Sprawozdanie z listy 1

Eksploracja danych

Marta Stankiewicz (282244)
Paweł Nowak (282223)

2025-03-31

Spis treści

Spis rysunków

Spis tabel

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

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 Określenie 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.

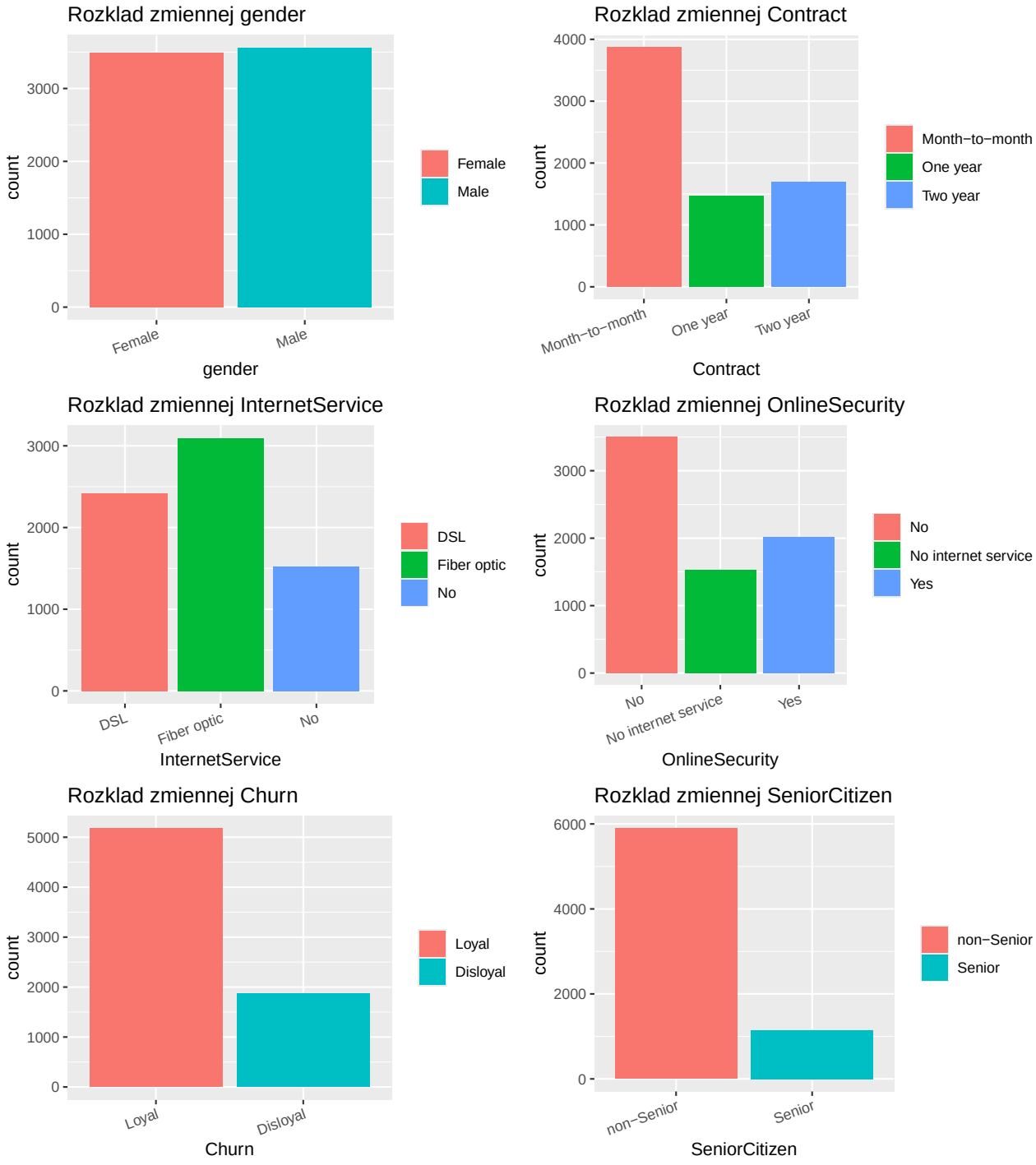
2 Etap 2. Analiza opisowa - wskaźniki sumaryczne i wykresy

