

# Analiza danych telekomunikacyjnych

## Eksploracja danych - Lista nr 1

Ksawery Józefowski, 277513

2025-03-31

### Spis treści

<b>1 Wprowadzenie</b>	<b>2</b>
<b>2 Opis danych</b>	<b>2</b>
<b>3 ETAP 1: Przygotowanie danych</b>	<b>2</b>
3.1 a) Wczytanie danych . . . . .	2
3.2 b) Podstawowe informacje o danych . . . . .	3
<b>4 ETAP 2: Analiza opisowa — wskaźniki sumaryczne i wykresy</b>	<b>5</b>
4.1 a) Wskaźniki sumaryczne . . . . .	5
4.2 b) Wykresy rozkładu zmiennych . . . . .	7
4.3 c) Analiza zależności między zmiennymi . . . . .	11
4.4 d) Interpretacja wyników . . . . .	13
4.5 Wnioski ogólne . . . . .	14
<b>5 ETAP 3: Analiza opisowa z podziałem na grupy</b>	<b>15</b>
5.1 a) Analiza z podziałem na grupy Churn . . . . .	15
5.2 b) Analiza z podziałem na grupy . . . . .	18
5.3 Wnioski dotyczące zróżnicowania: . . . . .	19
<b>6 ETAP 4: Podsumowanie — wnioski</b>	<b>20</b>
6.1 a) Podsumowanie wyników z etapów 1-3 . . . . .	20
6.2 b) Charakterystyka klientów firmy . . . . .	21
6.3 c) Przyczyny rezygnacji klientów i rekomendacje dla firmy . . . . .	21
6.4 Podsumowanie: . . . . .	22
<b>7 Spis literatury</b>	<b>22</b>

# 1 Wprowadzenie

Naszym celem jest zastosowanie metod analizy opisowej (wykresy i wskaźniki sumaryczne) do analizy danych telekomunikacyjnych dotyczących migracji klientów. Dane pochodzą ze zbioru WA\_Fn-UseC\_{Telco-Customer-Churn.csv, który zawiera informacje o klientach firmy telekomunikacyjnej. Analiza ma na celu zrozumienie, jakie czynniki wpływają na decyzję klientów o rezygnacji z usług.

## 2 Opis danych

Zbiór danych zawiera informacje o klientach firmy telekomunikacyjnej, w tym cechy demograficzne, informacje o usługach, z których korzystają, oraz informacje o tym, czy klient zrezygnował z usług.

## 3 ETAP 1: Przygotowanie danych

### 3.1 a) Wczytanie danych

```
# Wczytanie danych
dane <- read.csv("WA_Fn-UseC_-Telco-Customer-Churn.csv",
                  stringsAsFactors = TRUE)

# Sprawdzenie typów zmiennych
str(dane)
```

Po wywołaniu polecenia okazuje się, że kolumny `SeniorCitizen` i `tenure` są typu `integer`. ‘`tenure`’ zostanie zmienione na `numeric`, a `SeniorCitizen` na `factor`. Zostaną one poprawione za pomocą:

```
# Zmiana SeniorCitizen z integer na numeric
dane$SeniorCitizen <- as.factor(dane$SeniorCitizen)

# Zmiana tenure z integer na numeric
dane$tenure <- as.numeric(dane$tenure)
```

## **3.2 b) Podstawowe informacje o danych**

### **3.2.1 Rozmiar danych:**

Zbiór zawiera 7043 przypadków i 21 cech.

### **3.2.2 Typy zmiennych:**

#### **3.2.2.1 Zmienne jakościowe**

- **gender** – płeć klienta (Male, Female)
- **SeniorCitizen** – czy klient jest seniorem (0 = Nie, 1 = Tak)
- **Partner** – czy klient ma partnera (Yes, No)
- **Dependents** – czy klient ma osoby na utrzymaniu (Yes, No)
- **PhoneService** – czy klient ma usługę telefoniczną (Yes, No)
- **MultipleLines** – czy klient ma wiele linii (No phone service, No, Yes)
- **InternetService** – rodzaj usługi internetowej (DSL, Fiber optic, No)
- **OnlineSecurity** – czy klient ma ochronę online (No internet service, No, Yes)
- **OnlineBackup** – czy klient ma kopie zapasowe online (No internet service, No, Yes)
- **DeviceProtection** – czy klient ma ubezpieczenie urządzenia (No internet service, No, Yes)
- **TechSupport** – czy klient ma wsparcie techniczne (No internet service, No, Yes)
- **StreamingTV** – czy klient korzysta z telewizji internetowej (No internet service, No, Yes)
- **StreamingMovies** – czy klient korzysta z usług streamingowych (No internet service, No, Yes)
- **Contract** – rodzaj umowy (Month-to-month, One year, Two year)
- **PaperlessBilling** – czy klient korzysta z e-faktury (Yes, No)
- **PaymentMethod** – metoda płatności (Electronic check, Mailed check, Bank transfer, Credit card)
- **Zmienna docelowa: Churn** – czy klient zrezygnował (Yes, No)

#### **3.2.2.2 Zmienne ilościowe dyskretne**

- **tenure** – liczba miesięcy korzystania z usług

#### **3.2.2.3 Zmienne ilościowe ciągłe**

- **MonthlyCharges** – miesięczna opłata (wartości zmiennoprzecinkowe)
- **TotalCharges** – łączna kwota zapłacona przez klienta

### 3.2.3 Identyfikatory klientów:

Zmienna `customerID` pełni rolę identyfikatora i zostanie usunięta poleceniem:

```
# Usunięcie kolumny customerID  
dane <- dane[, -which(names(dane) == "customerID")]
```

### 3.2.4 Brakujące obserwacje:

Sprawdzamy czy istnieją brakujące obserwacje poleceniem:

```
# Sprawdzenie brakujących wartości  
any(is.na(dane))
```

Polecenie zwraca TRUE, więc w danych znajdują się brakujące wartości. Szukając głębiej brakujące wartości znajdują się w kolumnie `totalcharges`.

