Lista 1

Eksploracja danych

Igor Misterowicz, 282245

2025-03-31

Contents

1 Etap 1

1.1 a.

Wczytuję dane oraz upewniam się, że typy są prawidłowo rozpoznane.

```
setwd('C:/Users/igorm/Programowanie/data_mining/files')
data <- read.csv("WA_Fn-UseC_-Telco-Customer-Churn.csv",stringsAsFactors = T)
data$SeniorCitizen <- sapply(data$SeniorCitizen,
                             FUN = function(x) ifelse(x == 1, "Yes", "No"))
data$SeniorCitizen <- as.factor(data$SeniorCitizen)
attach(data)
```

Poprawy wymagała jedynie kolumna "SeniorCitizen", ponieważ kodowanie danych nie było spójne z resztą tabeli.

1.2 b.

Liczba przypadków wynosi:

```
nrow(data)
```

```
## [1] 7043
```

Liczba cech wynosi:

```
ncol(data)
```

```
## [1] 21
```

Rolę identyfikatora spełnia kolumna "customerID", będzie ona nieprzydatna w analizie, ponieważ nie interesują nas poszczególni klienci, tylko wnioski odnoszące się do ogółu.

```
data <- subset(data, select = -customerID)
```

Braki danych znajdziemy jedynie w kolumnie "TotalCharges".

```
anyNA(TotalCharges)
```

```
## [1] TRUE
```

```r
anyNA(subset(data, select = c(-TotalCharges)))
```

```
## [1] FALSE
```

"No internet service" występujący w dodatkach do usług internetowych też w pewnym sensie jest brakującą daną. Mimo to, postać ta jest przydana w analizie, ponieważ wiemy jak ją interpretować.

```r
levels(OnlineSecurity)
```

```
## [1] "No"                  "No internet service" "Yes"
```

2 Etap 2

2.1 a.

W tabeli 1 znajdujemy wkaźniki sumaryczne dla wybranych zmiennych ilościowych, rozkłady zmiennych ilościowych zostaną zilustrowane na wykresach.

```r
my.summary <- function(X)
{
  result <- c(min(X),quantile(X,0.25), median(X), mean(X), quantile(X,0.75),
              max(X), var(X), sd(X), IQR(X))
  return(result)
}

result <- data.frame(tenure = my.summary(tenure),
                     MonthlyCharges = my.summary(MonthlyCharges),
                     TotalCharges = my.summary(
                     TotalCharges[!is.na(TotalCharges)]),
                     row.names = c("min", "Q1", "median", "mean", "Q3", "max",
                                   "var", "sd", "IQR"))
result <- as.matrix(result)

tab1 <- xtable(result, digits = 2, row.names = TRUE,
               caption = "wybrane wskaźniki sumaryczne dla zmiennych ciągłych",
               label = "tab:tabela1")
print(tab1, type = "latex", table.placement = "H", comment=FALSE)
```

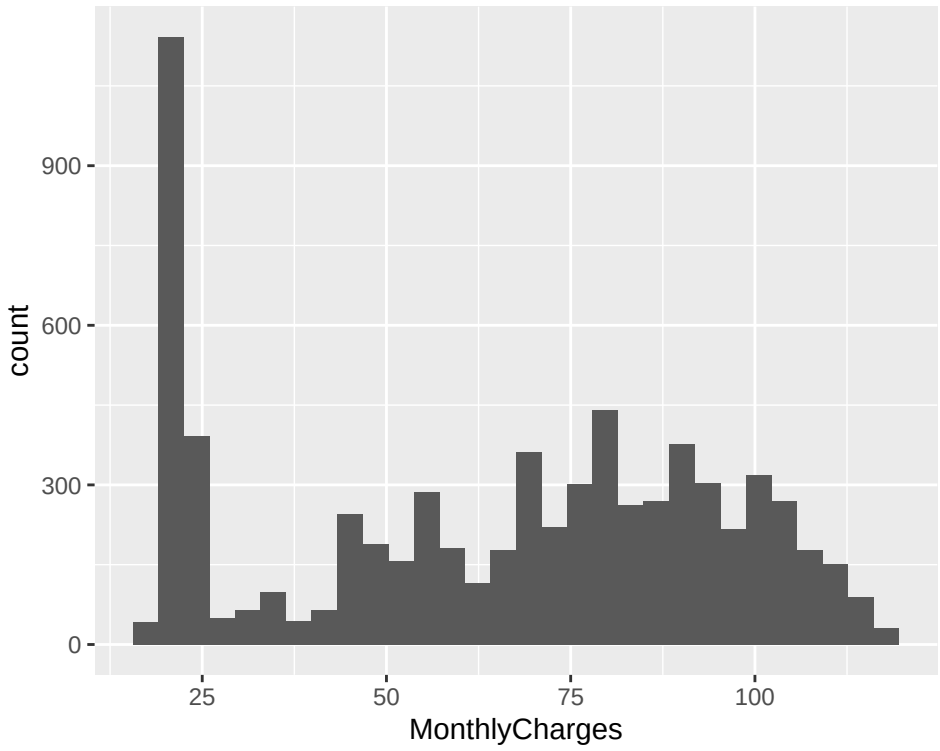
	tenure	MonthlyCharges	TotalCharges
min	0.00	18.25	18.80
Q1	9.00	35.50	401.45
median	29.00	70.35	1397.47
mean	32.37	64.76	2283.30
Q3	55.00	89.85	3794.74
max	72.00	118.75	8684.80
var	603.17	905.41	5138252.41
sd	24.56	30.09	2266.77
IQR	46.00	54.35	3393.29

Table 1: wybrane wskaźniki sumaryczne dla zmiennych ciągłych

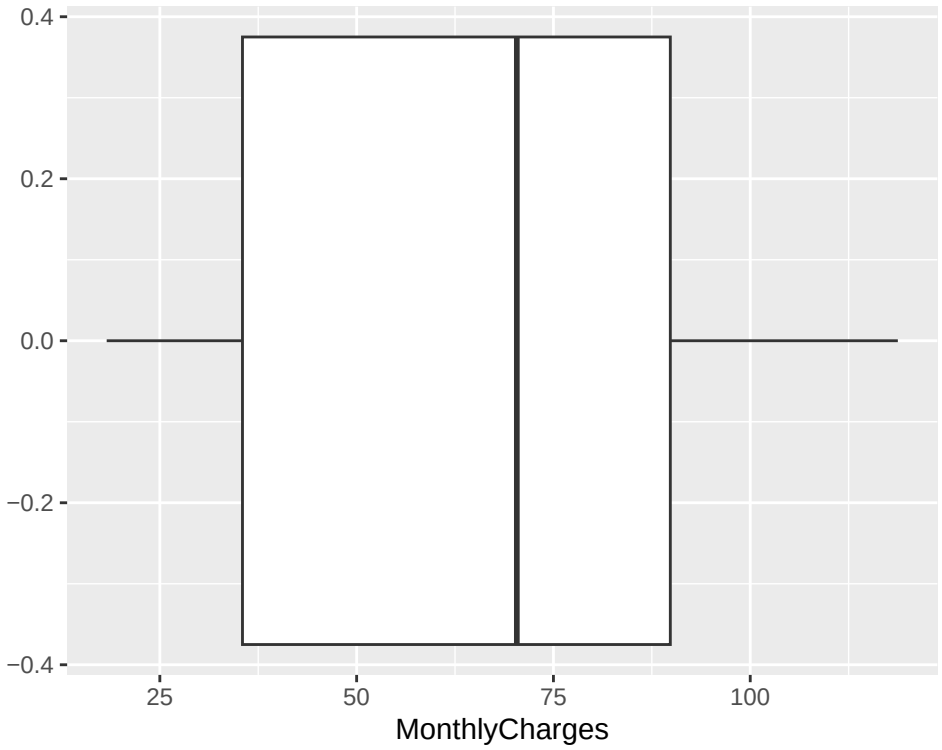
2.2 b.

Ilustrujemy teraz wykresy zmiennych ilościowych

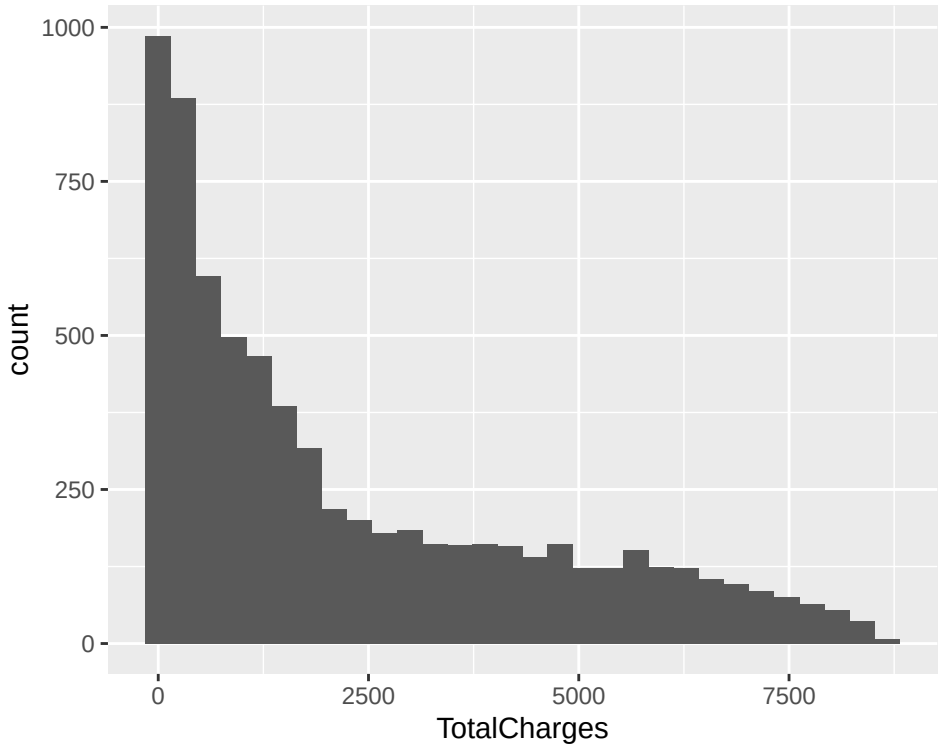
```
par(mfrow = c(1,2))
ggplot(data, aes(x = MonthlyCharges)) + geom_histogram()
```



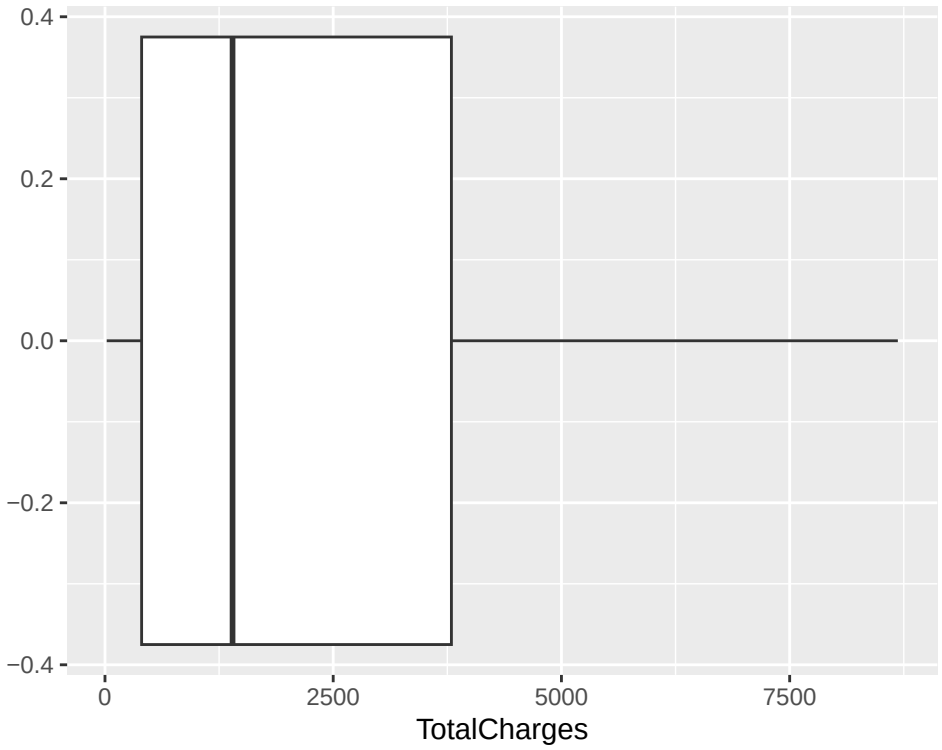

```
ggplot(data, aes(x = MonthlyCharges)) + geom_boxplot()
```



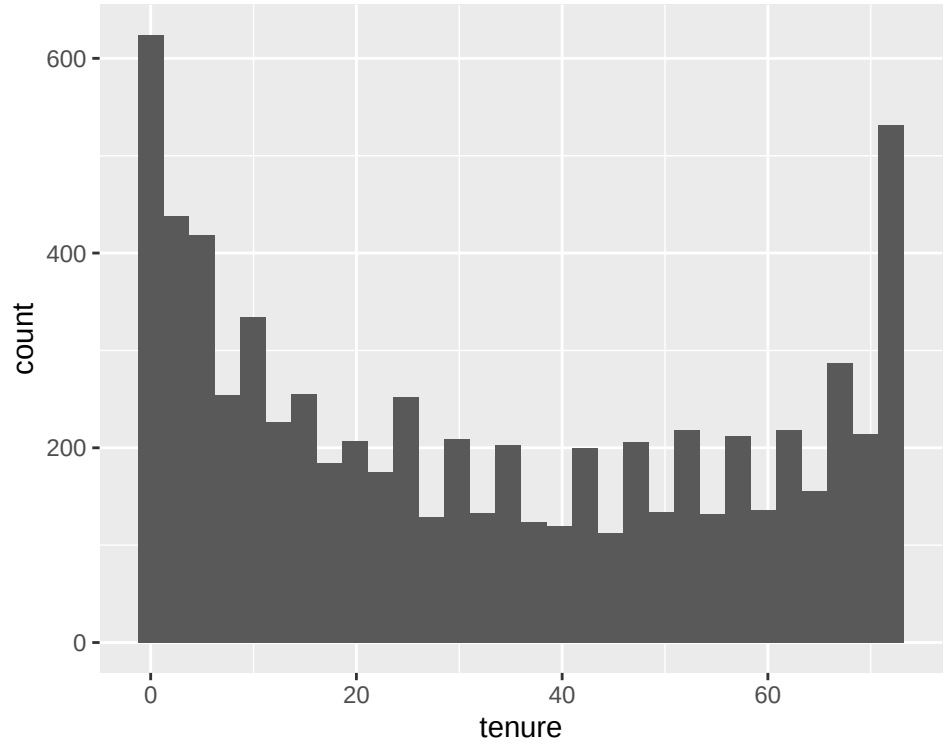

```
par(mfrow = c(1,2))
ggplot(data, aes(x = TotalCharges)) + geom_histogram()
```



