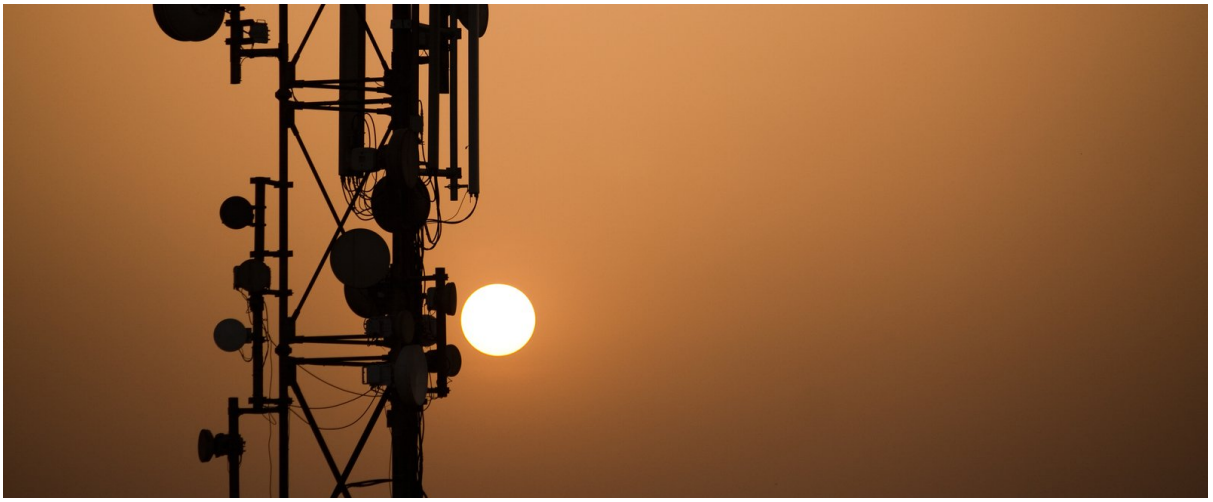


Telco Customer Churn Prediction

Introduction

This analysis focuses on the behavior of telecom customers who are more likely to leave the platform. I intend to find out the most prominent customer behaviors through EDA and later use some predictive analytics techniques to identify the customers most likely to churn.

Dataset

Source: Telco Customer Churn (kaggle.com)

Tools

1. Visual Studio Code (To predict the dataset by using R)

Analysis

```
# Import packages
install.packages("googlesheets4")
install.packages("ggplot2")
install.packages("cowplot")
library(tidyverse)
library(googlesheets4)
library(ggplot2)
library(cowplot)
library(randomForest)
library(caret)
library(pROC)
```

- import library that used for this project.

```
# Import Data
telco <- read.csv(file = 'C:/Data analytic/Project/R/Source/Telco_Customer_Churn.csv', header = TRUE)
```

- import data to Visual Studio Code

```
# Explore Data
head(telco)
glimpse(telco)
```

- Explore data for this project

```
> head(telco)
  customerID gender SeniorCitizen Partner Dependents tenure PhoneService
1 7590-VHVEG Female             0     Yes         No      1           No
2 5575-GNVDE   Male             0      No         No     34          Yes
3 3668-QPYBK   Male             0      No         No      2          Yes
4 7795-CFOCW   Male             0      No         No     45           No
5 9237-HQITU Female             0      No         No      2          Yes
6 9305-CDSKC Female             0      No         No      8          Yes
     MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
1 No phone service             DSL             No          Yes               No
2               No             DSL            Yes           No              Yes
3               No             DSL            Yes          Yes               No
4 No phone service             DSL            Yes           No              Yes
5               No     Fiber optic             No           No               No
6              Yes     Fiber optic             No           No              Yes
  TechSupport StreamingTV StreamingMovies       Contract PaperlessBilling
1          No          No              No Month-to-month              Yes
2          No          No              No       One year               No
3          No          No              No Month-to-month              Yes
4         Yes          No              No       One year               No
5          No          No              No Month-to-month              Yes
6          No         Yes             Yes Month-to-month              Yes
              PaymentMethod MonthlyCharges TotalCharges Churn
1          Electronic check          29.85        29.85    No
2              Mailed check          56.95      1889.50    No
3              Mailed check          53.85       108.15   Yes
4 Bank transfer (automatic)          42.30      1840.75    No
5          Electronic check          70.70       151.65   Yes
6          Electronic check          99.65       820.50   Yes
```

```
> glimpse(telco)
Rows: 7,043
Columns: 21
$ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW…
$ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female",…
$ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes…
$ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"…
$ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2…
$ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", …
$ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "…
$ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber opt…
$ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "…
$ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "N…
$ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Y…
$ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes…
$ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Ye…
$ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes…
$ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One …
$ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", …
$ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "…
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7…
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949…
$ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Y…
```

```
## DROP NA (missing values)
telco_clean_na <- na.omit(telco)
nrow(telco_clean_na)

glimpse(telco_clean_na)
```

