

Sprawozdanie z listy 1

Eksploracja danych

Marta Stankiewicz (282244)

Paweł Nowak (282223)

2025-03-31

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	Istotność zmiennych i eliminacja redundancji	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 wybranych 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
3	Etap 3. Analiza opisowa z podziałem na grupy	8
3.1	Podstawowe wskaźniki sumaryczne dla zmiennych ciągłych z podziałem na grupy klientów lojalnych i odchodzących	8
3.2	Wykresy pudełkowe dla zmiennych ilościowych z podziałem na grupy klientów lojalnych i odchodzących	9
3.3	Wykresy słupkowe dla wybranych zmiennych kategoriowych z podziałem na grupy klientów lojalnych i odchodzących	10
3.4	Analiza wyników	11
4	Etap 4. Podsumowanie – wnioski z przeprowadzonej analizy	11

Spis rysunków

1	Rozkłady wybranych zmiennych kategoriowych	4
2	Wykresy pudełkowe zmiennych ciągłych	5

3	Histogramy zmiennych ciągłych	6
4	Wykresy rozrzutu wraz z krzywą regresji liniowej	7
5	Wykresy pudełkowe zmiennych ciągłych z podziałem na grupy	9
6	Rozkłady wybranych zmiennych kategoriowych z podziałem na klientów lojalnych i nielojalnych	10

Spis tabel

1	Wskaźniki sumaryczne dla zmiennych ciągłych	3
2	Porównanie wskaźników sumarycznych dla grup klientów lojalnych i odchodzących na podstawie zmiennej <i>TotalCharges</i>	8
3	Porównanie wskaźników sumarycznych dla grup klientów lojalnych i odchodzących na podstawie zmiennej <i>MonthlyCharges</i>	8
4	Porównanie wskaźników sumarycznych dla grup klientów lojalnych i odchodzących na podstawie zmiennej <i>tenure</i>	8

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 *OnlineSecurity* 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

Wśród wszystkich zmiennych jedynie `r.n.missing[“variable”][1]` zawiera brakujące wartości, których liczba wynosi `r.n.missing[“na.count”][1]`. Aby zapewnić spójność danych, zastosujemy imputację tych braków, uzupełniając je średnią wartością. Brakujące dane są standardowo oznaczone jako NA, a w zbiorze nie występują niestandardowe reprezentacje braków.

1.3 Istotność zmiennych i eliminacja redundancji

Naszym celem jest przewidzenie, czy klient zrezygnuje z usług firmy na podstawie dostępnych cech. W tym celu usuniemy zmienne, które nie mają wpływu na decyzje konsumentów lub są

funkcją innych zmiennych. Przykładowo, atrybut customerID jest unikalnym identyfikatorem klienta i nie niesie informacji istotnych dla analizy zachowań konsumenckich, dlatego zostanie wykluczony ze zbioru danych.

