Sprawozdanie z listy 1

Eksploracja danych

Marta Stankiewicz, Paweł Nowak
numery albumów: 282244 282223

2025-03-22

Spis treści

1 Etap 1. Przygotowanie danych. Podstawowe informacje o danych.

1.1 Opis danych, rozmiar ramki danych, typy danych.

Zbiór danych, którym sie zajmujemy, zawiera informacje o **7043** klientach sieci sklepów **Telco**, która oferuje różne usługi z branży telekomunikacji, rozrywki, Internetu itp.
Każdy klient został opisany przy użyciu **21** zmiennych, wsród których znajdziemy te opisujące dane osobiste klienta (np. zmienna *Partner*, wskazująca, czy dana osoba ma partnera), jak i te określające, czy dany klient skorzystał z usług oferowanych przez firmę. Najwięcej cech pochodzi właśnie z tej drugiej grupy zmiennych.
Większość zmiennych są zmiennymi ilośćiowymi nieporządkowymi, określającymi między innymi, czy dany klient wykupił daną telekomunikacyjną. Przykładowo — zmienna *Online-Security* informuje, czy osoba korzysta z usługi bezpieczeństwa w sieci (*Yes*), nie korzysta (*No*) czy też w ogóle nie ma dostępu do Internetu (*No internet service*).

1.2 Brakujące wartości.

Ze wszystkich zmiennych dostępnych w ramce danych, jedynie zmienna *TotalCharges* zawiera brakujące wartości. Zawiera ich *11*.Dokonamy imputacji wartości tej zmiennej, opierając się na podejściu ze średnią. Wartości brakujące są kodowane standardowo, tj. jako *NA*. Nie znajdujemy w zbiorze danych niestandardowej reprezentacji wartośći brakujących.

1.3 Okreslenie istotności zmiennych, eliminacja rendundancji danych.

Naszym celem jest przewidzenie, czy dany klient zrezygnuje z usług firmy na podstawie dostępnych cech. W celu wyeliminowania redundancji danych, skasujemy te zmienne, które albo nie mają żadnego wpływu na decyzje klienta albo są funkcją pozostałych atrybutów. Atrybut **customerID** z pewnością nie ma wpływu na zachowanie konsumenckie klienta, bowiem jest jedynie jego unikalnym identyfikatorem.

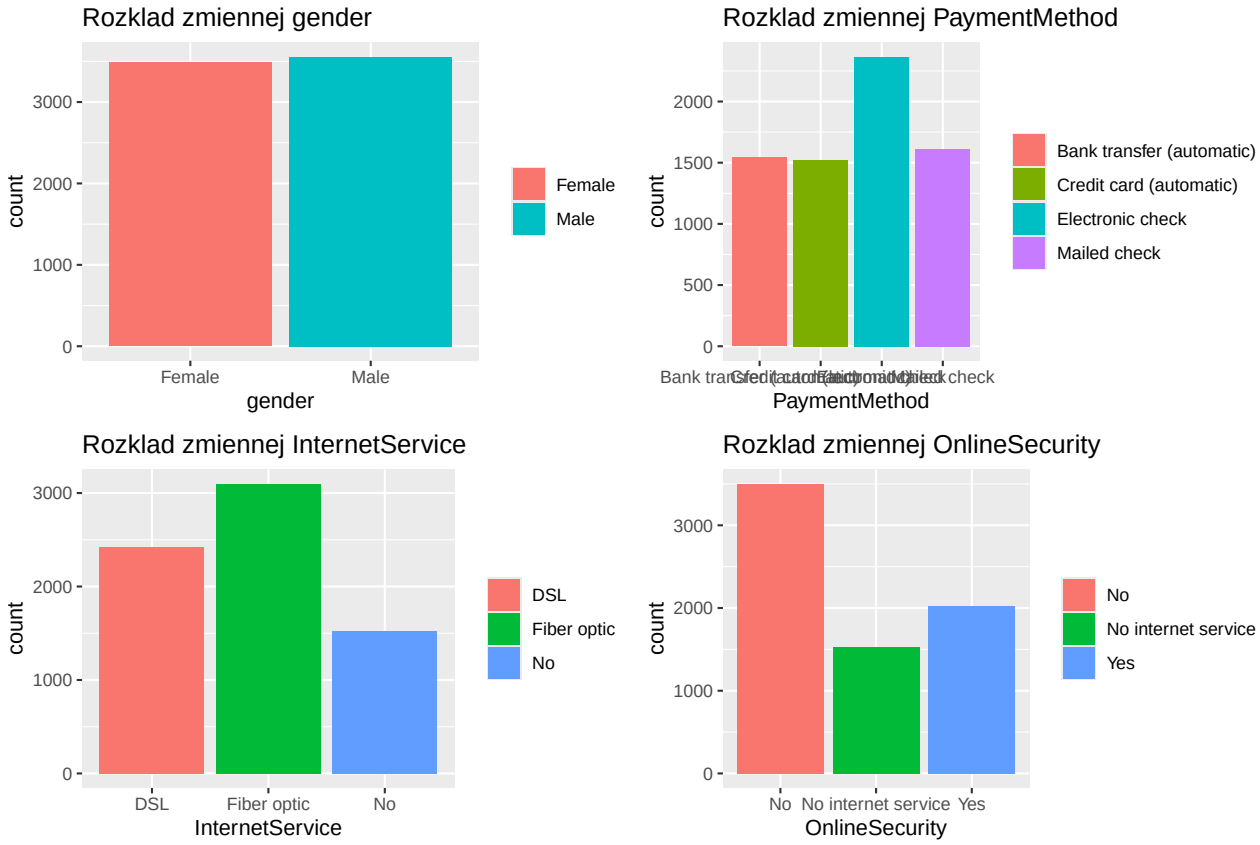
2 Etap 2. Analiza opisowa - wskaźniki sumaryczne i wykresy