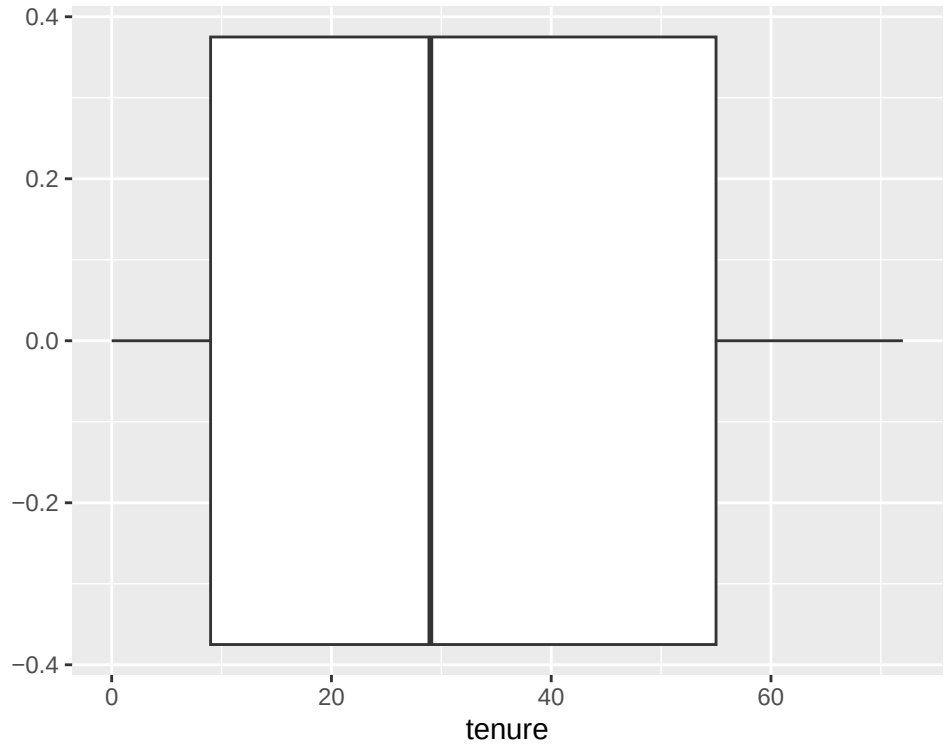
```
ggplot(data, aes(x = TotalCharges)) + geom_boxplot()
```




```
par(mfrow = c(1,2))
ggplot(data, aes(x = tenure)) + geom_histogram()
```

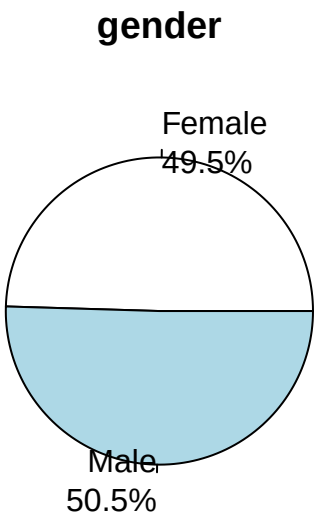


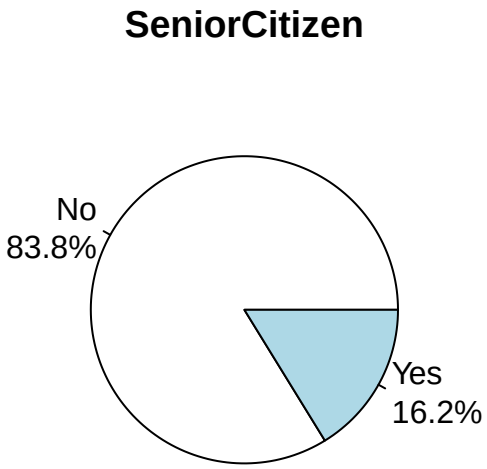

```
ggplot(data, aes(x = tenure)) + geom_boxplot()
```

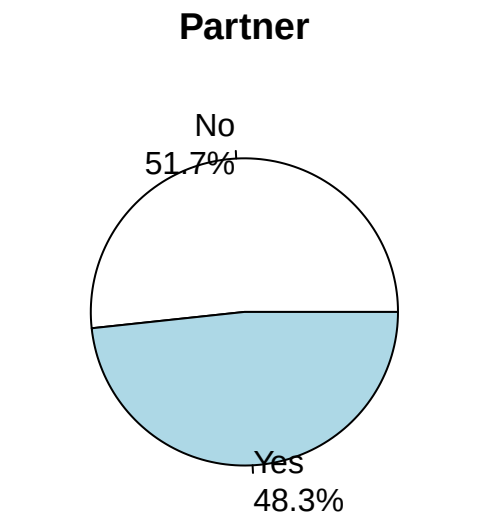


Wykresy kołowe dla zmiennych jakościowych

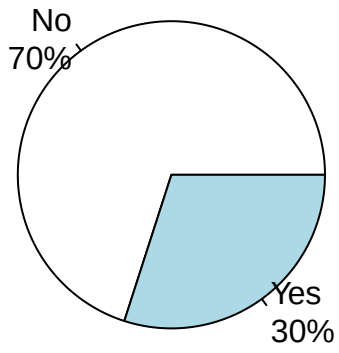
```r
for(i in 1:ncol(data)){
  if(is.factor(data[[i]])){
    counts <- table(data[[i]])
    percents <- round(100 * counts / sum(counts), 1)
    labels <- paste(names(counts), "\n", percents, "%", sep = "")
    pie(counts, labels = labels, main = colnames(data)[i])
  }
}
```