2.1 Podstawowe wskaźniki sumaryczne dla zmiennych ciągłych

Tabela 1: Wskaźniki sumaryczne dla zmiennych ciągłych

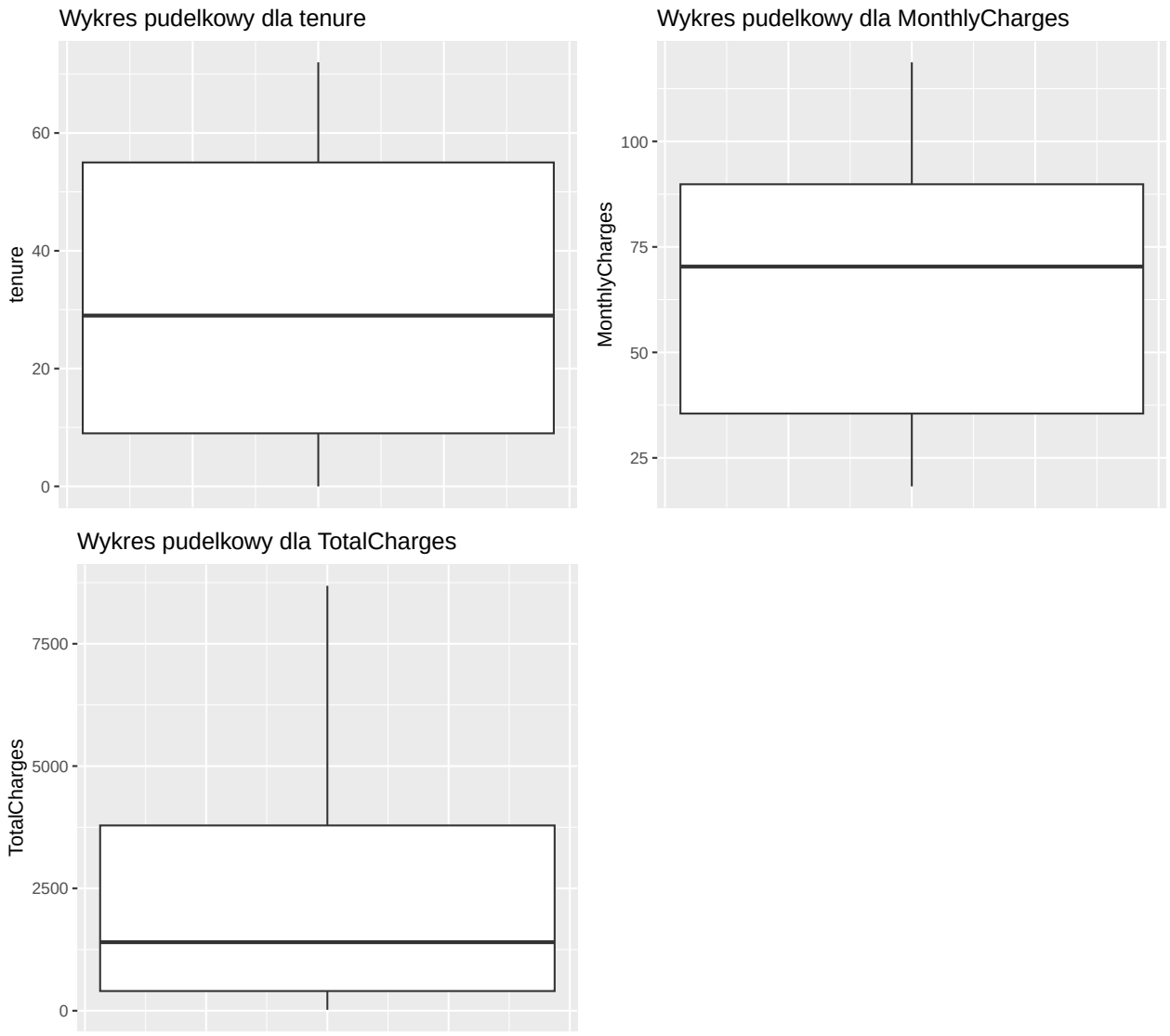
	tenure	MonthlyCharges	TotalCharges
Min	0.00	18.25	18.80
Mean	32.37	64.76	2283.30
Median	29.00	70.35	1400.55
SD	24.56	30.09	2265.00
IQR	46.00	54.35	3384.38
Max	72.00	118.75	8684.80

