

# Sprawozdanie z listy 1

## Eksploracja danych

Daria Grzelak, album 277533

2025-03-28

### Spis treści

<b>1</b>	<b>Krótki opis zagadnienia</b>	<b>1</b>
<b>2</b>	<b>Opis analiz</b>	<b>1</b>
2.1	Opis danych . . . . .	1
<b>3</b>	<b>Wyniki</b>	<b>3</b>
3.1	Zmienne liczbowe – analiza ogólna . . . . .	3
3.2	Zmienne jakościowe – analiza ogólna . . . . .	7
3.3	Zmienne liczbowe – analiza z podziałem . . . . .	12
3.4	Zmienne jakościowe – analiza z podziałem . . . . .	16
3.5	Podsumowanie . . . . .	21

## 1 Krótki opis zagadnienia

Moim celem jest analiza danych o klientach pewnej firmy telekomunikacyjnej. Znajdują się w nich informacje o danych klientów i usługach, z których korzystają, a także informacja o tym, czy nadal korzystają z usług firmy, czy odeszli. Na podstawie przeprowadzonych analiz zamierzam odpowiedzieć na następujące pytania:

- Co charakteryzuje klientów? Jakie różnice można zaobserwować pomiędzy klientami, którzy pozostali, a tymi, którzy odeszli?
- Jakie działania powinna wprowadzić firma, aby zachęcić klientów do pozostania?

## 2 Opis analiz

Aby znaleźć odpowiedzi na postawione pytania, wykorzystam metody analizy opisowej. Przeanalizuję rozkłady poszczególnych zmiennych (robiąc testy statystyczne, a także prezentując wyniki w formie graficznej), a także zbadam różnice w rozkładach poszczególnych zmiennych pomiędzy klientami, którzy odeszli, a tymi, którzy pozostali.

## 2.1 Opis danych

Analizowane dane zawierają informacje o klientach firmy telekomunikacyjnej – mają 21 cech oraz 7043 przypadki. W tabeli znajduje się bardziej szczegółowy opis cech.

Nazwa (oryginalna)	Typ	Opis
customerI	Jakościowa	Identyfikator klienta
gender	Jakościowa	Płeć klienta
SeniorCitizen	Jakościowa	Czy klient jest seniorem
Partner	Jakościowa	Czy klient ma partnera
Dependents	Jakościowa	Czy klient ma utrzymanków
tenure	Liczbową	Jak długo klient korzysta z usług firmy (w miesiącach)
PhoneService	Jakościowa	Czy klient korzysta z usług telefonicznych
MultipleLines	Jakościowa	Czy klient ma wiele linii telefonicznych
InternetService	Jakościowa	Czy klient korzysta z usługi internetowej
OnlineSecurity	Jakościowa	Czy klient korzysta z usługi bezpieczeństwa online
OnlineBackup	Jakościowa	Czy klient korzysta z usługi kopii zapasowej
DeviceProtection	Jakościowa	Czy klient korzysta z usługi ochrony urządzenia
TechSupport	Jakościowa	Czy klient korzysta ze wsparcia technicznego
StreamingTV	Jakościowa	Czy klient posiada telewizję streamingową
StreamingMovies	Jakościowa	Czy klient korzysta ze streamingu filmów
Contract	Jakościowa	Typ umowy (miesięczna, roczna, dwuletnia)
PaperlessBilling	Jakościowa	Czy klient rozlicza się przez rachunki bez użycia papieru
PaymentMethod	Jakościowa	W jaki sposób klient opłaca rachunki
MonthlyCharges	Liczbową	Kwota, którą klient płaci miesięcznie
TotalCharges	Liczbową	Łączna kwota, którą klient płaci
Churn	Jakościowa	Czy klient odszedł w ciągu ostatniego miesiąca

Tabela 1: Opis cech zawartych w danych

Z tabeli ?? widać, że cecha CustomerID pełni funkcję identyfikatora klienta. Będzie zbędna w dalszej analizie, zatem ją usunę.

Po usunięciu tej cechy pozostało nam 20 cech.

Dane w większości są jakościowe, numeryczne są tylko trzy cechy: tenure (liczba miesięcy, w których klient korzysta z usługi), MonthlyCharges (miesięczne opłaty) oraz TotalCharges (łączne opłaty).

Sprawdzenie brakujących wartości daje natomiast następujący wynik:

```
# Sprawdzenie i wyświetlenie, w których kolumnach występują
# brakujące wartości
colSums(is.na(churn))
```

```
##          gender SeniorCitizen Partner Dependents
##                0            0       0           0
```

```

##          tenure      PhoneService     MultipleLines   InternetService
##          0                  0                  0                  0
##  OnlineSecurity      OnlineBackup  DeviceProtection      TechSupport
##          0                  0                  0                  0
##      StreamingTV  StreamingMovies       Contract  PaperlessBilling
##          0                  0                  0                  0
##  PaymentMethod    MonthlyCharges     TotalCharges        Churn
##          0                  0                  11                  0

```

Jak widać, jedyną kolumną, w której występują brakujące wartości, jest TotalCharges. Brakujące wartości w tej kolumnie są równoważne z wartością tenure równą 0, co udowadnia odpowiednie porównanie dwóch podzbiorów (są takie same).

```
## [1] NA NA
```

Dalsza obserwacja danych wykazała, że nie występują nietypowe wartości, takie jak niestandardowe kodowania brakujących wartości.

### 3 Wyniki

Najpierw dokonuję analizy ogólnej, bez podziału na klientów, którzy odeszli i którzy pozostali. W następnej części analiza zostaje dokonana na odpowiednich podzbiorach, przedstawionych w poniższym kodzie.

```
# Podział na podzbiory względem osób, które odeszły i które zostały
churn.yes <- subset(churn, subset=Churn=="Yes")
churn.no <- subset(churn, subset=Churn=="No")

# usunięcie zmiennej churn z podzbiorów
# (bo w każdym podzbiorze odpowiednio wartość ta jest taka sama)
churn.yes <- churn.yes[,1:19]
churn.no <- churn.no[,1:19]
```

### 3.1 Zmienne liczbowe – analiza ogólna

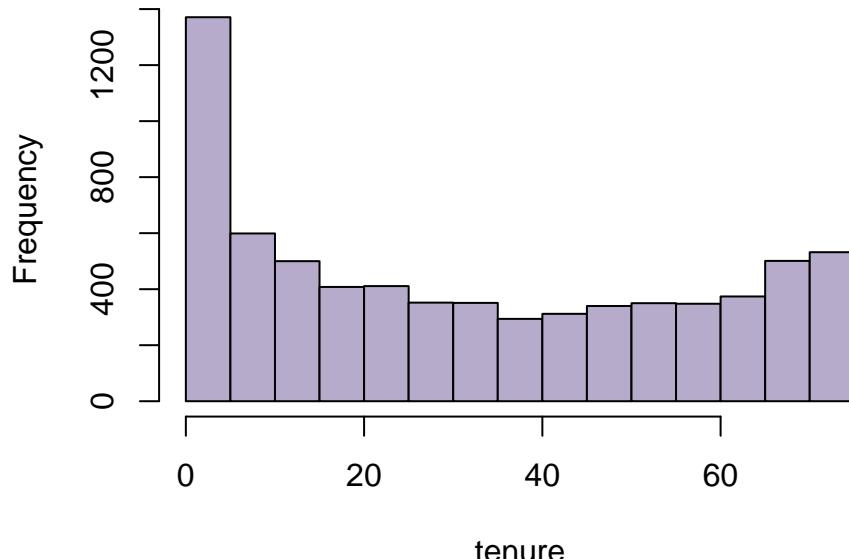
W analizowanych danych występują 3 zmienne liczbowe, dla których wyznaczamy podstawowe wskaźniki sumaryczne oraz wykresy rozkładu. Dla wszystkich korzystamy z histogramu, bo chociaż zmienna tenure jest dyskretna, możemy ją traktować jak ciągłą, ponieważ przyjmuje dużo wartości.

