# Exploratory Data Analysis for Telecom Churn Analysis

Malvika Singh

# EDA performed

We start with basic exploration and then move on to specific user demographics.We will see how does gender, age, having dependents or not and having a partner or not affects the Churn. We will also see how tenure and contract type affect the churn. Correlation between the variables will be checked through the correlation matrix.

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(miscset)
```

```
## 
## Attaching package: 'miscset'
```

```
## The following object is masked from 'package:dplyr':
## 
##     collapse
```

```
# reading in the data
df <- read_csv("C:/Users/esaminl/Downloads/TelCh.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   SeniorCitizen = col_double(),
##   tenure = col_double(),
##   MonthlyCharges = col_double(),
##   TotalCharges = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
# dimensions of the data
dim_desc(df)
```

```
## [1] "[7,043 x 21]"
```

```
names(df)
```

```
##  [1] "customerID"       "gender"           "SeniorCitizen"
##  [4] "Partner"          "Dependents"       "tenure"
##  [7] "PhoneService"     "MultipleLines"    "InternetService"
## [10] "OnlineSecurity"   "OnlineBackup"     "DeviceProtection"
## [13] "TechSupport"      "StreamingTV"      "StreamingMovies"
## [16] "Contract"         "PaperlessBilling" "PaymentMethod"
## [19] "MonthlyCharges"   "TotalCharges"     "Churn"
```

```
glimpse(df)
```

```
## Observations: 7,043
## Variables: 21
## $ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "77...
## $ gender           <chr> "Female", "Male", "Male", "Male", "Female", "...
## $ SeniorCitizen    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "N...
## $ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "N...
## $ tenure           <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 5...
## $ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes"...
## $ MultipleLines    <chr> "No phone service", "No", "No", "No phone ser...
## $ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "F...
## $ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", ...
## $ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", ...
## $ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", ...
## $ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "N...
## $ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "...
## $ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "N...
## $ Contract         <chr> "Month-to-month", "One year", "Month-to-month...
## $ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes"...
## $ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed c...
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89....
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820....
## $ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", ...
```

# Dataset and variable transformation

Converting to factor level for analysis. Null values in Total Charges column replaced with the mean of the total charges.

```
df <- df %>% mutate_if(is.character, as.factor)
df$SeniorCitizen <- as.factor(df$SeniorCitizen)
glimpse(df)
```

```
## Observations: 7,043
## Variables: 21
## $ customerID       <fct> 7590-VHVEG, 5575-GNVDE, 3668-QPYBK, 7795-CFOC...
## $ gender           <fct> Female, Male, Male, Male, Female, Female, Mal...
## $ SeniorCitizen    <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ Partner          <fct> Yes, No, No, No, No, No, No, No, Yes, No, Yes...
## $ Dependents       <fct> No, No, No, No, No, No, Yes, No, No, Yes, Yes...
## $ tenure           <dbl> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 5...
## $ PhoneService     <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, Yes...
## $ MultipleLines    <fct> No phone service, No, No, No phone service, N...
## $ InternetService  <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber optic,...
## $ OnlineSecurity   <fct> No, Yes, Yes, Yes, No, No, No, Yes, No, Yes, ...
## $ OnlineBackup     <fct> Yes, No, Yes, No, No, No, Yes, No, No, Yes, N...
## $ DeviceProtection <fct> No, Yes, No, Yes, No, Yes, No, No, Yes, No, N...
## $ TechSupport      <fct> No, No, No, Yes, No, No, No, No, Yes, No, No,...
## $ StreamingTV      <fct> No, No, No, No, No, Yes, Yes, No, Yes, No, No...
## $ StreamingMovies  <fct> No, No, No, No, No, Yes, No, No, Yes, No, No,...
## $ Contract         <fct> Month-to-month, One year, Month-to-month, One...
## $ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, No,...
## $ PaymentMethod    <fct> Electronic check, Mailed check, Mailed check,...
## $ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89....
## $ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820....
## $ Churn            <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No, N...
```

```
df %>% map(~ sum(is.na(.)))
```

```
## $customerID
## [1] 0
##
## $gender
## [1] 0
##
## $SeniorCitizen
## [1] 0
##
## $Partner
## [1] 0
##
## $Dependents
## [1] 0
##
## $tenure
## [1] 0
##
## $PhoneService
## [1] 0
##
## $MultipleLines
## [1] 0
##
## $InternetService
## [1] 0
##
## $OnlineSecurity
## [1] 0
##
## $OnlineBackup
## [1] 0
##
## $DeviceProtection
## [1] 0
##
## $TechSupport
## [1] 0
##
## $StreamingTV
## [1] 0
##
## $StreamingMovies
## [1] 0
##
## $Contract
## [1] 0
##
## $PaperlessBilling
## [1] 0
##
## $PaymentMethod
## [1] 0
```

```
##
## $MonthlyCharges
## [1] 0
##
## $TotalCharges
## [1] 11
##
## $Churn
## [1] 0
```

```
# imputing with the mean
df <- df %>%
  mutate(TotalCharges = replace(TotalCharges,
                                is.na(TotalCharges),
                                mean(TotalCharges, na.rm = T)))