2.2 Wykresy słupkowe dla wybranych zmiennych kategorycznych



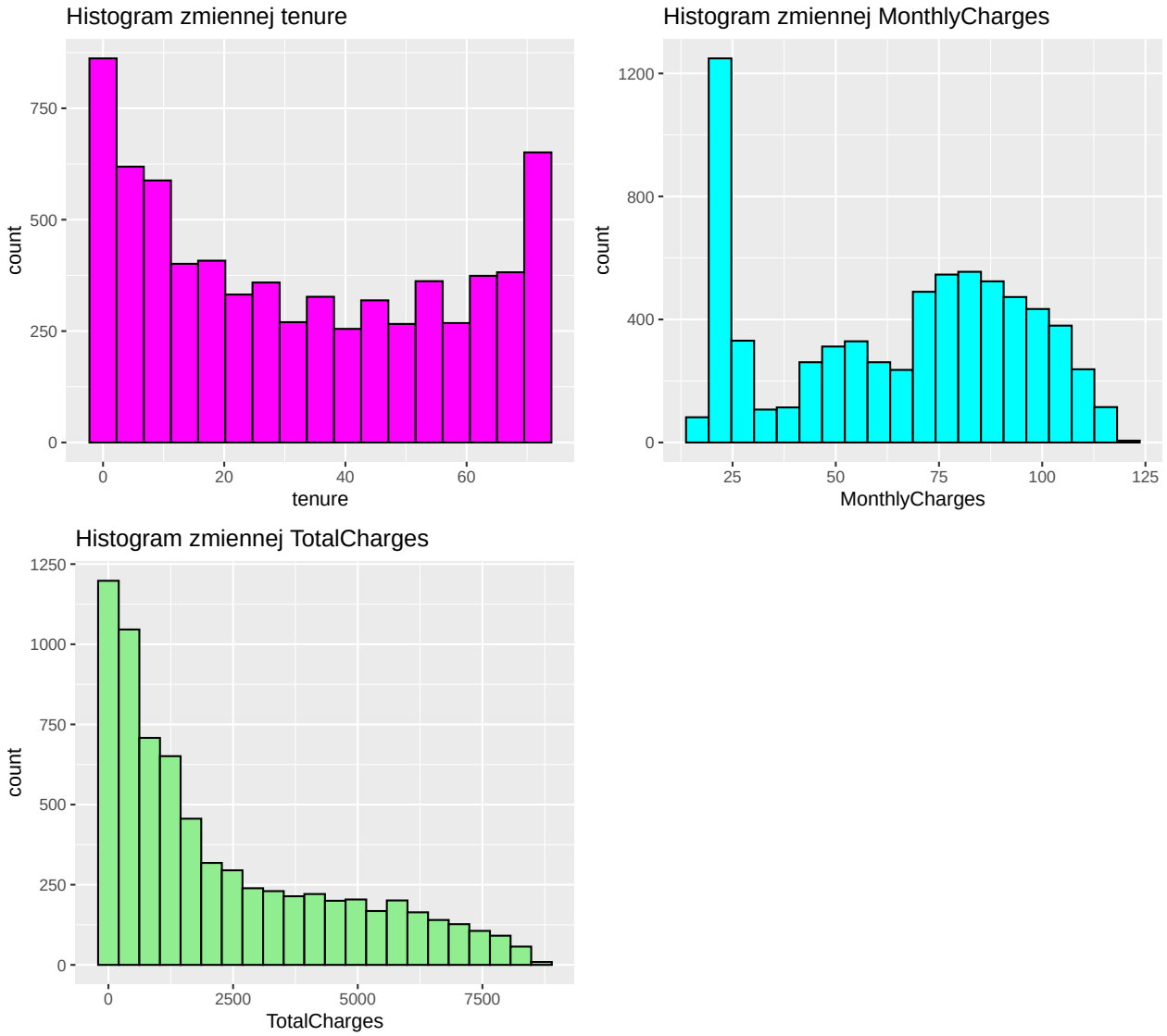
Rysunek 1: Rozkłady wybranych zmiennych kategorycznych