Tabela 2: Podstawowe wskaźniki sumaryczne dla zmiennych tenure, MonthlyCharges oraz TotalCharges

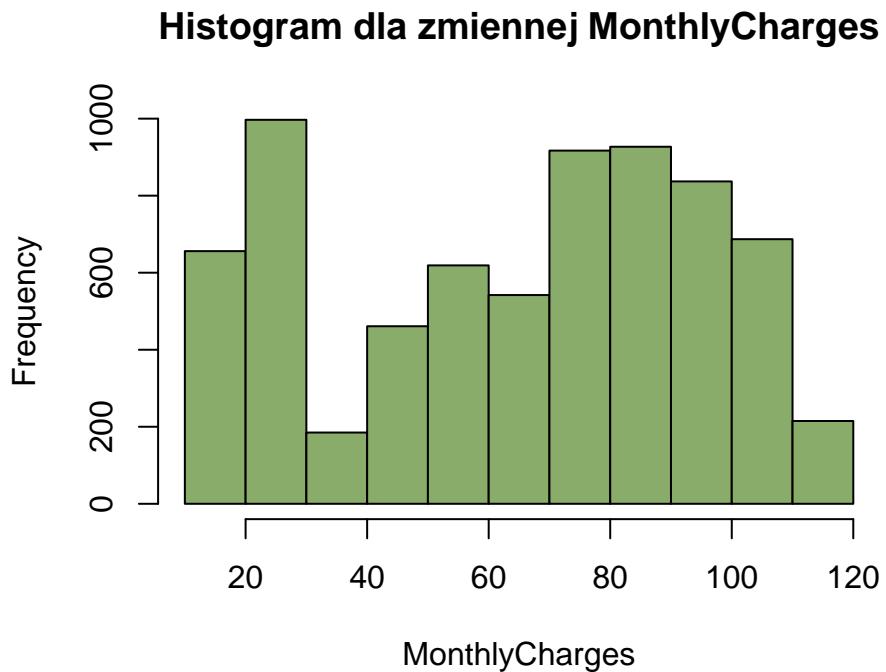
	tenure	MonthlyCharges	TotalCharges
Min	0.00	18.25	18.80
Q1	9.00	35.50	401.45
Median	29.00	70.35	1397.47
Mean	32.37	64.76	2283.30
Q3	55.00	89.85	3794.74
Max	72.00	118.75	8684.80
Var	603.17	905.41	5138252.41
SD	24.56	30.09	2266.77
R	72.00	100.50	8666.00
IQR	46.00	54.35	3393.29
Skewness	0.24	-0.22	0.96

Poniżej znajdują się również histogramy dla poszczególnych zmiennych.

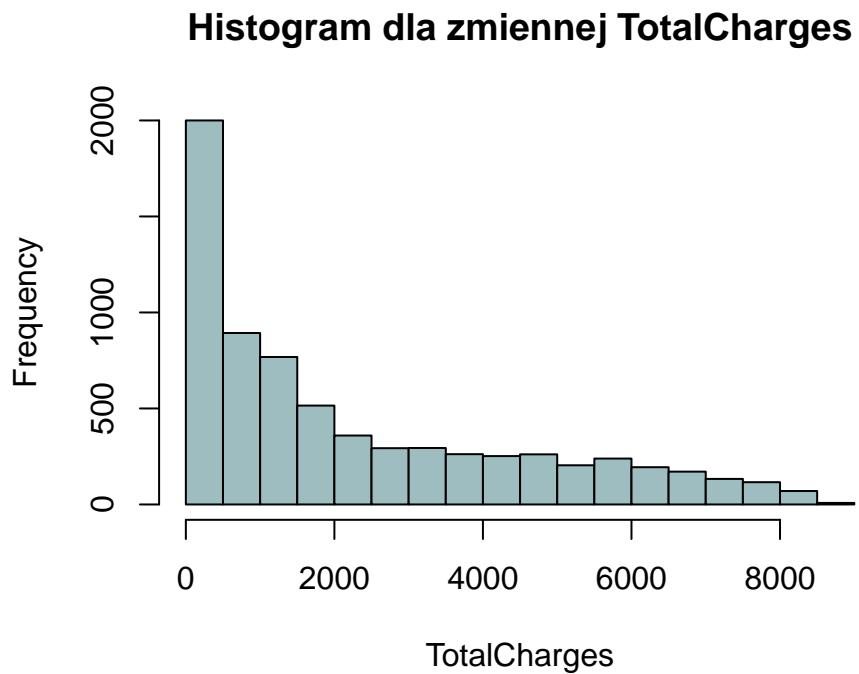
#### Histogram dla zmiennej tenure



Rysunek 1: Histogram dla zmiennej tenure



Rysunek 2: Histogram dla zmiennej MonthlyCharges



Rysunek 3: Histogram dla zmiennej TotalCharges

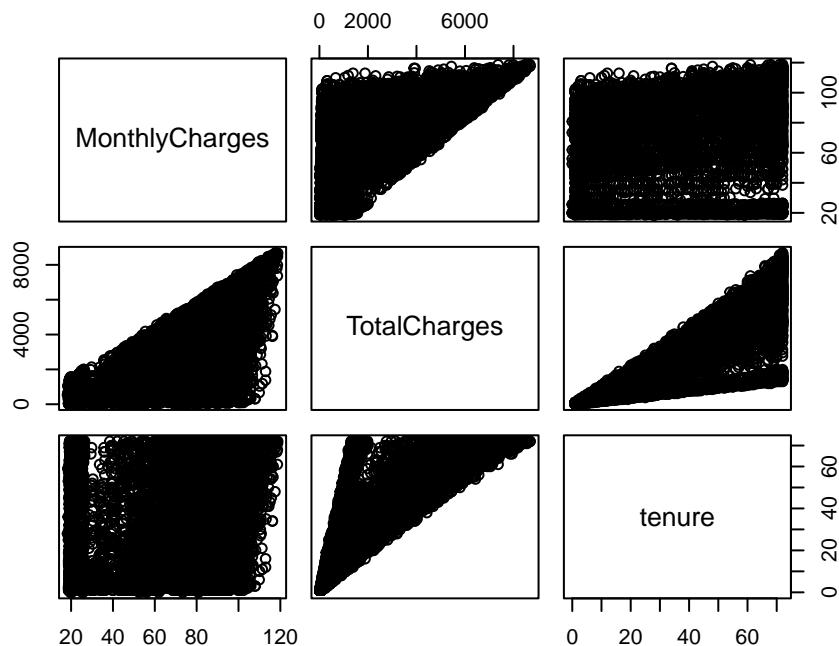
Na postawie wykresów oraz wskaźników można powiedzieć kilka rzeczy o tych zmiennych. W przypadku zmiennej tenure (wykres 1) można łatwo zauważyc, że najwięcej klientów zostaje tylko przez kilka pierwszych miesięcy – aż 25% zostaje jedynie przez 9 miesięcy! Połowa

klientów natomiast zostaje przez około 2,5 roku (dokładnie przez 29 miesięcy). To prowadzi do wniosku, że firma powinna skupić się na zachęcaniu przede wszystkim nowych klientów.

Zmienna TotalCharges (wykres 3) ma wykres nieco podobny do wykresu tenure, charakteryzujący się podobnymi cechami – połowa klientów zapłaciła łącznie nie tak wiele w porównaniu do maksymalnej możliwej kwoty.

Natomiast w przypadku zmiennej MonthlyCharges (wykres 2) można dostrzec, że jej rozkład jest wielomodalny – największą częstotliwością występowania wartości zauważam w przedziałach 20–30 oraz 70–90. Ponadto więcej wartości skupia się raczej wokół tej drugiej mody – potwierdza to miara skośności, która jest ujemna i wskazuje raczej na dłuższy lewy ogon.