- Drop missing values for this data

```
> glimpse(telco_clean_na)
Rows: 7,032
Columns: 21
$ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW…
$ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female",…
$ SeniorCitizen    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes…
$ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"…
$ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2…
$ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", …
$ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "…
$ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber opt…
$ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "…
$ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "N…
$ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Y…
$ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes…
$ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Ye…
$ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes…
$ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One …
$ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", …
$ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "…
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7…
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949…
$ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Y…
```

```
## Change factor for Visual
telco_clean_na$SeniorCitizen <- as.factor(ifelse(telco_clean_na$SeniorCitizen==1, 'YES', 'NO'))

glimpse(telco_clean_na)
```

- Change factor in column SeniorCitizen for Visual

```
> glimpse(telco_clean_na)
Rows: 7,032
Columns: 21
$ customerID       <chr> "7590-VHVEG", "5575-GNVDE", "3668-QPYBK", "7795-CFOCW…
$ gender           <chr> "Female", "Male", "Male", "Male", "Female", "Female",…
$ SeniorCitizen    <fct> NO, NO, NO, NO, NO, NO, NO, NO, NO, NO, NO, NO, NO, N…
$ Partner          <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "Yes…
$ Dependents       <chr> "No", "No", "No", "No", "No", "No", "Yes", "No", "No"…
$ tenure           <int> 1, 34, 2, 45, 2, 8, 22, 10, 28, 62, 13, 16, 58, 49, 2…
$ PhoneService     <chr> "No", "Yes", "Yes", "No", "Yes", "Yes", "Yes", "No", …
$ MultipleLines    <chr> "No phone service", "No", "No", "No phone service", "…
$ InternetService  <chr> "DSL", "DSL", "DSL", "DSL", "Fiber optic", "Fiber opt…
$ OnlineSecurity   <chr> "No", "Yes", "Yes", "Yes", "No", "No", "No", "Yes", "…
$ OnlineBackup     <chr> "Yes", "No", "Yes", "No", "No", "No", "Yes", "No", "N…
$ DeviceProtection <chr> "No", "Yes", "No", "Yes", "No", "Yes", "No", "No", "Y…
$ TechSupport      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "Yes…
$ StreamingTV      <chr> "No", "No", "No", "No", "No", "Yes", "Yes", "No", "Ye…
$ StreamingMovies  <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "Yes…
$ Contract         <chr> "Month-to-month", "One year", "Month-to-month", "One …
$ PaperlessBilling <chr> "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes", "No", …
$ PaymentMethod    <chr> "Electronic check", "Mailed check", "Mailed check", "…
$ MonthlyCharges   <dbl> 29.85, 56.95, 53.85, 42.30, 70.70, 99.65, 89.10, 29.7…
$ TotalCharges     <dbl> 29.85, 1889.50, 108.15, 1840.75, 151.65, 820.50, 1949…
$ Churn            <chr> "No", "No", "Yes", "No", "Yes", "Yes", "No", "No", "Y…
```