2.3 Wykresy pudełkowe dla zmiennych ilościowych



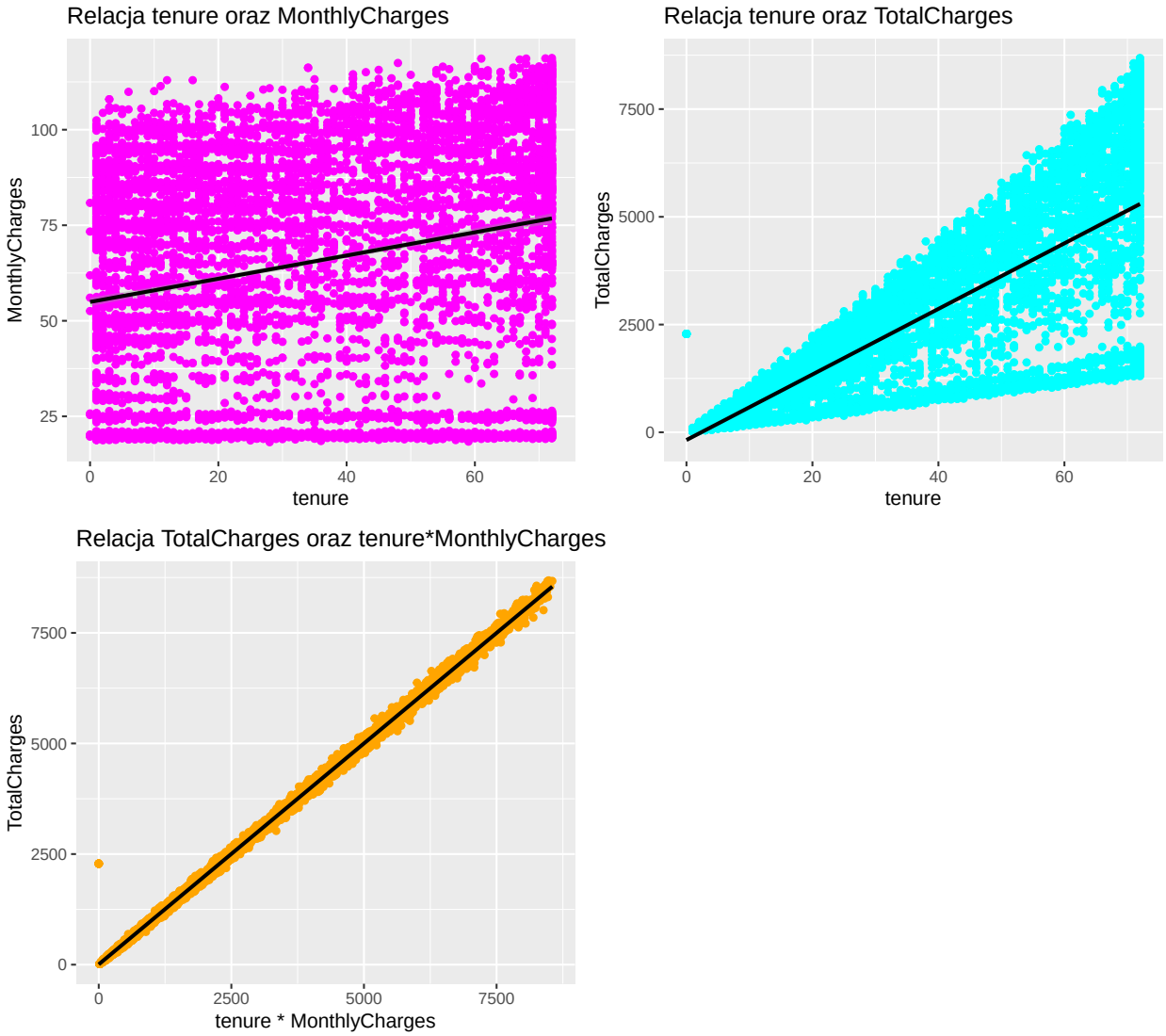
Rysunek 2: Wykresy pudełkowe zmiennych ciągłych

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.

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.

