

Sprawozdanie z listy 1

Eksploracja danych

Marta Stankiewicz, Paweł Nowak

numery albumów: 282244 282223

2025-03-25

Spis treści

1	Etap 1. Przygotowanie danych. Podstawowe informacje o danych.	2
1.1	Opis danych, rozmiar ramki danych, typy danych.	2
1.2	Brakujące wartości.	2
1.3	Określenie istotności zmiennych, eliminacja redundancji danych.	2
2	Etap 2. Analiza opisowa - wskaźniki sumaryczne i wykresy	3
2.1	Podstawowe wskaźniki sumaryczne dla zmiennych ciągłych	3
2.2	Wykresy słupkowe dla zmiennych kategoriowych	4
2.3	Wykresy pudełkowe dla zmiennych ilościowych	5
2.4	Histogramy dla zmiennych ilościowych	6
2.5	Wykresy rozrzutu wraz z krzywą regresji dla zmiennych ilościowych	7
2.6	Interpretacja wykresów.	7

Spis rysunków

1	Rozkłady zmiennych kategoriowych	4
2	Wykresy pudełkowe zmiennych ciągłych	5
3	Histogramy zmiennych ciągłych	6
4	Wykresy rozrzutu wraz z krzywą regresji liniowej	7

Spis tabel

1	Wskaźniki sumaryczne dla zmiennych ciągłych	3
---	---	---

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

1.1 Opis danych, rozmiar ramki danych, typy danych.

Zbiór danych, którym si 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**. 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 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 konsumentów klienta, bowiem jest jedynie jego unikalnym identyfikatorem.

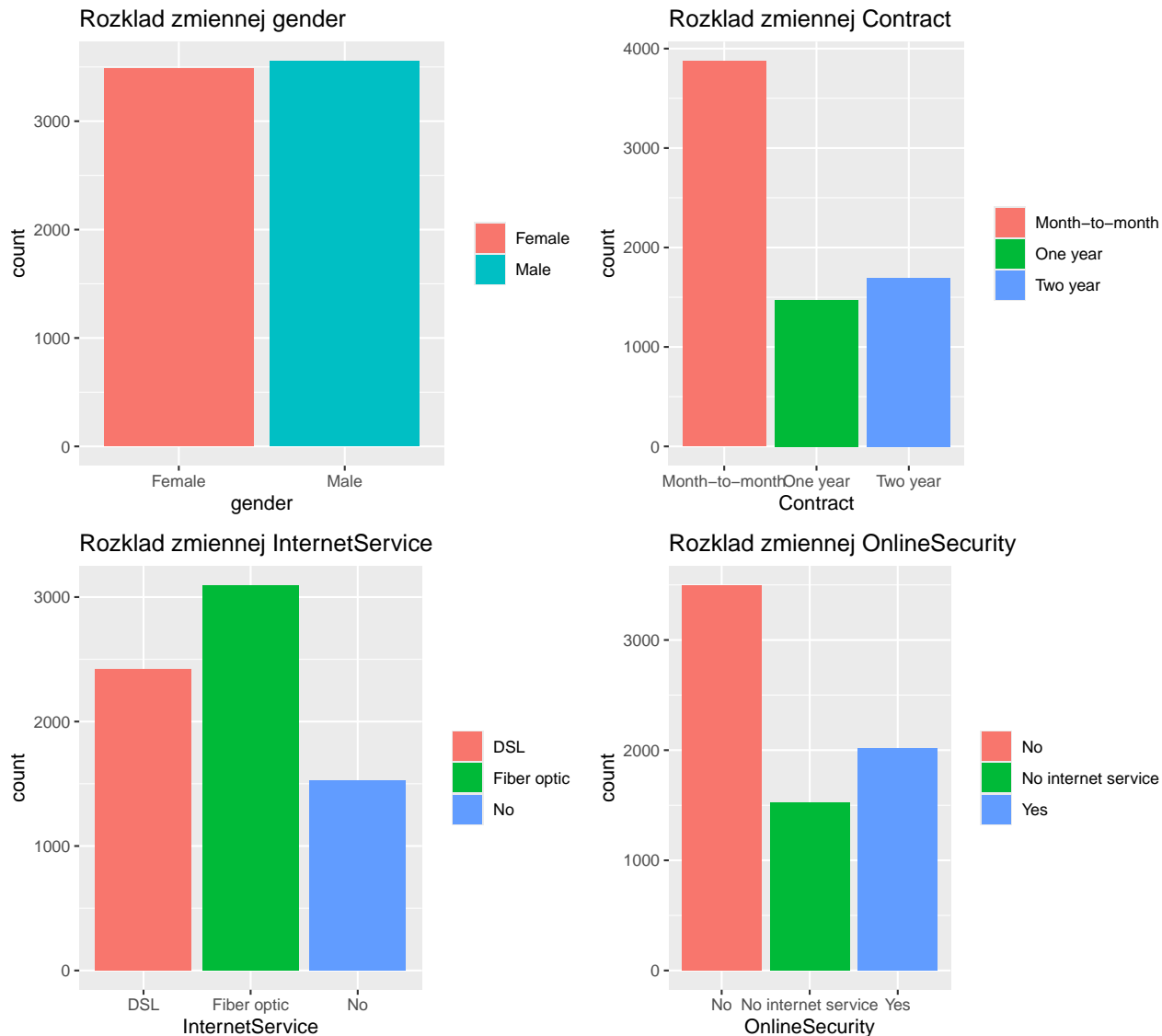
2 Etap 2. Analiza opisowa - wskaźniki sumaryczne i wykresy