# checking that the imputation worked
sum(is.na(df$TotalCharges))
```

```
## [1] 0
```

Checking proportion of churned or not in the dataset. 64% of the data consists of people who have not churned.

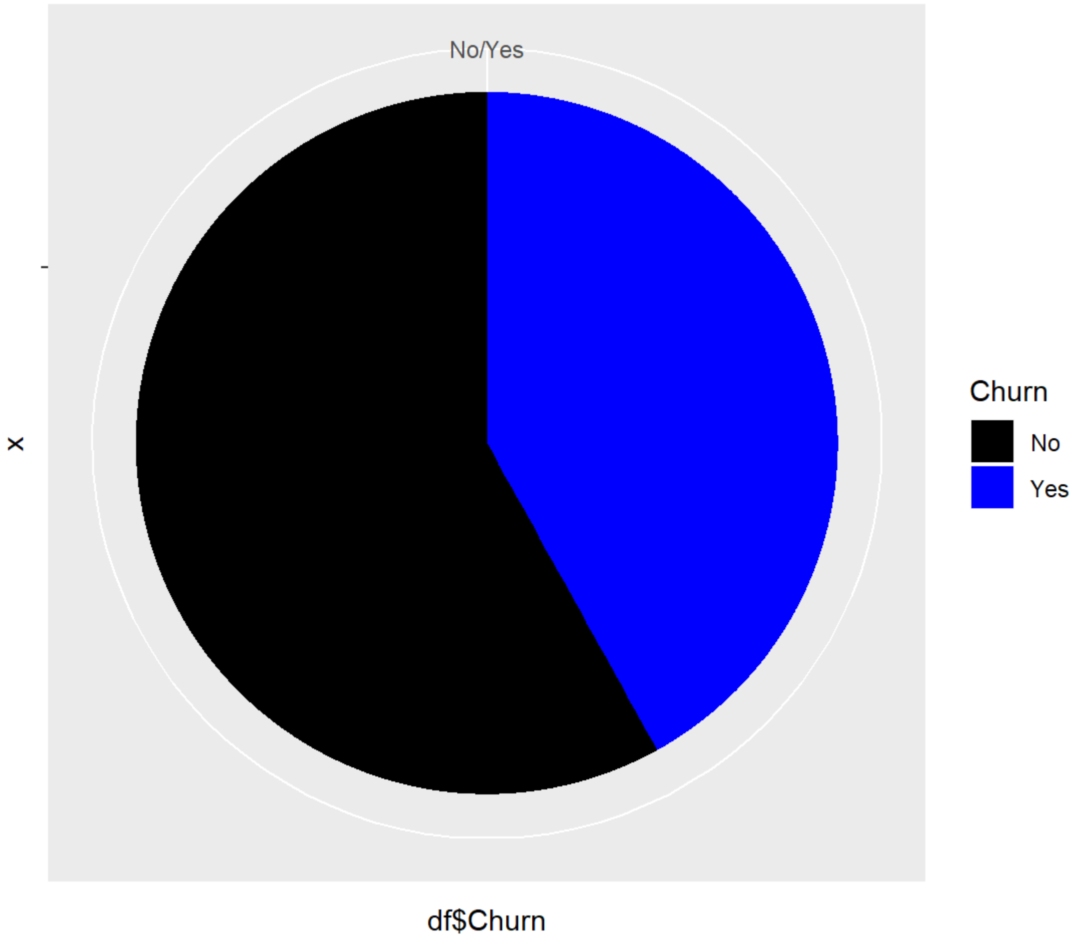
```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

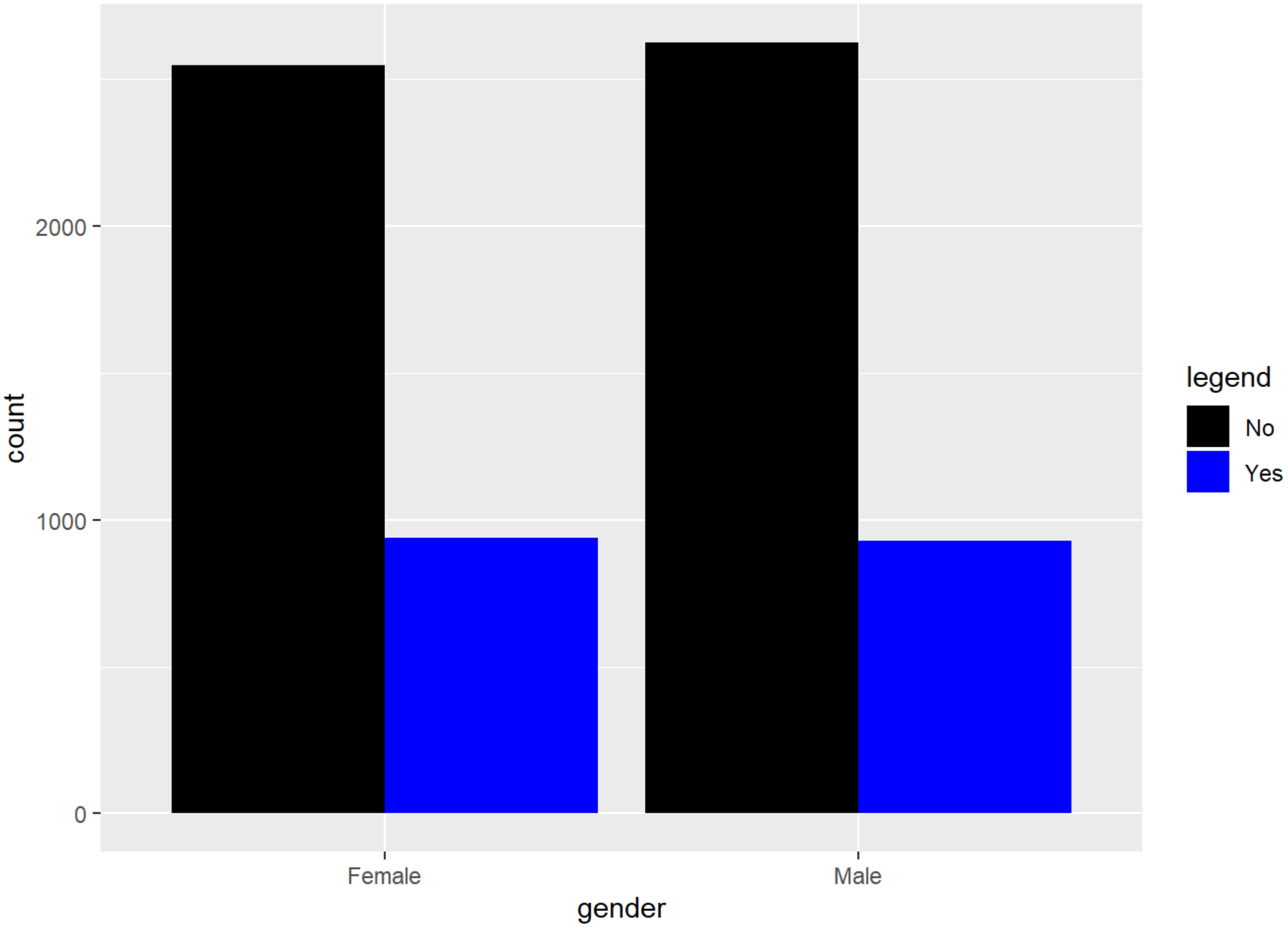
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
bp<- ggplot(df, aes(x="", y=df$Churn, fill=Churn))+
geom_bar(width = 1, stat = "identity")