3 Etap 3. Analiza opisowa z podziałem na grupy

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

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

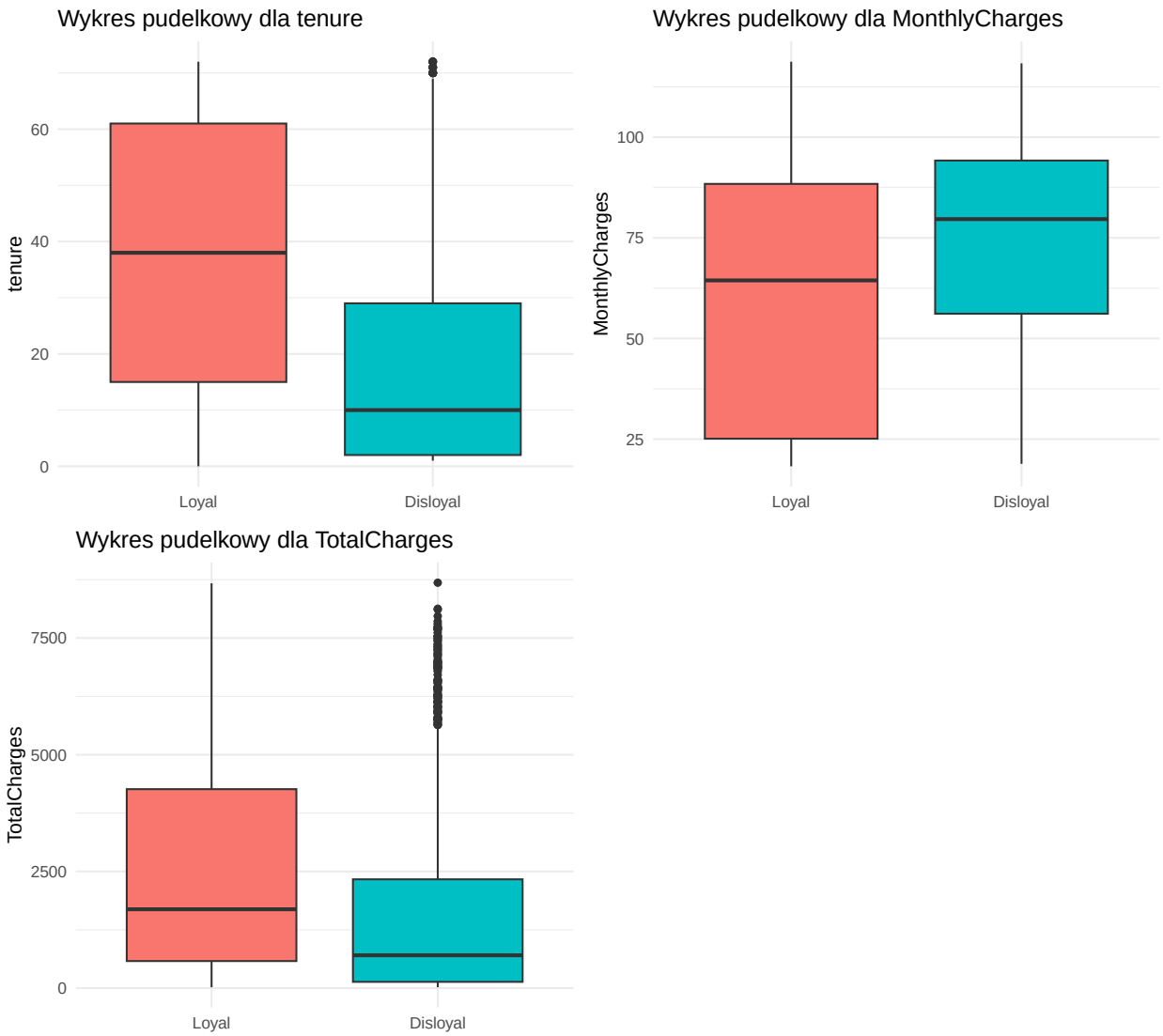
Tabela 3: Porównanie wskaźników sumarycznych dla grup klientów lojalnych i odchodzących na podstawie zmiennej MonthlyCharges