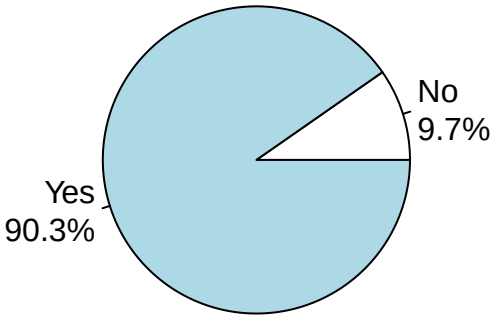

SeniorCitizen
No
83.8%
Yes
16.2%


Partner
No
51.7%
Yes
48.3%

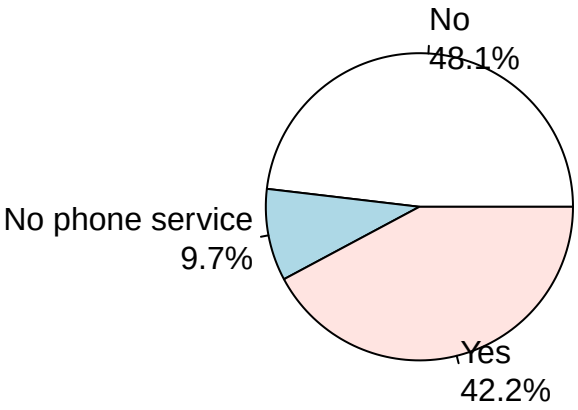
Dependents



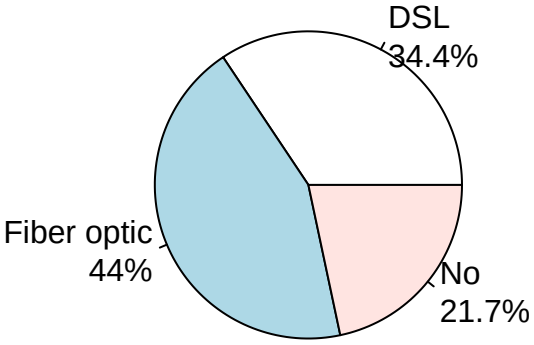
PhoneService

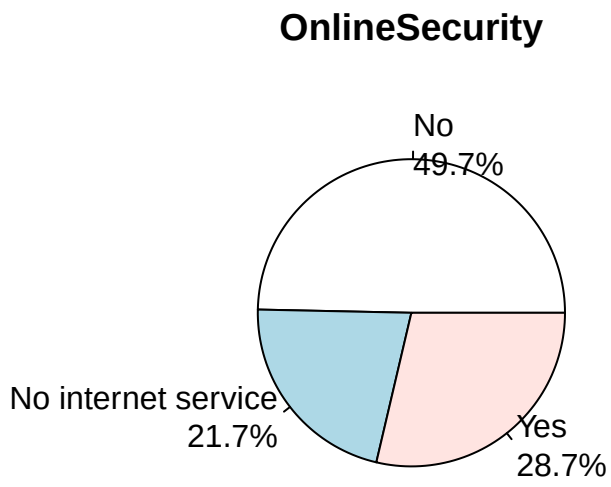


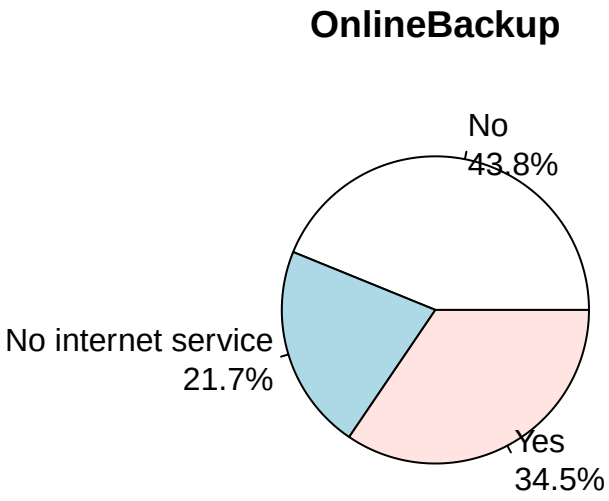
MultipleLines

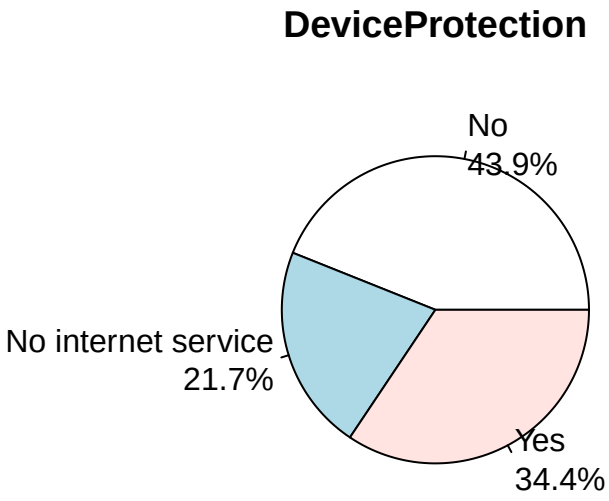


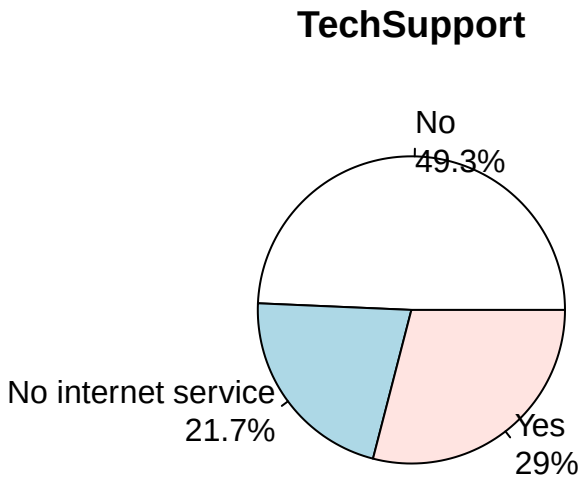
InternetService

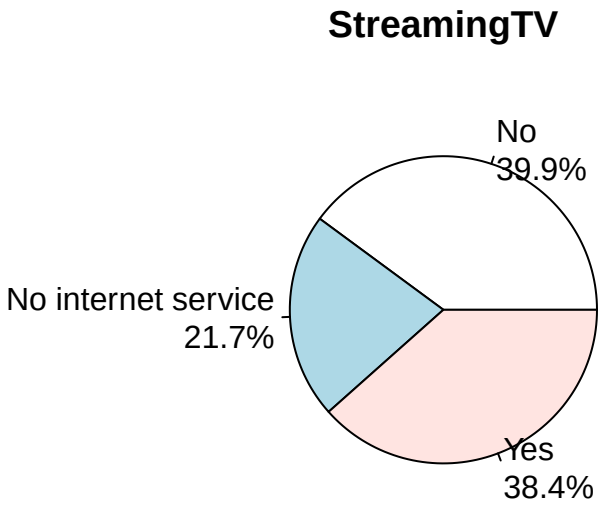


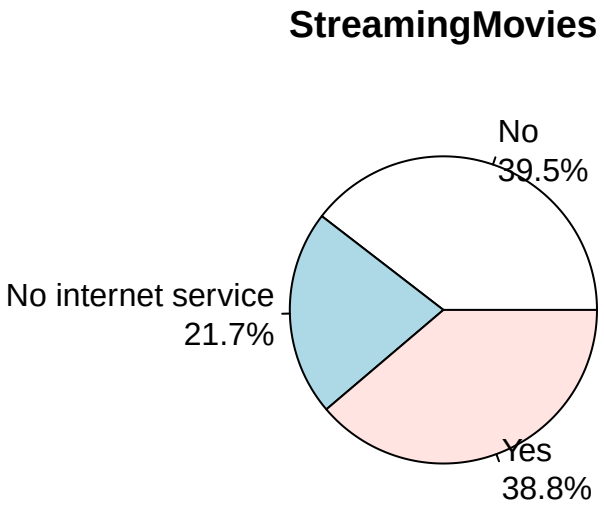

OnlineSecurity
No
49.7%
No internet service
21.7%
Yes
28.7%

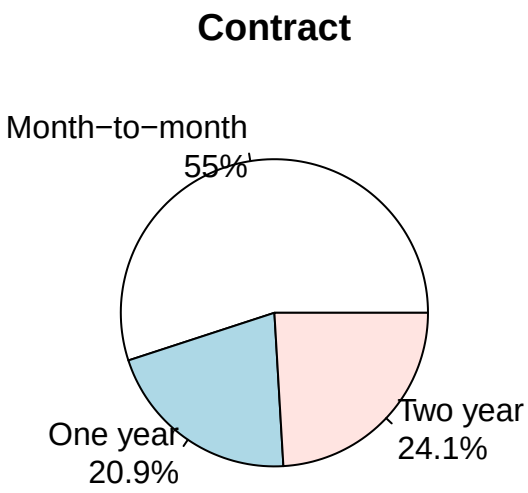

OnlineBackup
No
43.8%
No internet service
21.7%
Yes
34.5%

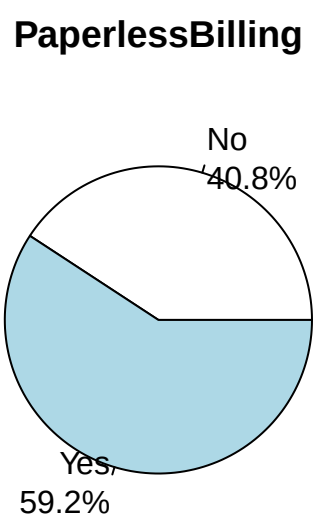

DeviceProtection
No
43.9%
No internet service
21.7%
Yes
34.4%

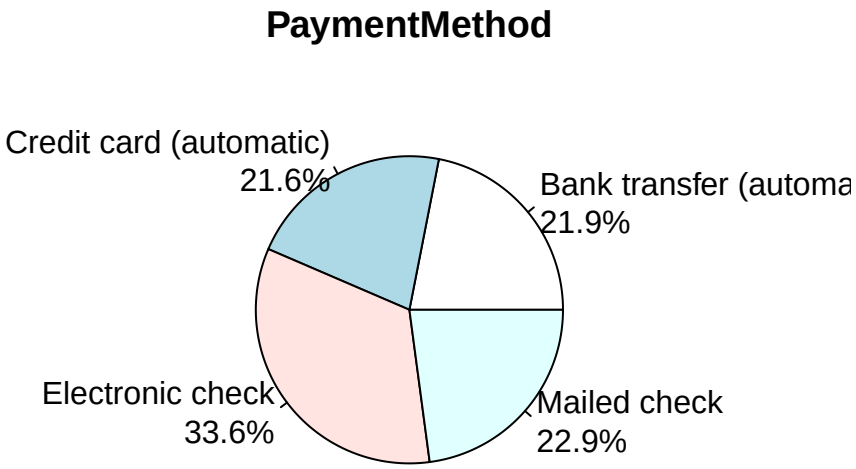

TechSupport
No
49.3%
No internet service
21.7%
Yes
29%

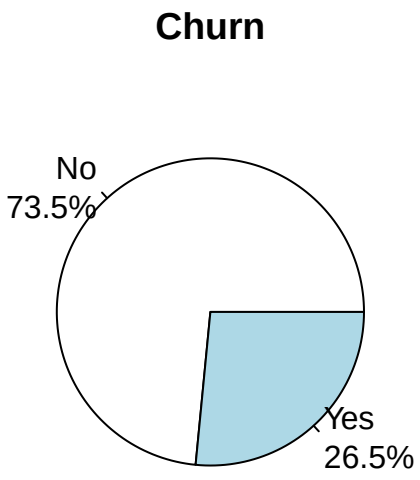

StreamingTV
No
39.9%
No internet service
21.7%
Yes
38.4%


StreamingMovies
No
39.5%
No internet service
21.7%
Yes
38.8%


Contract
Month−to−month
55%
One year
20.9%
Two year
24.1%


PaperlessBilling
No
40.8%
Yes
59.2%

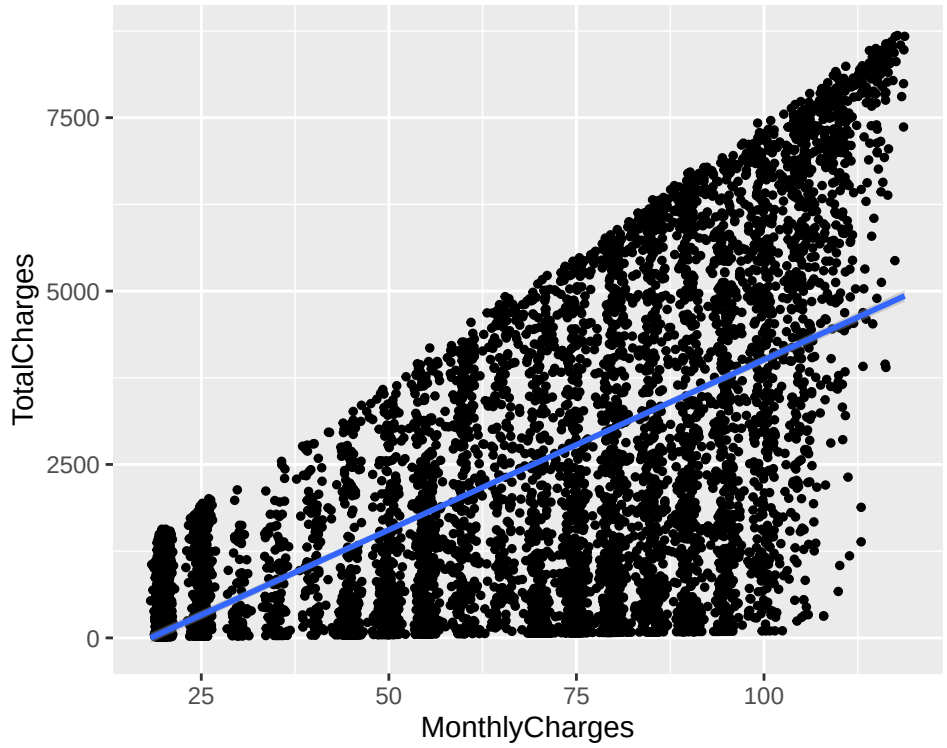

PaymentMethod
Credit card (automatic)
21.6%
Bank transfer (automa
21.9%
Electronic check
33.6%
Mailed check
22.9%


Churn
No
73.5%
Yes
26.5%

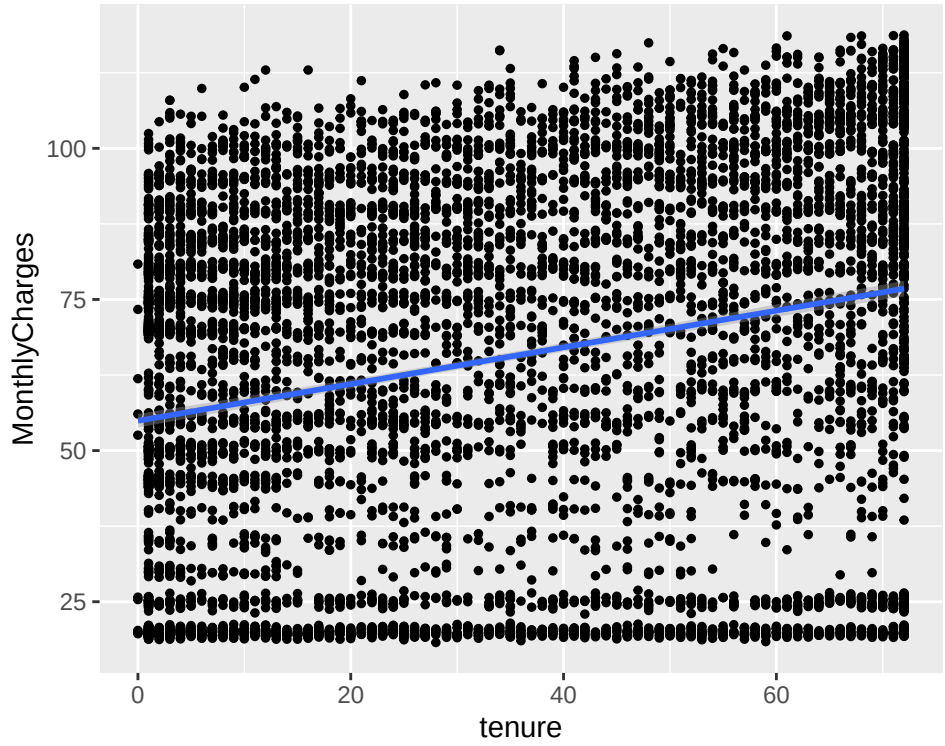
2.3 c.

Wykresy rozrzutu dla zmiennych ciągłych

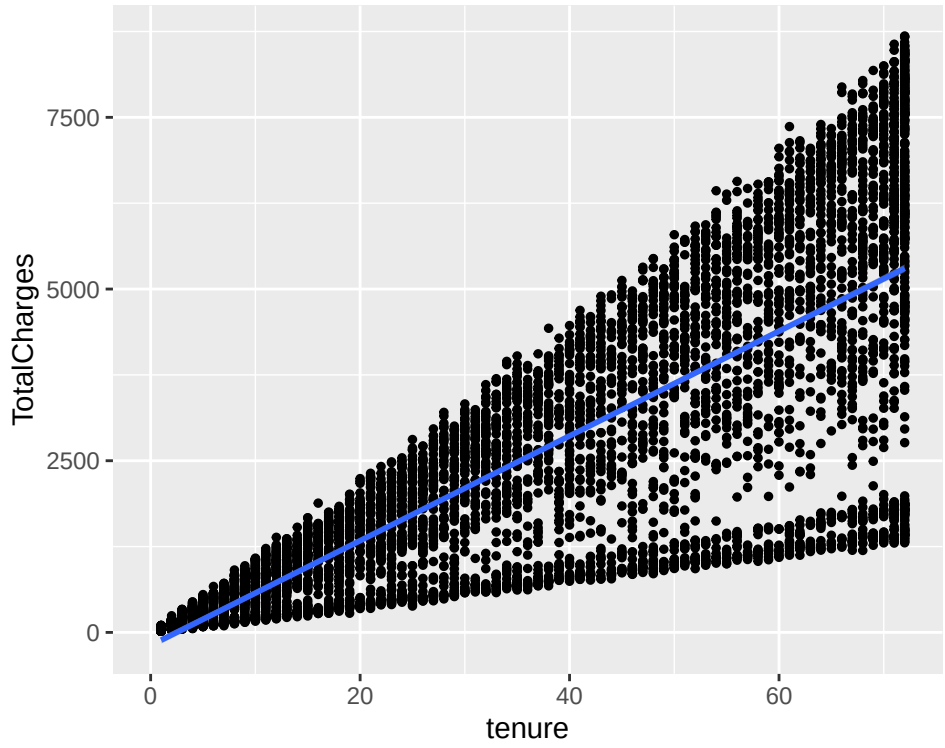
```
ggplot(data, aes(x = MonthlyCharges, y = TotalCharges)) + geom_point(size = 1) +
  geom_smooth(method = "lm")
```




```
ggplot(data, aes(x = tenure, y = MonthlyCharges)) + geom_point(size = 1) +
  geom_smooth(method = "lm")
```




```
ggplot(data, aes(x = tenure, y = TotalCharges)) + geom_point(size = 1) +
  geom_smooth(method = "lm")
```



2.4 d.

2.4.1 Zależności między zmiennymi ciągłymi

Zależności między “MonthlyCharges”, a “TotalCharges” oraz między “tenure”, a “TotalCharges” nie są zaskakujące. Ciekawą obserwacją jest natomiast, że “tenure” jest skorelowany z “MonthlyCharges”, tzn. klienci z długim starzem mają większe opłaty miesięczne.

2.4.2 Zakresy zmiennych ilościowych

Dla zmiennych “tenure”, “MonthlyCharges”, “TotalCharges” zakresy zmienności wynoszą odpowiednio (0,72), (18.25,118.75) oraz (18.8,8684.8). “TotalCharges” ma największe odchylenie standardowe, wynoszące 2266.77.

2.4.3 Opis rozkładów

Zmienna “TotalCharges” jest silnie zagęszczona przy zerze, obserwacje “MonthlyCharges” rzadko pojawiają się w przedziale (25,40). Rozkład “Tenure” jest mniej więcej symetryczny, przypomina literę “U” z niewielkim zagęszczeniem z lewej strony.

2.4.4 Częstość przyjmowania kategorii zmiennych jakościowych

Mężczyzn jest tyle samo co kobiet, połowa osób posiada partnera. Seniorzy stanowią niewielką część klientów. Większość nie posiada osób zależnych. Klienci prawie zawsze korzystają z usług telefonicznych. Większą popularnością od DSL cieszy się internet światłowodowy. Pośród każdej z dodatkowych usług tj. bezpieczeństwo online, kopia zapasowa online, ochrona urządzenia i wsparcie techniczne, aż trzy czwarte klientów z niej nie korzysta. Niewiele wiekszą popularnością cieszą się usługi streamingowe. Najczęściej wybieranym okresem płatności jest miesięczny, a klienci rzadziej wybierają rachunek na papierze. Wszystkie dostępne formy płatności są równie atrakcyjne. Do tej pory tylko jedna czwarta klientów odeszła od firmy.

3 Etap 3

3.1 a.