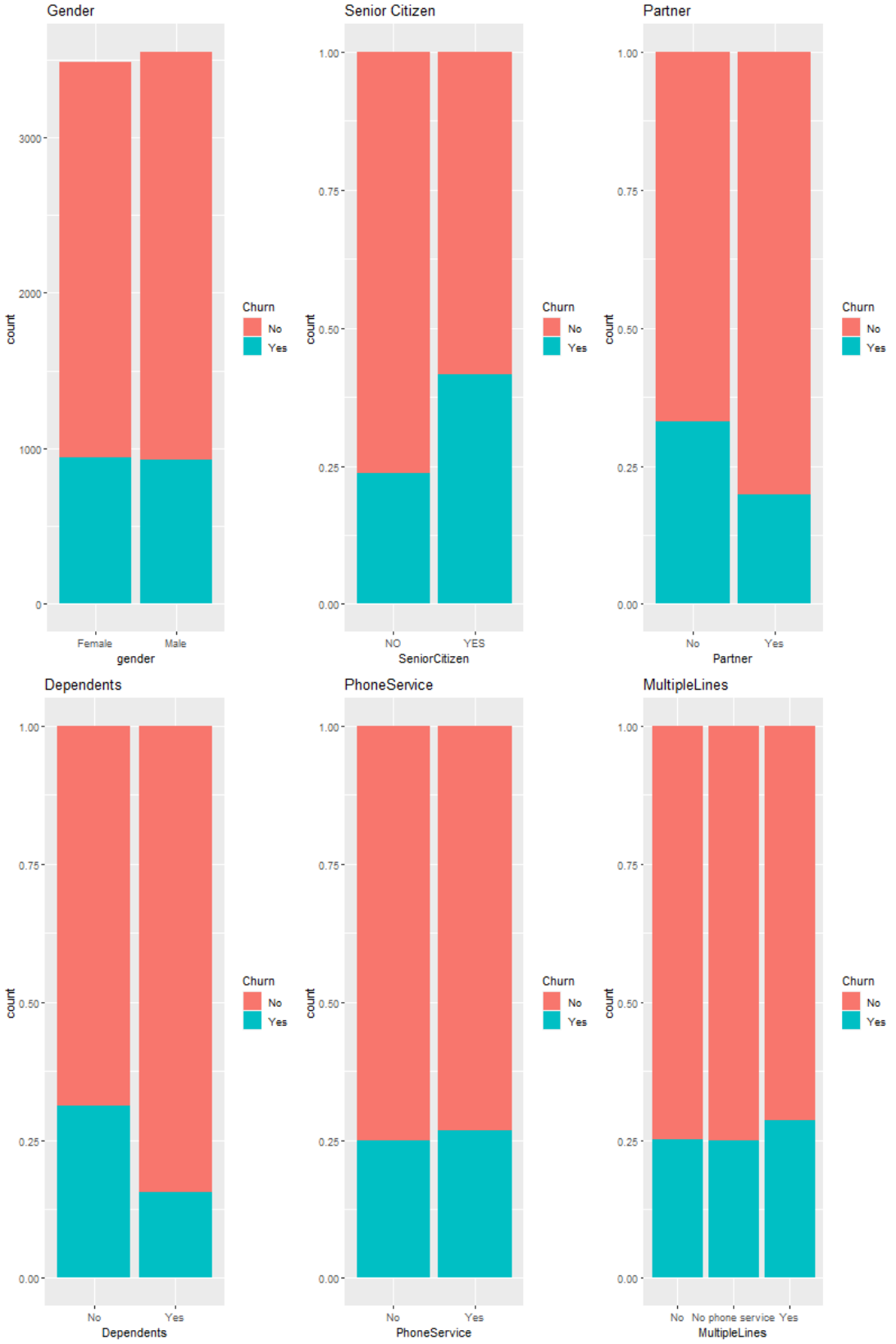
- create data visual for explore this data

```
# create data visual for explore this data
plot1 <- ggplot(telco_clean_na, aes(x = gender, fill = Churn)) +
  geom_bar() +
  labs(title = "Gender")

plot2 <- ggplot(telco_clean_na, aes(x = SeniorCitizen, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "Senior Citizen")

plot3 <- ggplot(telco_clean_na, aes(x = Partner, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "Partner")

plot4 <- ggplot(telco_clean_na, aes(x = Dependents, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "Dependents")

plot5 <- ggplot(telco_clean_na, aes(x = PhoneService, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "PhoneService")
```