### 3.2.5 Nietypowe wartości

W danych nie znaleziono żadnych nietypowych wartości.

## 4 ETAP 2: Analiza opisowa — wskaźniki sumaryczne i wykresy

### 4.1 a) Wskaźniki sumaryczne

#### 4.1.1 Zmienne ilościowe

Tabela 1: Wskaźniki sumaryczne zmiennych ilościowych

Wskaźnik	Wartość
tenure_srednia	32.37
tenure_mediania	29.00
tenure_min	0.00
tenure_max	72.00
tenure_sd	24.56
MonthlyCharges_srednia	64.76
MonthlyCharges_mediania	70.35
MonthlyCharges_min	18.25
MonthlyCharges_max	118.75
MonthlyCharges_sd	30.09
TotalCharges_srednia	2283.30
TotalCharges_mediania	1397.47
TotalCharges_min	18.80
TotalCharges_max	8684.80
TotalCharges_sd	2266.77

Tabela przedstawia podstawowe wskaźniki sumaryczne dla trzech zmiennych:

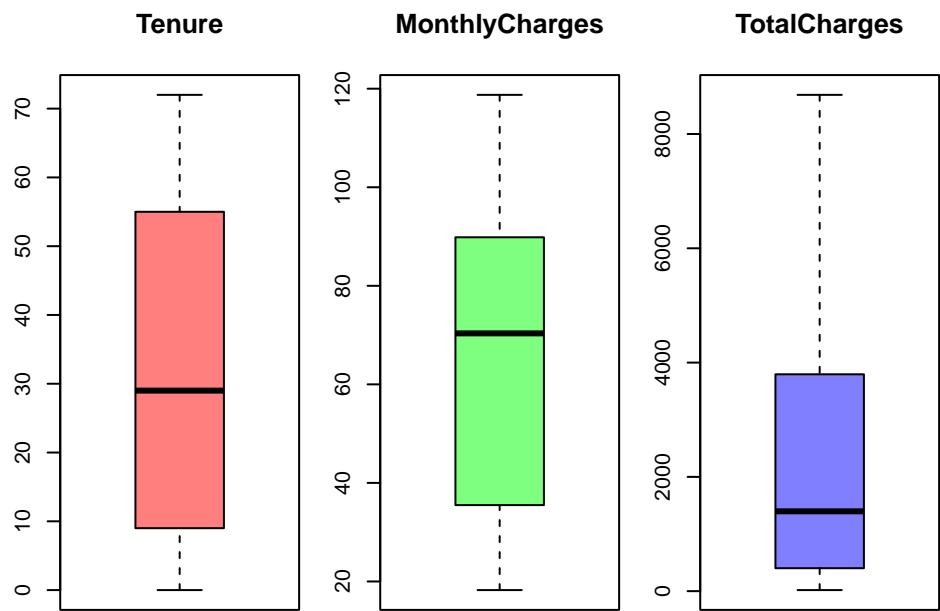
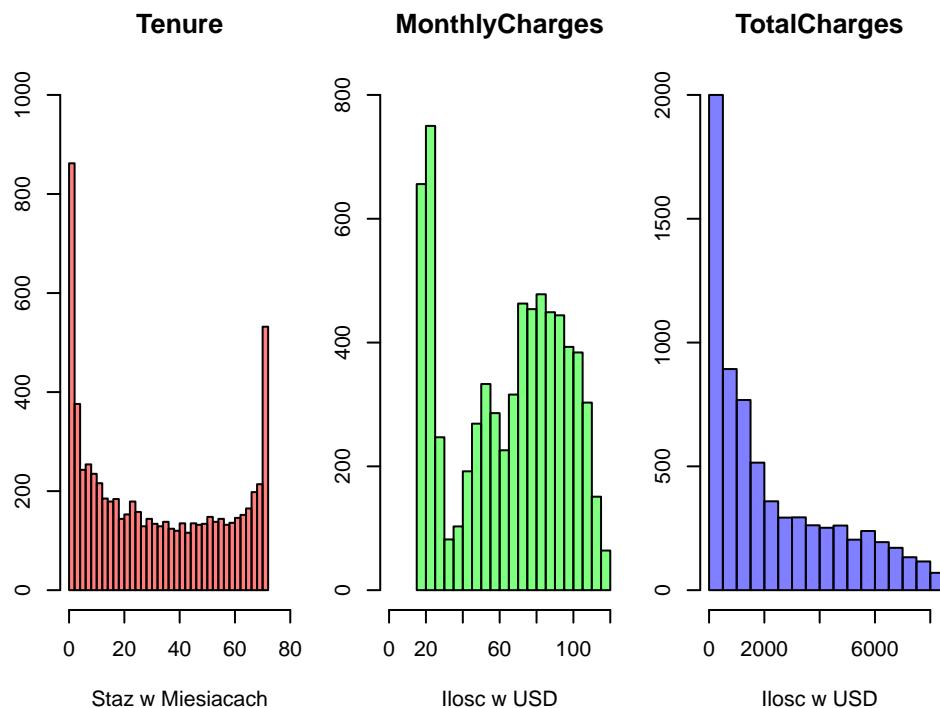
- **tenure:** Średnia wynosi 32,37 miesięcy, a mediana 29 miesięcy, co sugeruje, że większość klientów pozostaje z firmą przez około 2-3 lata. Najkrótszy okres to 0 miesięcy (nowi klienci), a najdłuższy to 72 miesiące (ponad 6 lat). Duże odchylenie standardowe (24,56) wskazuje na duże zróżnicowanie w długości trwania umowy.
- **MonthlyCharges:** Średnia miesięczna opłata to 64,76 USD, ale mediana (70,35 USD) jest wyższa, co może wskazywać na to, że wiele klientów płaci więcej niż średnia. Najniższa opłata to 18,25 USD, a najwyższa to 118,75 USD, co pokazuje dużą różnorodność w kwotach miesięcznych.
- **TotalCharges:** Średnia całkowita kwota zapłacona przez klientów wynosi 2283,30 USD, ale mediana (1397,47 USD) jest znacznie niższa, co sugeruje, że wielu klientów zapłaciło mniejsze sumy. Z dużym odchyleniem standardowym (2266,77 USD) wskazuje to na znaczną różnorodność w całkowitych płatnościach.