Wskaźniki sumaryczne z uwzględnieniem podziału na grupy

```
my.summary <- function(X)
{
  result <- c(min(X),quantile(X,0.25), median(X), mean(X), quantile(X,0.75),
              max(X), var(X), sd(X), IQR(X))
  return(result)
}

data_ChurnN <- subset(data, subset = (Churn == "No"))
data_ChurnY <- subset(data, subset = (Churn == "Yes"))


result_N <- data.frame(tenure = my.summary(data_ChurnN$tenure),
                       MonthlyCharges = my.summary(data_ChurnN$MonthlyCharges),
                       TotalCharges = my.summary(
                       data_ChurnN$TotalCharges[
                       !is.na(data_ChurnN$TotalCharges)]),
                       row.names = c("min", "Q1", "median", "mean", "Q3", "max",
                                     "var", "sd", "IQR"))

tab1 <- xtable(result_N, digits = 2, row.names = TRUE,
               caption = "wybrane wskaźniki dla obecnych klientów",
               label = "tab:tabela2")
print(tab1, type = "latex", table.placement = "H", comment=FALSE)
```

	tenure	MonthlyCharges	TotalCharges
min	0.00	18.25	18.80
Q1	15.00	25.10	577.83
median	38.00	64.43	1683.60
mean	37.57	61.27	2555.34
Q3	61.00	88.40	4264.12
max	72.00	118.75	8672.45
var	581.47	966.75	5426369.84
sd	24.11	31.09	2329.46
IQR	46.00	63.30	3686.30

Table 2: wybrane wskaźniki dla obecnych klientów

```
result_Y <- data.frame(tenure = my.summary(data_ChurnY$tenure),
                       MonthlyCharges = my.summary(data_ChurnY$MonthlyCharges),
                       TotalCharges = my.summary(
                       data_ChurnY$TotalCharges[
                       !is.na(data_ChurnY$TotalCharges)]),
                       row.names = c("min", "Q1", "median", "mean", "Q3", "max",
```

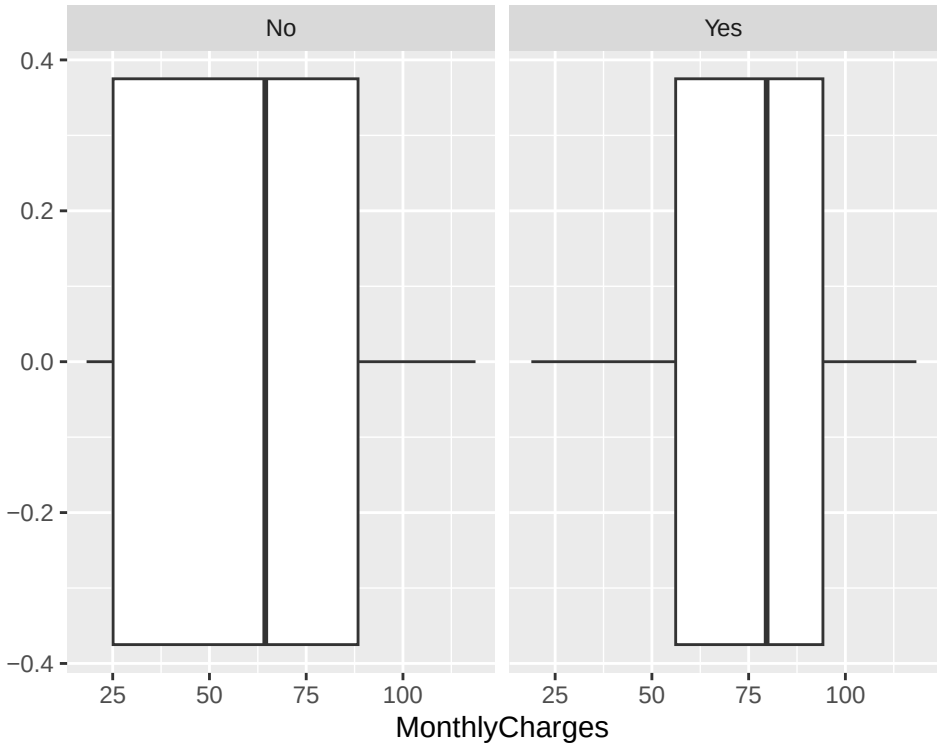
```
                                    "var", "sd", "IQR"))

tab1 <- xtable(result_Y, digits = 2, row.names = TRUE,
               caption = "wybrane wskaźniki dla klientów, którzy odeszli",
               label = "tab:tabela3")
print(tab1, type = "latex", table.placement = "H", comment=FALSE)
```

	tenure	MonthlyCharges	TotalCharges
min	1.00	18.85	18.85
Q1	2.00	56.15	134.50
median	10.00	79.65	703.55
mean	17.98	74.44	1531.80
Q3	29.00	94.20	2331.30
max	72.00	118.35	8684.80
var	381.46	608.41	3575211.60
sd	19.53	24.67	1890.82
IQR	27.00	38.05	2196.80

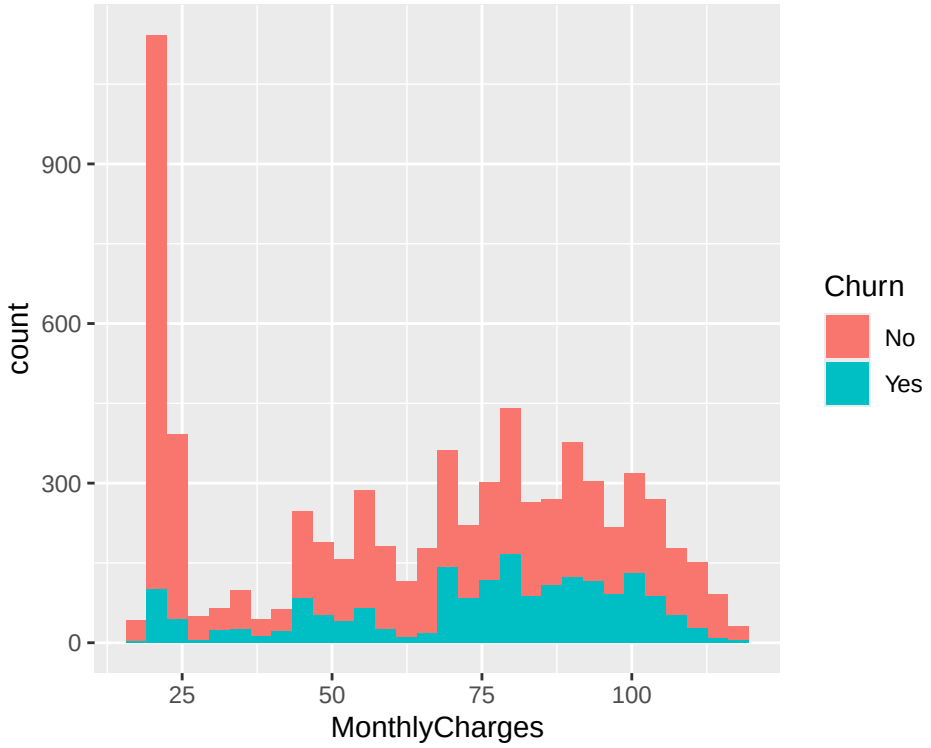
Table 3: wybrane wskaźniki dla klientów, którzy odeszli

Wykresy zmiennych ilościowych pogrupowane ze względu na to czy klient odszedł

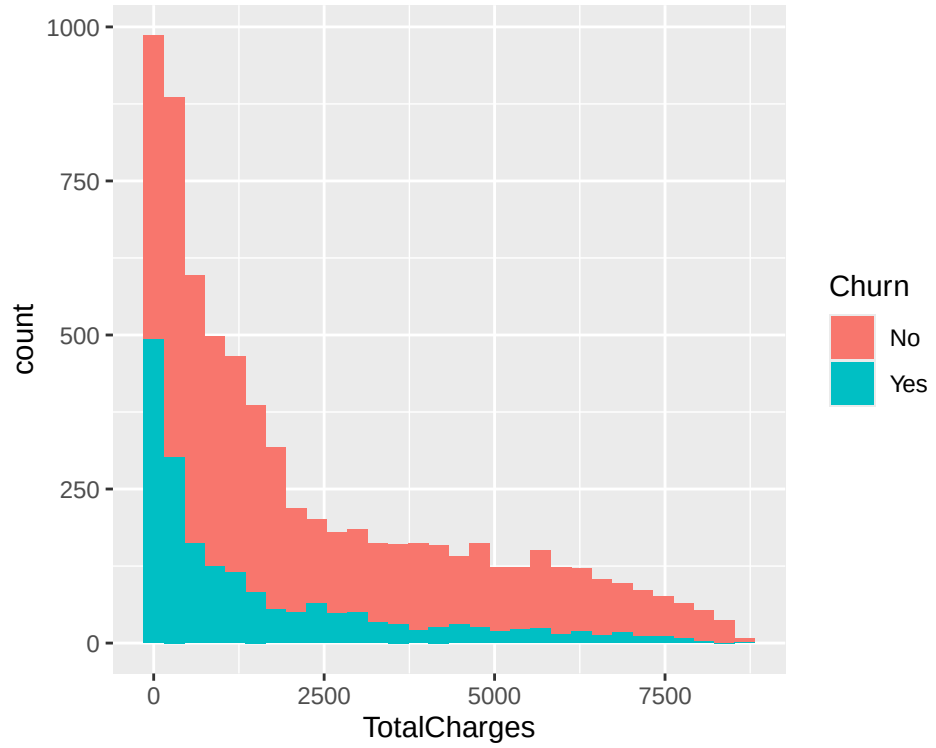
```
par(mfrow = c(1,2))
ggplot(data, aes(x = MonthlyCharges)) + geom_boxplot() + facet_wrap(Churn)
```



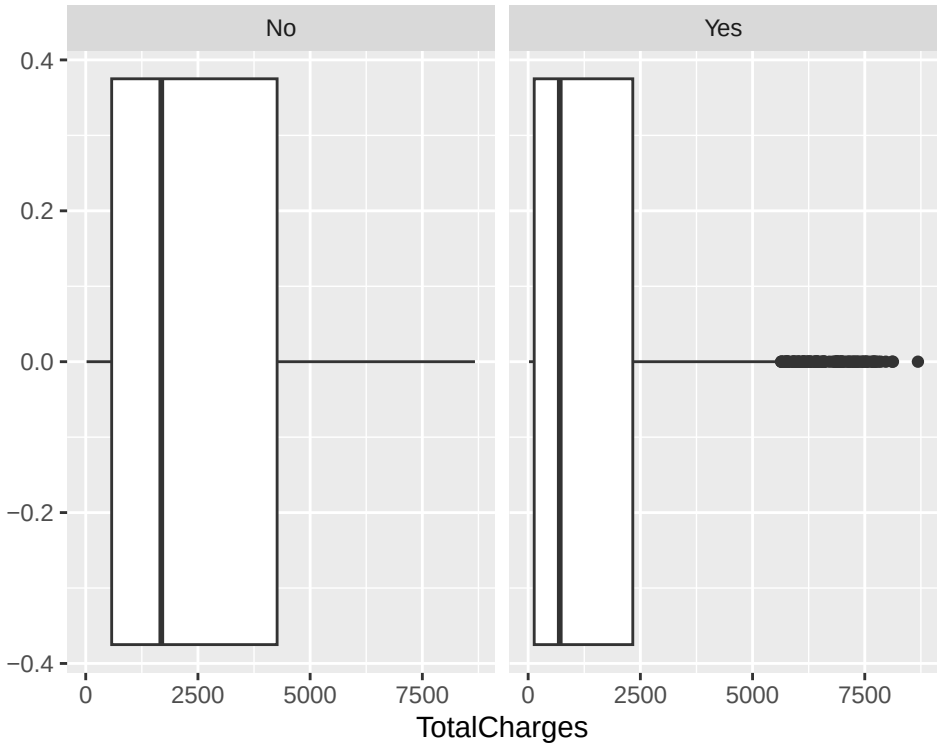

```
ggplot(data, aes(x = MonthlyCharges, fill = Churn)) + geom_histogram()
```



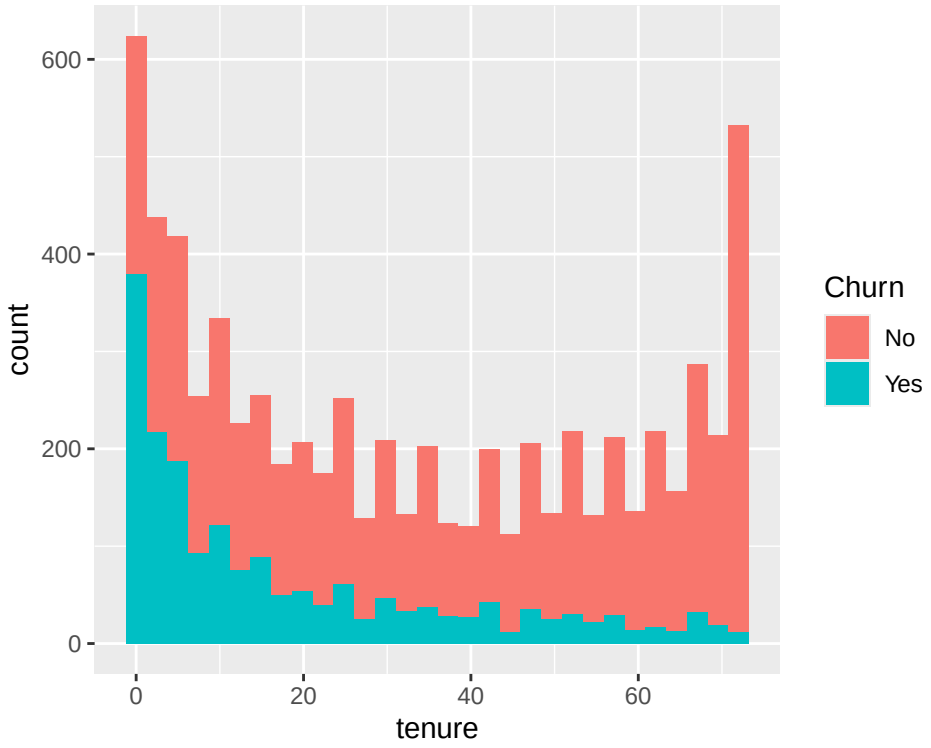
```
par(mfrow = c(1,2))
ggplot(data, aes(x = TotalCharges, fill = Churn)) + geom_histogram()
```