2.1 Podstawowe wskaźniki sumaryczne dla zmiennych ciągłych

Tabela 1: Wskaźniki sumaryczne dla zmiennych ciągłych

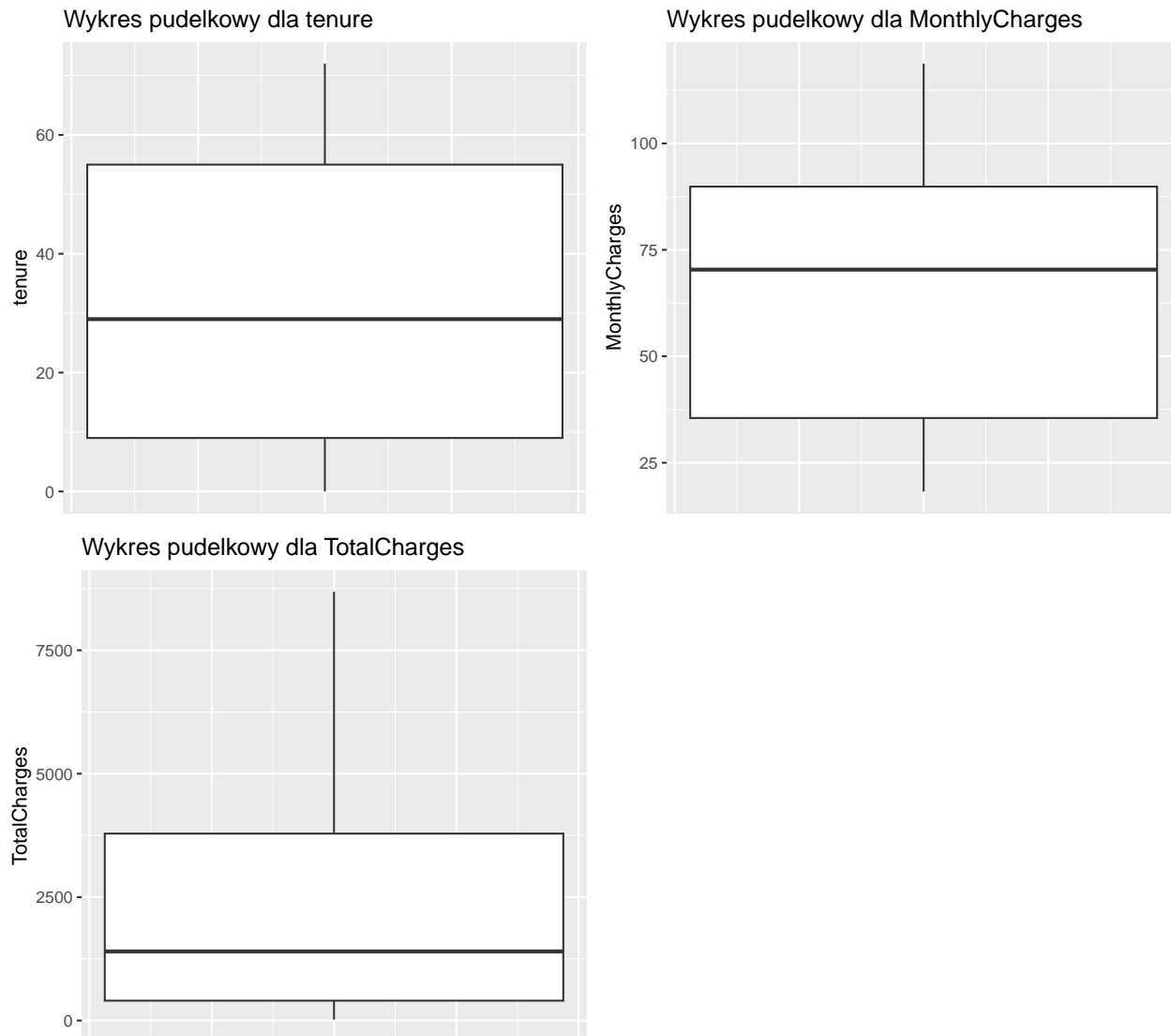
	tenure	MonthlyCharges	TotalCharges
Min	0.00	18.25	18.80
Mean	32.37	64.76	2283.30
Median	29.00	70.35	1400.55
SD	24.56	30.09	2265.00
IQR	46.00	54.35	3384.38
Max	72.00	118.75	8684.80