#### 4.1.2 Zmienne jakościowe

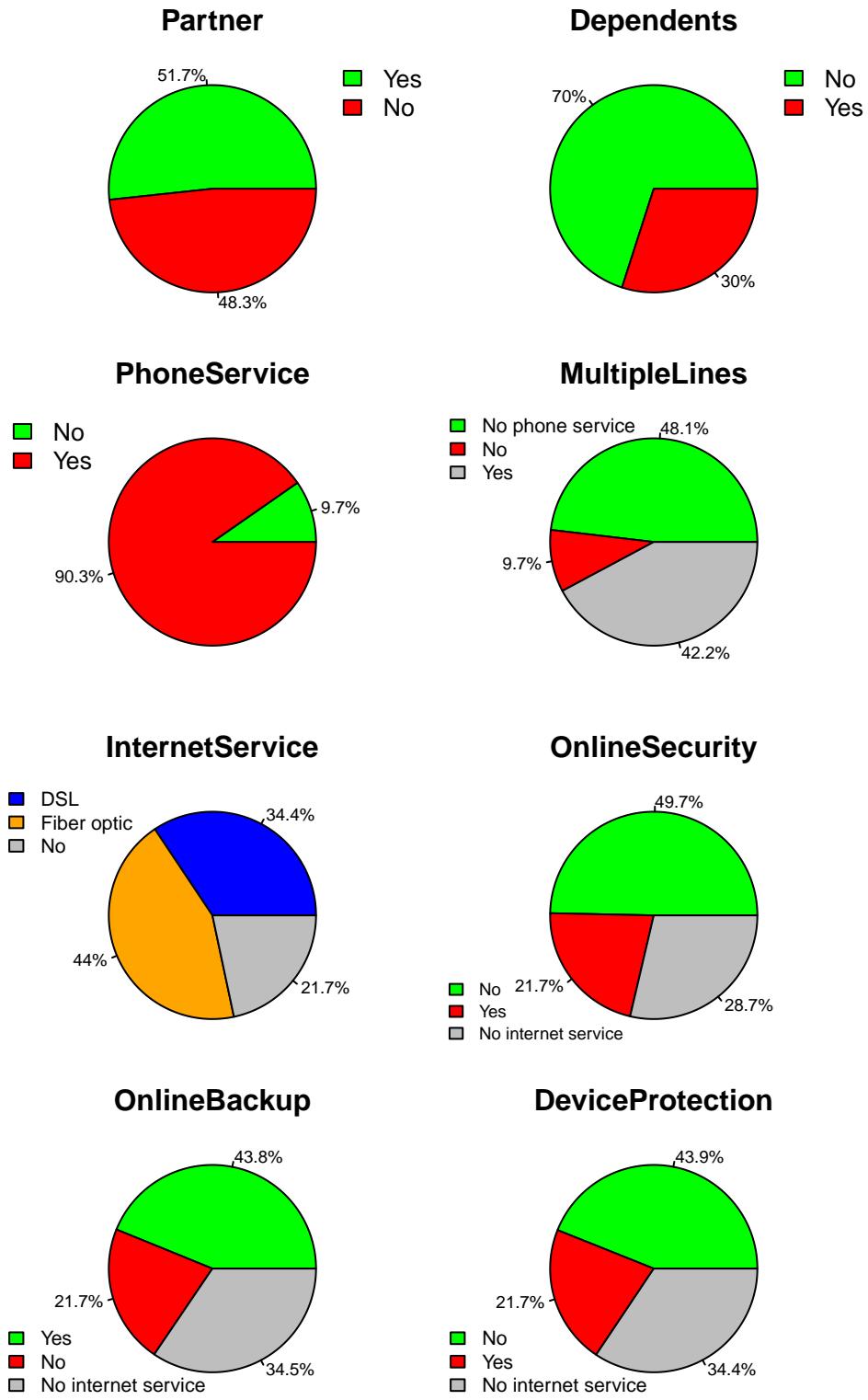
Zmienna	Wartość	Liczność	%
gender	Female	3488	49.5
	Male	3555	50.5
SeniorCitizen	0	5901	83.8
	1	1142	16.2
Partner	No	3641	51.7
	Yes	3402	48.3
Dependents	No	4933	70.0
	Yes	2110	30.0
PhoneService	No	682	9.7
	Yes	6361	90.3
MultipleLines	No	3390	48.1
	No phone service	682	9.7
	Yes	2971	42.2
InternetService	DSL	2421	34.4
	Fiber optic	3096	44.0
	No	1526	21.7
OnlineSecurity	No	3498	49.7
	No internet service	1526	21.7
	Yes	2019	28.7
OnlineBackup	No	3088	43.8
	No internet service	1526	21.7
	Yes	2429	34.5
DeviceProtection	No	3095	43.9
	No internet service	1526	21.7
	Yes	2422	34.4
TechSupport	No	3473	49.3
	No internet service	1526	21.7
	Yes	2044	29.0
StreamingTV	No	2810	39.9
	No internet service	1526	21.7
	Yes	2707	38.4
StreamingMovies	No	2785	39.5
	No internet service	1526	21.7
	Yes	2732	38.8
Contract	Month-to-month	3875	55.0
	One year	1473	20.9
	Two year	1695	24.1
PaperlessBilling	No	2872	40.8
	Yes	4171	59.2
PaymentMethod	Bank transfer (automatic)	1544	21.9
	Credit card (automatic)	1522	21.6
	Electronic check	2365	33.6
	Mailed check	1612	22.9
Churn	No	5174	73.5
	Yes	1869	26.5

