

Sprawozdanie z listy 1

Eksploracja danych

Marta Stankiewicz, Paweł Nowak

numery albumów: 282244 282223

2025-03-18

Spis treści

1	Etap 1. Przygotowanie danych. Podstawowe informacje o danych.	1
1.1	Opis danych, rozmiar ramki danych, typy danych.	1
1.2	Brakujące wartości.	2
1.3	Określenie istotności zmiennych, eliminacja redundancji danych.	2
2	Etap 2. Analiza opisowa - wskaźniki sumaryczne i wykresy	2
2.1	Podstawowe wskaźniki sumaryczne dla zmiennych ciągłych	2
2.2	Wykresy słupkowe dla zmiennych kategoriowych	3

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

1.1 Opis danych, rozmiar ramki danych, typy danych.

Zbiór danych, którym się 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, wśród 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 ilościowymi 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. Wiersze, które nie przyjmują dla tej zmiennej żadnej wartości, zostaną porzucone. Wartości brakujące są kodowane standardowo, tj. jako *NA*. Nie znajdujemy w zbiorze danych niestandardowej reprezentacji wartości brakujących.

1.3 Okreslenie istotności zmiennych, eliminacja redundancji 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.

Racjonalne wydaje się również skasowanie kolumny **TotalCharges**. Zlicza ona całkowity koszt poniesiony przez klienta od początku zawarcia umowy z firmą. Tę wartość możemy bez straty informacji otworzyć, mnożąc ze sobą zmienne **tenure** oraz **MonthlyCharges** określające odpowiednio długość trwania umowy (wyrażona w miesiącach) i miesięczną taryfę za usługi.

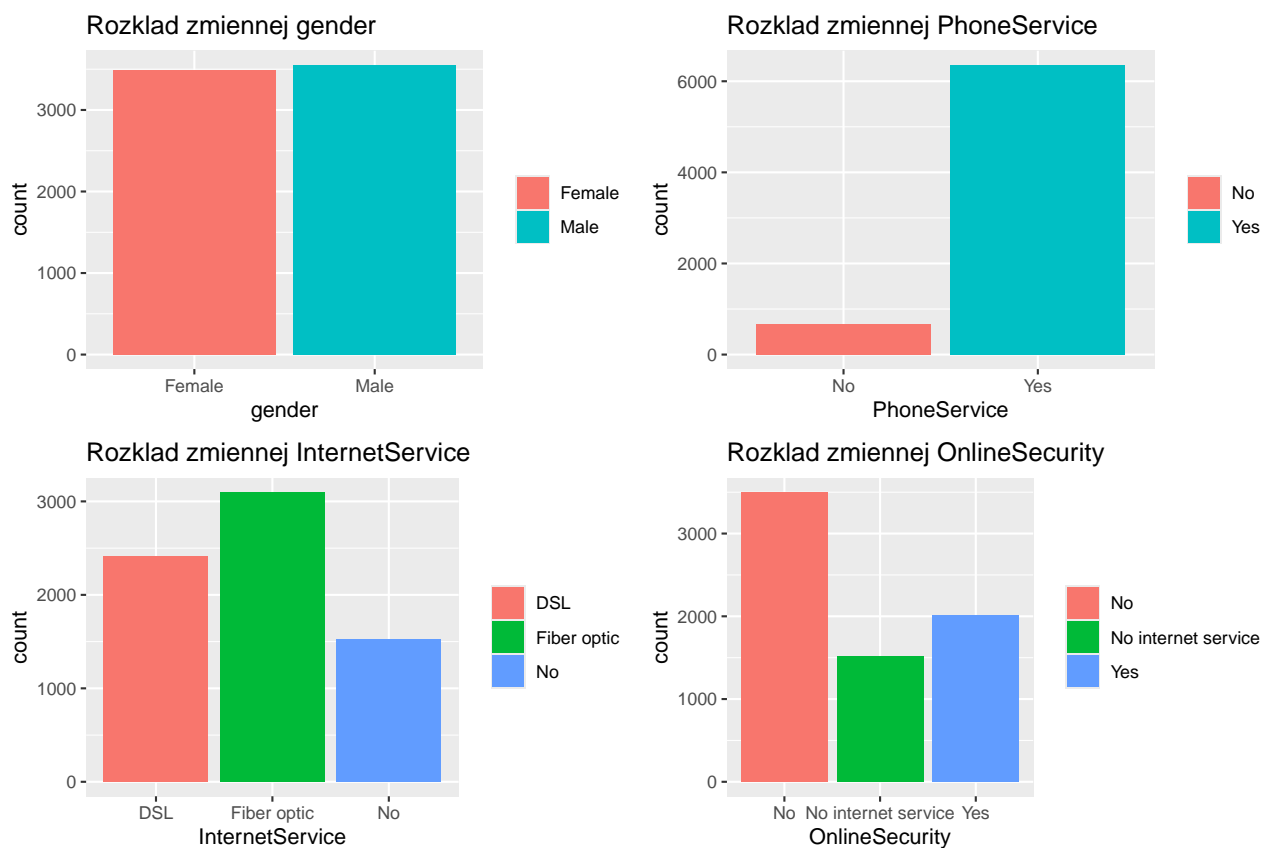
2 Etap 2. Analiza opisowa - wskaźniki sumaryczne i wykresy

2.1 Podstawowe wskaźniki sumaryczne dla zmiennych ciągłych

Tabela 1: Wskaźniki sumaryczne dla zmiennych ciągłych

	tenure	MonthlyCharges
Min	1.00	18.25
Mean	32.42	64.80
Median	29.00	70.35
SD	24.55	30.09
IQR	46.00	54.27
Max	72.00	118.75

2.2 Wykresy słupkowe dla zmiennych kategorycznych



Rysunek 1: Rozkłady zmiennych kategorycznych