2.2 Wykresy słupkowe dla zmiennych kategorycznych



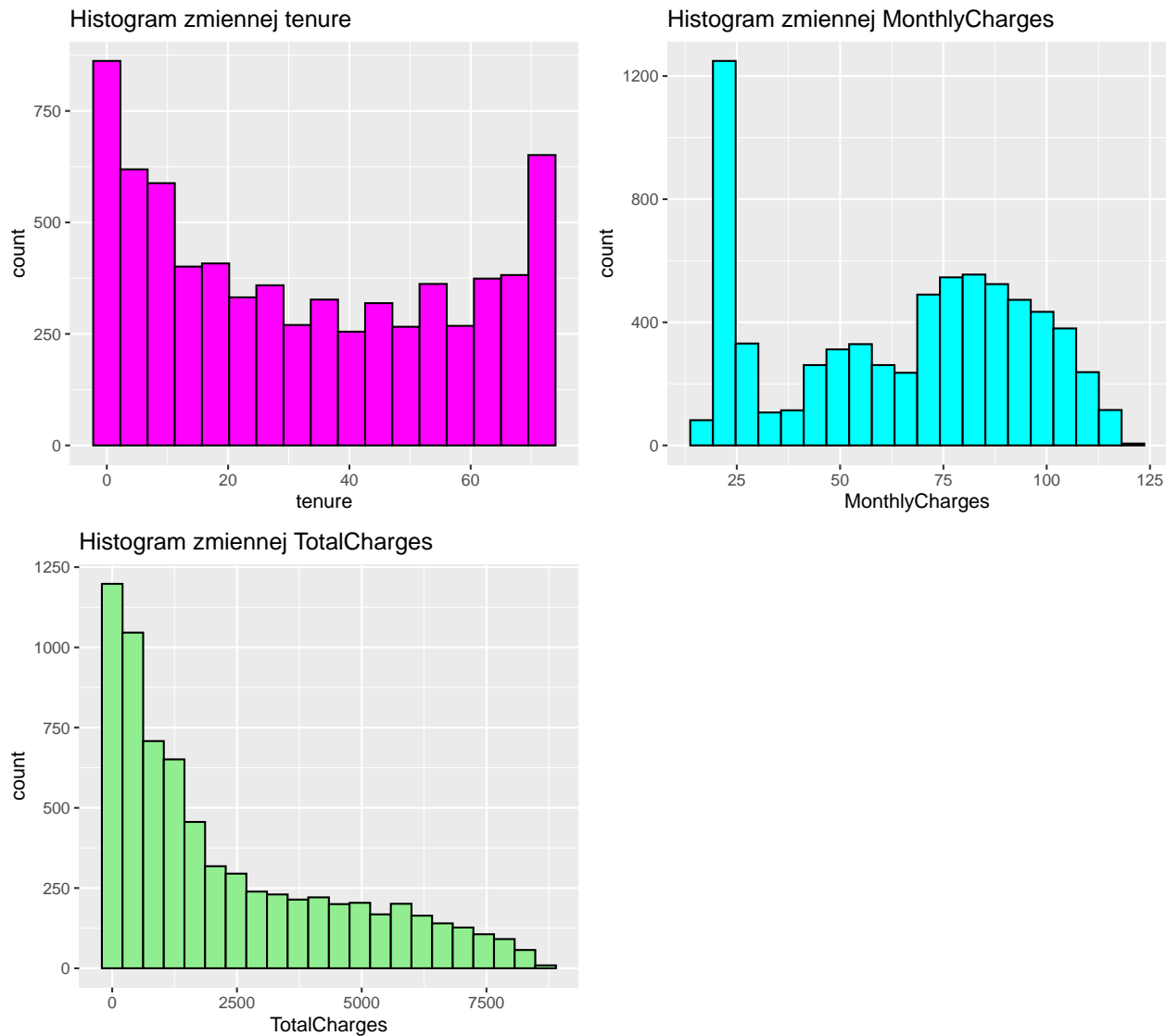
Rysunek 1: Rozkłady zmiennych kategorycznych