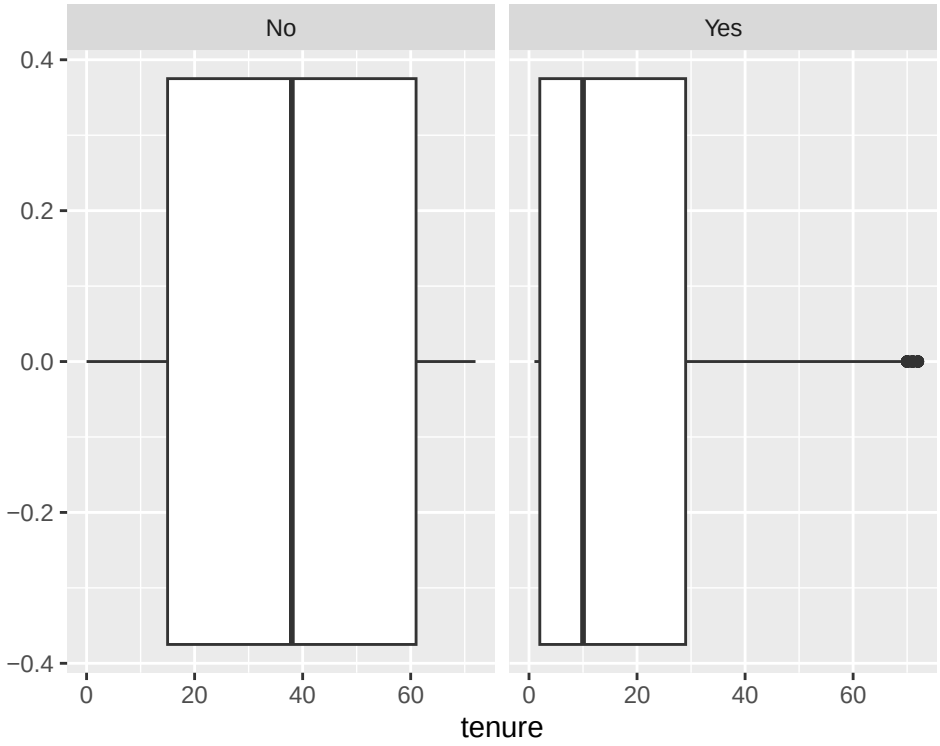

```
ggplot(data, aes(x = TotalCharges)) + geom_boxplot() + facet_wrap(Churn)
```




```
par(mfrow = c(1,2))
ggplot(data, aes(x = tenure, fill = Churn)) + geom_histogram()
```

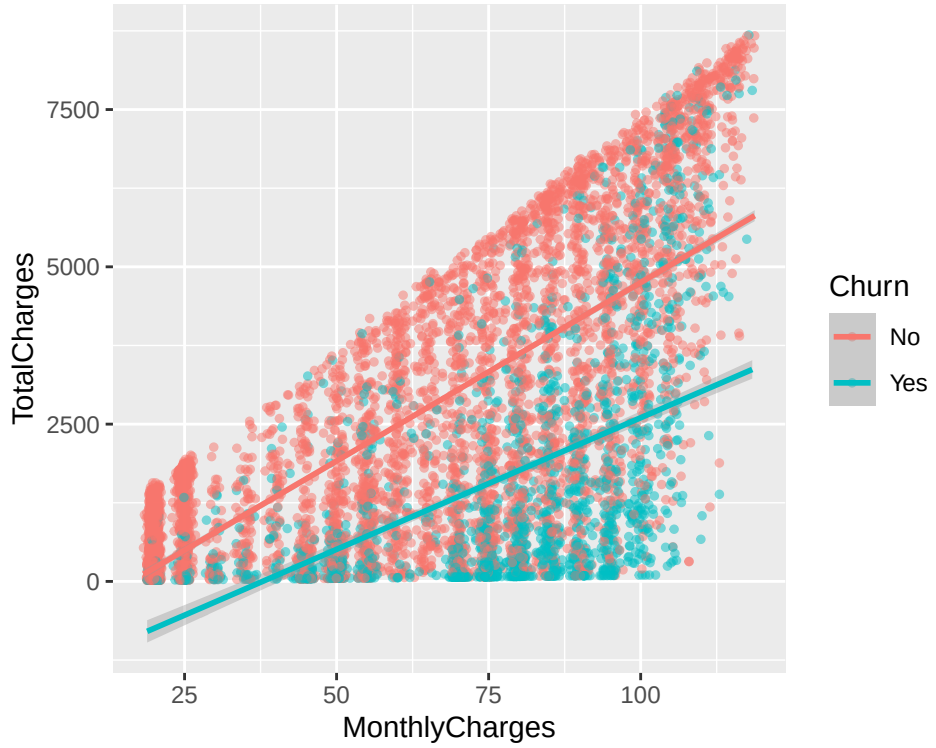


```
ggplot(data, aes(x = tenure)) + geom_boxplot() + facet_wrap(Churn)
```

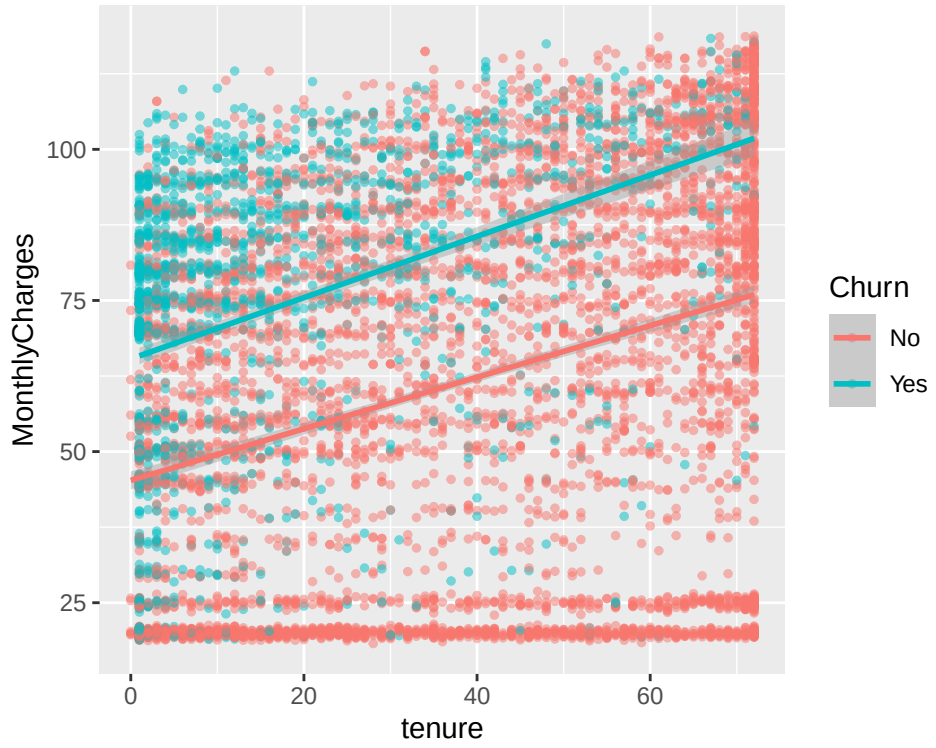


Wykresy rozrzutu dla zmiennych ciągłych w podziale na grupy

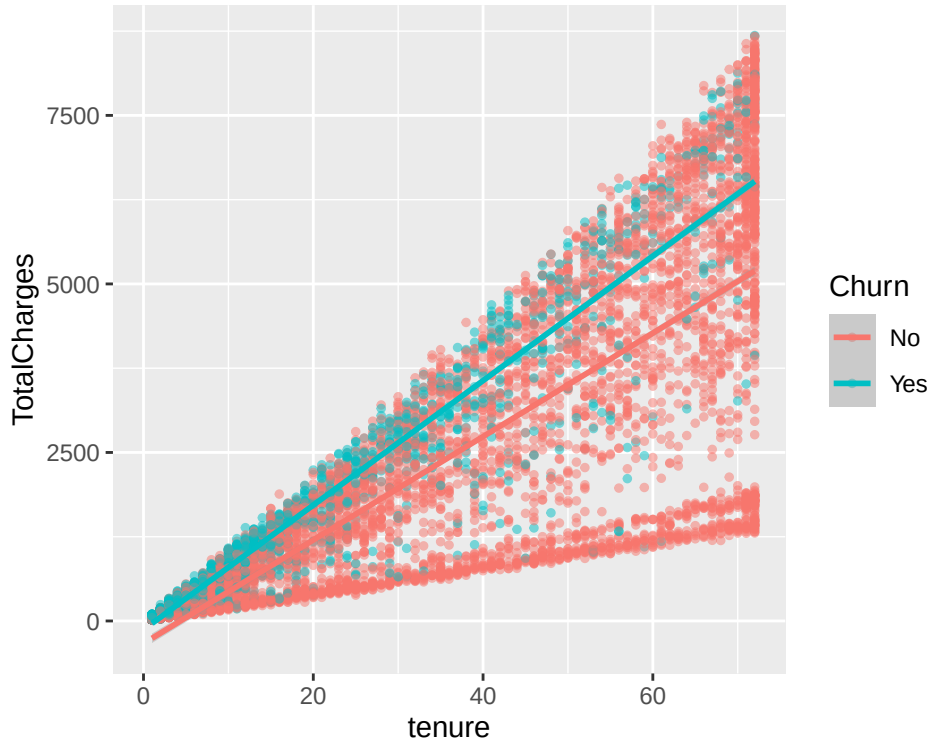
```
ggplot(data, aes(x = MonthlyCharges, y = TotalCharges, colour = Churn)) +
  geom_point(alpha = 0.5, size = 1) + geom_smooth(method = "lm")
```




```
ggplot(data, aes(x = tenure, y = MonthlyCharges, colour = Churn)) +
  geom_point(alpha = 0.5, size = 1) + geom_smooth(method = "lm")
```

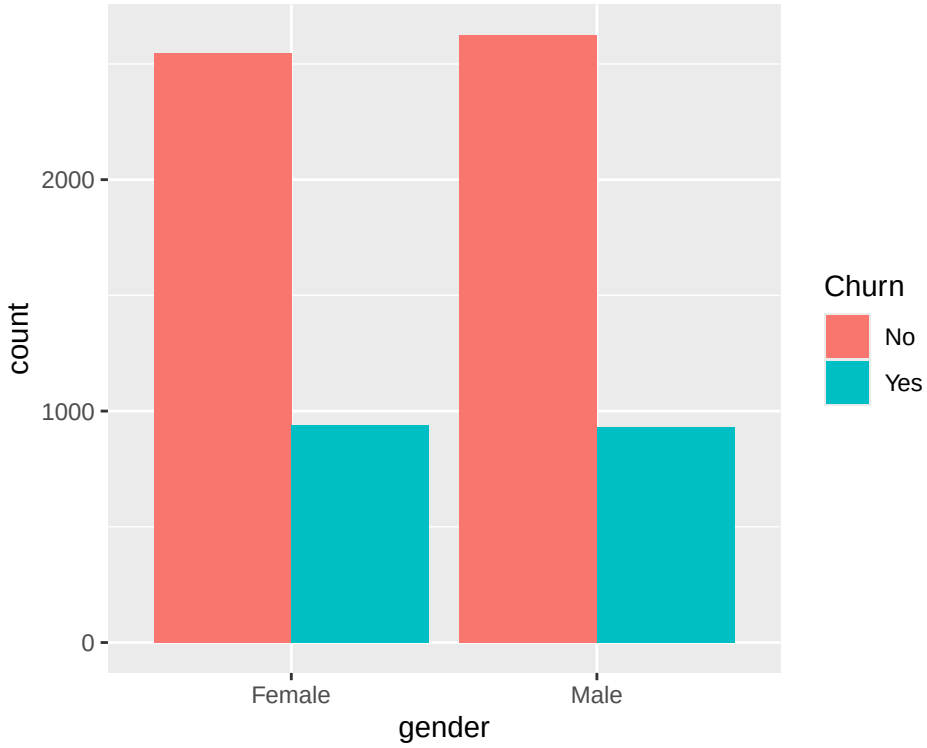


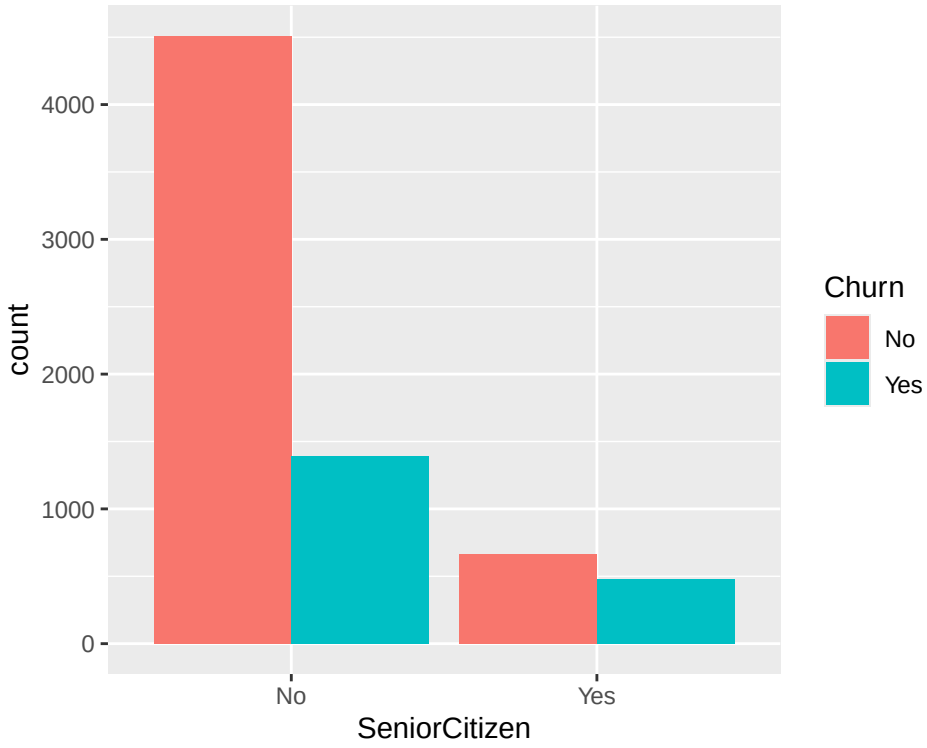
```
ggplot(data, aes(x = tenure, y = TotalCharges, colour = Churn)) +
  geom_point(alpha = 0.5, size = 1) + geom_smooth(method = "lm")
```