pie <- bp + coord_polar("y", start=0)+ scale_fill_manual(values=c("No" = "black", "Yes" = "blue"
))
pie
```



```
ggplot(df) +
  geom_bar(aes(x = gender, fill = Churn), position = "dodge")+
scale_fill_manual("legend", values = c("No" = "black", "Yes" = "blue"))
```

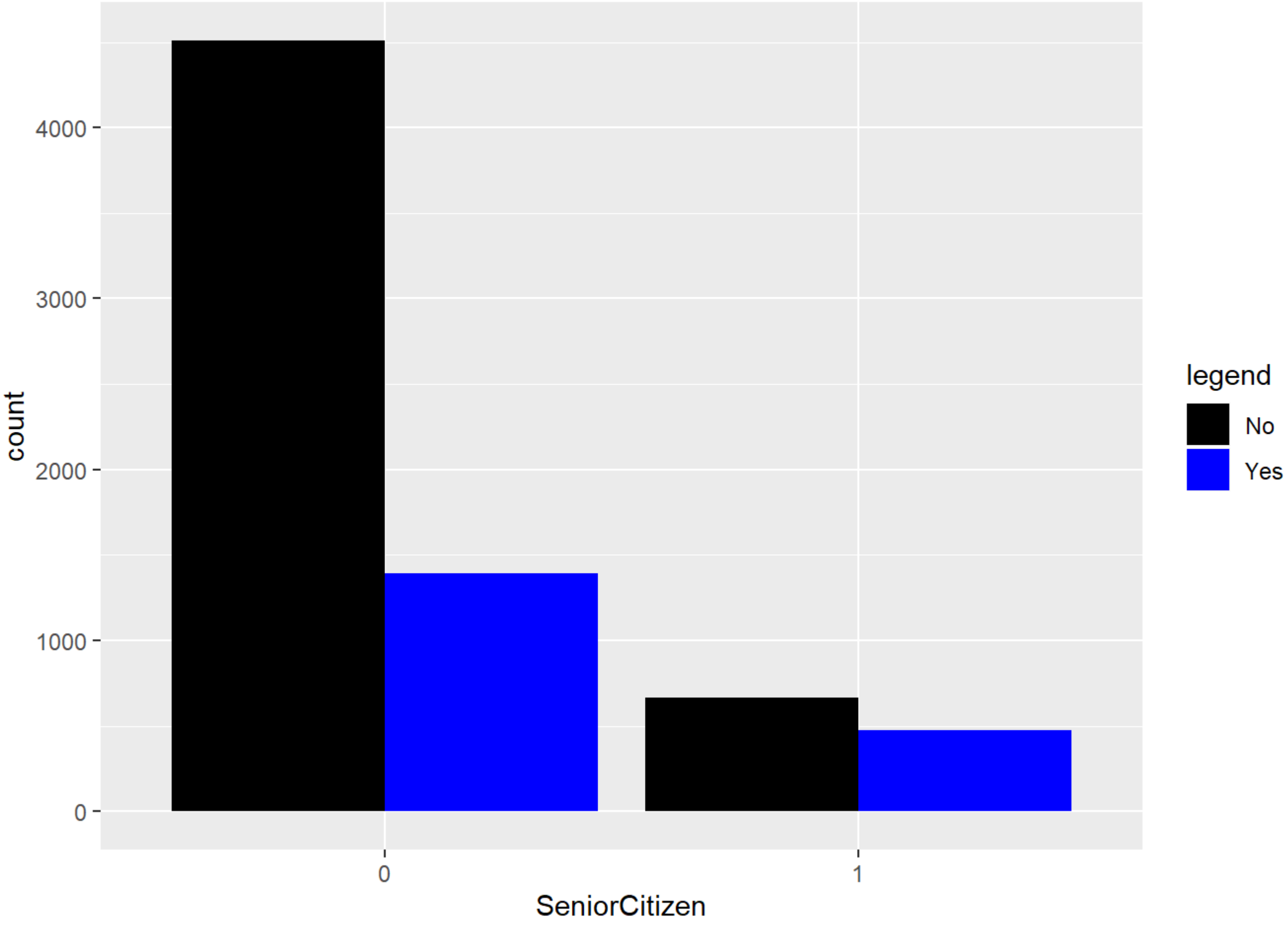


