TELECOM CHURN ANALYSIS

About the Analysis:

This analysis focuses on the behaviour of telecom customers who are more likely to leave the platform. I intend to find out the most striking behaviour of customers through EDA and later on use some of the predictive analytics techniques to determine the customers who are most likely to churn.

About the Analysis:

This analysis focuses on the behaviour of telecom customers who are more likely to leave the platform. I intend to find out the most striking behaviour of customers through.

- Loading the library

```
library(dplyr)
library(ggplot2)
```

- Importing the file:

```
telecom=read.csv(file.choose(),header = T,stringsAsFactors = T)
```

- Getting the structure of the dataset

```
> glimpse(telecom)
Rows: 7,032
Columns: 20
$ gender           <fct> Female, Male, Male, Male, Female, Female, ~
$ SeniorCitizen    <chr> "No", "No", "No", "No", "No", "No", "No", ~
$ Partner          <fct> Yes, No, No, No, No, No, No, No, Yes, No, ~
$ Dependents       <fct> No, No, No, No, No, No, Yes, No, No, Yes, ~
$ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16~
$ PhoneService     <fct> No, Yes, Yes, No, Yes, Yes, Yes, No, Yes, ~
$ MultipleLines    <fct> No phone service, No, No, No phone service~
$ InternetService  <fct> DSL, DSL, DSL, DSL, Fiber optic, Fiber opt~
$ OnlineSecurity   <fct> No, Yes, Yes, Yes, No, No, No, Yes, No, Ye~
$ OnlineBackup     <fct> Yes, No, Yes, No, No, No, No, No, No, Yes~
$ DeviceProtection <fct> No, Yes, No, Yes, No, Yes, No, No, Yes, No~
$ TechSupport      <fct> No, No, No, Yes, No, No, No, No, Yes, No, ~
$ StreamingTV      <fct> No, No, No, No, No, Yes, Yes, No, Yes, No,~
$ StreamingMovies  <fct> No, No, No, No, No, Yes, No, No, Yes, No, ~
$ Contract         <fct> Month-to-month, One year, Month-to-month, ~
$ PaperlessBilling <fct> Yes, No, Yes, No, Yes, Yes, Yes, No, Yes, ~
$ PaymentMethod    <fct> Electronic check, Mailed check, Mailed che~
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, ~
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 8~
$ Churn            <fct> No, No, Yes, No, Yes, Yes, No, No, Yes, No~
>
```

```
> str(telecom)
'data.frame':   7032 obs. of  20 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : chr  "No" "No" "No" "No" ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
 $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
 $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
 $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...
 $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...
 $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...
 $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...
 $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...
 $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
 - attr(*, "na.action")= 'omit' Named int [1:11] 489 754 937 1083 1341 3332 3827 4381 5219 6671 ...
  ..- attr(*, "names")= chr [1:11] "489" "754" "937" "1083" ...
```

- Finding the missing values:

```
> sum(is.na(telecom))
[1] 11
> sapply(telecom, function(x) sum(is.na(x)))
      customerID           gender    SeniorCitizen
               0                0                0
         Partner       Dependents           tenure
               0                0                0
   PhoneService    MultipleLines  InternetService
               0                0                0
  OnlineSecurity     OnlineBackup DeviceProtection
               0                0                0
     TechSupport      StreamingTV  StreamingMovies
               0                0                0
        Contract PaperlessBilling    PaymentMethod
               0                0                0
  MonthlyCharges     TotalCharges            Churn
               0               11                0
>
```

- There are only 11 missing data in the Total Charges field.

```
> sum(is.na(telecom$TotalCharges))/nrow(telecom)
[1] 0.001561834
```

- As the percentage of the missing value is very less, hence removing them from the dataset.

```
> telecom=na.omit(telecom)
> sapply(telecom, function(x) sum(is.na(x)))
      customerID           gender    SeniorCitizen
               0                0                0
         Partner       Dependents           tenure
               0                0                0
    PhoneService    MultipleLines  InternetService
               0                0                0
  OnlineSecurity     OnlineBackup DeviceProtection
               0                0                0
     TechSupport      StreamingTV  StreamingMovies
               0                0                0
        Contract PaperlessBilling    PaymentMethod
               0                0                0
  MonthlyCharges     TotalCharges            Churn
               0                0                0
>
```