### 3.1.1 Zależności pomiędzy zmiennymi liczbowymi



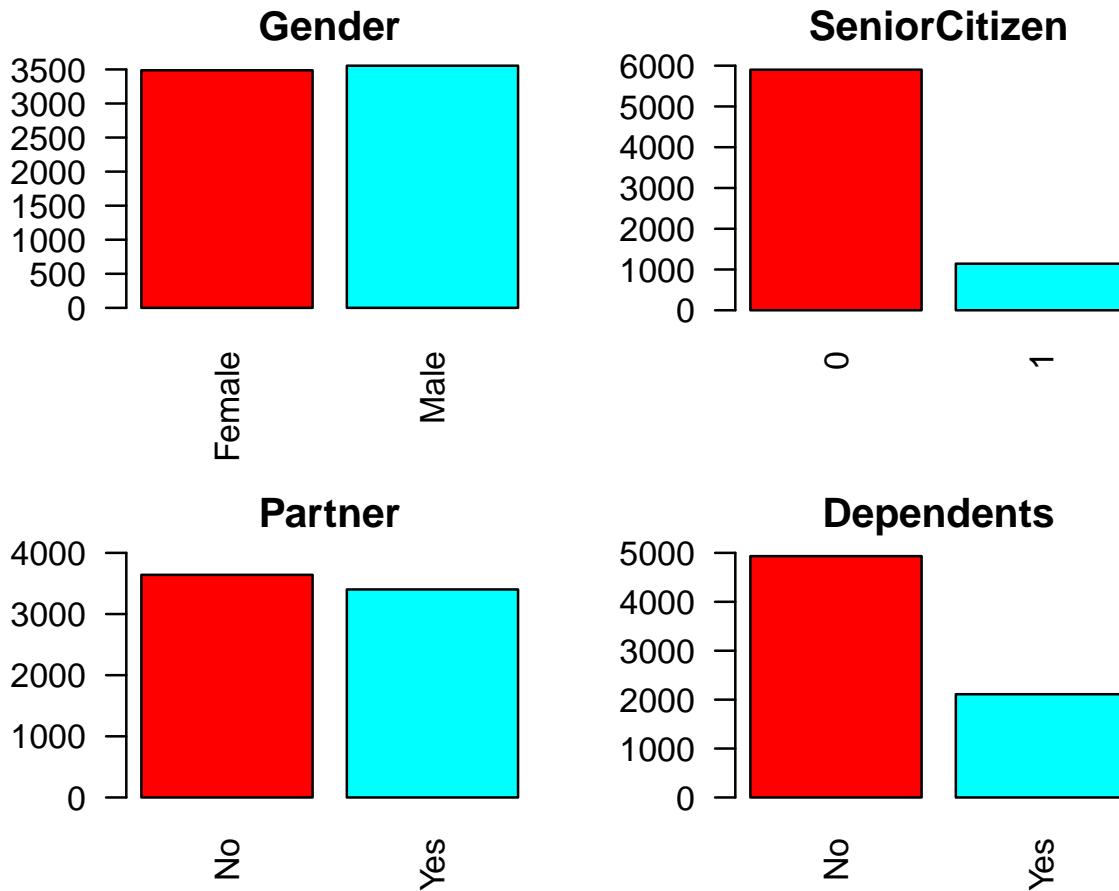
Rysunek 4: Wykresy rozrzutu dla zmiennych liczbowych

Współczynniki korelacji pomiędzy poszczególnymi zmiennymi prezentują się następująco:

- tenure i MonthlyCharges: 0.2478999,
- tenure i TotalCharges: 0.8258805,
- MonthlyCharges i TotalCharges: 0.6510648.

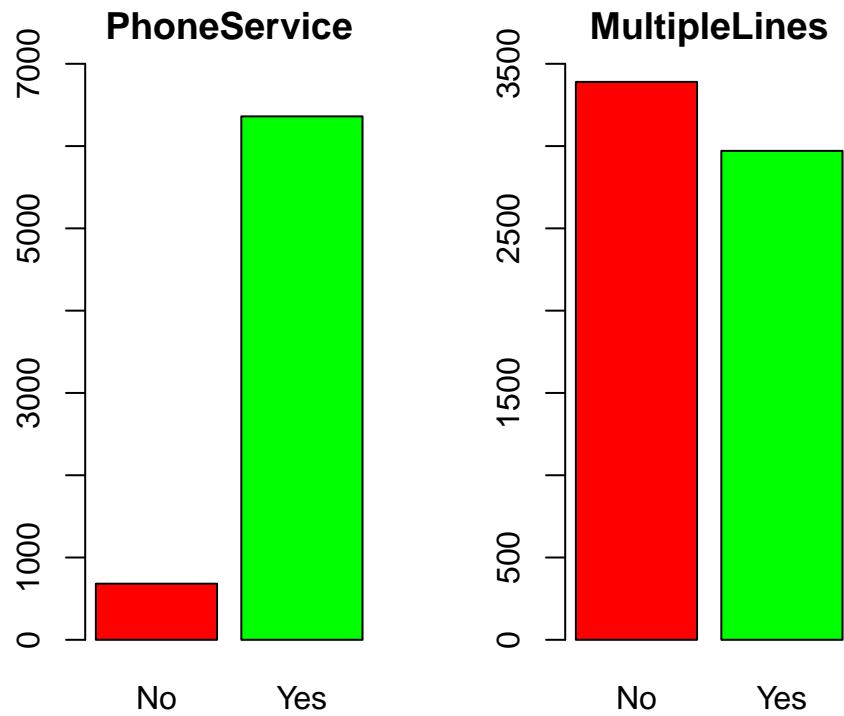
Największa korelacja występuje pomiędzy zmiennymi tenure i TotalCharges – jak można się spodziewać, łączne opłaty będą powiązane z czasem korzystania z usług firmy. Słabsza korelacja występuje pomiędzy opłatami miesięcznymi i łącznymi, natomiast praktycznie nie ma związku pomiędzy czasem korzystania z firmy a miesięcznymi opłatami.

### 3.2 Zmienne jakościowe – analiza ogólna



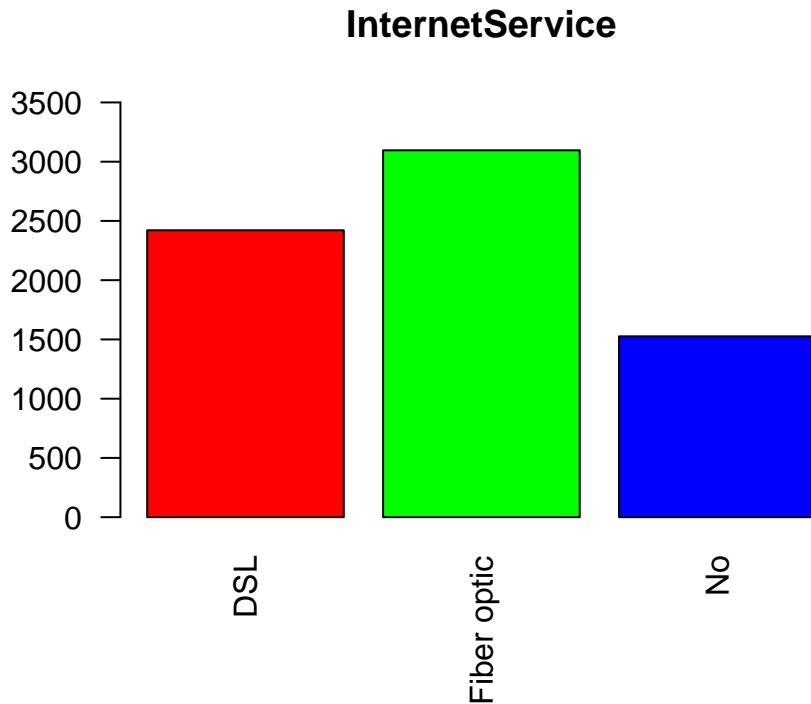
Rysunek 5: Wykresy słupkowe przedstawiające podstawowe informacje o klientach

Z podstawowych informacji o klientach (rysunek 5) można dowiedzieć się, że są to głównie osoby niebędące seniorami i nie mające nikogo na swoim utrzymaniu. Odrobinę więcej osób nie ma partnera niż ma, jednak nie jest to ogromna różnica. Nie ma dominacji którejś z płci.