```
df %>%
  group_by(gender,Churn) %>%
  summarise(n=n())
```

```
## # A tibble: 4 x 3
## # Groups:   gender [2]
##   gender Churn     n
##   <fct>  <fct> <int>
## 1 Female No     2549
## 2 Female Yes     939
## 3 Male   No     2625
## 4 Male   Yes     930
```

Gender does not significantly affect the Churn.

```
#SeniorCitizen
ggplot(df) +
  geom_bar(aes(x = SeniorCitizen, fill = Churn), position = "dodge")+
scale_fill_manual("legend", values = c("No" = "black", "Yes" = "blue"))
```



```
df %>%
  group_by(SeniorCitizen) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

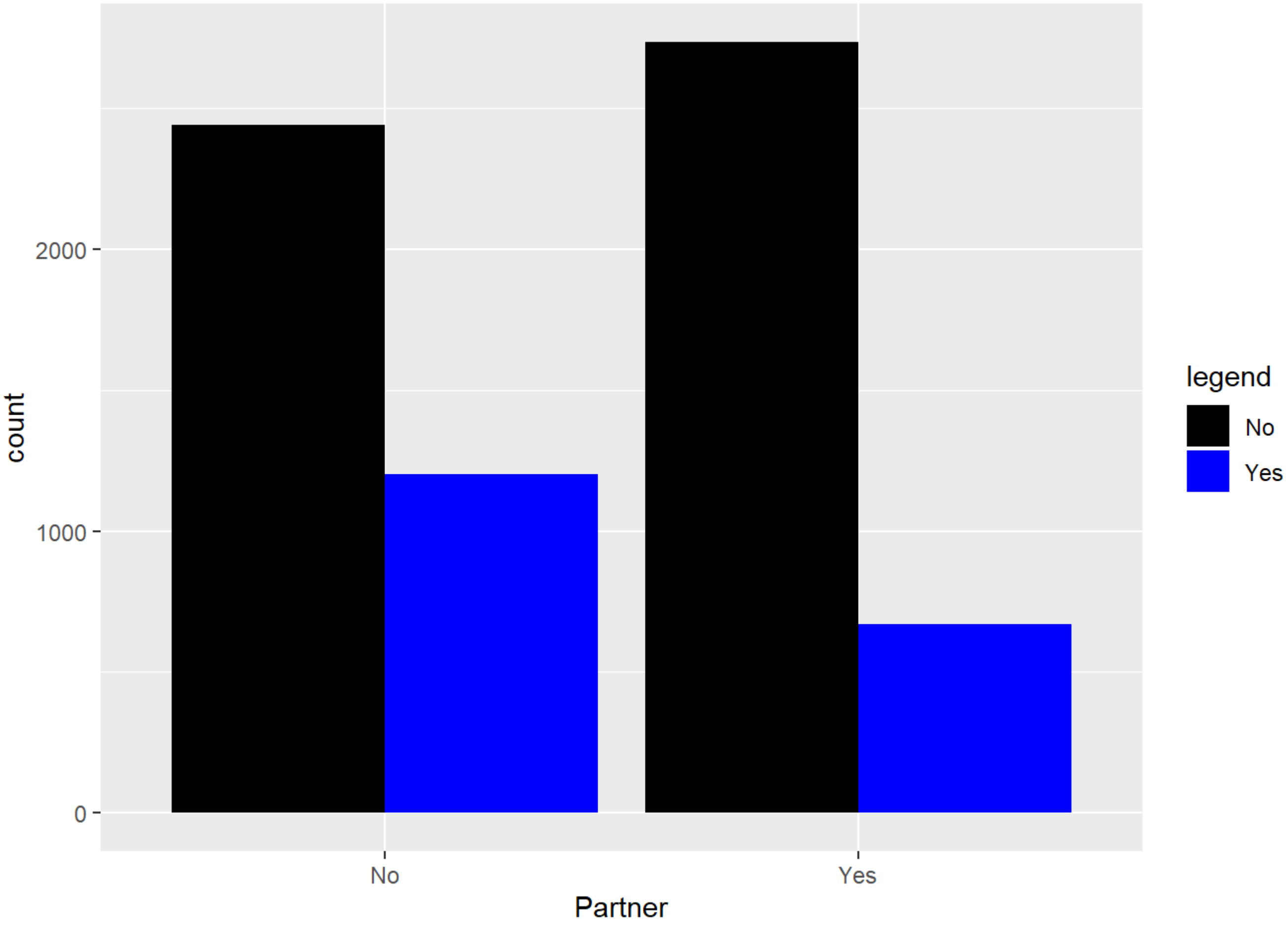
```
## # A tibble: 2 x 3
##   SeniorCitizen     n  freq
##   <fct>         <int> <dbl>
## 1 0              5901 0.838
## 2 1              1142 0.162
```

```
df %>%
  group_by(SeniorCitizen, Churn) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   SeniorCitizen [2]
##   SeniorCitizen Churn     n  freq
##   <fct>         <fct> <int> <dbl>
## 1 0             No     4508 0.764
## 2 0             Yes    1393 0.236
## 3 1             No      666 0.583
## 4 1             Yes     476 0.417
```