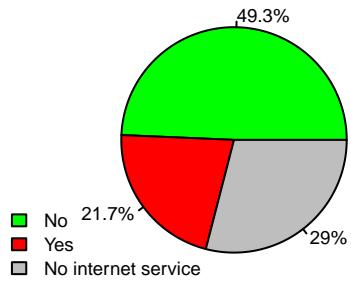
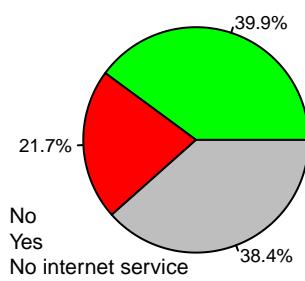
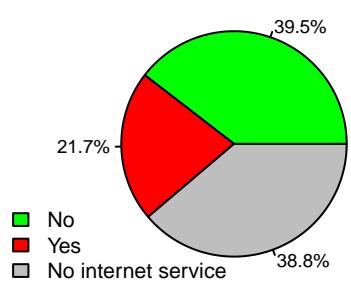
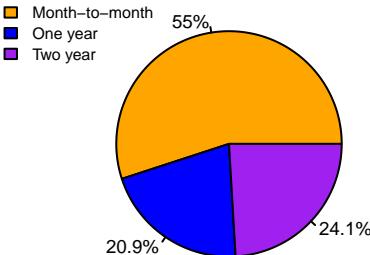
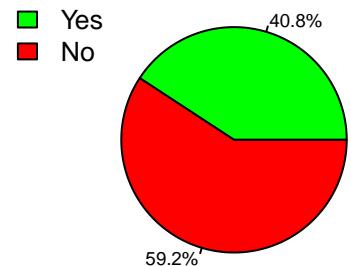
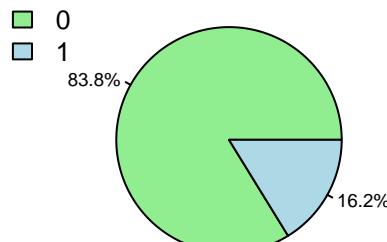
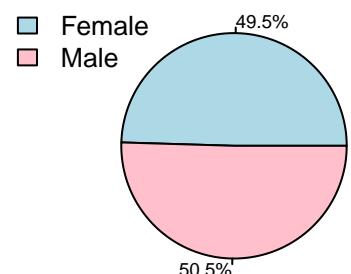
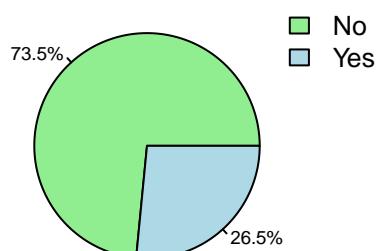
## 4.2 b) Wykresy rozkładu zmiennych

### 4.2.1 Zmienne ilościowe

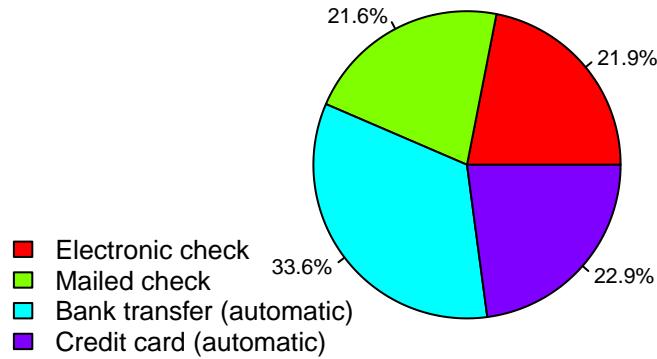


#### 4.2.2 Zmienne jakościowe



**TechSupport****StreamingTV****StreamingMovies****Contract****PaperlessBilling****SeniorCitizen****Gender****Churn**

## PaymentMethod



### 4.3 c) Analiza zależności między zmiennymi

**Macierz wykresów rozrzutu dla zmiennych ilościowych**

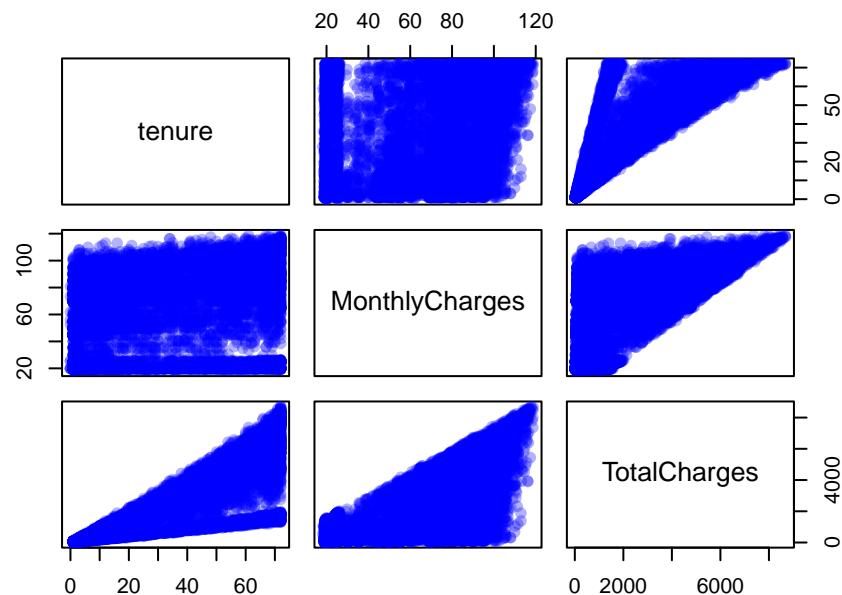


Tabela 2: Macierz korelacji

	tenure	MonthlyCharges	TotalCharges
tenure	1.00	0.25	0.83
MonthlyCharges	0.25	1.00	0.65
TotalCharges	0.83	0.65	1.00