2.1 Podstawowe wskaźniki sumaryczne dla zmiennych ciągłych

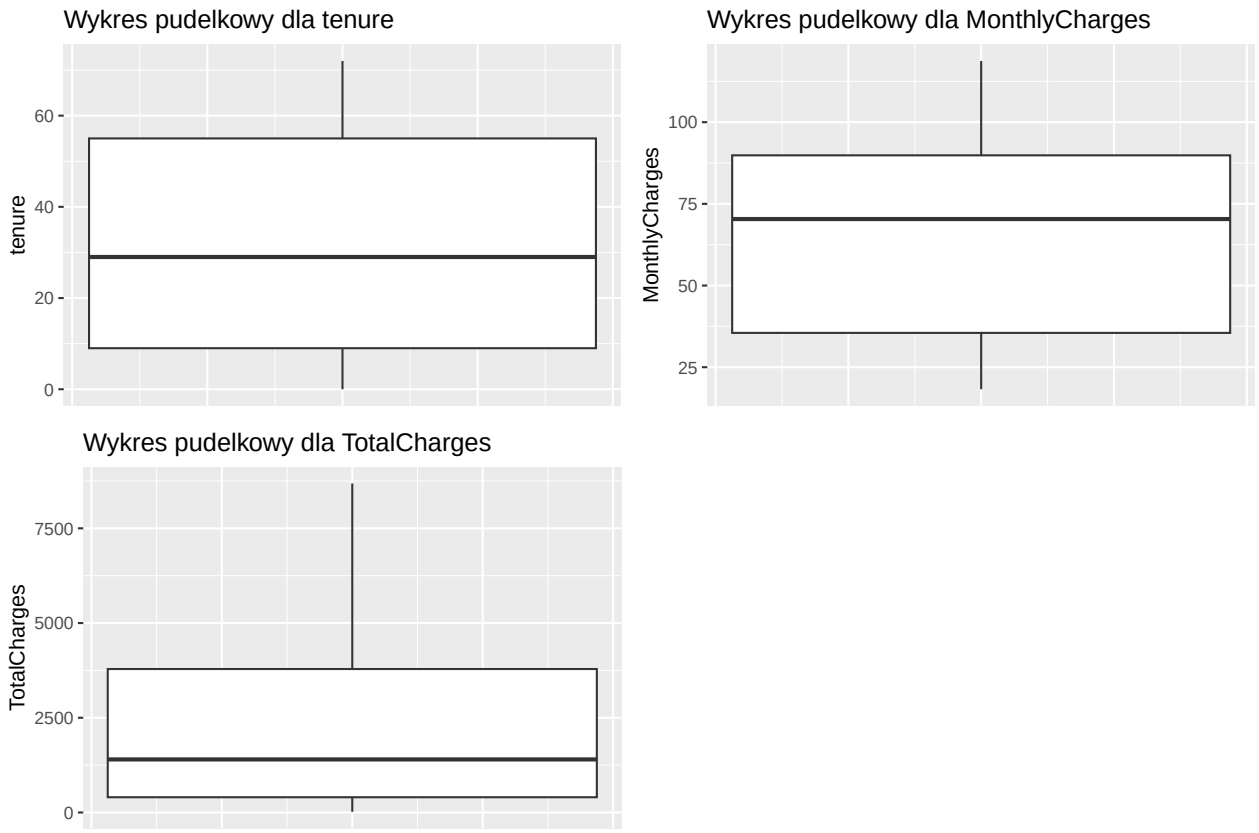
Tabela 1: Wskaźniki sumaryczne dla zmiennych ciągłych