Approximately 16% of the customers are senior citizens, and roughly 42% of those senior citizens churn. On the other hand, of the 84% of customers that are not senior citizens, only 24% churn. These results show that senior citizens are much more likely to churn.

```
#Partner
ggplot(df) +
  geom_bar(aes(x=Partner, fill = Churn), position = "dodge")+
scale_fill_manual("legend", values = c("No" = "black", "Yes" = "blue"))
```



```
df %>%
  group_by(Partner) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

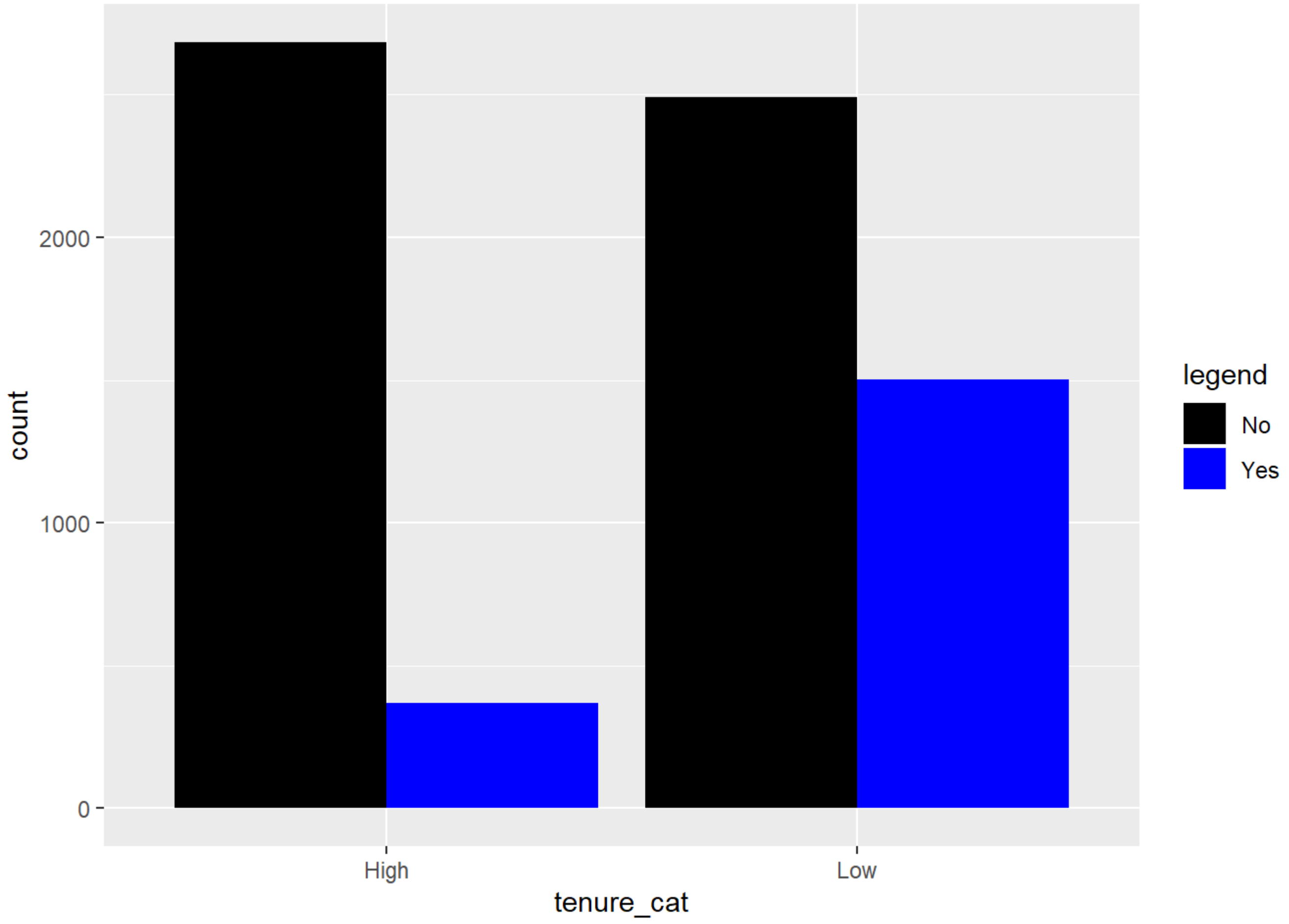
```
## # A tibble: 2 x 3
##   Partner     n  freq
##   <fct>   <int> <dbl>
## 1 No       3641 0.517
## 2 Yes      3402 0.483
```

```
df %>%
  group_by(Partner, Churn) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   Partner [2]
##   Partner Churn     n  freq
##   <fct>   <fct> <int> <dbl>
## 1 No      No     2441 0.670
## 2 No      Yes    1200 0.330
## 3 Yes     No     2733 0.803
## 4 Yes     Yes     669 0.197
```

Roughly half of the people have partners. Of the people with partners, 20% churn. For people without partners, approximately 33% churn.