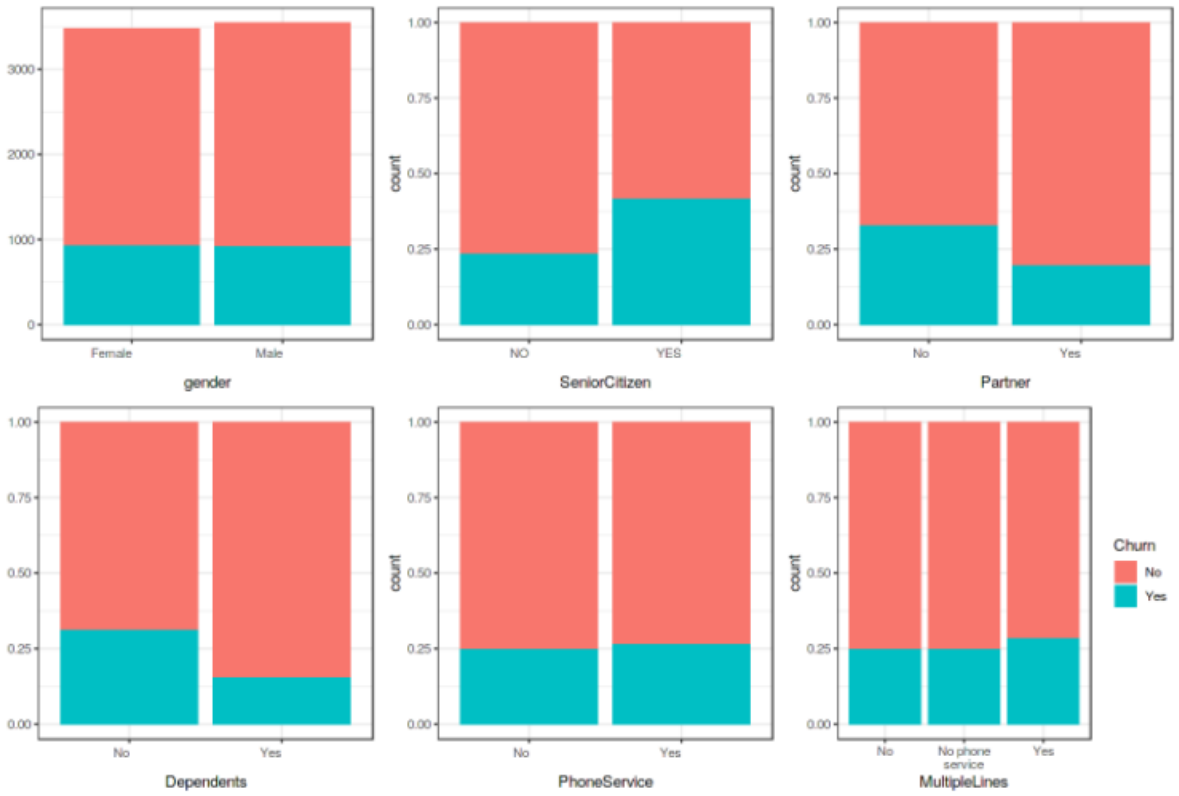
```
> table(telecom$Churn)

  No  Yes
5163 1869
```

- So here the number of non-churn user is more than churn user.
- Converting the senior citizen from 1,0 to yes and no

```
telecom$SeniorCitizen=ifelse(telecom$SeniorCitizen==1,"Yes","No")
```