2.3 Wykresy pudełkowe dla zmiennych ilościowych



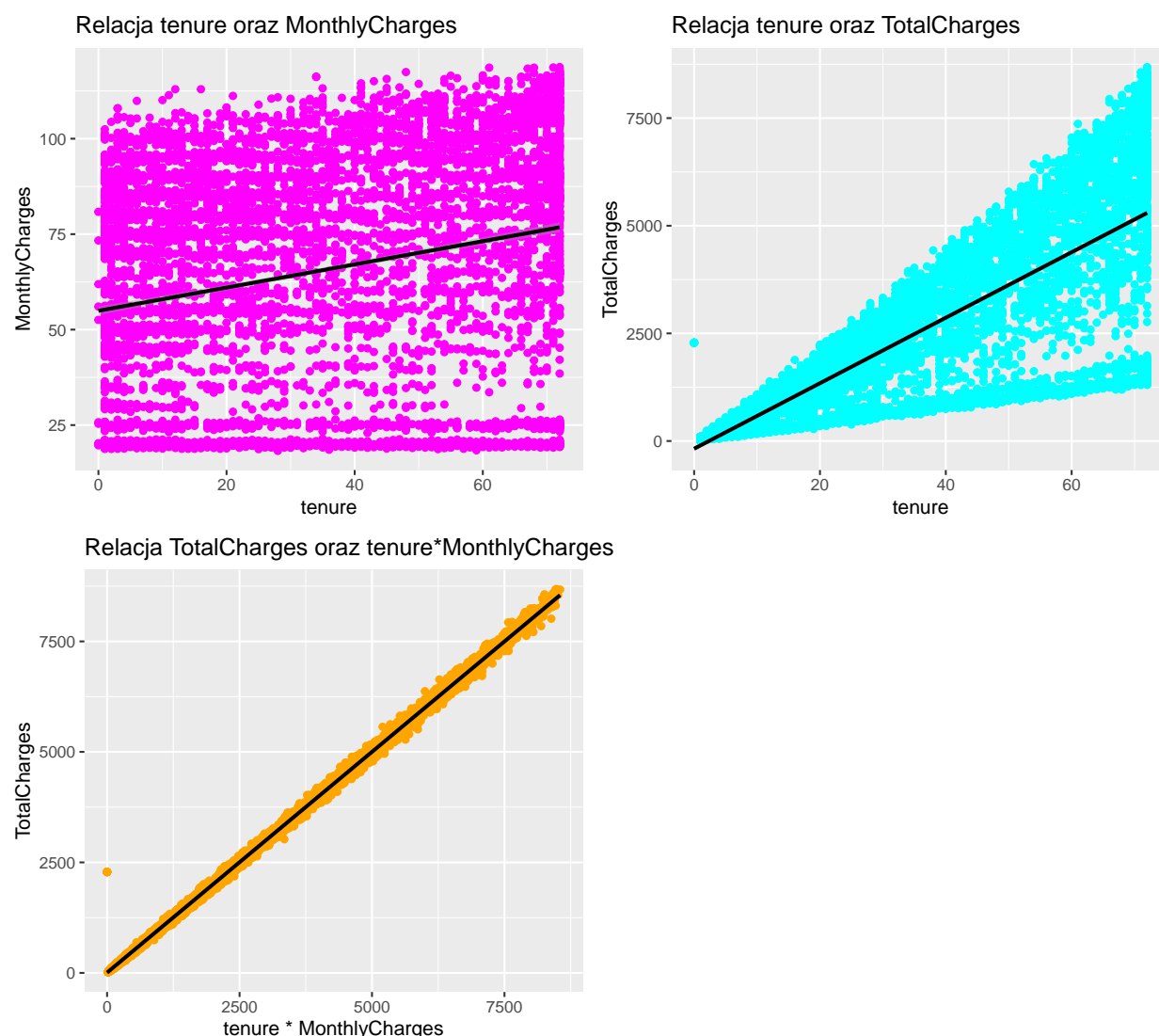
Rysunek 2: Wykresy pudełkowe zmiennych ciągłych

2.4 Histogramy dla zmiennych ilościowych



Rysunek 3: Histogramy zmiennych ciągłych

2.5 Wykresy rozrzutu wraz z krzywą regresji dla zmiennych ilościowych



Rysunek 4: Wykresy rozrzutu wraz z krzywą regresji liniowej

2.6 Interpretacja wykresów.

Przyglądając się rozkładowi zmiennych jakościowych, możemy dojść do wielu ciekawych wniosków. Przede wszystkim rozkład płci klientów jest jednostajny. Spośród wszystkich typów kontraktu (miesięczny, roczny, dwuletni) zdecydowanie największą popularnością cieszy się kontrakt miesięczny. Klienci korzystający z usług internetowych najchętniej korzystają ze światłowodu, chociaż druga najczęstsza opcja (tj. DSL) ma również spore grono odbiorców. Największy niepokój budzi kompletny brak zainteresowania usługami z zakresu cyberbezpieczeństwa. Przeważająca większość konsumentów nie korzysta z tych rozwiązań mimo dostępu do łącza internetowego. Rozkłady zmiennych ciągłych wykazują różne ciekawe właściwości.

Patrząc na wykres 3 obserwujemy rozkład U-modalny dla zmiennej *tenure*, który jest w przybliżeniu rozkładem symetrycznym. Z kolei zmienna **TotalCharges** wyróżnia się rozkładem prawostronnie skośnym jednomodalnym. Najciekawszy rozkład wykazuje zmienna **MonthlyChargess**, który jest jednomodalny oraz prawostronnie skośny. W oczy rzuca się najwyższy słupek znajdujący się na lewo od środka histogramu. Największą zmiennością charakteryzuje się zmienna **TotalCharges**, której większość wartości kumuluje się wokół wartości 0.