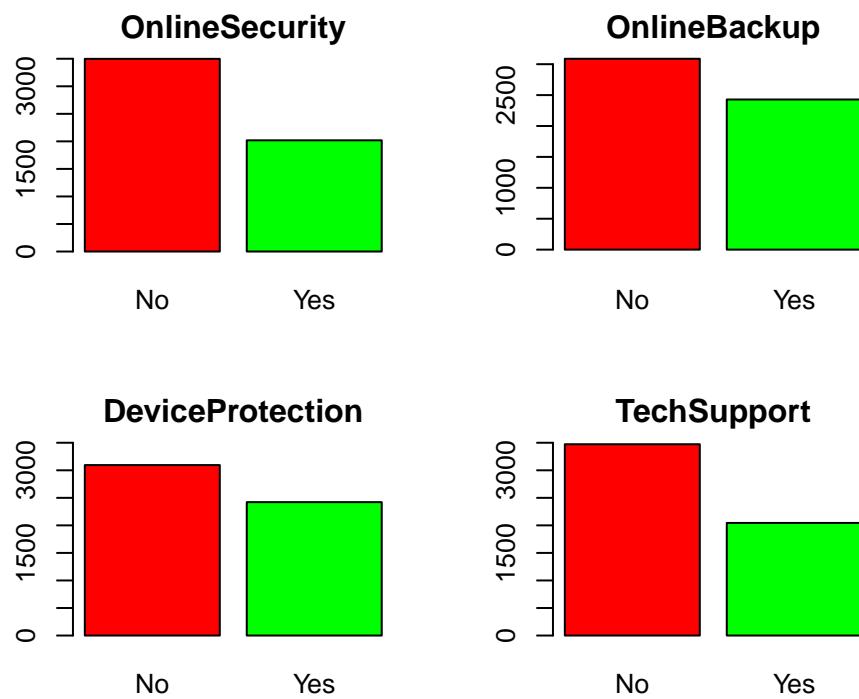


Rysunek 6: Wykresy słupkowe przedstawiające informacje o usługach telefonicznych

Rysunek 6 pokazuje, że zdecydowana większość klientów korzysta z usług telefonicznych, a wśród użytkowników tych usług większa część korzysta z jednej linii, lecz nie w przytaczającej większości, ponieważ całkiem sporo osób korzysta z wielu linii.



Rysunek 7: Wykres słupkowy przedstawiający informacje ogólne o usłudze internetowej

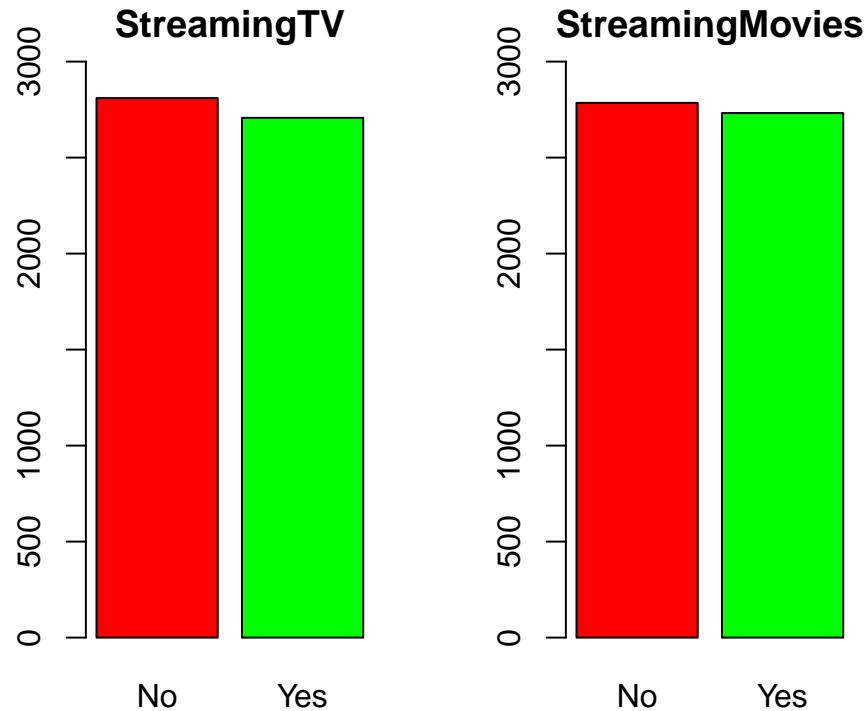


Rysunek 8: Wykresy słupkowe przedstawiające informacje o dodatkowych usługach internetowych

Jeśli chodzi o usługi internetowe, najwięcej klientów korzysta ze światłowodu (wykres 7),

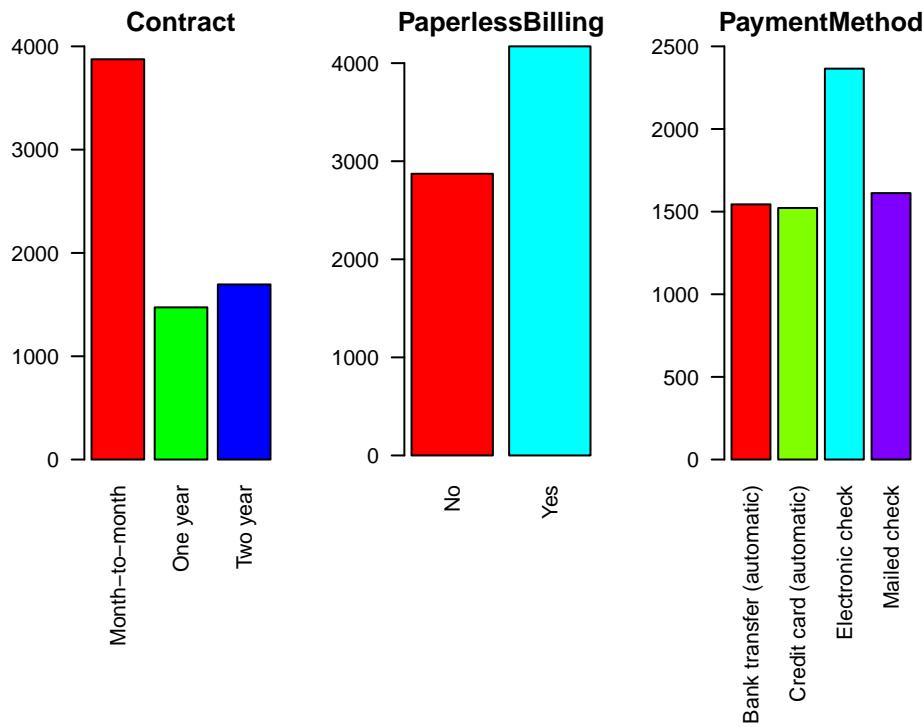
natomiaszt w drugiej kolejności plasuje się skrótka miedziana. Mniejsza część klientów nie korzysta z usług internetowych, jednak liczba ta jest większa niż w przypadku osób nie korzystających z usług telefonicznych.

Na rysunku 8 da się łatwo dostrzec, że osoby korzystające z usług internetowych w większości nie korzystają z dodatkowych związkanych usług, w szczególności najmniej zainteresowaniem cieszy się usługa bezpieczeństwa online.



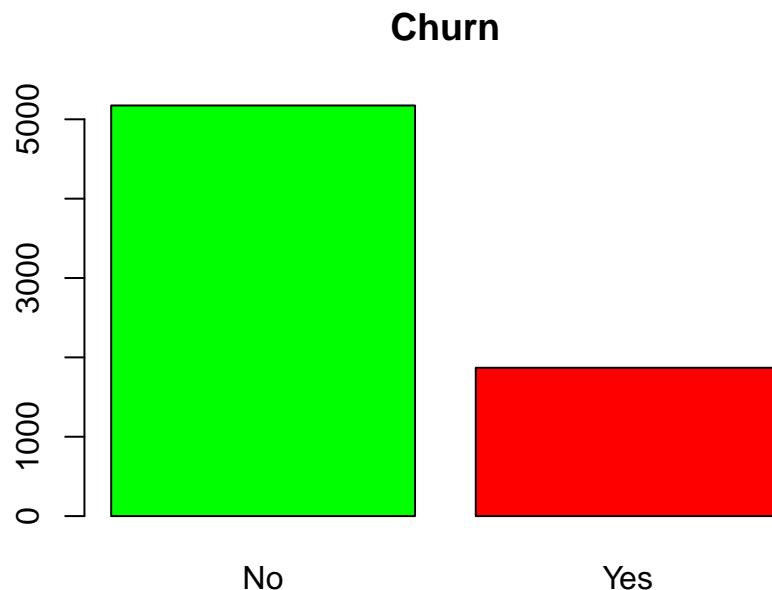
Rysunek 9: Wykresy słupkowe przedstawiające informacje o usługach streamingowych

Jeśli chodzi o usługi streamingowe, wyniki wśród osób z aktywną usługą internetową są raczej wyrównane (rysunek 9).