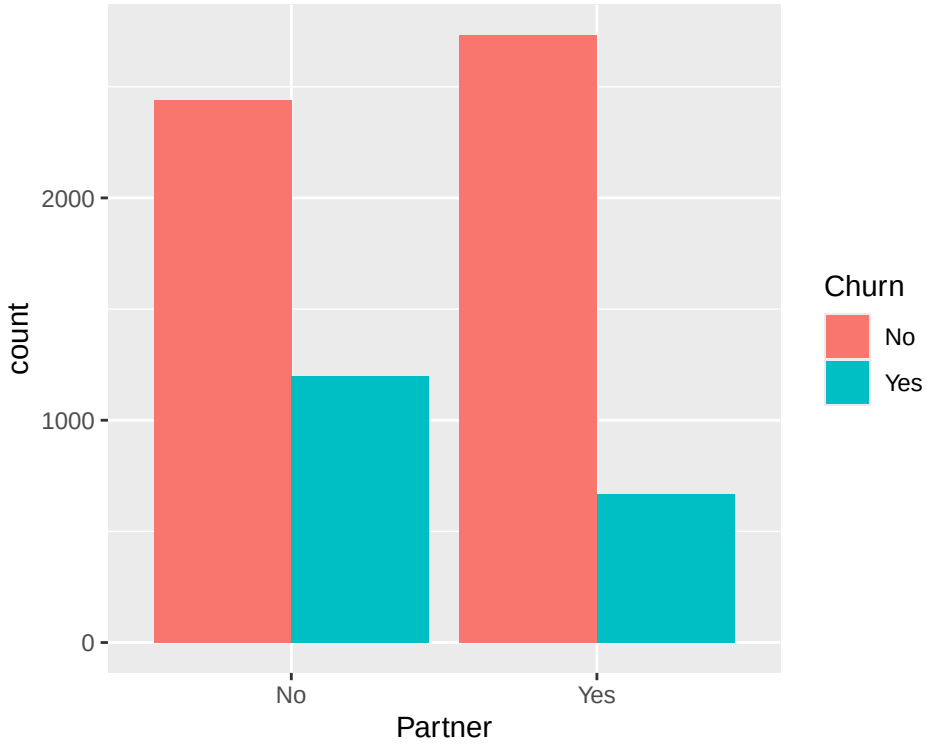


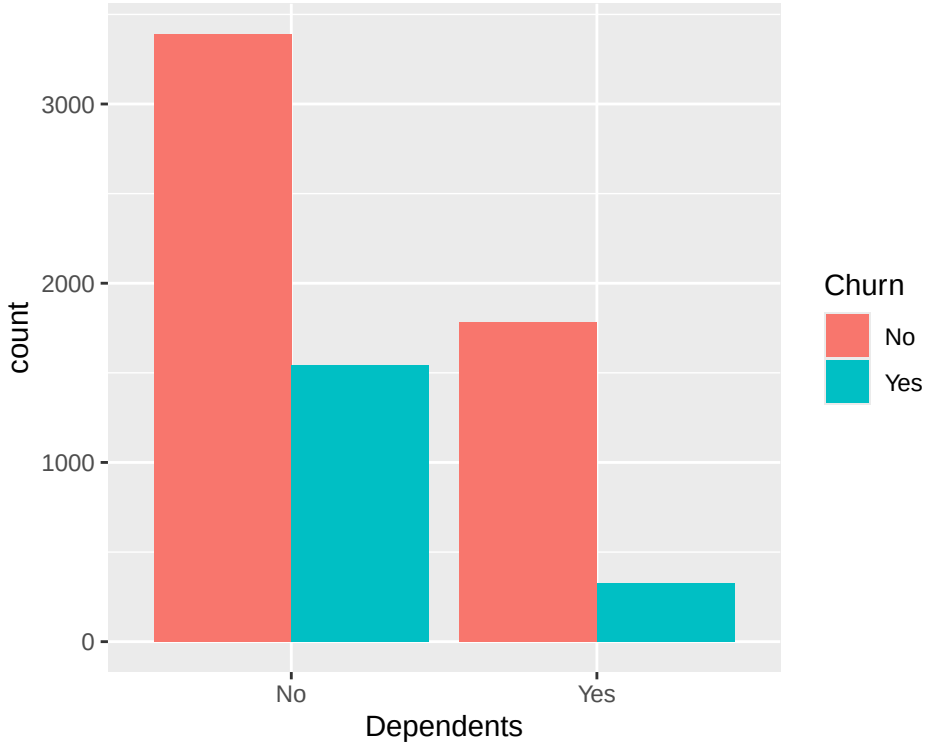
Zmienne ilościowe w podziale na grupy

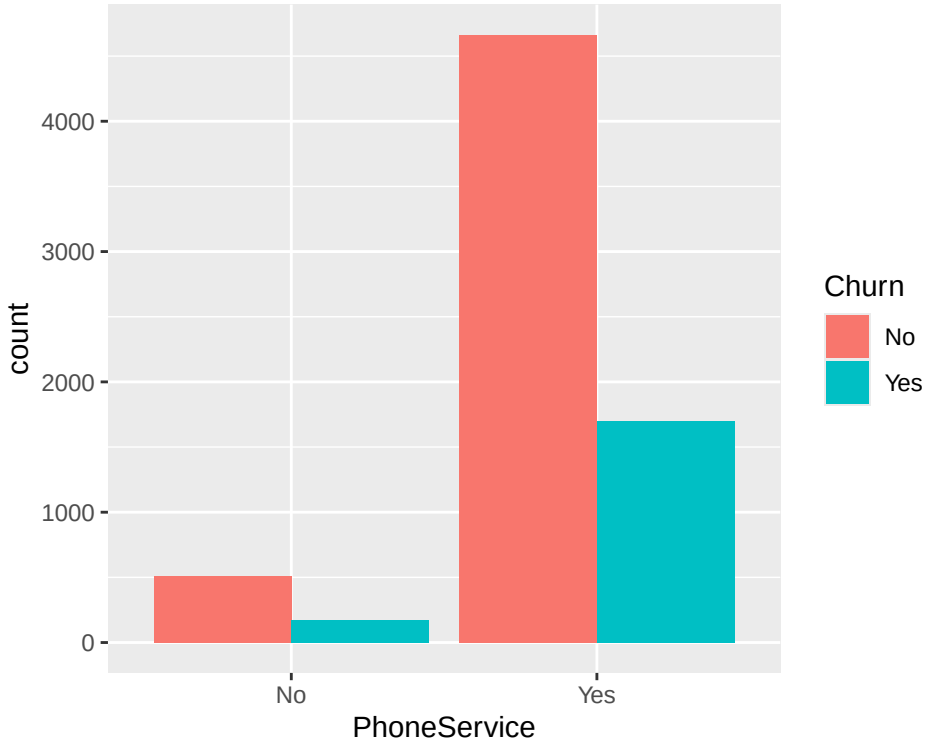
```
for(i in 1:ncol(data)){
  if(is.factor(data[[i]]) & colnames(data)[i] != "Churn"){
    chart <- ggplot(data, aes(x = data[[i]], fill = Churn)) +
      geom_bar(position = "dodge") + xlab(colnames(data)[i])
    print(chart)
  }
}
```