```
tenure_cat <- with(df, ifelse(tenure <= 35, "Low", "High"))
df <- data.frame(df, tenure_cat)
ggplot(df) +
  geom_bar(aes_string(x="tenure_cat", fill="Churn"), position = "dodge")+
scale_fill_manual("legend", values = c("No" = "black", "Yes" = "blue"))
```



```
df %>% group_by(tenure_cat, Churn) %>%
  summarise(n=n()) %>%
  mutate(freq = n / sum(n))
```

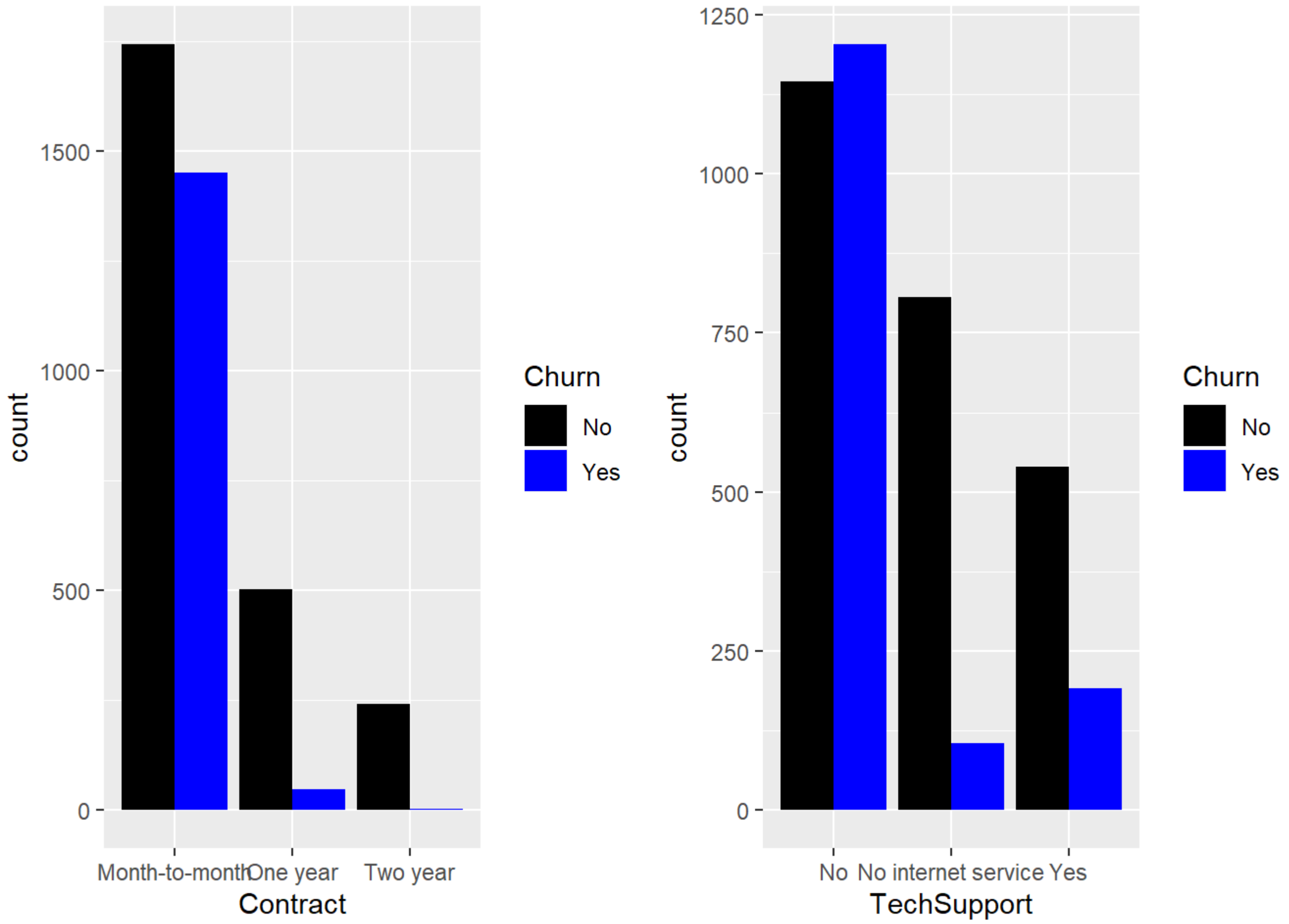
```
## # A tibble: 4 x 4
## # Groups:   tenure_cat [2]
##   tenure_cat Churn     n  freq
##   <fct>      <fct> <int> <dbl>
## 1 High       No     2683 0.879
## 2 High       Yes     368 0.121
## 3 Low        No     2491 0.624
## 4 Low        Yes    1501 0.376
```

```
df %>% group_by(tenure_cat) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

```
## # A tibble: 2 x 3
##   tenure_cat     n  freq
##   <fct>      <int> <dbl>
## 1 High        3051 0.433
## 2 Low         3992 0.567
```