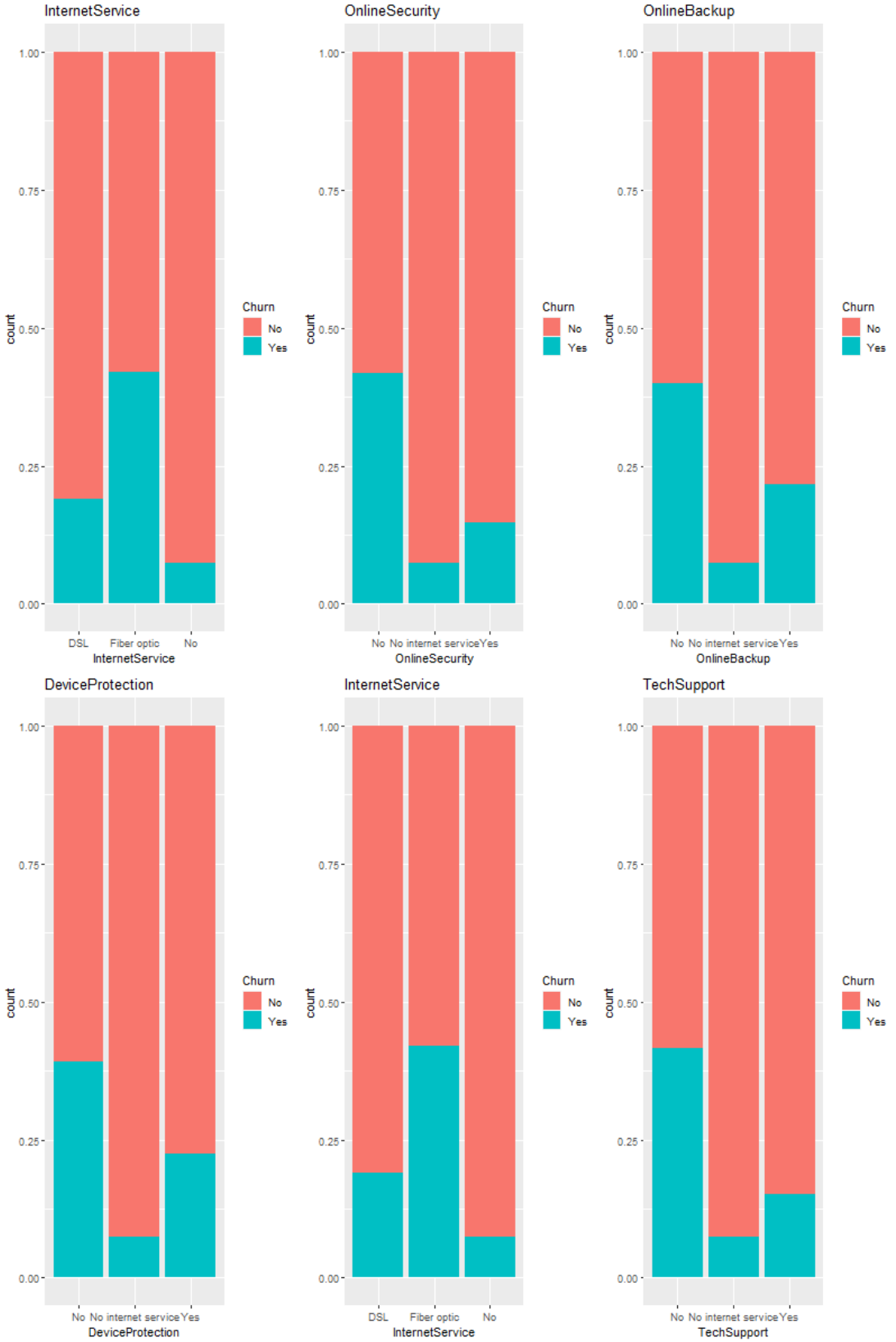
```
plot6 <- ggplot(telco_clean_na, aes(x =MultipleLines, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "MultipleLines")

plot7 <- ggplot(telco_clean_na, aes(x =InternetService, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "InternetService")

plot8 <- ggplot(telco_clean_na, aes(x =OnlineSecurity, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "OnlineSecurity")

plot9 <- ggplot(telco_clean_na, aes(x =OnlineBackup, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "OnlineBackup")

plot10 <- ggplot(telco_clean_na, aes(x =DeviceProtection, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "DeviceProtection")

plot11 <- ggplot(telco_clean_na, aes(x =InternetService, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "InternetService")

plot12 <- ggplot(telco_clean_na, aes(x =TechSupport, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "TechSupport")

plot13 <- ggplot(telco_clean_na, aes(x =StreamingTV, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "StreamingTV")

plot14 <- ggplot(telco_clean_na, aes(x =StreamingMovies, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "StreamingMovies")

plot15 <- ggplot(telco_clean_na, aes(x =Contract, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "Contract")

plot16 <- ggplot(telco_clean_na, aes(x =PaperlessBilling, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "PaperlessBilling")

plot17 <- ggplot(telco_clean_na, aes(x =PaymentMethod, fill = Churn)) +
  geom_bar(position = 'fill') +
  labs(title = "PaymentMethod")

# create graph plot
grid_1 <- plot_grid(plot1, plot2, plot3, plot4, plot5, plot6, ncol = 3)
grid_2 <- plot_grid(plot7, plot8, plot9, plot10, plot11, plot12, ncol = 3)
grid_3 <- plot_grid(plot13, plot14, plot15, plot16, ncol = 2)
grid_4 <- plot_grid(plot17)

# show graph
grid_1
grid_2
grid_3
grid_4
```