Rysunek 10: Wykresy słupkowe przedstawiające informacje o płatnościach

Z rysunku 10 można odczytać, że zdecydowana większość klientów podpisała miesięczne umowy, natomiast wśród pozostałych nieznacznie większym zainteresowaniem cieszą się dwuletnie umowy względem rocznych. Większość klientów korzysta także z elektronicznych rachunków. Jeśli chodzi o metody płatności, największą popularnością cieszy się elektroniczny czek, natomiast wśród pozostałych metod płatności żadna się szczególnie nie wyróżnia.



Rysunek 11: Wykres słupkowy przedstawiający informację o odejściach

Większość klientów, jak się okazuje, pozostaje lojalna wobec firmy (wykres 11).

### 3.3 Zmienne liczbowe – analiza z podziałem

Analizy dokonuję w taki sam sposób jak wcześniej, tyle że z podziałem na podzbiory.

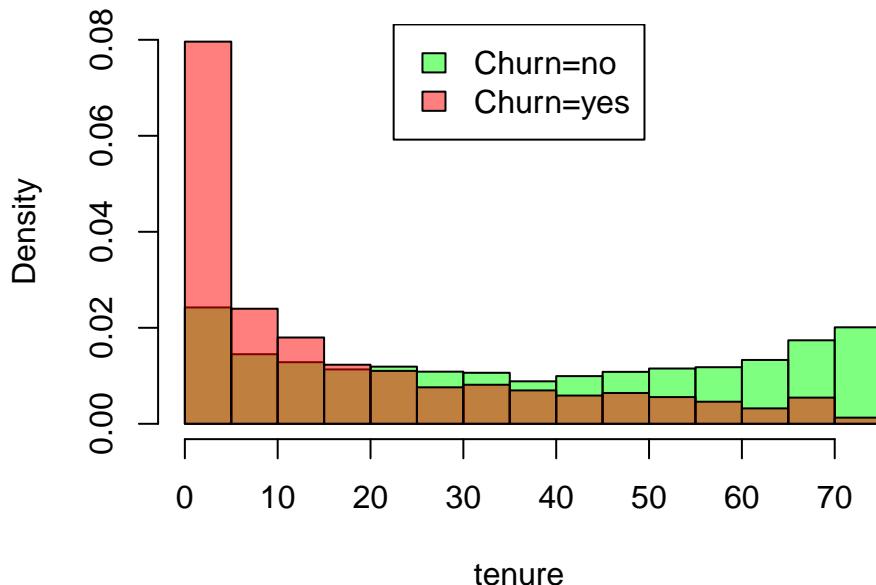
Najpierw wyznaczę wskaźniki sumaryczne.

Tabela 3: Podstawowe wskaźniki sumaryczne dla zmiennych tenure, MonthlyCharges (w tabeli MonthlyC) oraz TotalCharges (w tabeli TotalC) z podziałem względem parametru Churn

	tenure.no	tenure.yes	MonthlyC.no	MonthlyC.yes	TotalC.no	TotalC.yes
Min	0.00	0.00	18.25	18.85	18.80	18.85
Q1	15.00	15.00	25.10	56.15	577.82	134.50
Median	38.00	38.00	64.43	79.65	1683.60	703.55
Mean	37.57	37.57	61.27	74.44	2555.34	1531.80
Q3	61.00	61.00	88.40	94.20	4264.12	2331.30
Max	72.00	72.00	118.75	118.35	8672.45	8684.80
Var	581.47	581.47	966.75	608.41	5426369.84	3575211.60
SD	24.11	24.11	31.09	24.67	2329.46	1890.82
R	72.00	72.00	100.50	99.50	8653.65	8665.95
IQR	46.00	46.00	63.30	38.05	3686.30	2196.80
Skewness	-0.03	-0.03	-0.03	-0.73	0.81	1.51

W następnej kolejności przeanalizuję wykresy.

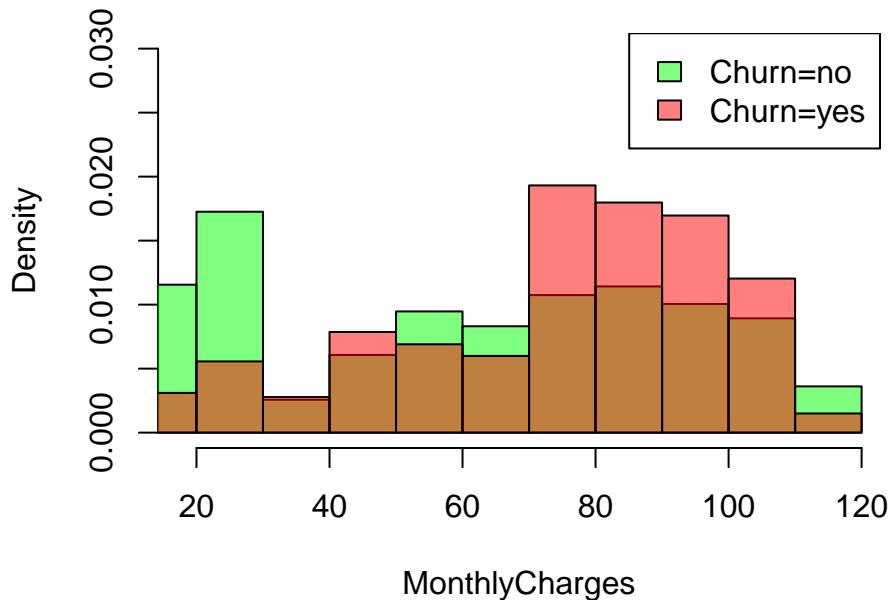
### Histogram tenure wzgledem churn



Rysunek 12: Histogram tenure z podziałem na klientów, którzy zostali i którzy odeszli

Z wykresu 12 można dowiedzieć się, że osoby, które odchodzą, najczęściej zostają krócej – odchodzą w pierwszych miesiącach.

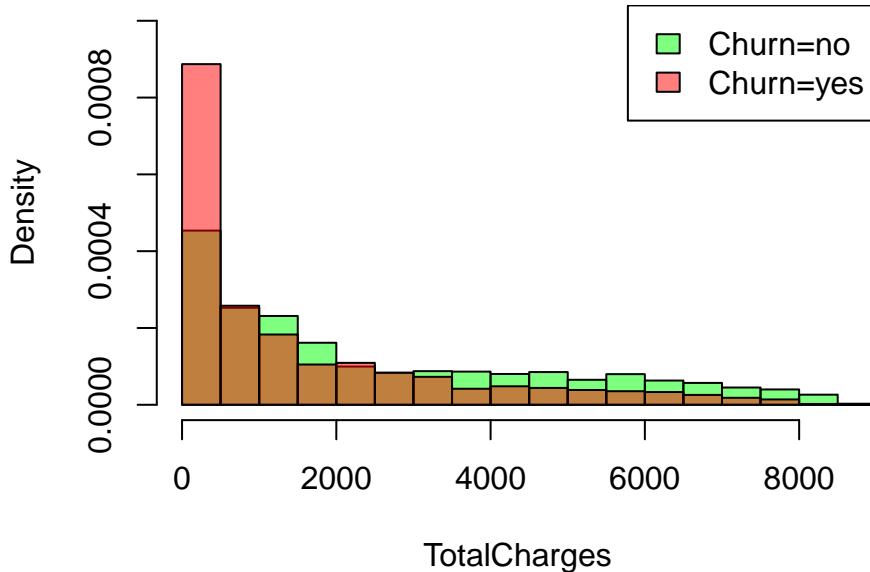
### Histogram MonthlyCharges wzgledem churn



Rysunek 13: Histogram MonthlyCharges z podziałem na klientów, którzy zostali i którzy odeszli

Wykres 13 pokazuje, że na ogół klienci, którzy odeszli, mieli wyższe opłaty od tych, którzy zostali, choć, co ciekawe, wśród osób, które mają najwyższe ze wszystkich opłat, więcej osób nadal jest lojalnych wobec firmy.