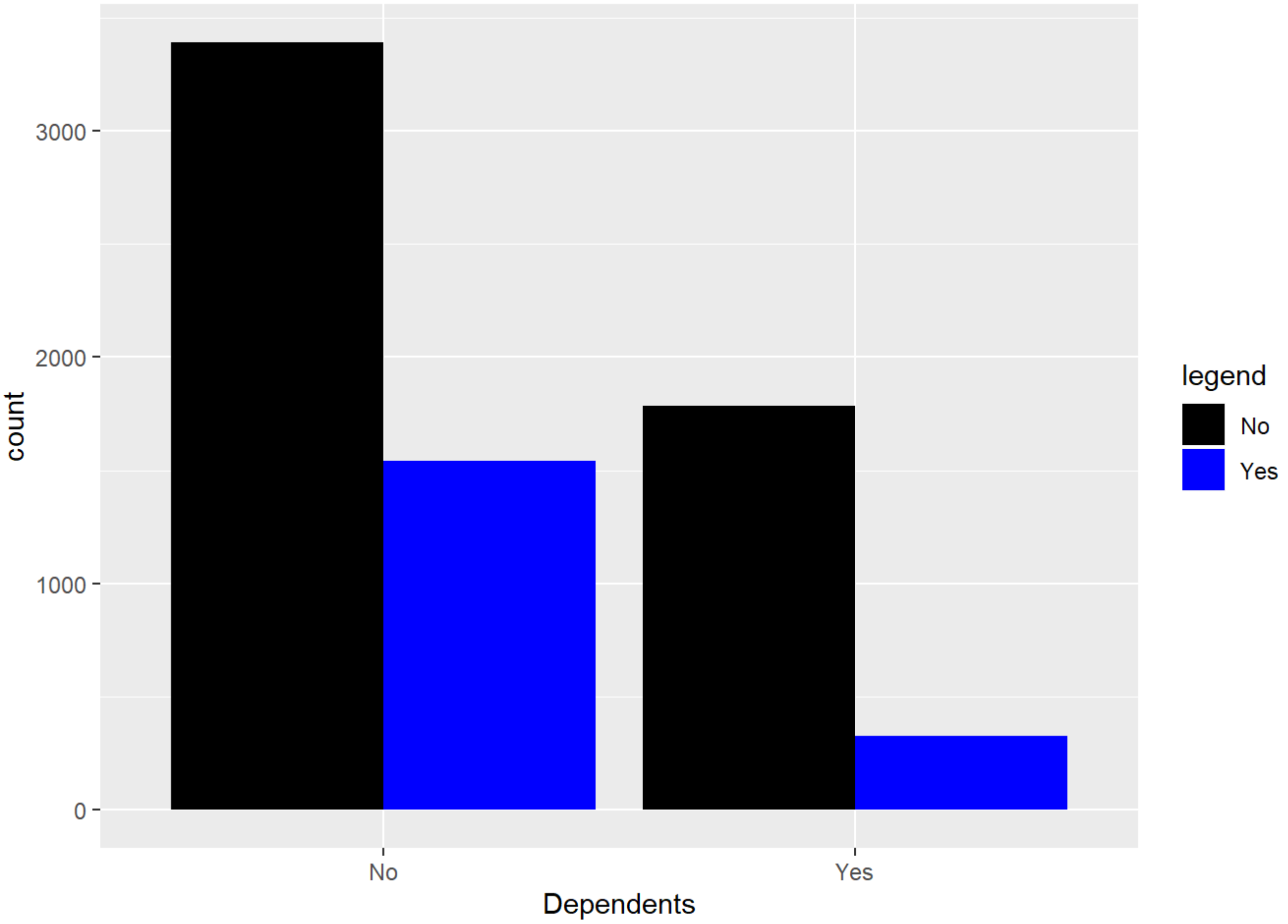
```
ten_cat <- df %>% filter(tenure_cat == "Low")

ggplotGrid(ncol=2,
lapply(c("Contract","TechSupport"),
       function(col){
         ggplot(ten_cat,aes_string(col)) + geom_bar(aes(fill=Churn),position="dodge")+
scale_fill_manual("Churn", values = c("No" = "black", "Yes" = "blue"))
       }))
```



About 38% of the users with low tenures churn out. In the low tenure people, the ones with month-to-month contract churn out in larger numbers. But they are also the ones who do not churn out a lot. So we look at the tech support that they have received during this time. The ones with no tech support churned out the most during this period. So for people who have not spent a lot of time with the company value Tech Support more.

```
ggplot(df) +
  geom_bar(aes_string(x="Dependents", fill="Churn"), position = "dodge")+
scale_fill_manual("legend", values = c("No" = "black", "Yes" = "blue"))
```



```
df %>% group_by(Dependents, Churn) %>%
  summarise(n=n()) %>%
  mutate(freq = n / sum(n))
```

```
## # A tibble: 4 x 4
## # Groups:   Dependents [2]
##   Dependents Churn     n  freq
##   <fct>      <fct> <int> <dbl>
## 1 No         No     3390 0.687
## 2 No         Yes    1543 0.313
## 3 Yes        No     1784 0.845
## 4 Yes        Yes     326 0.155
```

```
df %>% group_by(Dependents) %>%
  summarise(n = n()) %>%
  mutate(freq = n / sum(n))
```

```
## # A tibble: 2 x 3
##   Dependents     n  freq
##   <fct>      <int> <dbl>
## 1 No          4933 0.700
## 2 Yes         2110 0.300
```