### Macierz wykresów rozrzutu dla zmiennych ilościowych wg Churn

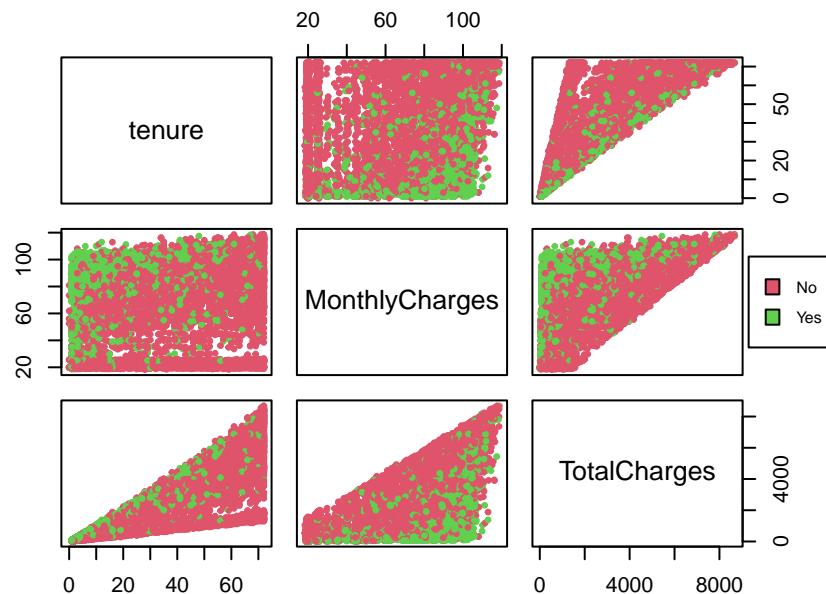


Tabela 3: Macierz korelacji (Churn = No)

	tenure	MonthlyCharges	TotalCharges
tenure	1.00	0.33	0.79
MonthlyCharges	0.33	1.00	0.76
TotalCharges	0.79	0.76	1.00

Tabela 4: Macierz korelacji (Churn = Yes)

	tenure	MonthlyCharges	TotalCharges
tenure	1.00	0.40	0.95
MonthlyCharges	0.40	1.00	0.55
TotalCharges	0.95	0.55	1.00

## 4.4 d) Interpretacja wyników

### 4.4.1 Zakres wartości dla poszczególnych zmiennych

#### 4.4.1.1 Zmienne ilościowe:

- **tenure:**
  - Zakres: 0-72 miesięcy
  - Rozpiętość: 72 miesiące
- **MonthlyCharges:**
  - Zakres: 18.25-118.75 USD
  - Rozpiętość: 100.5 USD
- **TotalCharges:**
  - Zakres: 18.80-8684.80 USD
  - Rozpiętość: 8666 USD

#### 4.4.1.2 Zmienne jakościowe:

Wszystkie zmienne jakościowe mają wartości kategoryczne np:

- **gender:** Female/Male
- **SeniorCitizen:** 0/1
- **Contract:** Month-to-month/One year/Two year
- **Churn:** No/Yes

### 4.4.2 Analiza zmiennych ilościowych

#### 4.4.2.1 Symetria rozkładów:

- **tenure:**
  - Rozkład prawo skośny
  - (średnia 32.37 vs mediana 29.00)
- **MonthlyCharges:**
  - Rozkład lekko lewo skośny
  - (średnia 64.76 vs mediana 70.35)
- **TotalCharges:**
  - Rozkład silnie prawostronnie skośny
  - (średnia 2283.30 vs mediana 1397.47)

#### 4.4.2.2 Zmienna:

Największą zmienność (na podstawie odchylenia standardowego) wykazuje:

1. TotalCharges ( $sd = 2266.77$ )
2. MonthlyCharges ( $sd = 30.09$ )
3. tenure ( $sd = 24.56$ )

#### 4.4.3 Analiza zmiennych jakościowych

##### 4.4.3.1 Częstość kategorii:

- **gender:** prawie równy podział (Female 49.5%, Male 50.5%)
- **SeniorCitizen:** większość to młodsi klienci (83.8% vs 16.2%)
- **Partner:** lekka przewaga osób bez partnera (51.7% vs 48.3%)
- **Dependents:** większość bez osób na utrzymaniu (70% vs 30%)
- **PhoneService:** zdecydowana większość ma usługę telefoniczną (90.3%)
- **InternetService:**
  - Fiber optic (44%)
  - DSL (34.4%)
  - Brak (21.7%)
- **Contract:**
  - Miesięczne (55%)
  - Roczne (20.9%)
  - Dwuletnie (24.1%)
- **PaperlessBilling:** preferowane (59.2% vs 40.8%)
- **PaymentMethod:**
  - Electronic check (33.6%)
  - Pozostałe metody (~22% każda)
- **Churn:** większość nie rezygnuje (73.5% vs 26.5%)

#### 4.4.4 Analiza zależności między zmiennymi

Macierz korelacji pokazuje:

- Silną dodatnią korelację między **tenure** a **TotalCharges** (0.83)
- Umiarkowaną korelację między **MonthlyCharges** a **TotalCharges** (0.65)
- Słabą korelację między **tenure** a **MonthlyCharges** (0.25)

### 4.5 Wnioski ogólne

1. Długość współpracy (tenure) silnie wpływa na wydatki klientów – im dłuższy okres współpracy, tym wyższe całkowite opłaty (TotalCharges), co wskazuje na lojalność klientów i kumulację kosztów.
2. Wysoka zmienność wydatków klientów – największe różnice między klientami dotyczą TotalCharges (od 18.80 do 8684.80 USD), co może wynikać z różnic w długości współpracy i wybranych usługach.
3. Rozkład opłat są asymetryczne – TotalCharges i tenure są prawoskośne (więcej nowych klientów), a MonthlyCharges lewoskośny (częstsze wyższe miesięczne opłaty).
4. Większość klientów nie rezygnuje z usług (Churn: 73.5%) – ale istotne są czynniki jak typ umowy (55% miesięczne) i usługi dodatkowe (np. 44% Fiber optic).
5. Profil klienta: młody, bez osób na utrzymaniu – większość to osoby niebędące seniorami (83.8%), bez dependents (70%), często korzystające z paperless billing (59.2%) i usług internetowych (78.3%).