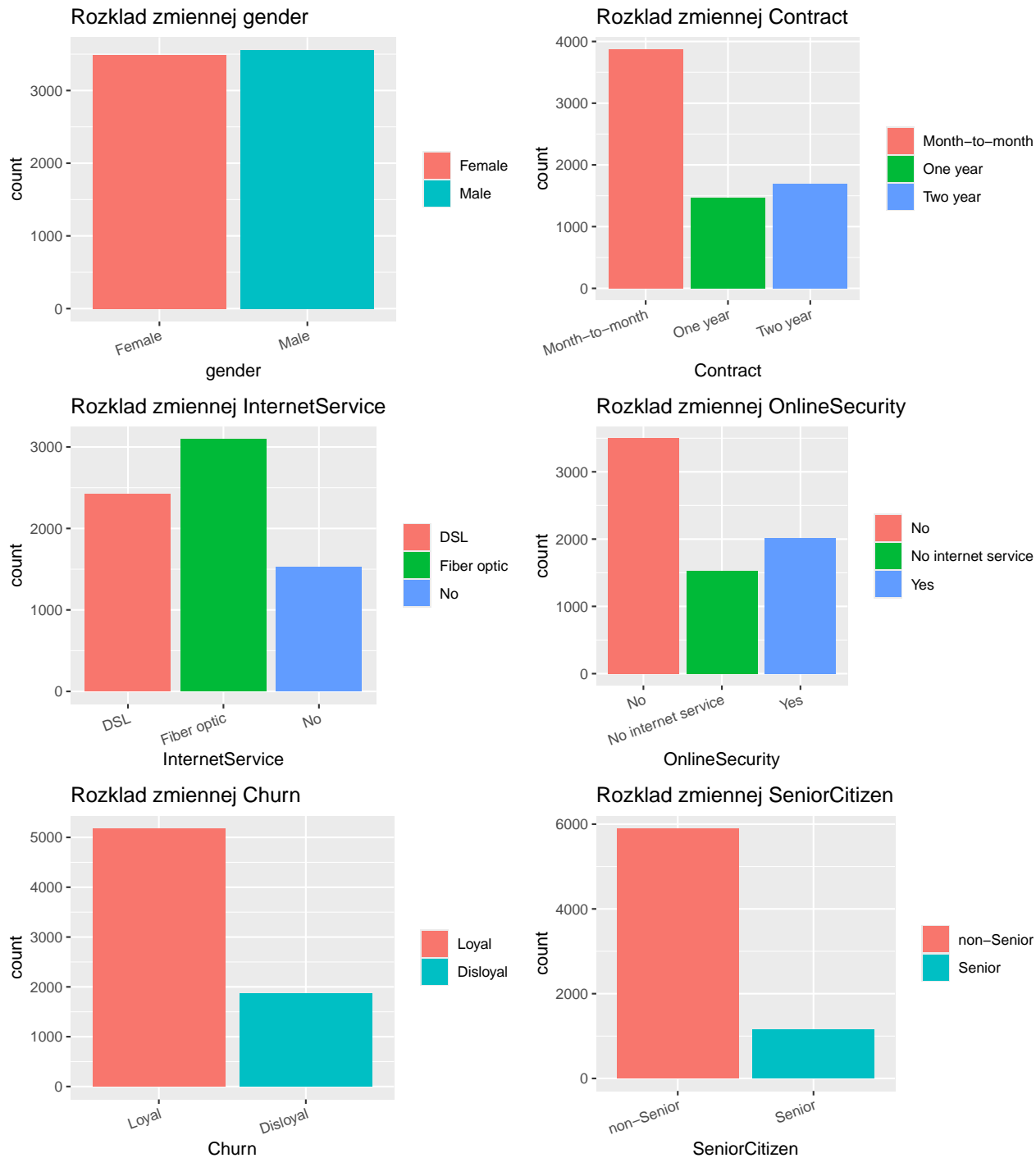
2 Etap 2. Analiza opisowa - wskaźniki summaryczne i wykresy