	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

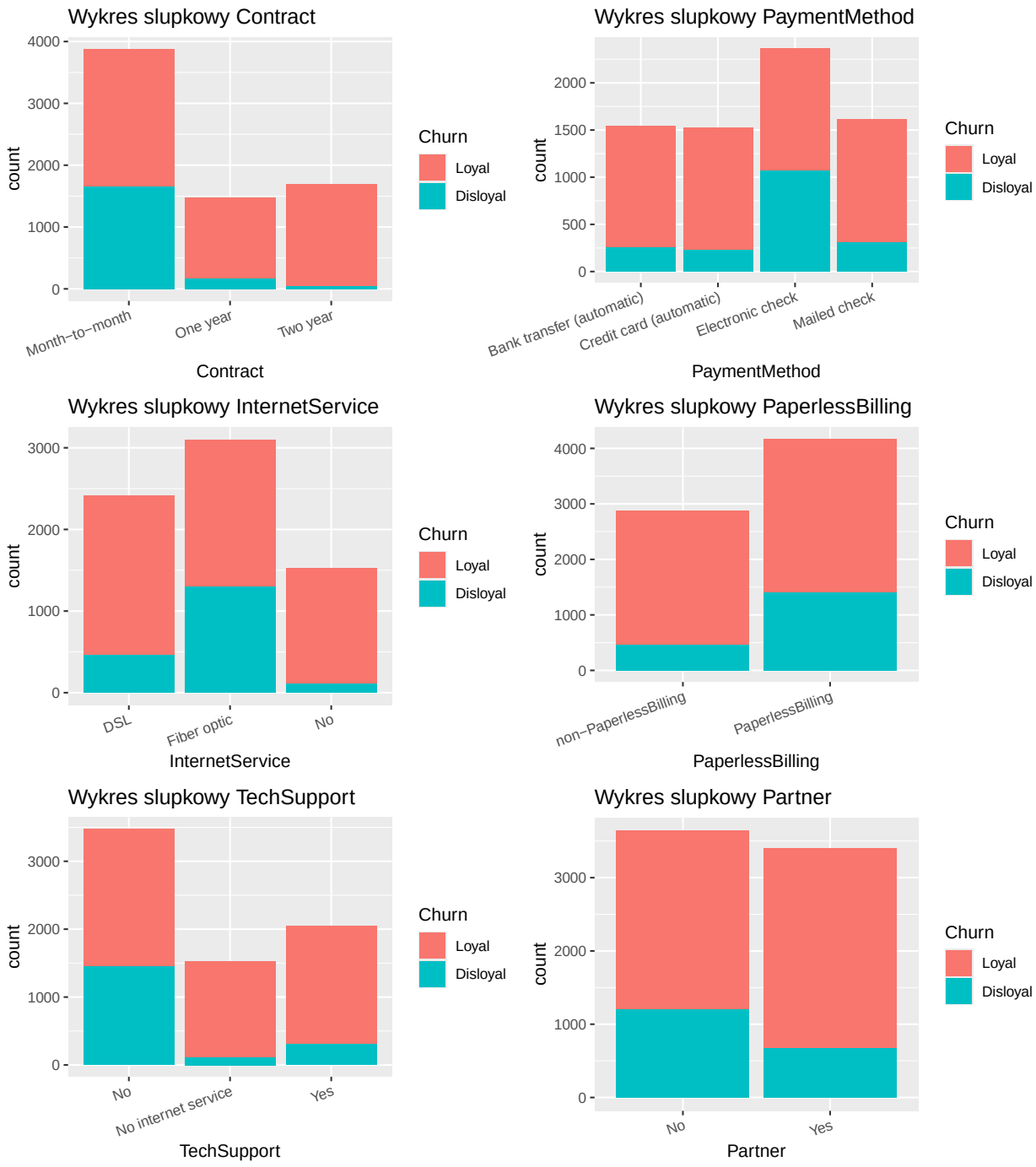
	min	Q1	median	mean	Q3	max	sd	IQR
lojalni	0	15	38	37.57	61	72	24.11	46
nielojalni	1	2	10	17.98	29	72	19.53	27

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



Rysunek 5: Wykresy pudełkowe zmiennych ciągłych z podziałem na grupy

3.3 Wykresy słupkowe dla wybranych zmiennych kategorycznych z podziałem na grupy klientów lojalnych i odchodzacych



Rysunek 6: Rozkłady wybranych zmiennych kategorycznych z podziałem na klientów lojalnych i nielojalnych

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, niezaprzeczalnie 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.