	tenure	MonthlyCharges	TotalCharges
Min	0.00	18.25	18.80
Mean	32.37	64.76	2283.30
Median	29.00	70.35	1400.55
SD	24.56	30.09	2265.00
IQR	46.00	54.35	3384.38
Max	72.00	118.75	8684.80

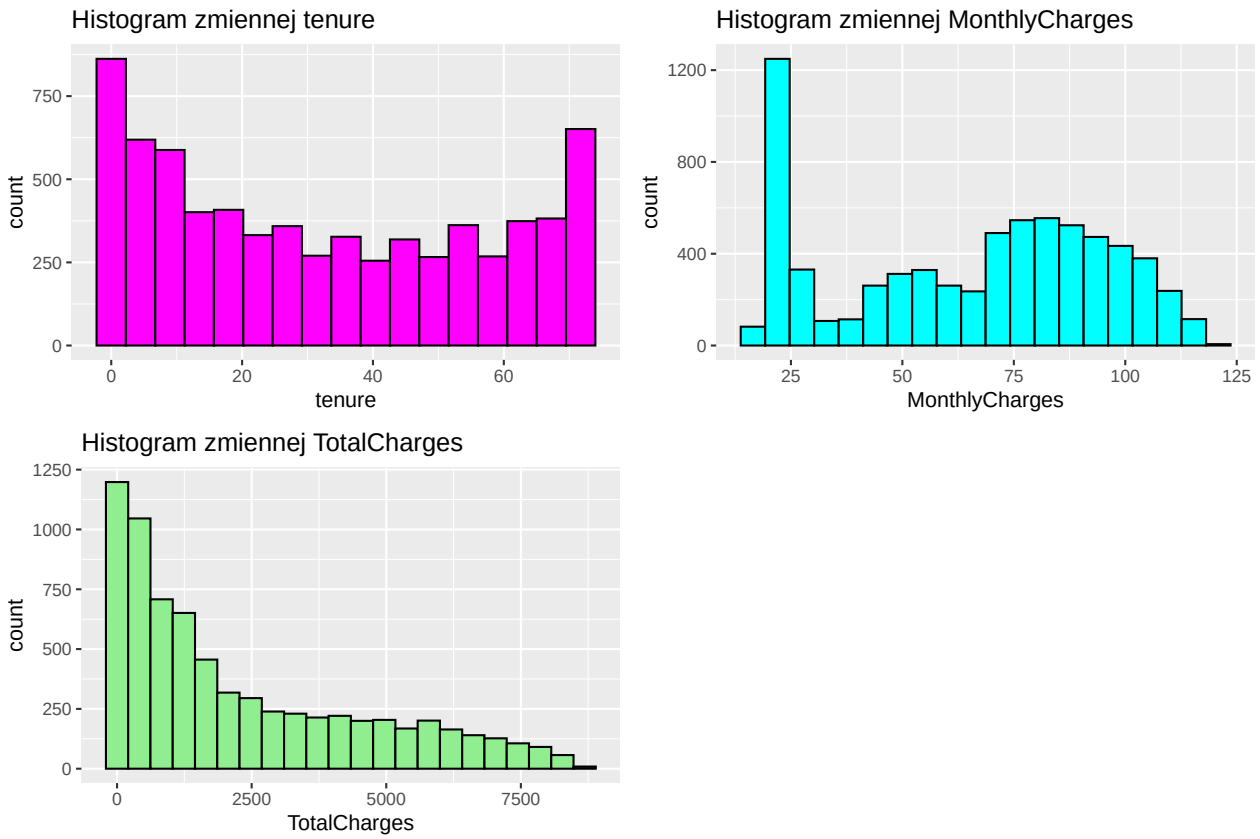
2.2 Wykresy słupkowe dla zmiennych kategorycznych