- **Gender** The churn percent is almost equal in case of Male and Females

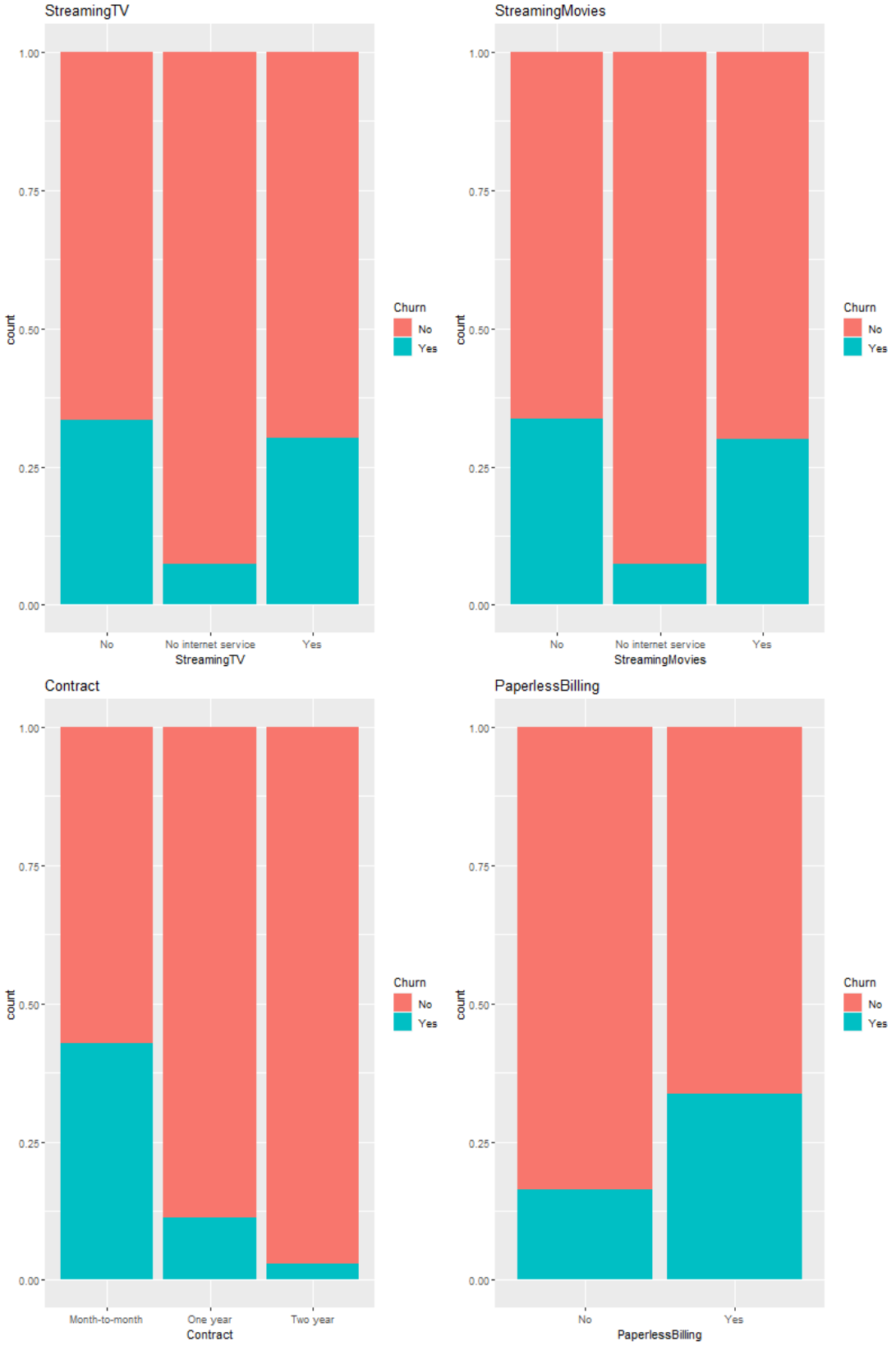
- The percent of churn is higher in case of **senior citizens**
- Customers with **Partners** and **Dependents** have lower churn rate as compared to those who don't have partners & Dependents.