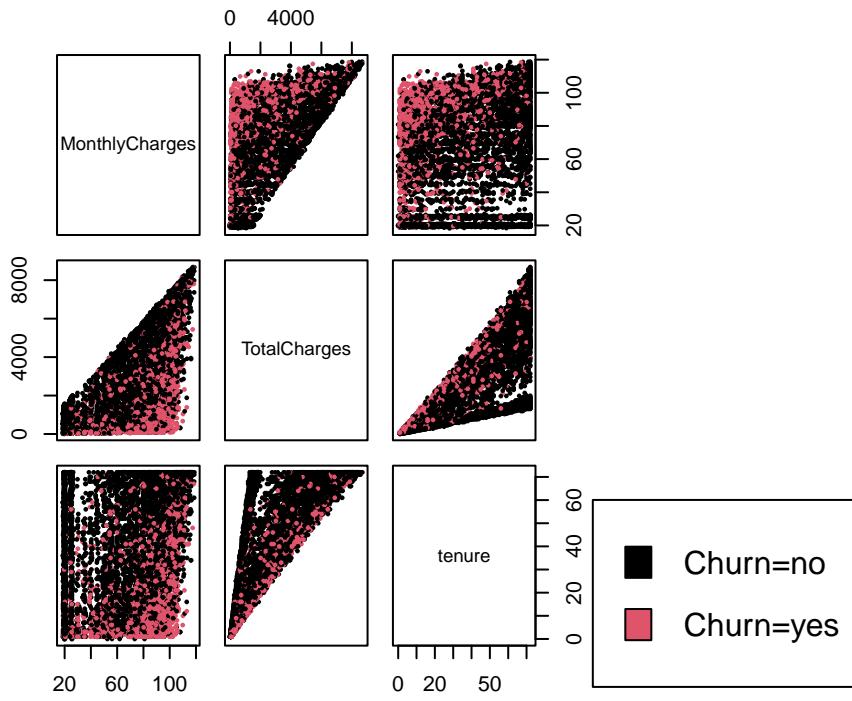
### Histogram TotalCharges względem churn



Rysunek 14: Histogram TotalCharges z podziałem na klientów, którzy zostali i którzy odeszli

W przypadku łącznych opłat osoby, które odeszły, częściej wpadały w zakres najniższych opłat, natomiast dla wyższych opłat wyniki dla klientów lojalnych i nielojalnych prezentują się podobnie (wykres 14).

Teraz sprawdzę jeszcze wykresy rozrzutu i zależności dla zmiennych liczbowych z podziałem na klientów, którzy odeszli i którzy pozostali.



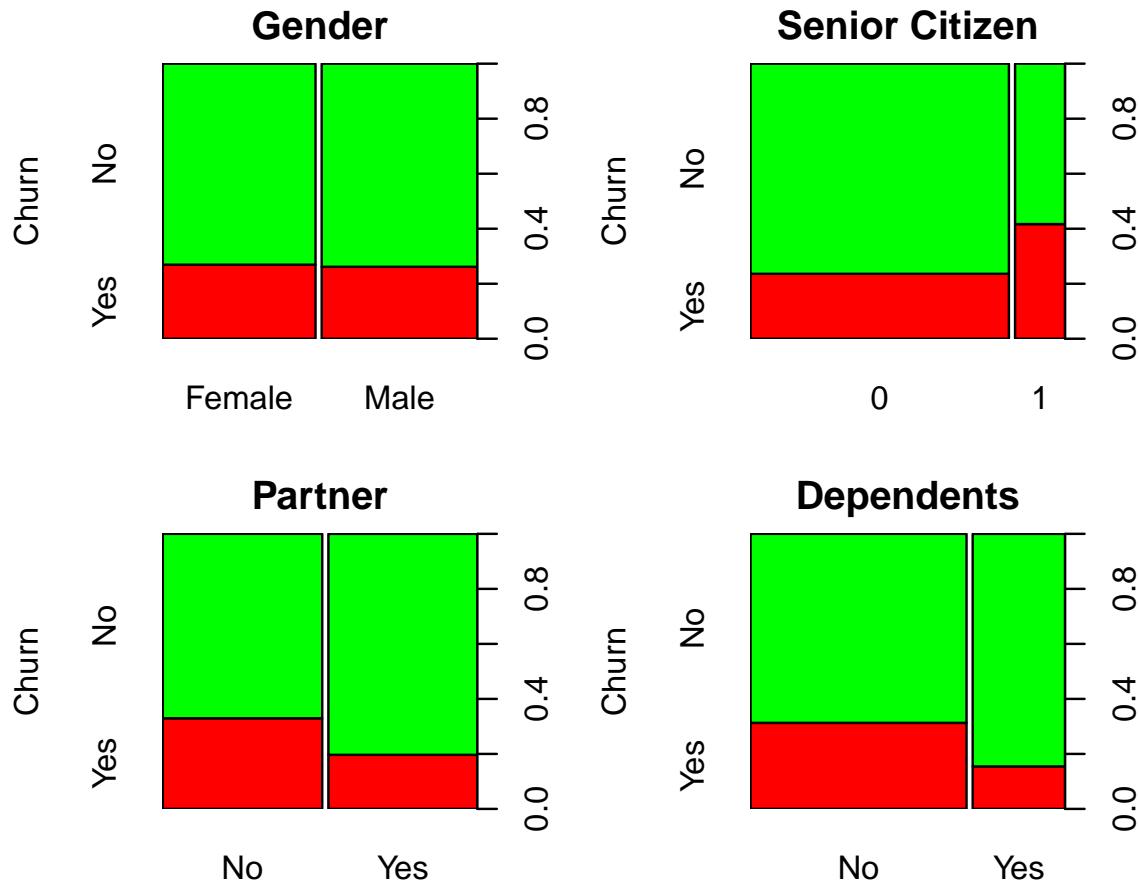
Rysunek 15: Wykresy rozrzutu dla zmiennych liczbowych z podziałem na klientów, którzy zostali i którzy odeszli

Tabela 4: Współczynniki korelacji dla zmiennych liczbowych z uwzględnieniem parametru churn

	Churn=no	Churn=yes
tenure i MonthlyCharges	0.33	0.40
tenure i TotalCharges	0.79	0.95
MonthlyCharges i TotalCharges	0.76	0.55

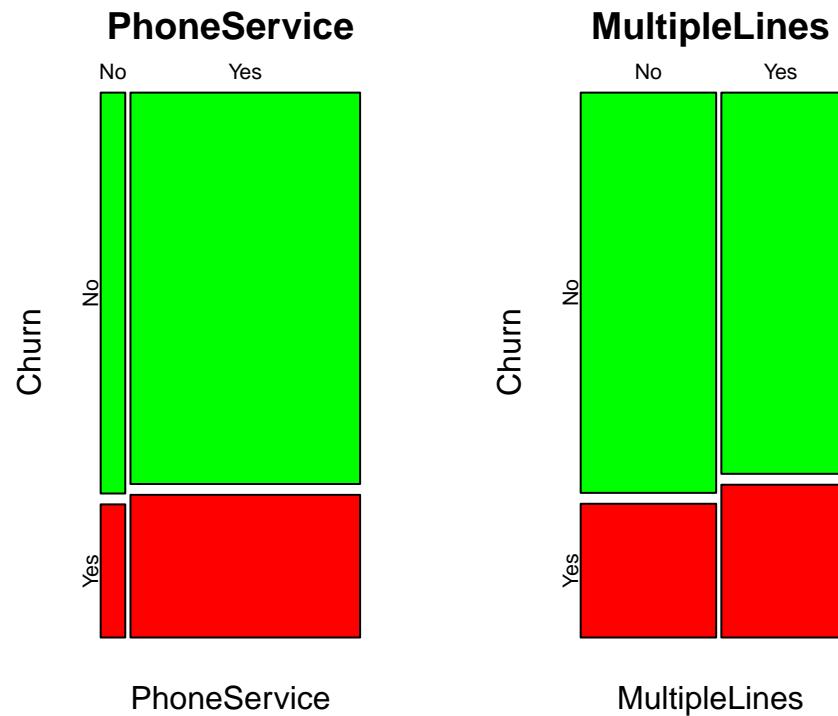
Największe powiązanie można zauważyć dla zmiennych tenure i TotalCharges w przypadku klientów, którzy odeszli (rysunek 15, tabela 4).

