku oraz numerami siatki geograficznej miejsca wypadku.



Obsługa drugiego źródła danych w programie Knime