2.1 Podstawowe wskaźniki summaryczne dla zmiennych ciągłych

Tabela 1: Wskaźniki summaryczne dla zmiennych ciągłych

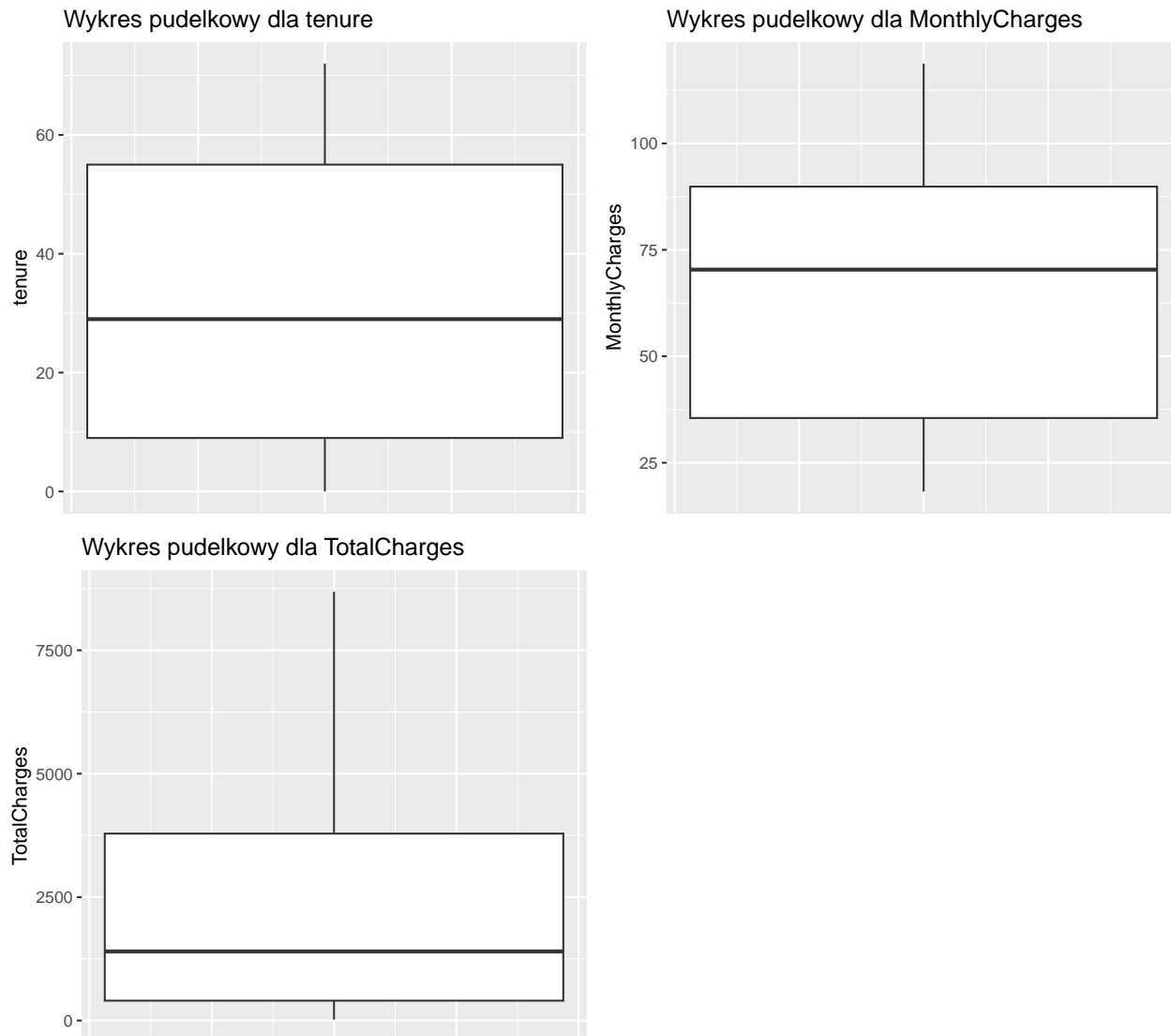
	tenure	MonthlyCharges	TotalCharges
Min	0.00	18.25	18.80
Mean	32.37	64.76	2283.30
Median	29.00	70.35	1400.55
SD	24.56	30.09	2265.00
IQR	46.00	54.35	3384.38
Max	72.00	118.75	8684.80