### 3.4 Zmienne jakościowe – analiza z podziałem



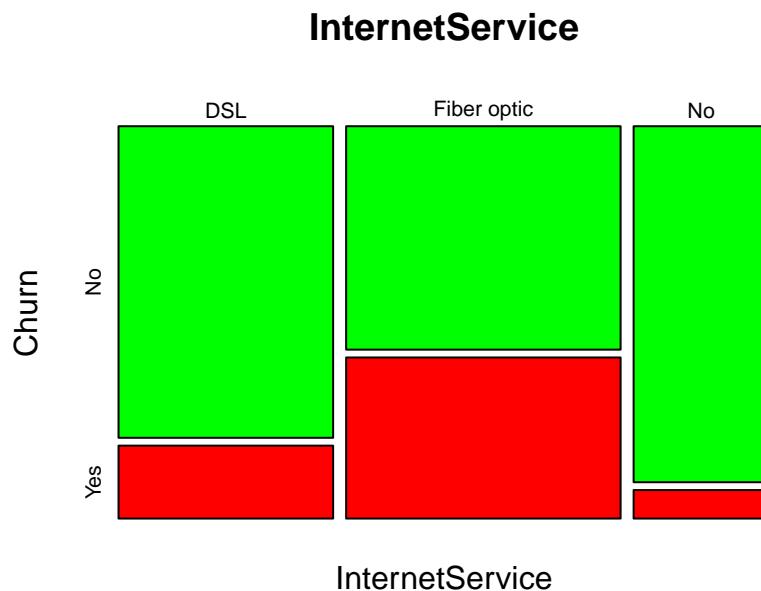
Rysunek 16: Wykresy mozaikowe przedstawiające podstawowe informacje o klientach z podziałem na tych, którzy zostali i którzy odeszli

Na podstawie wykresów dotyczących ogólnych danych o klientach z rysunku 16 nie ma związku pomiędzy płcią a decyzją o korzystaniu z usług. Można natomiast zauważyc, że nieco częściej odchodzą seniorzy, osoby samotne oraz takie, które nie mają utrzymanków.

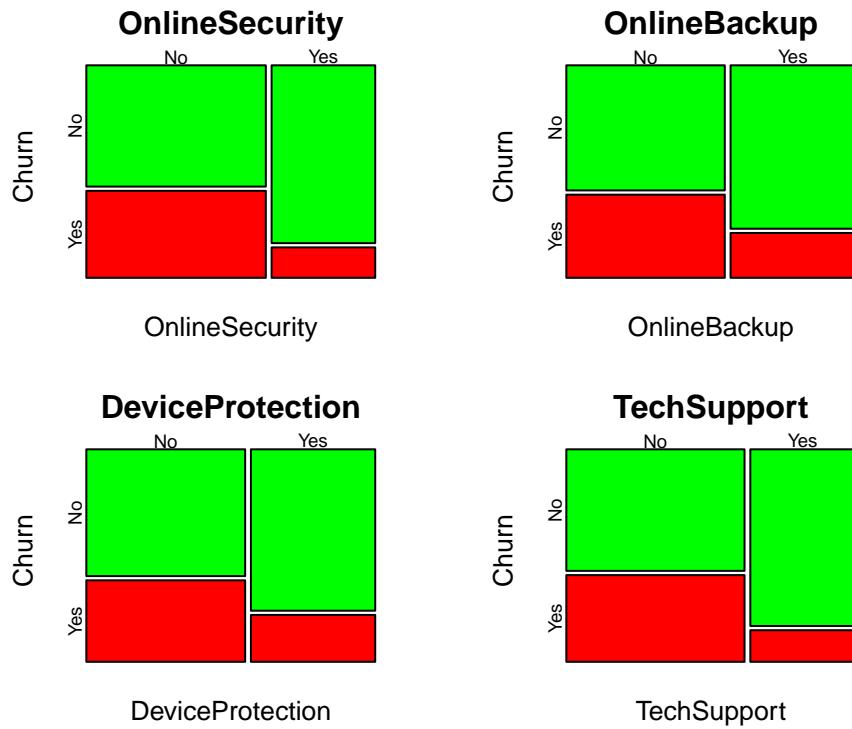


Rysunek 17: Wykresy mozaikowe przedstawiające informacje o usługach telefonicznych z podziałem na tych, którzy zostali i którzy odeszli

Rysunek 17 pokazuje, że usługi telefoniczne nie mają istotnego wpływu na to, czy klienci odchodzą, czy nie.

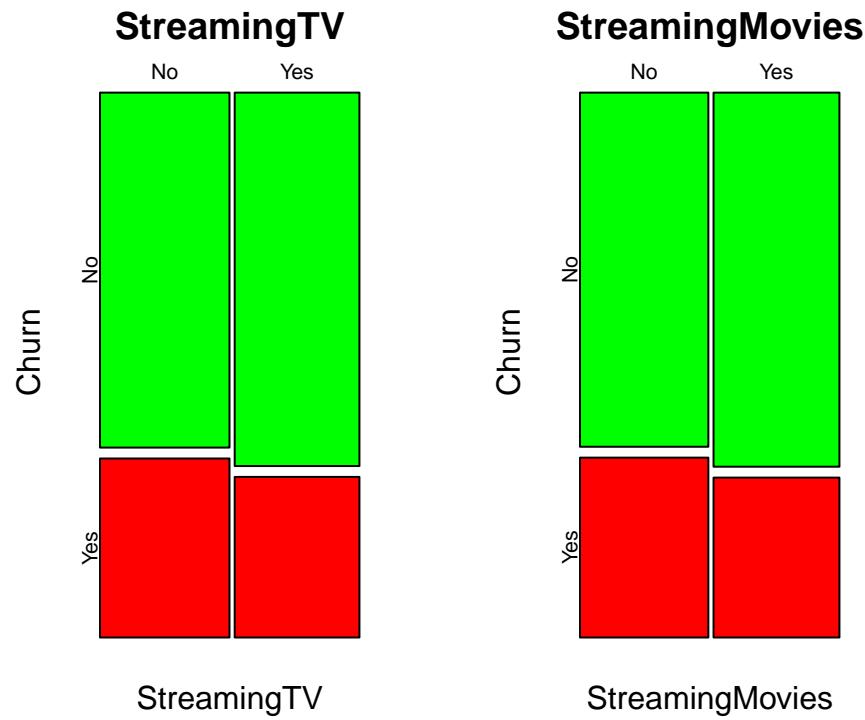


Rysunek 18: Wykres mozaikowy przedstawiający informacje o usługach internetowych z podziałem na tych, którzy zostali i którzy odeszli