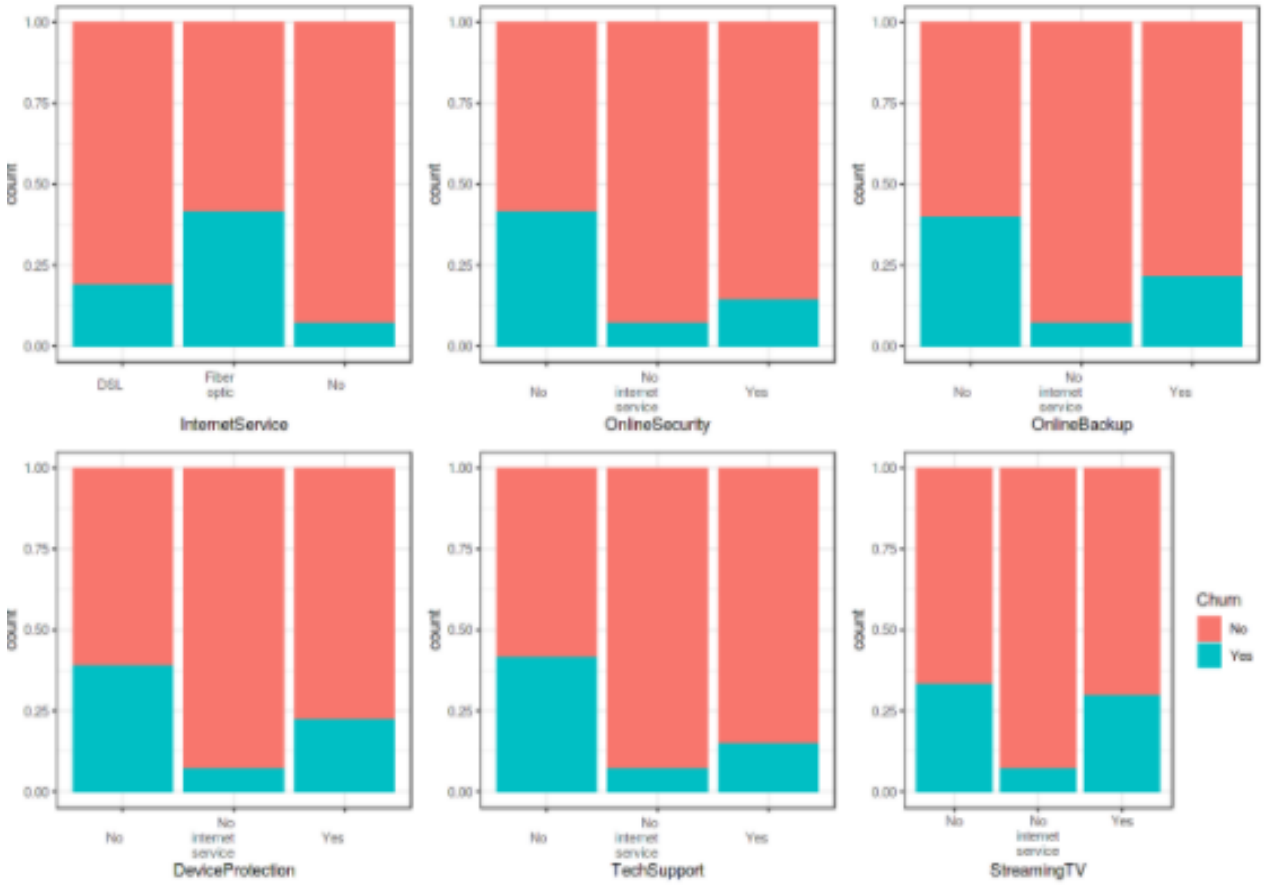
Exploratory Data Analysis:

```
ggplot(telecom, aes(x=gender,fill=Churn))+ geom_bar()

ggplot(telecom, aes(x=SeniorCitizen,fill=Churn))+ geom_bar(position = 'fill')

ggplot(telecom, aes(x=Partner,fill=Churn))+ geom_bar(position = 'fill')

ggplot(telecom, aes(x=Dependents,fill=Churn))+ geom_bar(position = 'fill')

ggplot(telecom, aes(x=PhoneService,fill=Churn))+ geom_bar(position = 'fill')

ggplot(telecom, aes(x=MultipleLines,fill=Churn))+ geom_bar(position = 'fill')
```