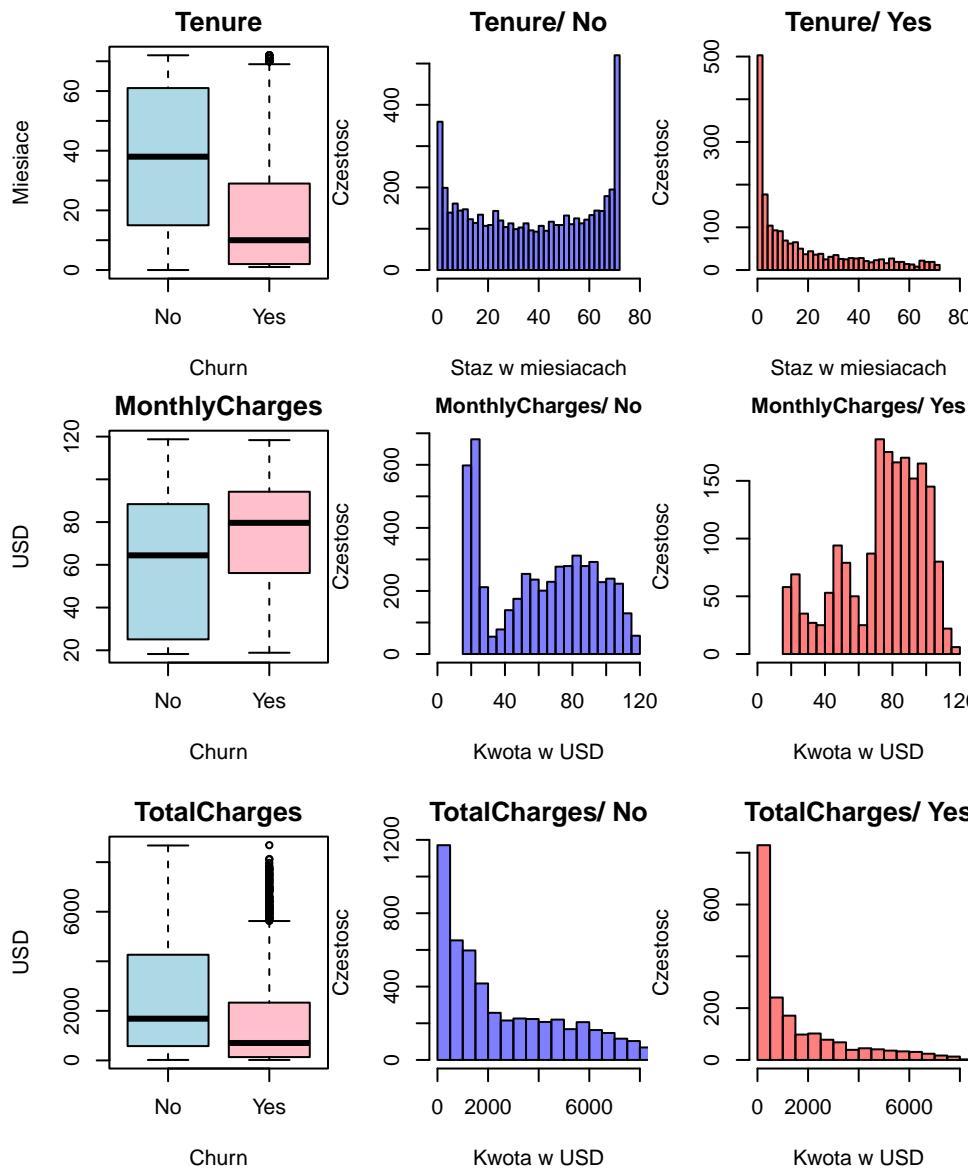
## 5 ETAP 3: Analiza opisowa z podziałem na grupy