2.2 Wykresy słupkowe dla wybranych zmiennych kategorycznych



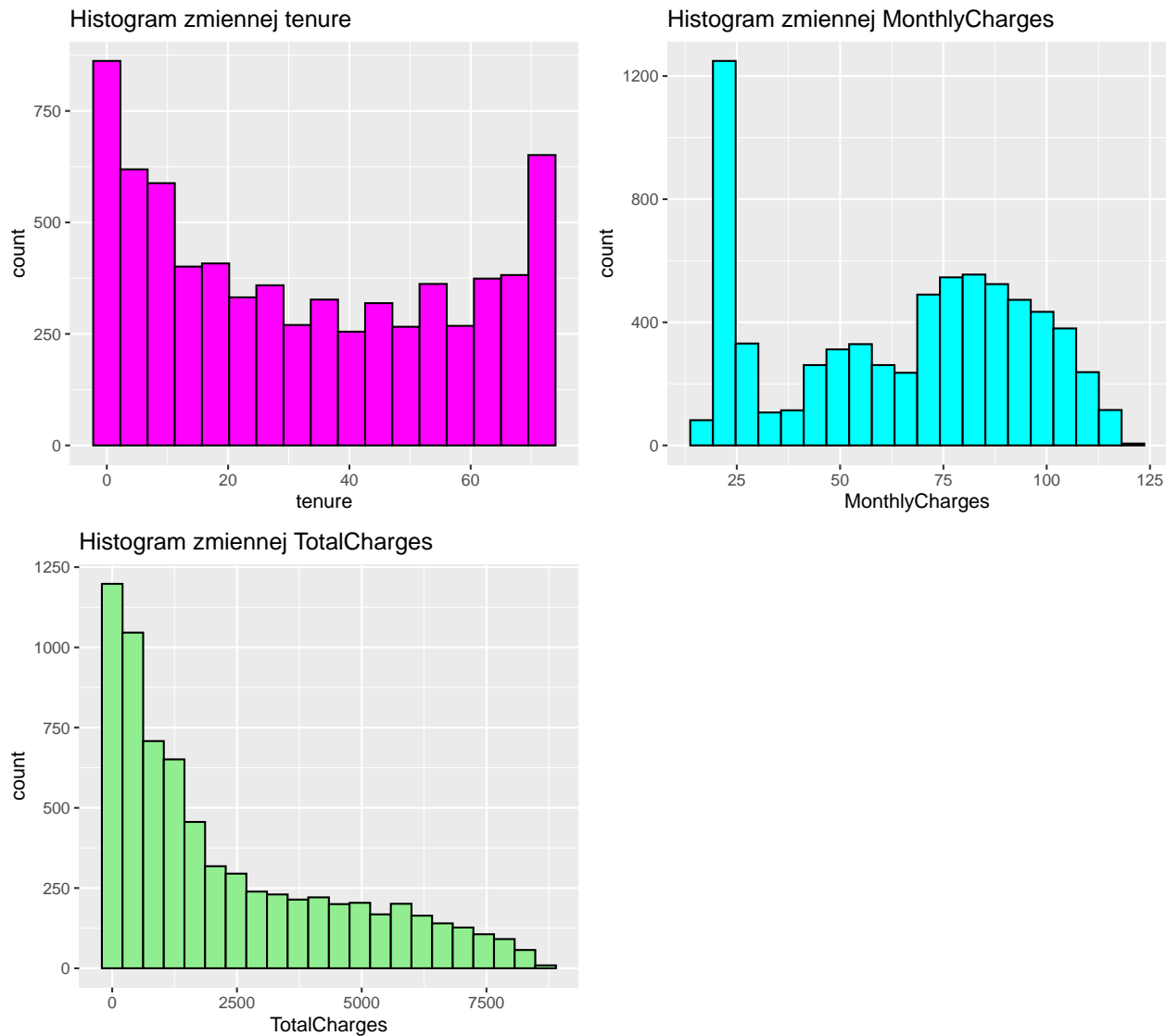
Rysunek 1: Rozkłady wybranych zmiennych kategorycznych

2.3 Wykresy pudełkowe dla zmiennych ilościowych



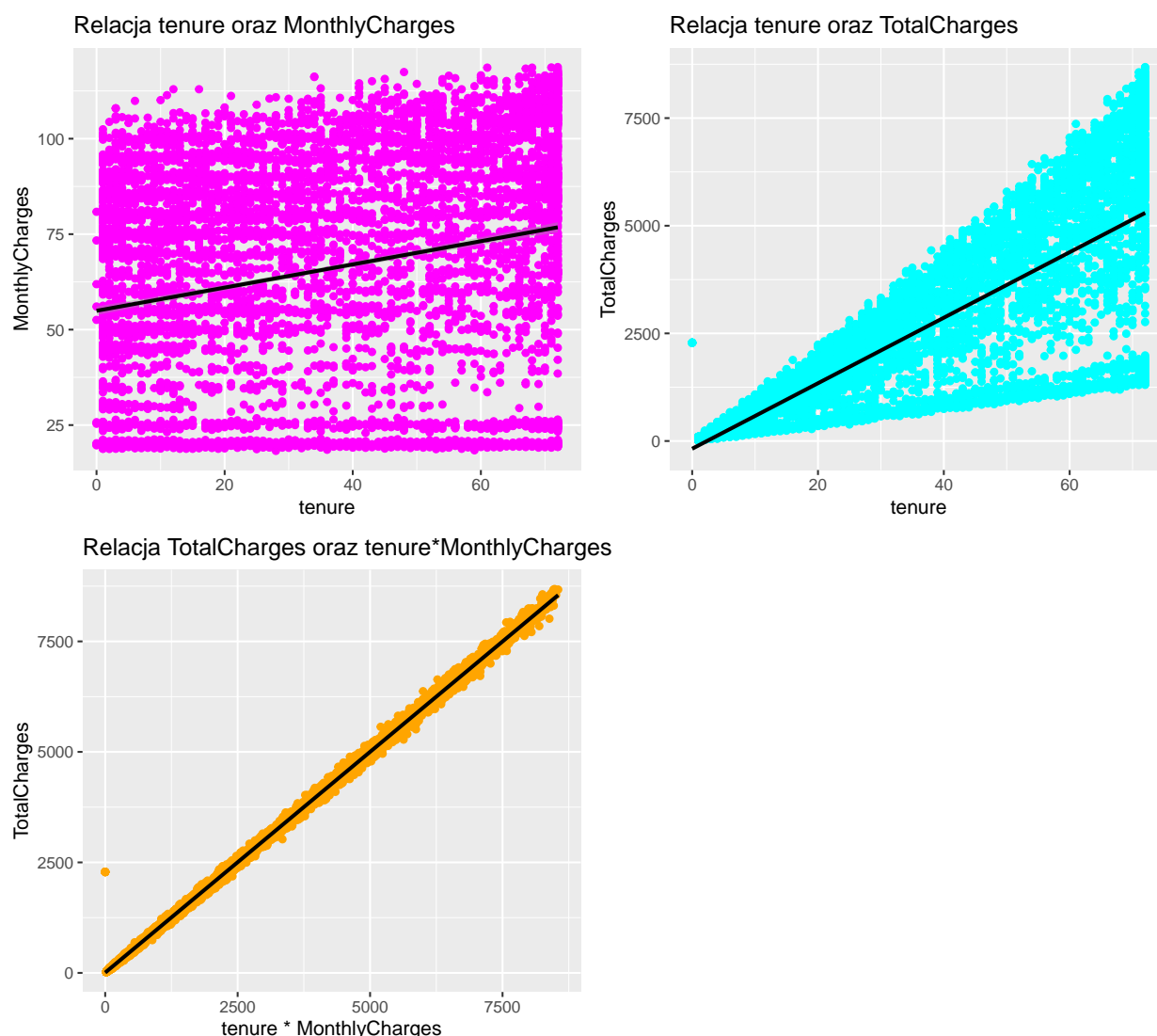
Rysunek 2: Wykresy pudełkowe zmiennych ciągłych