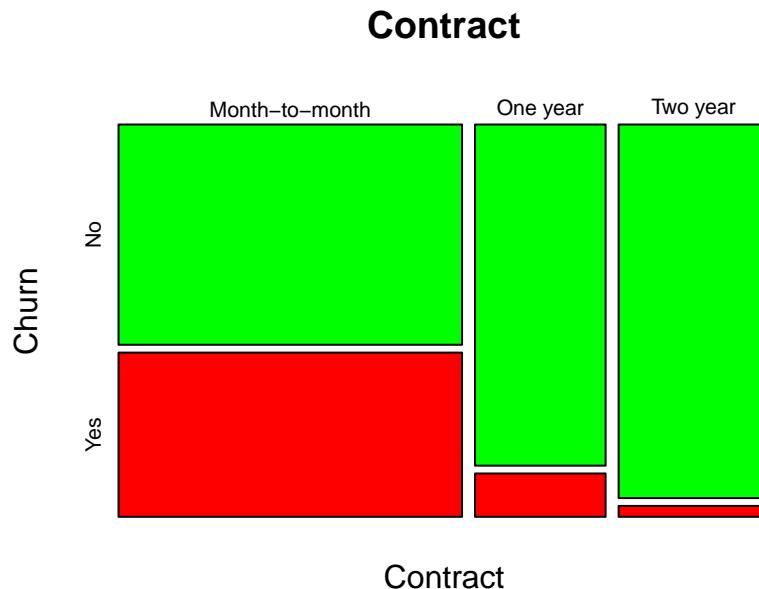
Rysunek 19: Wykresy mozaikowe przedstawiające informacje o dodatkowych usługach internetowych z podziałem na tych, którzy zostali i którzy odeszli

Użytkownicy korzystający z Internetu odchodzą częściej niż ci, którzy w ogóle nie korzystają z tej usługi. Najczęściej rezygnują użytkownicy światłowodu (wykres 18). Z rysunku 19 można wywnioskować natomiast, że korzystanie z dodatkowych usług związanych z Internetem jednak nieco zmniejsza prawdopodobieństwo odejścia – najbardziej skuteczne w tej kwestii są usługi bezpieczeństwa online i wsparcia technicznego, mniejsze znaczenie mają backup i ochrona urządzenia.

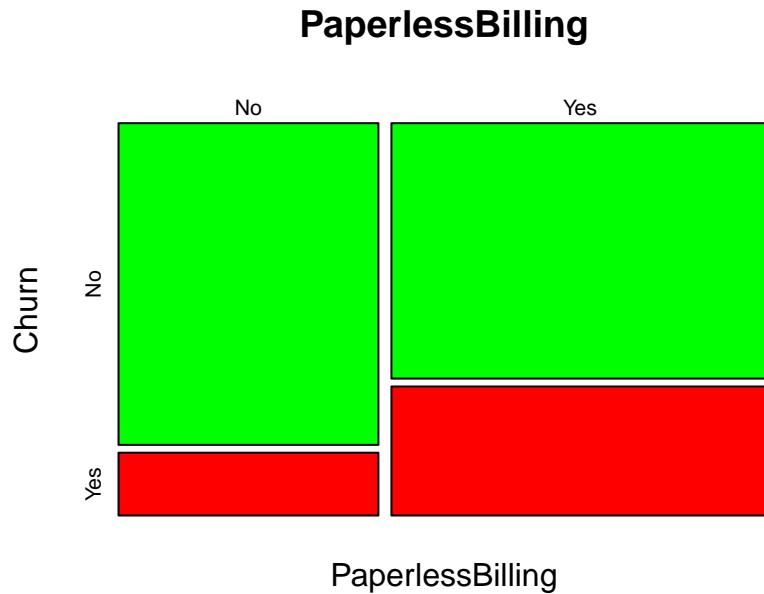


Rysunek 20: Wykresy mozaikowe przedstawiające informacje o usługach streamingowych z podziałem na tych, którzy zostali i którzy odeszli

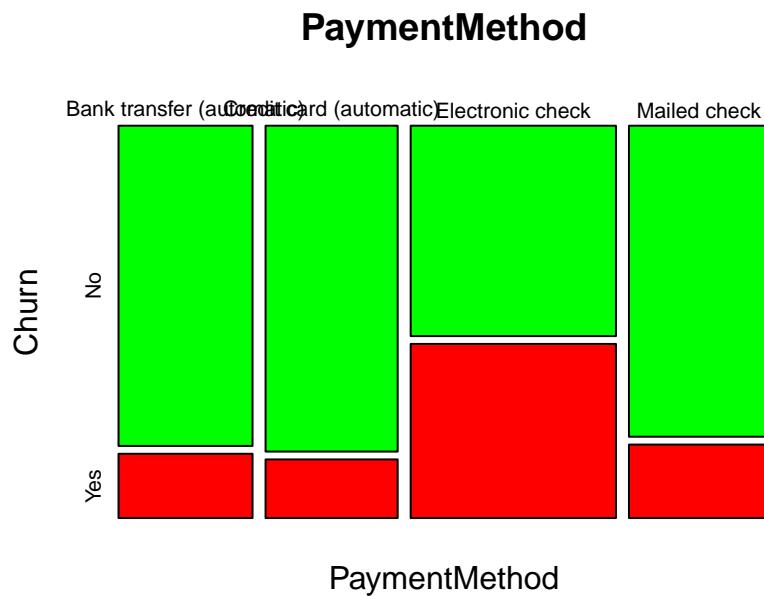
Wśród osób korzystających z Internetu usługi streamingowe nie zdają się zbytnio zmieniać sytuacji – odrobinę zmniejszają szansę odejścia, jednak różnica pomiędzy odejściami osób mających te usługi a nie mających jest niewielka (rysunek 20).



Rysunek 21: Wykres mozaikowy przedstawiający informacje o typie umowy z podziałem na tych, którzy zostali i którzy odeszli



Rysunek 22: Wykres mozaikowy przedstawiający informacje o rachunkach z podziałem na tych, którzy zostali i którzy odeszli



Rysunek 23: Wykres mozaikowy przedstawiający informacje o metodzie płatności z podziałem na tych, którzy zostali i którzy odeszli

Wykresy 21, 22 oraz 23 pokazują, że detale umowy i płatności mają spore znaczenie. Najczęściej odchodzą osoby o umowach miesięcznych – dłuższe umowy zmniejszają ryzyko odejścia. Jesli chodzi o rozliczanie płatności, można zauważyć, że częściej odchodzą osoby, które rozliczają się elektronicznie, a także te, które płacą przy pomocy elektronicznego czeku. Inne metody płatności nie różnią się szczególnie pomiędzy sobą.

### 3.5 Podsumowanie

Dzięki powyższej analizie można zauważać kilka rzeczy.

1. Najwięcej klientów odchodzi w pierwszych miesiącach od rozpoczęcia korzystania z usług firmy.
2. Częściej odchodzą osoby mające wyższe miesięczne opłaty.
3. Nieco częściej odchodzą seniorzy, osoby samotne i nie posiadające utrzymanków.
4. Częściej odchodzą użytkownicy usług internetowych, zwłaszcza klienci korzystający ze światłowodu (a warto zauważać, że najwięcej klientów w ogóle korzysta ze światłowodu).
5. Dodatkowe usługi związane z Internetem nieco zmniejszają prawdopodobieństwo odejścia.
6. Najczęściej odchodzą osoby mające miesięczne umowy, odchodzi znacznie mniej osób z rocznymi umowami, a najmniej odchodzi tych, którzy mają umowy dwuletnie. Warto zauważać, że najwięcej osób decyduje się na miesięczne umowy!
7. Częściej odchodzą osoby bez papierowych rachunków. Jednocześnie większość klientów woli rozliczać się bez papieru.
8. Najczęściej odchodzą osoby korzystające z e-mailowego czeku – czyli metody używanej przez największą liczbę klientów.

Aby zapobiec odchodzeniu klientów, moim zdaniem firma może podjąć następujące działania:

- bardziej dbać o nowych klientów, a także proponować im dłuższe umowy,
- jeśli to możliwe, zmniejszyć nieco miesięczne opłaty,
- poprawić jakość internetowych usług, zwłaszcza światłowodu, gdyż mimo jego popularności jego użytkownicy dość często odchodzą,
- przy zawieraniu umów włączających usługę internetową proponować dodatkowe powiązane usługi,
- przyjrzeć się elektronicznemu rozliczaniu umów, a także e-mailowym czekom.