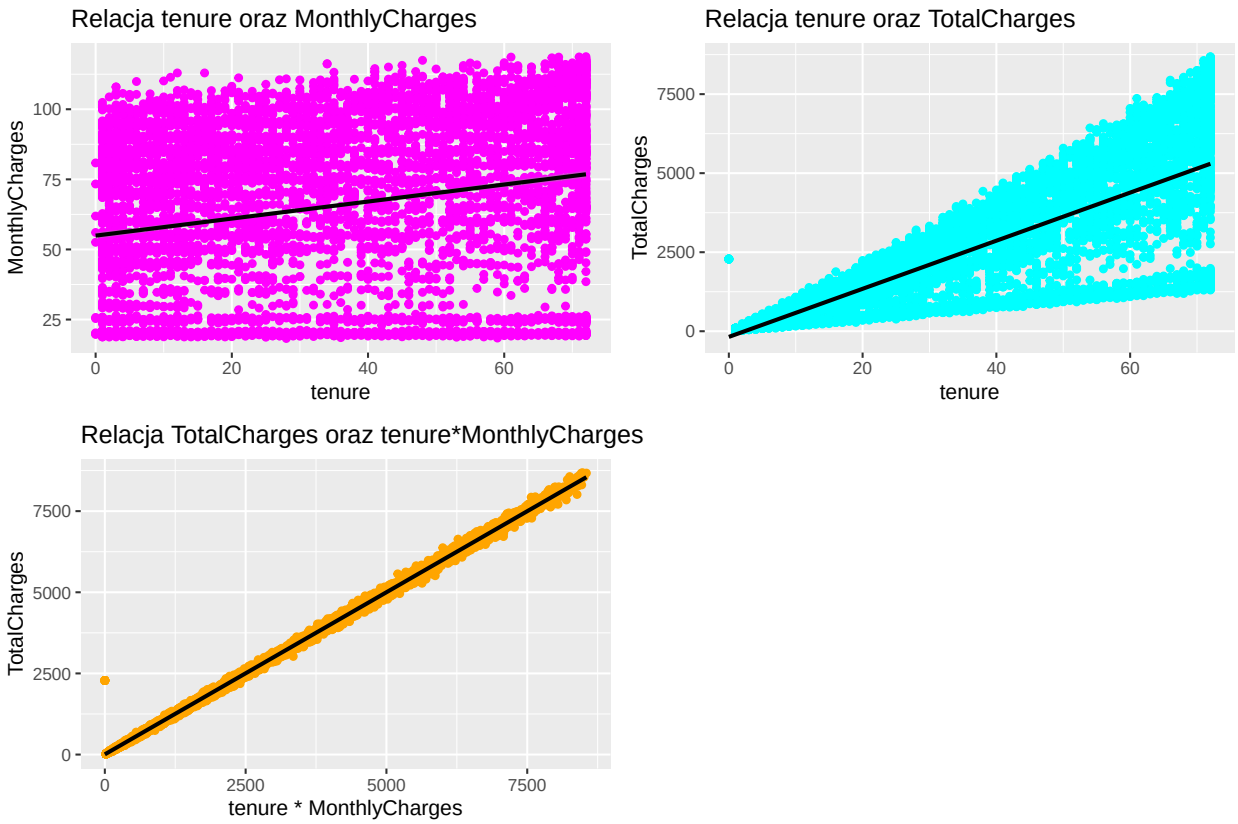
Rysunek 1: Rozkłady zmiennych kategorycznych



Rysunek 2: Wykresy pudełkowe zmiennych ciągłych



Rysunek 3: Histogramy zmiennych ciągłych



Rysunek 4: Wykresy rozrzutu wraz z krzywą regresji liniowej