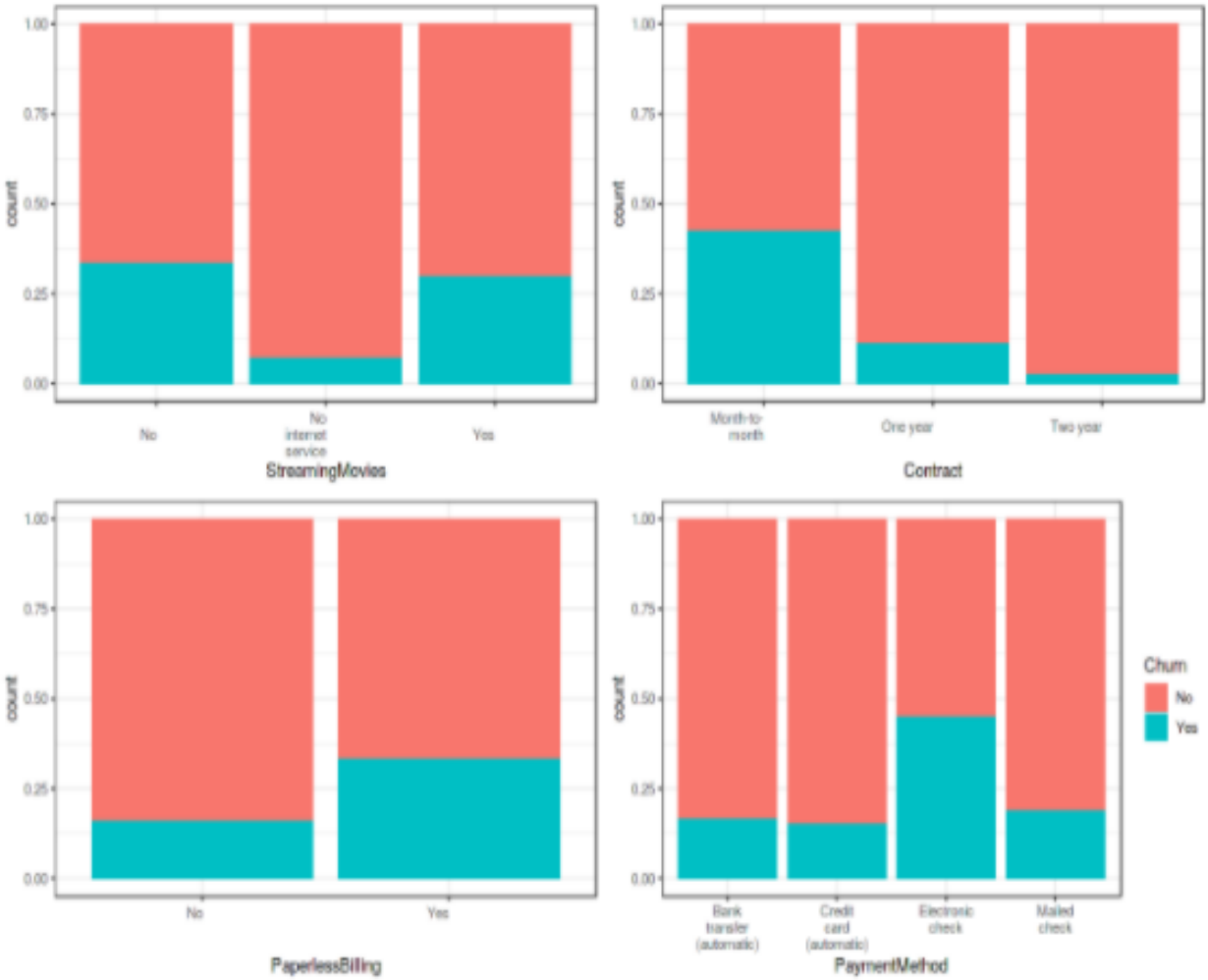
- **Gender** - The churn percent is almost equal in case of Male and Females
- The percent of churn is higher in case of **senior citizens**
- **Customers with Partners and Dependents** have lower churn rate as compared to those who don't have partners & Dependents.

```
> ggplot(telecom, aes(x=InternetService,fill=Churn))+ geom_bar()
> ggplot(telecom, aes(x=OnlineSecurity,fill=Churn))+ geom_bar(position = 'fill')
> ggplot(telecom, aes(x=OnlineBackup,fill=Churn))+ geom_bar(position = 'fill')
> ggplot(telecom, aes(x=DeviceProtection,fill=Churn))+ geom_bar(position = 'fill')
> ggplot(telecom, aes(x=TechSupport,fill=Churn))+ geom_bar(position = 'fill')
> ggplot(telecom, aes(x=StreamingTV,fill=Churn))+ geom_bar(position = 'fill')
>
```