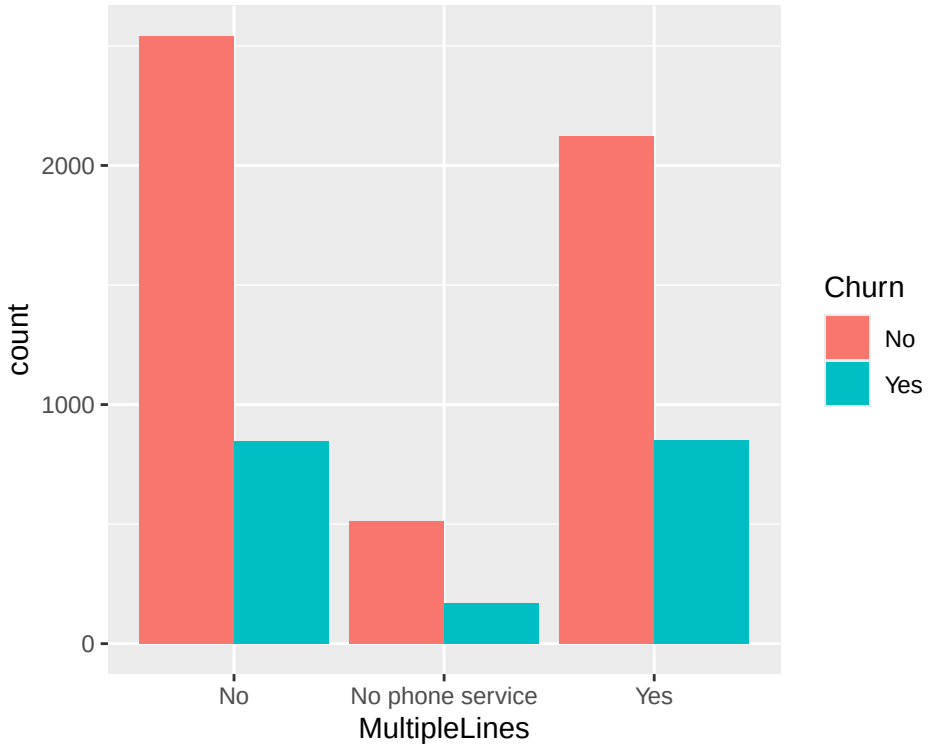

2000
1000
0
count
Female
Male
gender
Churn
No
Yes

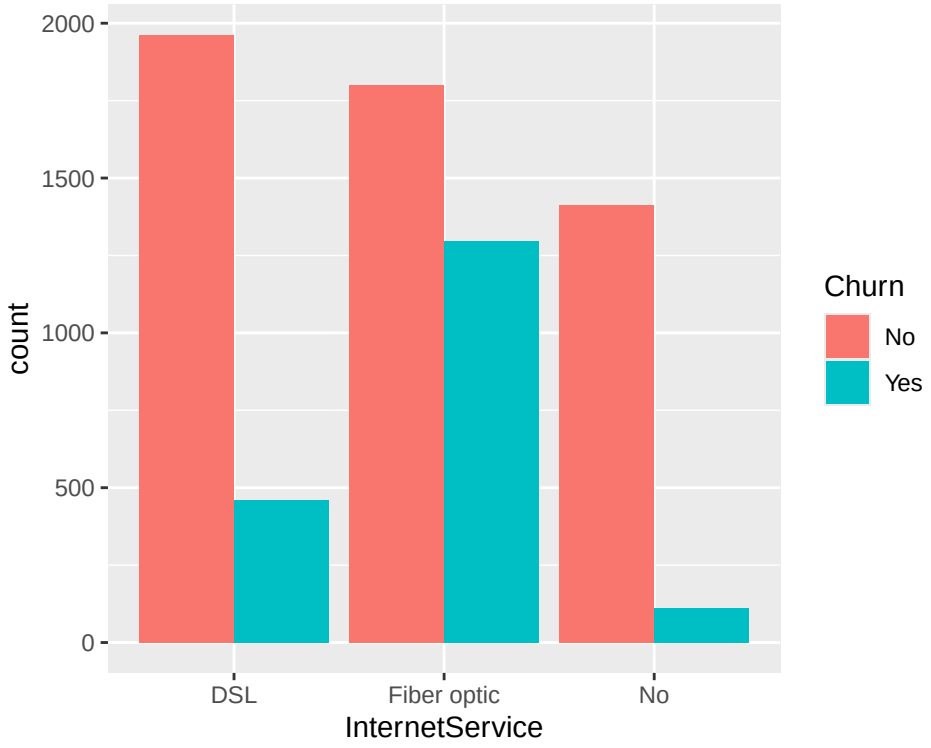

4000
3000
2000
1000
0
count
No
Yes
SeniorCitizen
Churn
No
Yes

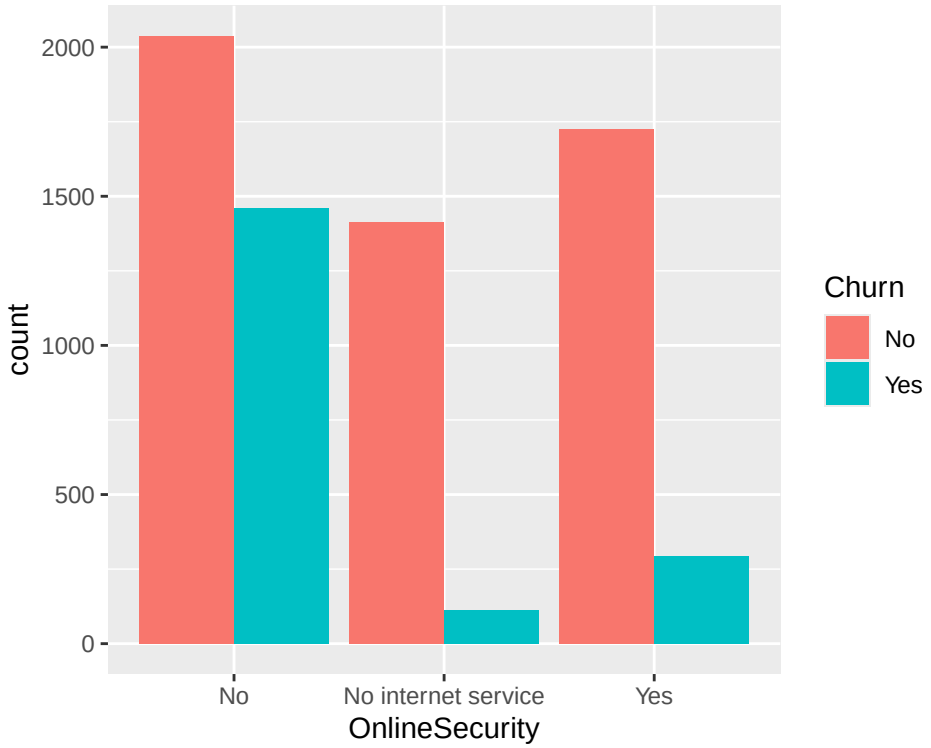

count
2000
1000
0
No
Yes
Partner
Churn
No
Yes

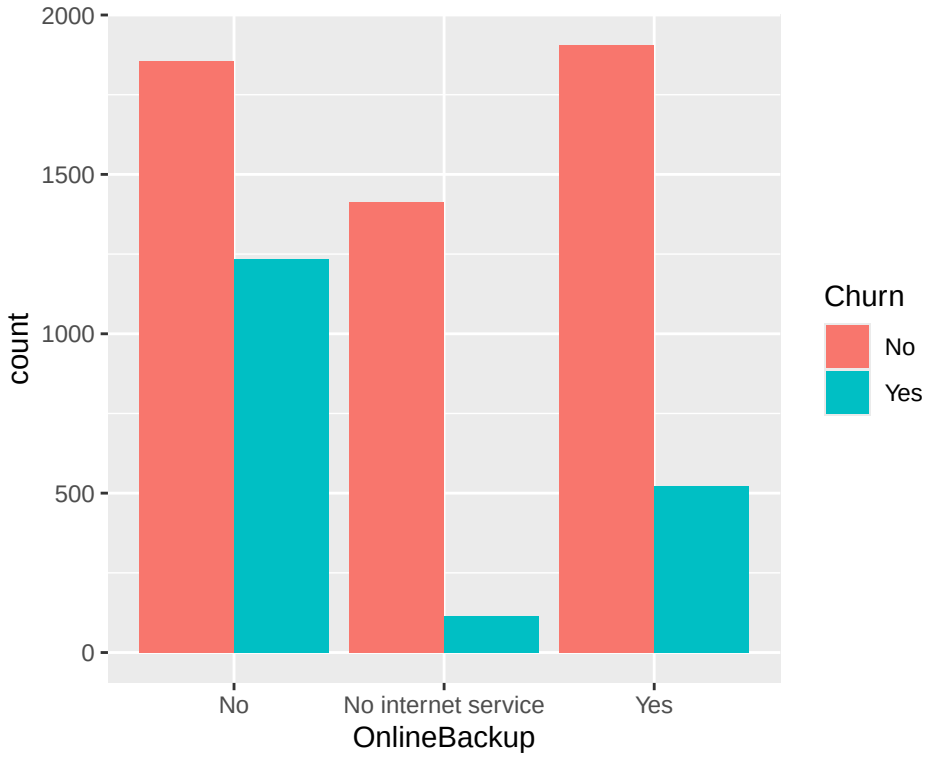

count
3000
2000
1000
0
No
Yes
Dependents
Churn
No
Yes

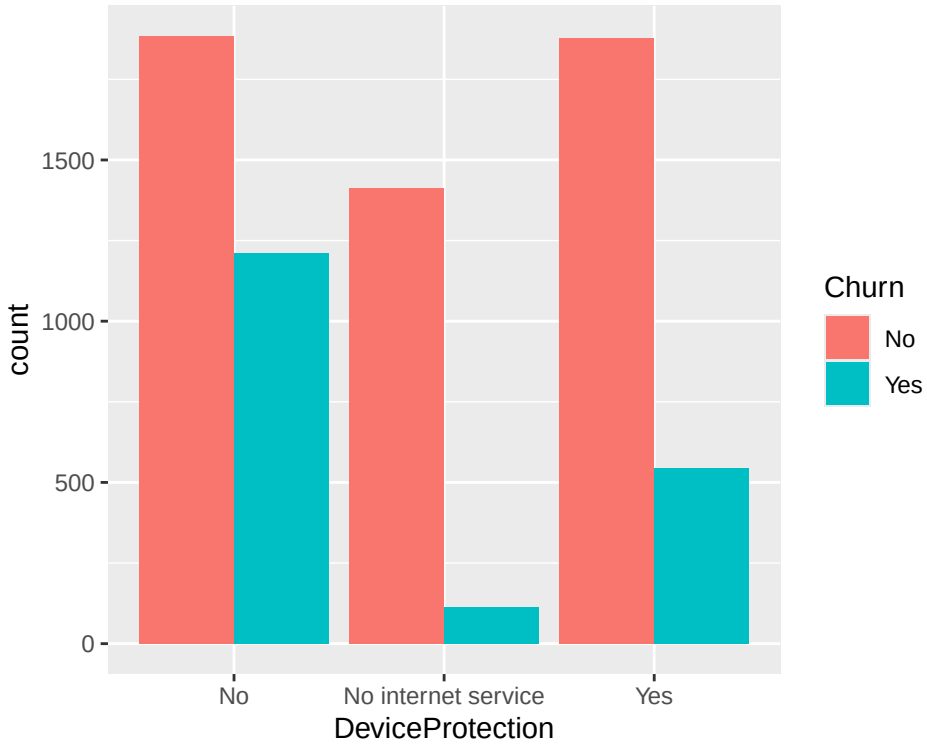

4000
3000
2000
1000
0
count
No
Yes
PhoneService
Churn
No
Yes

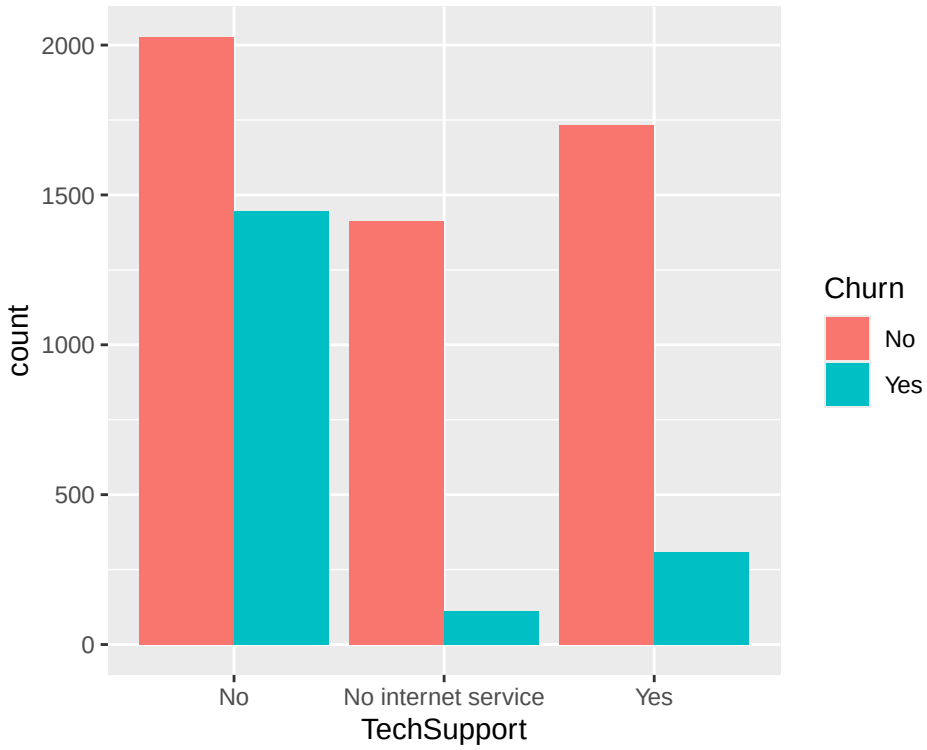

2000
1000
0
count
No
No phone service
Yes
MultipleLines
Churn
No
Yes

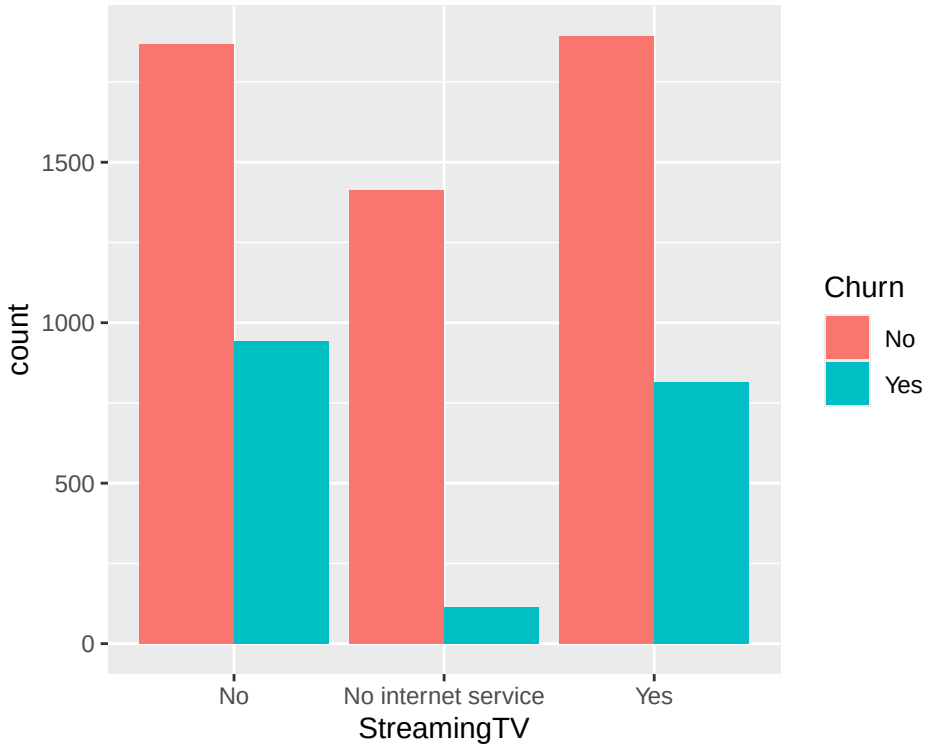

2000
1500
1000
500
0
count
DSL
Fiber optic
No
InternetService
Churn
No
Yes

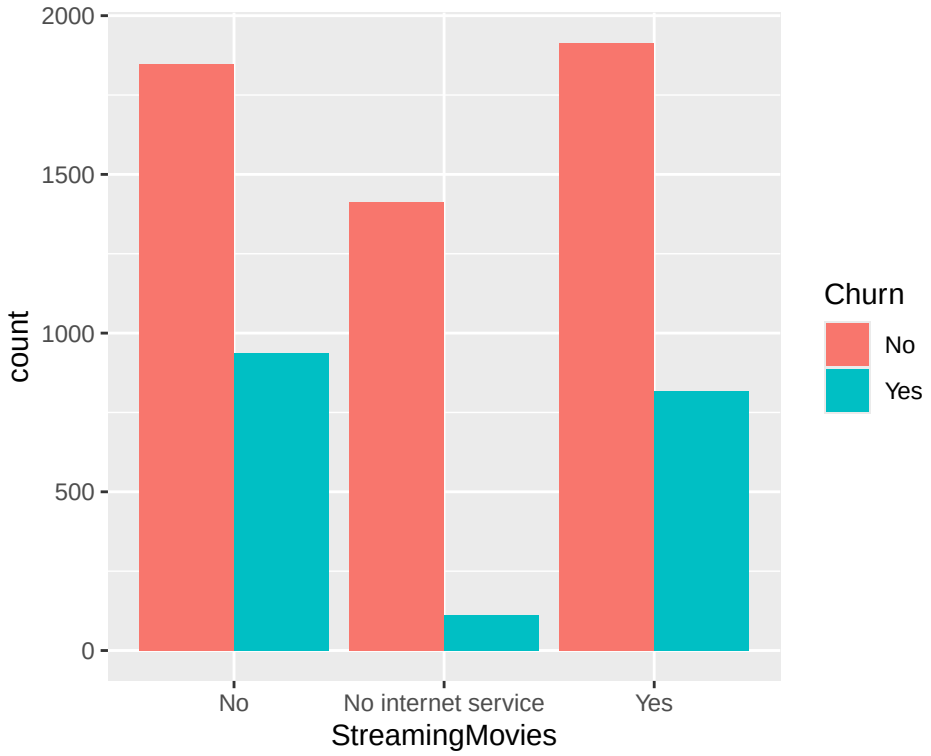

2000
1500
1000
500
0
count
No
No internet service
Yes
OnlineSecurity
Churn
No
Yes