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.

Analizując rozkład zmiennych jakościowych, można dostrzec kilka interesujących zależności. Przede wszystkim liczba klientów obu płci jest niemal identyczna, co wskazuje na równomierny podział wśród użytkowników. Wśród dostępnych typów kontraktu zdecydowanie najczęściej wybierany jest wariant miesięczny, wyraźnie dominując nad umowami rocznymi i dwuletnimi. Jeśli chodzi o rodzaj połączenia internetowego, najwięcej klientów korzysta ze światłowodu, choć technologia DSL również cieszy się sporą popularnością. Niepokojącym zjawiskiem jest natomiast niemal całkowity brak zainteresowania usługami cyberbezpieczeństwa. Mimo posiadania dostępu do internetu, większość użytkowników nie

decyduje się na skorzystanie z tych rozwiązań.

Przechodząc do rozkładów zmiennych ciągłych, wykres 3 ujawnia kilka ciekawych właściwości. Zmienna tenure charakteryzuje się rozkładem U-modalnym, który jest w przybliżeniu symetryczny. TotalCharges wyróżnia się natomiast rozkładem jednomodalnym o wyraźnej skośności prawostronnej. Najbardziej interesujący wydaje się rozkład MonthlyCharges, który łączy w sobie wcześniej wspomniane cechy. Szczególnie rzuca się w oczy najwyższy słupek po lewej stronie histogramu. Spośród analizowanych cech, największą zmienność wykazuje TotalCharges, przy czym większość jej wartości koncentruje się blisko zera.

Analizując zmienne ciągłe pod kątem zależności liniowych, zauważamy, że współczynnik korelacji między tenure a MonthlyCharges jest bliski zeru, co sugeruje brak istotnej zależności liniowej między tymi cechami. Podobny wniosek można wysnuć z wykresu rozrzutu tych zmiennych, choć przy nieco wyższym współczynniku korelacji. Najsilniejszą zależność liniową obserwujemy natomiast między TotalCharges a iloczynem tenure i MonthlyCharges, który stanowi ich interakcję. W tym przypadku regresja liniowa doskonale odwzorowuje zależność w danych, co potwierdza dobre dopasowanie modelu do obserwacji.