- **Churn rate is much higher in case of Fibre Optic Internet Services.**
- **Customers who do not have services like No Online Security, Online Backup and Tec Support have left the platform in the past month.**

```
> ggplot(telecom, aes(x=StreamingMovies,fill=Churn))+ geom_bar()
> ggplot(telecom, aes(x=Contract,fill=Churn))+ geom_bar(position = 'fill')
> ggplot(telecom, aes(x=PaperlessBilling,fill=Churn))+ geom_bar(position =
 'fill')
> ggplot(telecom, aes(x=PaymentMethod,fill=Churn))+ geom_bar(position = 'fil
l')
>
```

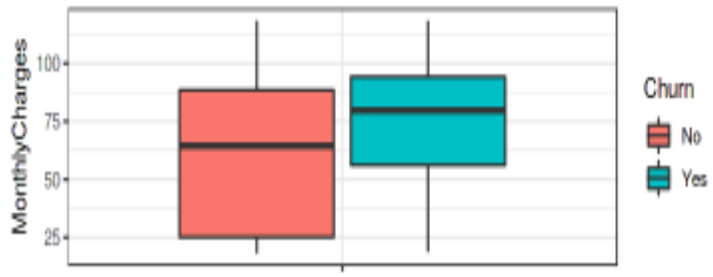


- A larger percent of Customers with monthly subscription have left when compared to Customers with one- or two-year contract.
- Churn percent is higher in case of customer's having paperless billing option.
- Customers who have Electronic Check Payment Method tend to leave the platform more when compared to other options.

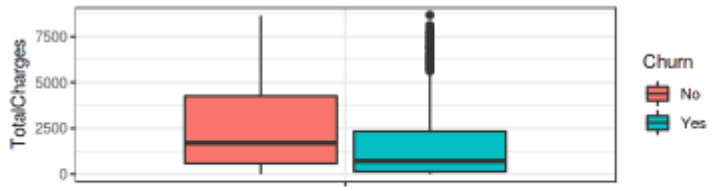
- **Analysing the continuous variable wrt churn:**

```
> ggplot(telecom, aes(y= tenure, x = "", fill = Churn)) +
+   geom_boxplot()
> ggplot(telecom, aes(y= MonthlyCharges, x = "", fill = Churn)) +
+   geom_boxplot()
> ggplot(telecom, aes(y= TotalCharges, x = "", fill = Churn)) +
+   geom_boxplot()
>
```

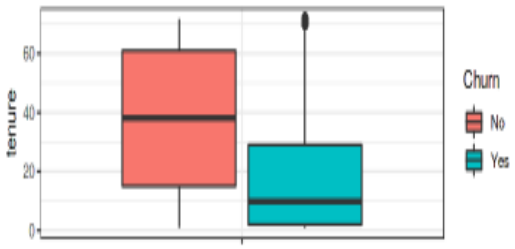
- **Monthly Charges**: **Customers who have churned, have high monthly charges. The median is above 75.**



- **Total Charges:** The median Total charges of customers who have churned is low.



- **Tenure:** The median tenure for customers who have left is around 10 months.



<u>Checking the correlation between continuous variables:</u>

```
> correaltion=telecom[,c('tenure','MonthlyCharges','TotalCharges')]
>
> as.data.frame(cor(correaltion))
                tenure MonthlyCharges TotalCharges
tenure          1.0000000      0.2468618    0.8258805
MonthlyCharges  0.2468618      1.0000000    0.6510648
TotalCharges    0.8258805      0.6510648    1.0000000
>
```

- **Total Charges has positive correlation with Monthly Charges and tenure.**