Approximately 30% of the people have dependents, of which 15% churn. For the other 70% that don't have dependents, 31% churn

```
# Total charges and tenure of senior citizens
df %>%
  select(SeniorCitizen, Churn, TotalCharges, tenure) %>%
  filter(SeniorCitizen == 1, Churn == "Yes") %>%
  summarize(n = n(),
            total = sum(TotalCharges),
            avg_tenure = sum(tenure)/n)
```

```
##     n    total avg_tenure
## 1 476 882405.2   21.03361
```

```
# Total charges and tenure of people without a partner
df %>%
  select(Partner, Churn, TotalCharges, tenure) %>%
  filter(Partner == "No", Churn == "Yes") %>%
  summarise(n = n(),
            total = sum(TotalCharges),
            avg_tenure = sum(tenure)/n)
```

```
##      n   total avg_tenure
## 1 1200 1306776   13.17667
```

```
# Total charges and tenure of people without dependents
df %>%
  select(Dependents, Churn, TotalCharges, tenure) %>%
  filter(Dependents == "No", Churn == "Yes") %>%
  summarise(n = n(),
            total = sum(TotalCharges),
            avg_tenure = sum(tenure)/n)
```

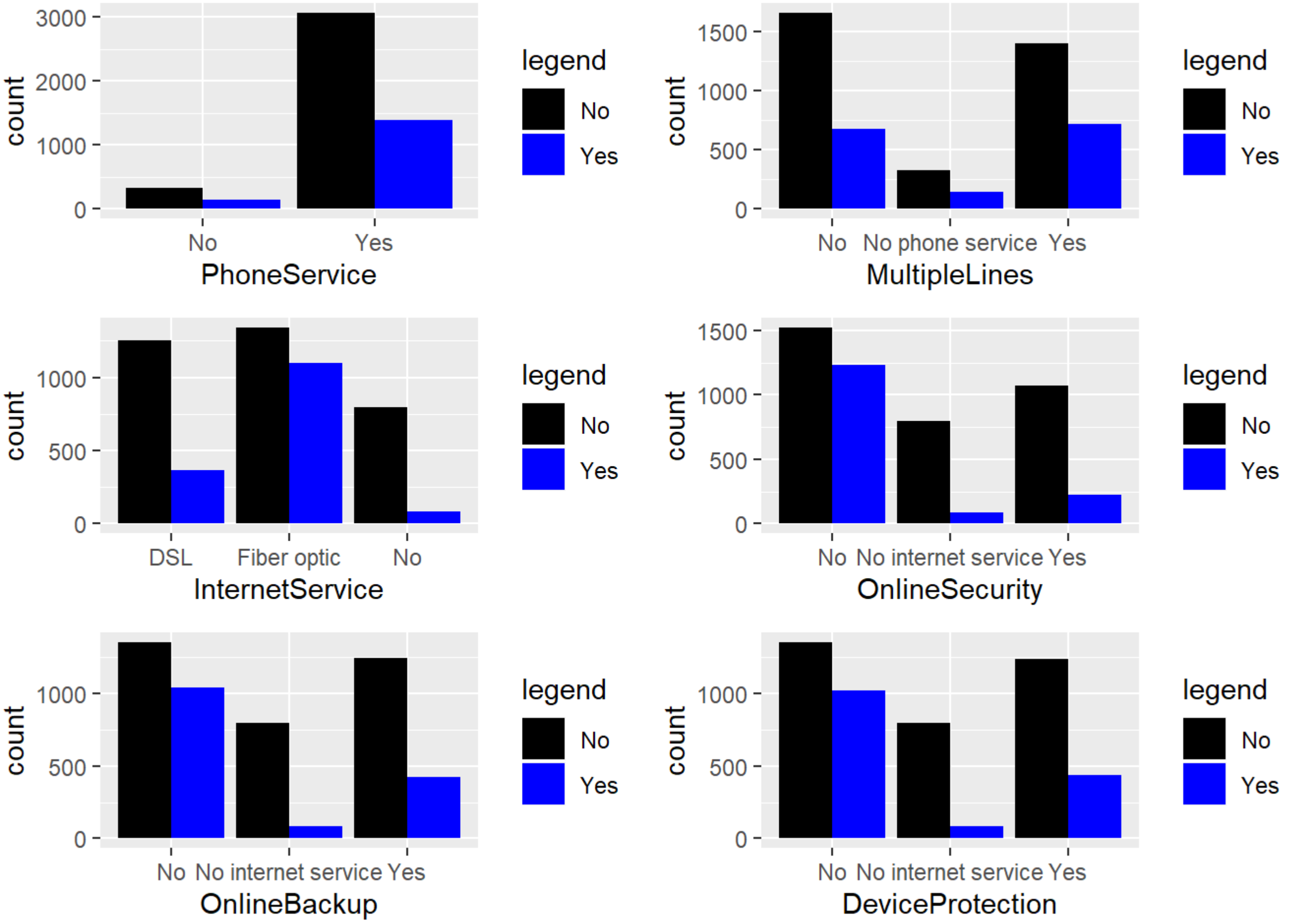
```
##      n   total avg_tenure
## 1 1543 2261840   17.12314
```

Based on the results, we should focus our efforts on people without dependents. This customer segment that churned had nearly 2.3MM in total charges compared to 1.3MM for people without partners, and only 900K for senior citizens.

Let's dig a little deeper and see what services that customer segment uses.

```
dependents <- df %>% filter(Dependents == "No")

ggplotGrid(ncol=2,
lapply(c("PhoneService","MultipleLines","InternetService","OnlineSecurity","OnlineBackup",
         "DeviceProtection"),
       function(col){
         ggplot(dependents,aes_string(col)) + geom_bar(aes(fill=Churn),position="dodge")+
scale_fill_manual("legend", values = c("No" = "black", "Yes" = "blue"))
       }))
```



People are not happy with phone service provided, so looking into those complaints would have been more helpful. Providing Online Security for people without dependents. We can provide a free trial period for this segment with their plan.