3 Etap 3. Analiza opisowa z podziałem na grupy

3.1 Podstawowe wskaźniki sumaryczne dla zmiennych ciągłych z podziałem na grupy klientów lojalnych i odchodzących

Tabela 2: Porównanie wskaźników sumarycznych dla grup klientów lojalnych i odchodzących na podstawie zmiennej TotalCharges

	min	Q1	median	mean	Q3	max	sd	IQR
lojalni	18.80	579.57	1689.18	2554.77	4262.85	8672.45	2327.01	3683.27
nielojalni	18.85	134.50	703.55	1531.80	2331.30	8684.80	1890.82	2196.80

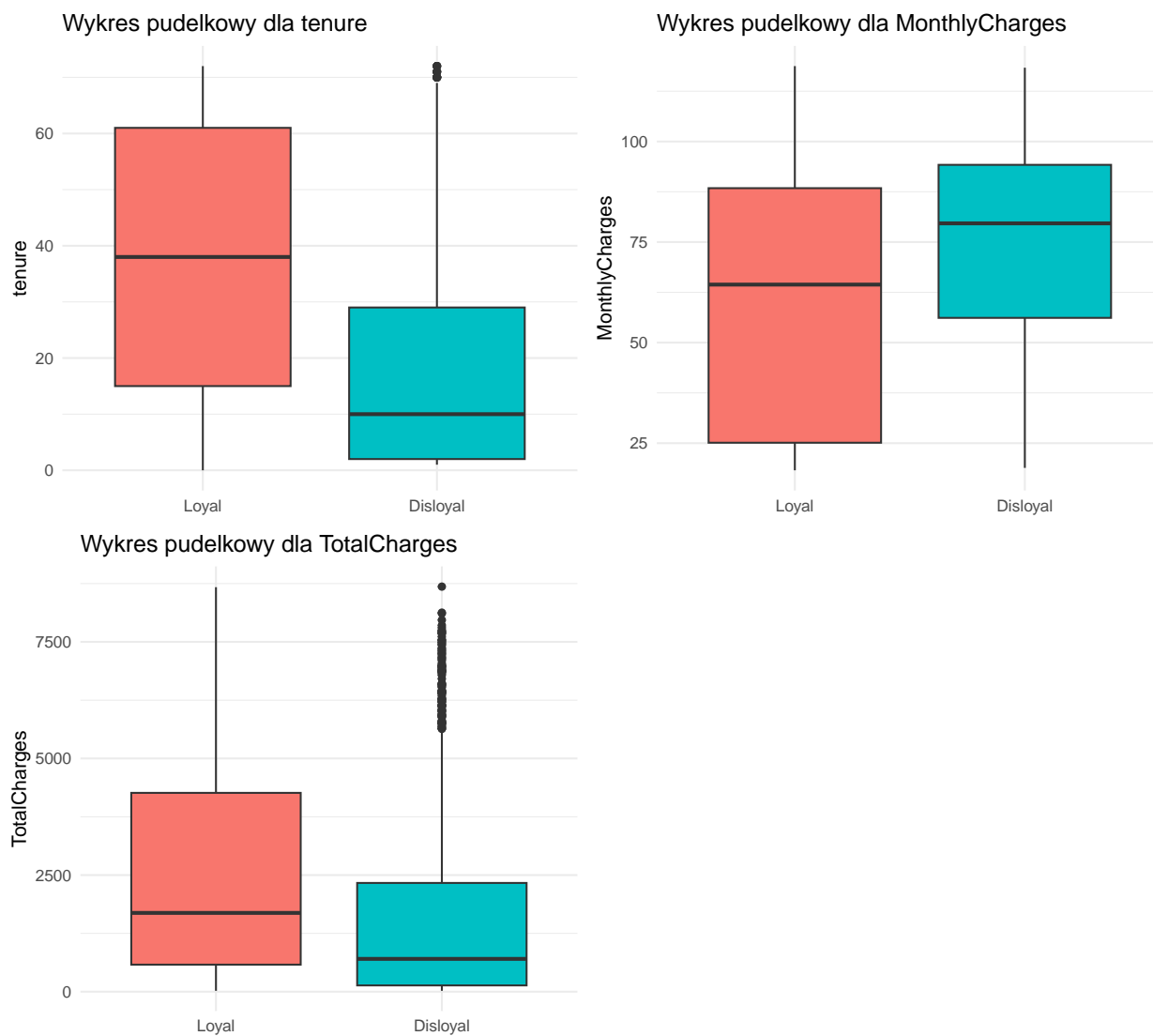
Tabela 3: Porównanie wskaźników sumarycznych dla grup klientów lojalnych i odchodzących na podstawie zmiennej MonthlyCharges

	min	Q1	median	mean	Q3	max	sd	IQR
lojalni	18.25	25.10	64.43	61.27	88.4	118.75	31.09	63.30
nielojalni	18.85	56.15	79.65	74.44	94.2	118.35	24.67	38.05