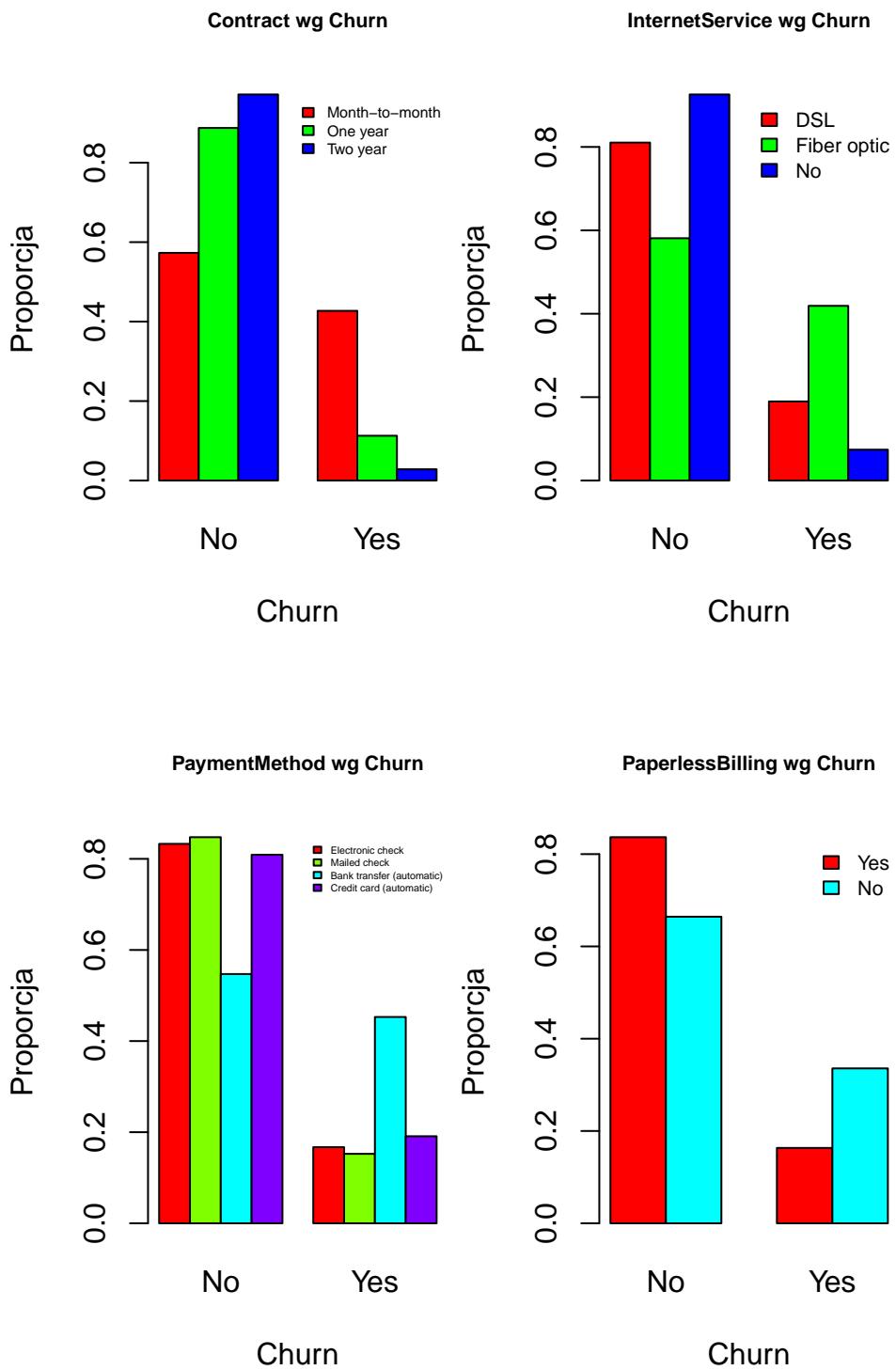
### 5.1 a) Analiza z podziałem na grupy Churn

#### 5.1.1 Zmienne ilościowe

Tabela 5: Porównanie wskaźników sumarycznych między grupami

Wskaźnik	Churn = Yes	Churn = No
tenure_srednia	17.98	37.57
tenure_mediania	10.00	38.00
tenure_min	1.00	0.00
tenure_max	72.00	72.00
tenure_sd	19.53	24.11
MonthlyCharges_srednia	74.44	61.27
MonthlyCharges_mediania	79.65	64.43
MonthlyCharges_min	18.85	18.25
MonthlyCharges_max	118.35	118.75
MonthlyCharges_sd	24.67	31.09
TotalCharges_srednia	1531.80	2555.34
TotalCharges_mediania	703.55	1683.60
TotalCharges_min	18.85	18.80
TotalCharges_max	8684.80	8672.45
TotalCharges_sd	1890.82	2329.46





## 5.2 b) Analiza z podziałem na grupy

Analizując wyniki, można zauważyc, że następujące zmienne wykazują największe zróżnicowanie między grupami klientów (Churn = Yes vs Churn = No):

### 5.2.1 Zmienne ilościowe:

- **tenure:**
  - Średni staż klientów, którzy odeszli: 17.98 miesięcy
  - Średni staż lojalnych klientów: 37.57 miesięcy
  - Różnica: 19.59 miesięcy (54.8% wyższa wartość dla lojalnych klientów)
- **MonthlyCharges:**
  - Średnia opłata miesięczna dla klientów, którzy odeszli: 74.44 USD
  - Średnia opłata miesięczna dla lojalnych klientów: 61.27 USD
  - Różnica: 13.17 USD (21.5% wyższa wartość dla klientów, którzy odeszli)
- **TotalCharges:**
  - Średnia opłata całkowita dla klientów, którzy odeszli: 1531.80 USD
  - Średnia opłata całkowita dla lojalnych klientów: 2555.34 USD
  - Różnica: 1023.54 USD (66.8% wyższa wartość dla lojalnych klientów)

### 5.2.2 Zmienne jakościowe:

- **Contract:**
  - Umowa miesięczna: 88.6% wśród klientów, którzy odeszli vs 49.3% wśród lojalnych
  - Umowa roczna: 7.1% vs 24.9%
  - Umowa dwuletnia: 4.3% vs 25.8%
- **InternetService:**
  - Fiber optic: 51.7% wśród klientów, którzy odeszli vs 41.4% wśród lojalnych
  - DSL: 29.9% vs 35.8%
  - Brak usługi: 18.4% vs 22.8%
- **PaymentMethod:**
  - Electronic check: 45.3% wśród klientów, którzy odeszli vs 29.8% wśród lojalnych
  - Mailed check: 15.7% vs 18.6%
  - Bank transfer: 19.9% vs 24.8%
  - Credit card: 19.1% vs 26.8%
- **PaperlessBilling:**
  - Tak: 69.1% wśród klientów, którzy odeszli vs 55.7% wśród lojalnych

### **5.3 Wnioski dotyczące zróżnicowania:**

Największe różnice między grupami obserwujemy dla:

- **Tenure:** klienci lojalni mają znacznie dłuższy staż
- **Contract:** klienci z umowami miesięcznymi znacznie częściej odchodzą
- **PaymentMethod:** klienci płacący elektronicznym czekiem częściej odchodzą
- **MonthlyCharges:** klienci, którzy odchodzą, płacą wyższe miesięczne opłaty

Zmienne, które najlepiej różnicują grupy to:

- **Contract:** (szczególnie umowa miesięczna vs długoterminowe)
- **Tenure:** (długość stażu)
- **PaymentMethod:** (electronic check vs inne metody)
- **MonthlyCharges:** (wysokość opłat miesięcznych)

Interesujące jest, że choć klienci z usługą Fiber optic częściej odchodzą, różnice nie są tak znaczące jak w przypadku innych zmiennych.