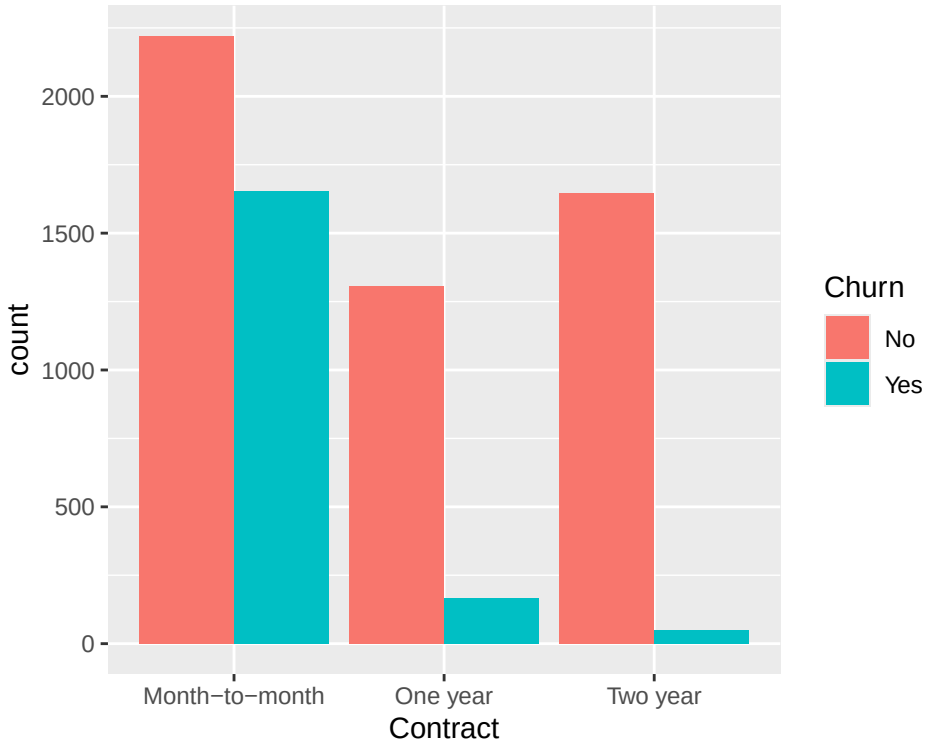

2000
1500
1000
500
0
count
No
No internet service
Yes
OnlineBackup
Churn
No
Yes

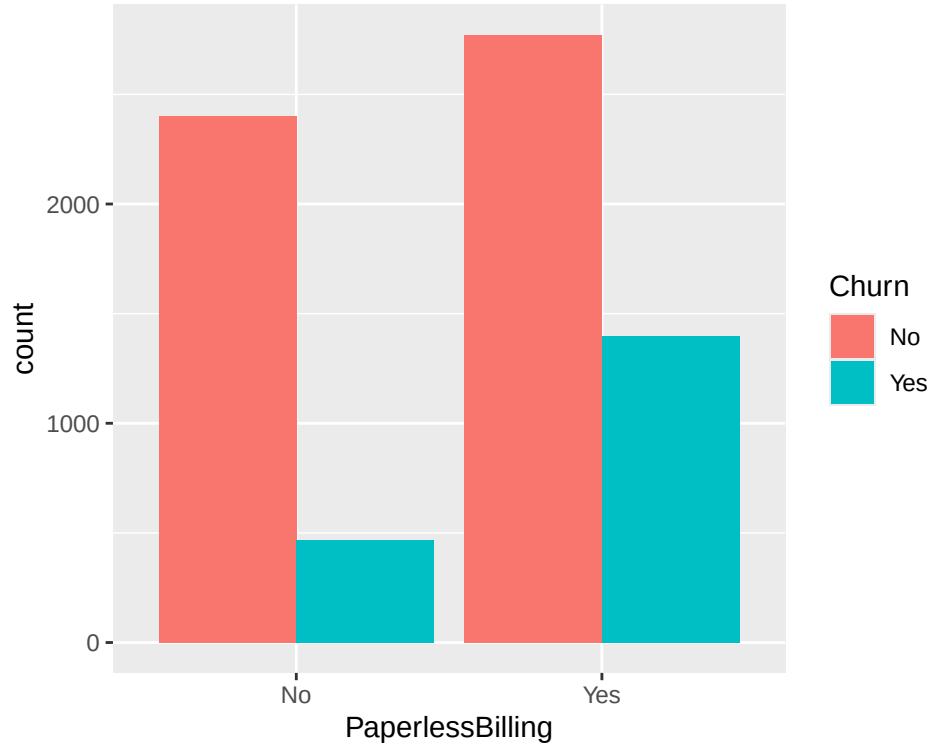

1500
1000
500
0
count
No
No internet service
Yes
DeviceProtection
Churn
No
Yes

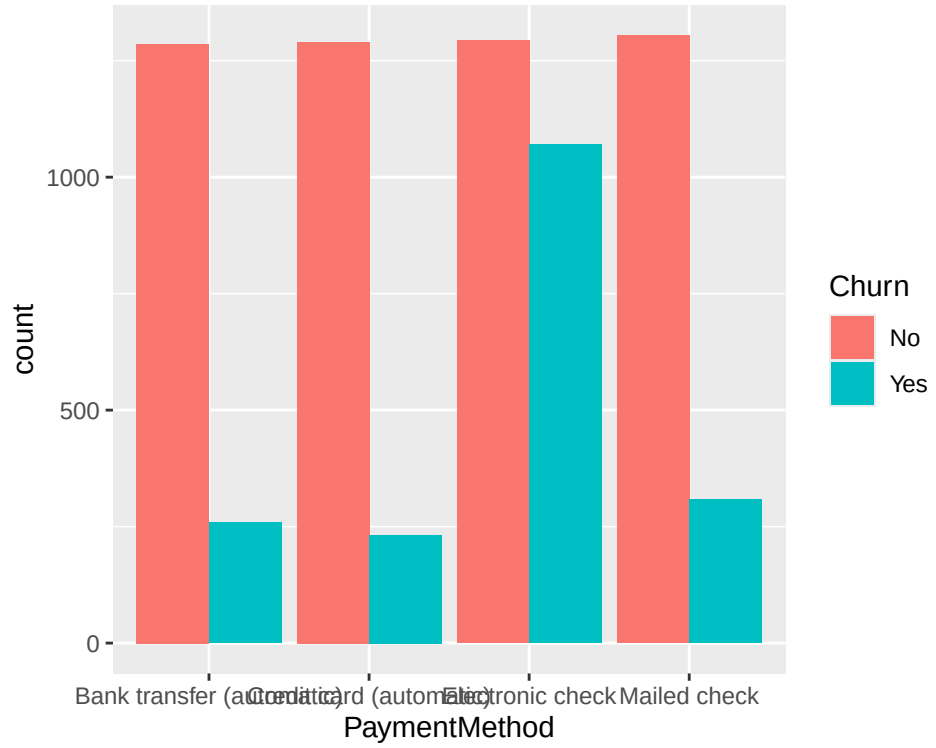

2000
1500
1000
500
0
count
No
No internet service
Yes
TechSupport
Churn
No
Yes


1500
1000
500
0
count
No
No internet service
Yes
StreamingTV
Churn
No
Yes


2000
1500
1000
500
0
count
No
No internet service
Yes
StreamingMovies
Churn
No
Yes

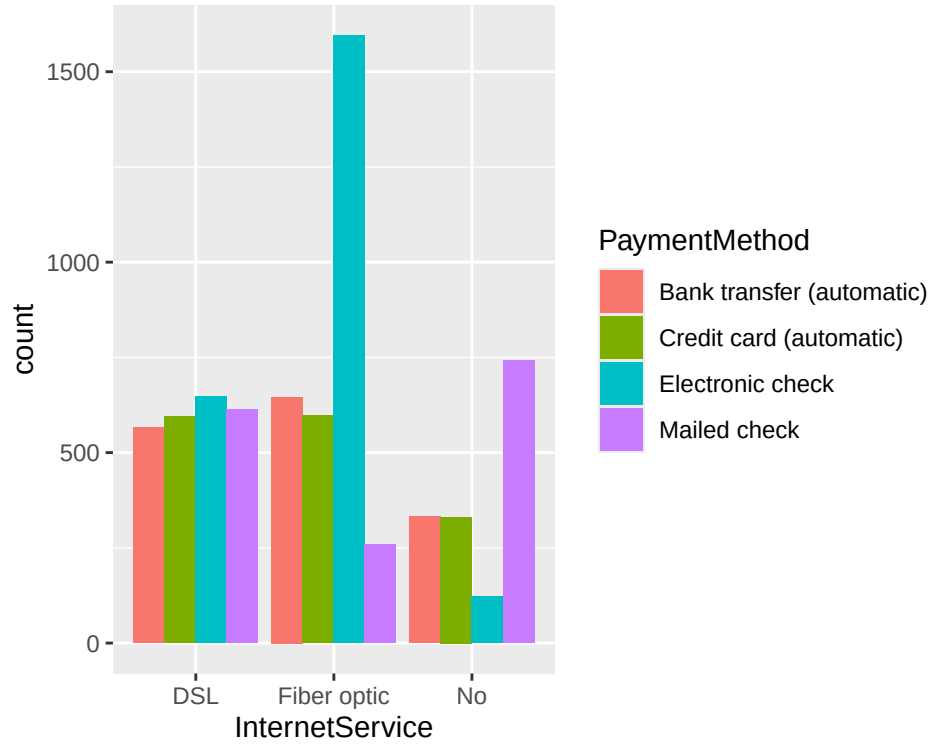

2000
1500
1000
500
0
count
Month−to−month
One year
Two year
Contract
Churn
No
Yes



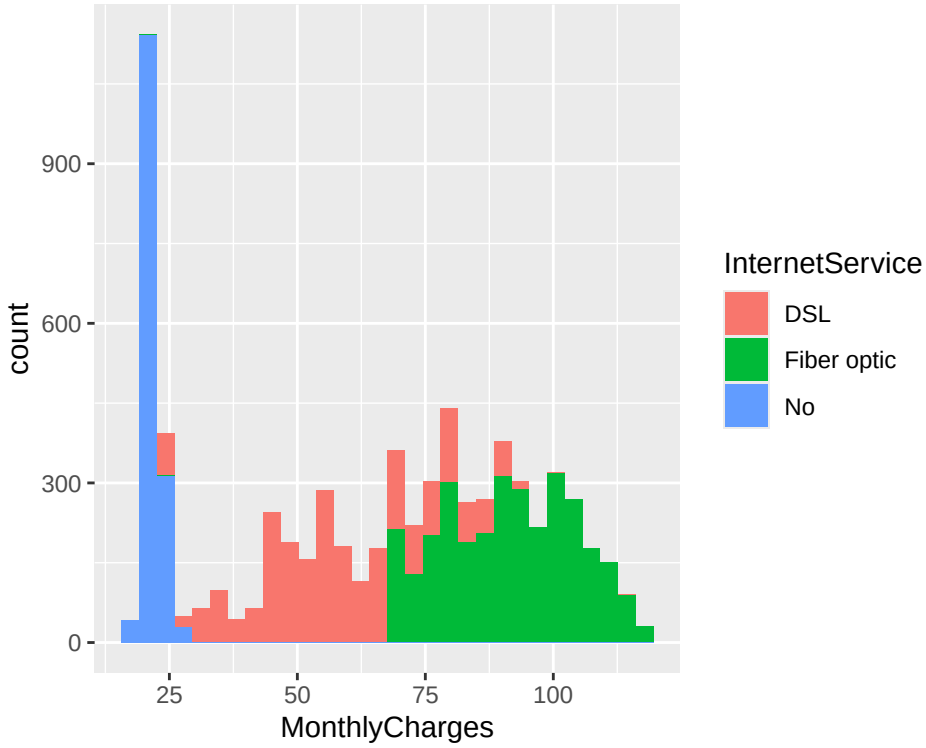


Ze względu na to, że często odchodzą klienci korzystający z czeku elektronicznego. Interesujące mogą być relacje metody płatności, a innymi zmiennymi. Warto spojrzeć również na zależność między internetem a opłatą miesięczną

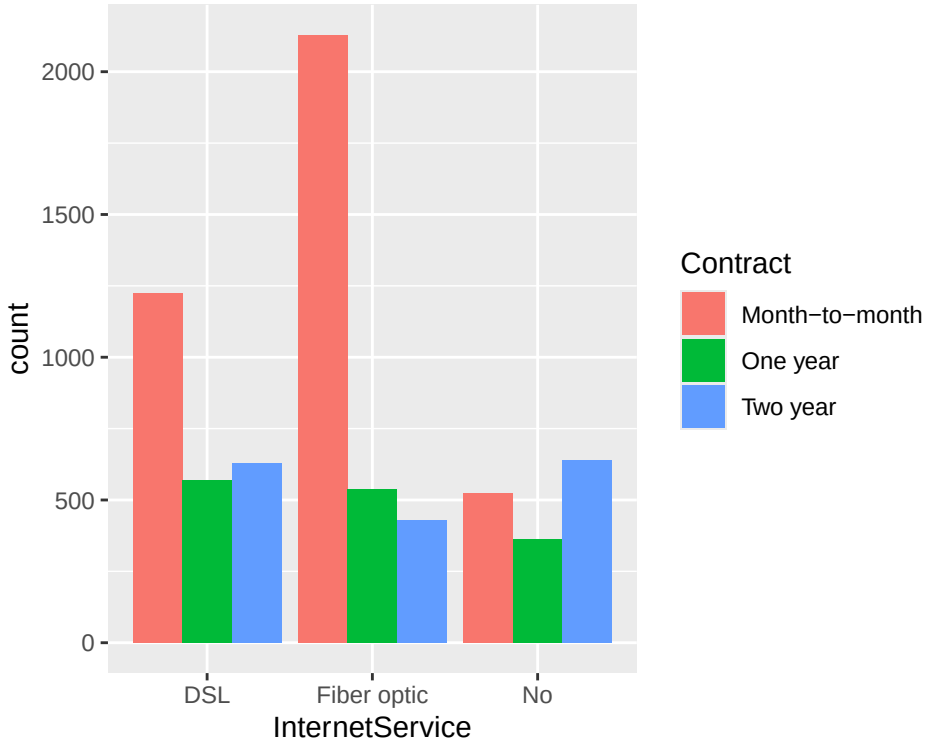
```r
ggplot(data, aes(x = InternetService, fill = PaymentMethod)) +
      geom_bar(position = "dodge")
```




```r
ggplot(data, aes(x = MonthlyCharges, fill = InternetService)) + geom_histogram()
```




```
ggplot(data, aes(x = InternetService, fill = Contract)) + geom_bar(position = "dodge")
```



3.2 b.

Spośród zmiennych jakościowych największe różnice znajdujemy w zmiennej “tenure”, widoczne są zarówno na wykresach rozrzutu jak i histogramach, jednak można dostrzec takowe również przy “TotalCharges” oraz “MonthlyCharges”. Pośród zmiennych ilościowych największe zróżnicowanie wykazują “InternetService”, “OnlineSeciurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “StreamingTV”, “StreamingMovies”. W pewnym sensie interesująca jest tylko zmienna internet service, ponieważ różnice , które znajdują się w kolejnych są konsekwencją tego, czy klient ma wykupiony internet czy nie. Ponadto możemy dostrzec zależność od zmiennej “Churn” w “Dependents”, “Contract”, “PaperlessBilling” oraz “PaymentMethod”.

4 Etap 4

4.1 a.

Dane wymagały jedynie drobnej poprawki w przygotowaniu do analiz, braków było niewiele. Mamy 17 zmiennych ilościowych oraz 3 interesujące nas jakościowe, wszyskie z nich dają się łatwo interpretować. Część wykazuje asymetryczność rozkładów oraz znaczne różnice w podziale na grupy, zatem możemy próbować wyciągać z nich różne wnioski.

4.2 b.

Większość klientów nie kwalifikuje się jako senior. Zazwyczaj mają osoby od siebie zależne. Ok. 90% korzysta z usług internetowych oraz ok. 80% z telefonicznych, zatem klienci często korzystają z obu jednocześnie. Zdecydowanie preferują kontrakty miesięczne. Do tej pory firma straciła jedną czwartą klientów. Ciekawą obeserwacją jest, że klienci z długim stażem mają tendencje do posiadania wyższych opłat miesięcznych.

4.3 c.

Najbardzie rzucającym się w oczy czynnikiem, wpływającym na to czy klient pozostał w firmie jest rodzaj internetu, z którego korzysta. Możemy podejrzewać że niezadowolenie spowodowane jest głównie światłowodem. Zauważamy, że często odchodzą klienci, którzy mają wysokie opłaty miesięczne, rozliczają się czekiem elektronicznym oraz płacą za usługi co miesiąc, jednak jak widać na wykresach są to właśnie użytkownicy światłowodu. Ponadto odchodzi wielu seniorów, jednak ci stanowią niewielki procent klienteli.

Najważniejsze jest rozstrzygnięcie kwestii czterech skorelowanych ze sobą cech tj. światłowodu, miesięcznego okresu rozliczania, płatności czekiem elektronicznym oraz opłat miesięcznych. W pierwszej kolejności należy przyjrzeć się jakości i cenie internetu światłowodowego, który zdaje się być głównym powodem niezadowolenia. Warto pomyśleć

nad zniżkami i/lub udogonieniemi dla seniorów, jedak ze względu na ich niewielką liczbę nie jest to sprawa najwyższej wagi.