Tabela 4: Porównanie wskaźników sumarycznych dla grup klientów lojalnych i odchodzących na podstawie zmiennej tenure

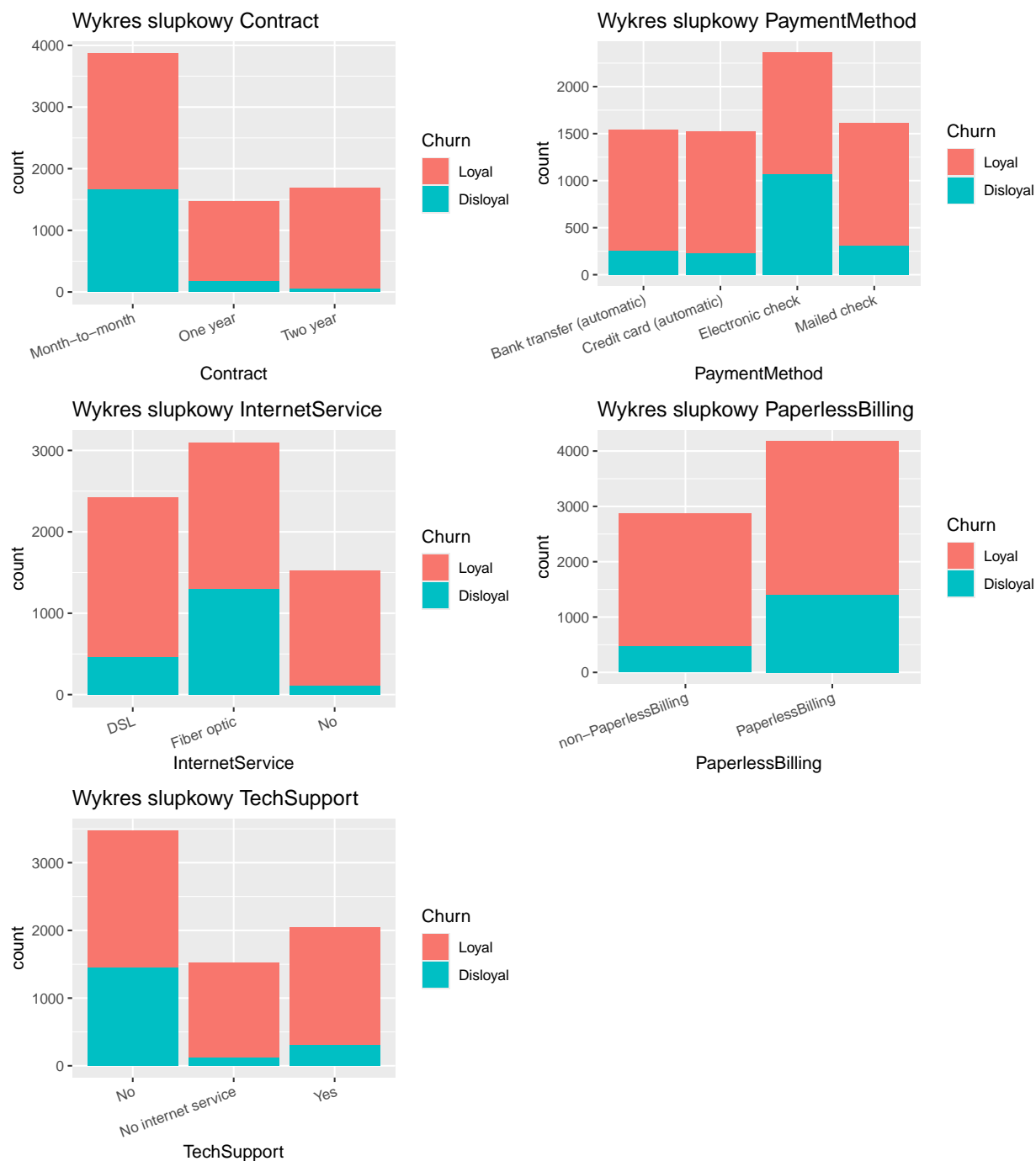
	min	Q1	median	mean	Q3	max	sd	IQR
lojalni	0	15	38	37.57	61	72	24.11	46
nielojalni	1	2	10	17.98	29	72	19.53	27