Te zmienne powinny być szczególniebrane pod uwagę przy budowaniu modeli predykcyjnych churnu.

## 6 ETAP 4: Podsumowanie — wnioski

### 6.1 a) Podsumowanie wyników z etapów 1-3

#### 6.1.1 Przygotowanie danych (Etap 1):

- Zbiór danych zawierał informacje o 7043 klientach i 20 zmiennych (ilościowych i jakościowych).
- Wykryto brakujące wartości w kolumnie TotalCharges, które zostały pominięte w analizie.
- Zmienne zostały odpowiednio przekonwertowane (np. SeniorCitizen na factor, tenure na numeric).

#### 6.1.2 Analiza opisowa (Etap 2):

- **Zmienne ilościowe:**
  - Średni staż klientów (tenure) wynosił 32,4 miesiące, ale rozkład był prawoskońny (większość klientów miała krótszy staż).
  - Średnia miesięczna opłata (MonthlyCharges) wynosiła 64,76 USD, przy czym klienci płacący więcej częściej rezygnowali.
  - Łączne opłaty (TotalCharges) były silnie skorelowane z długością stażu (korelacja 0,83).
- **Zmienne jakościowe:**
  - Większość klientów korzystała z usług telefonicznych (90%) i internetowych (78%), głównie DSL (34%) i Fiber optic (44%).
  - Najpopularniejszą umową była umowa miesięczna (55%), a najczęściej wybieraną metodą płatności – e-check (34%).
  - 26,5% klientów zrezygnowało z usług (Churn = Yes).

#### 6.1.3 Analiza z podziałem na grupy (Etap 3):

- **Klienci, którzy odeszli:**
  - Mieli krótszy staż (średnio 17,98 miesięcy vs. 37,57 miesięcy u lojalnych).
  - Płacili wyższe opłaty miesięczne (średnio 74,44 USD vs. 61,27 USD).
  - Częściej mieli umowy miesięczne (88,6% vs. 49,3%) i płacili e-checkiem (45,3% vs. 29,8%).
- **Klienci lojalni:**
  - Częściej mieli umowy długoterminowe (roczne/dwuletnie).
  - Korzystali z DSL (35,8% vs. 29,9%), a mniej z Fiber optic (41,4% vs. 51,7%).

## **6.2 b) Charakterystyka klientów firmy**

### **6.2.1 Profil demograficzny:**

- Prawie równy podział płci (50,5% mężczyzn, 49,5% kobiet).
- Głównie młodsi klienci (tylko 16,2% to seniorzy).
- 51,7% nie ma partnera, a 70% nie ma osób na utrzymaniu.

### **6.2.2 Korzystanie z usług:**

- 90% ma usługę telefoniczną, a 78% ma internet (głównie Fiber optic – 44%).

### **6.2.3 Najpopularniejsze dodatkowe usługi:**

- Streaming TV (38%) i Streaming Movies (39%).
- Znacznie mniej klientów korzysta z ochrony online (29%) czy wsparcia technicznego (29%).
- 59,2% korzysta z e-faktur.

### **6.2.4 Umowy i płatności:**

- 55% klientów ma umowy miesięczne, co może zwiększać ryzyko rezygnacji.
- 33,6% płaci e-checkiem, co wiąże się z wyższym wskaźnikiem rezygnacji.

## **6.3 c) Przyczyny rezygnacji klientów i rekomendacje dla firmy**

### **6.3.1 Główne przyczyny rezygnacji:**

- Wysokie miesięczne opłaty – Klienci, którzy odchodzą, płacą średnio 13,17 USD więcej niż lojalni.
- Krótki staż i umowy miesięczne – Nowi klienci (średnio 18 miesięcy vs. 38 miesięcy) częściej rezygnują.
- Metoda płatności (e-check) – 45,3% rezygnujących płaciło elektronicznym czekiem (vs. 29,8% lojalnych).
- Usługa Fiber optic – Wiąże się z wyższymi opłatami i większą rezygnacją (51,7% vs. 41,4%).

### **6.3.2 Rekomendacje dla firmy:**

- Obniżenie cen dla nowych klientów – Wprowadzenie promocji dla klientów z krótkim stażem, aby zwiększyć ich lojalność.
- Zachęta do umów długoterminowych – Np. zniżki za podpisanie umowy rocznej/dwuletniej.
- Programy lojalnościowe – Nagradzanie klientów za długolatnią współpracą (np. dodatkowe GB internetu, darmowe usługi).
- Poprawa jakości usługi Fiber optic – Jeśli klienci rezygnują z powodu wysokich cen, można wprowadzić tańsze pakiety.
- Zachęta do innych metod płatności – Np. rabaty za automatyczne przelewy/karty kredytowe, aby zmniejszyć odsetek płatności e-checkiem.

### **6.4 Podsumowanie:**

Firma powinna skupić się na obniżeniu kosztów dla nowych klientów, promowaniu umów długoterminowych oraz zwiększeniu wartości usług, aby zmniejszyć wskaźnik rezygnacji. Analiza wskazuje, że kluczowe czynniki wpływające na odejście to wysokie ceny, krótki staż i umowy miesięczne.

## **7 Spis literatury**

### **Literatura**

- [1] Kaggle, *Telco Customer Churn*, <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>.