3.2 Wykresy pudełkowe dla zmiennych ilościowych z podziałem na grupy klientów lojalnych i odchodzących



Rysunek 5: Wykresy pudełkowe zmiennych ciągłych z podziałem na grupy

3.3 Wykresy słupkowe dla wybranych zmiennych kategoriycznych z podziałem na grupy klientów lojalnych i odchodzących



Rysunek 6: Rozkłady wybranych zmiennych kategoriycznych z podziałem na klientów lojalnych i niełojalnych

3.4 Analiza wyników

Analizując wskaźniki sumaryczne dla zmiennych ciągłych oraz interpretując wykresy pudełkowe przedstawione na rysunku, można zauważyć, że najlepsze rozróżnienie między klientami, którzy nadal korzystają z usług firmy, a tymi, którzy zrezygnowali, zapewnia zmienna **tenure**, czyli liczba miesięcy, przez które klient korzystał z usług firmy. To właśnie w tej zmiennej występuje największa różnica między medianami dla obu grup – lojalnych i nielojalnych klientów. Warto również zwrócić uwagę na pozostałe dwie zmienne, **MonthlyCharges** oraz **TotalCharges**. Klienci, którzy zrezygnowali z usług, ponosili średnio wyższe miesięczne opłaty niż ci, którzy pozostali.

W przypadku rozkładu zmiennej **TotalCharges** warto zauważyć, że choć średnia wartość całkowitych wydatków klientów nielojalnych jest mniejsza, to próbki odstające w tej grupie znacznie przewyższają koszty ponoszone przez klientów lojalnych.

Przyjrzyjmy się teraz wykresom słupkowym. Jeżeli podzielimy klientów ze względu na typ umowy, którą zawarli z firmą, stanie się jasne, że największy odsetek klientów odchodzących występuje w grupie osób podlegających miesięcznemu typowi umowy.

Analizując inne cechy, obserwujemy, że największą grupę klientów odchodzących z firmy stanowią osoby, które swoje usługi opłacają, korzystając z czeku elektronicznego; zdecydowały się na kupno światłowodu; dostają elektroniczny rachunek za wybrane usługi, lub które nie wybrały usługi wsparcia technicznego.

4 Etap 4. Podsumowanie – wnioski z przeprowadzonej analizy

Nasza baza klientów cechuje się równomiernym podziałem pod względem płci – kobiety i mężczyźni są reprezentowani w jednakowych proporcjach. Wśród dostępnych opcji umownych największym zainteresowaniem cieszy się umowa miesięczna. Jeśli chodzi o internet, niezaprzecalnie dominuje światłowód.

Pod względem wieku przeważają osoby młode, choć starsi użytkownicy również chętnie korzystają z oferty. Natomiast rozwiązania z zakresu cyberbezpieczeństwa nie wzbudzają dużego zainteresowania. Zauważamy również pewien poziom rezygnacji, jednak większość klientów pozostaje z nami na dłużej.

Rezygnacja z usług firmy jest ściśle powiązana z typem zawartego kontraktu. Klienci posiadający umowę miesięczną częściej decydują się na odejście. Jest to w pełni zrozumiałe – krótkoterminowe zobowiązania ułatwiają rozwiązanie umowy, co sprawia, że decyzja o rezygnacji jest prostsza do podjęcia. Zadowoleni klienci chętniej wybierają długoterminowe kontrakty, co znacząco zmniejsza prawdopodobieństwo ich rezygnacji w przyszłości.

Aby zwiększyć lojalność klientów, firma powinna skupić się na promowaniu wsparcia technicznego (TechSupport). Problemy z działaniem produktów często prowadzą do frustracji, dlatego szybka i skuteczna pomoc w ich naprawie może znacząco poprawić satysfakcję użytkowników i zmniejszyć liczbę rezygnacji.

Wprowadzenie zniżek dla stałych klientów mogłoby znacząco podnieść poziom ich zadowolenia. Jak wynika z wykresu, lojalni użytkownicy ponoszą znacznie niższe miesięczne koszty w porównaniu do osób rezygnujących, co sugeruje, że atrakcyjna polityka cenowa sprzyja

długoterminowej współpracy.