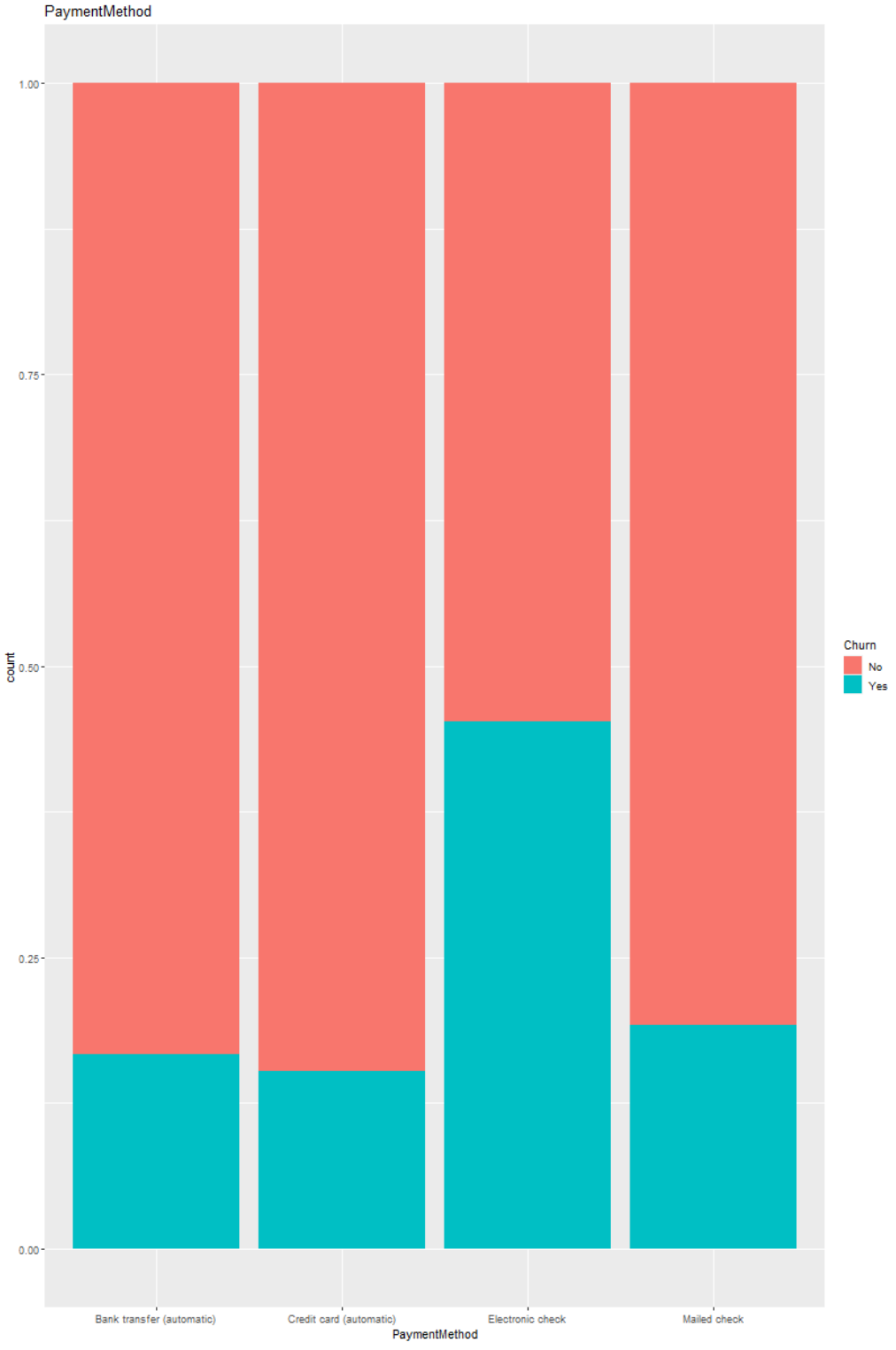
- Churn rate is much **higher** in case of **Fiber Optic InternetServices.**
- Customers who do not have services like **No OnlineSecurity** , **OnlineBackup** and **TechSupport** have left the platform in the past month.



- 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 cutsomers having **paperless billing option**.



- Customers who have **ElectronicCheck** PaymentMethod tend to **leave** the platform more when compared to other options.

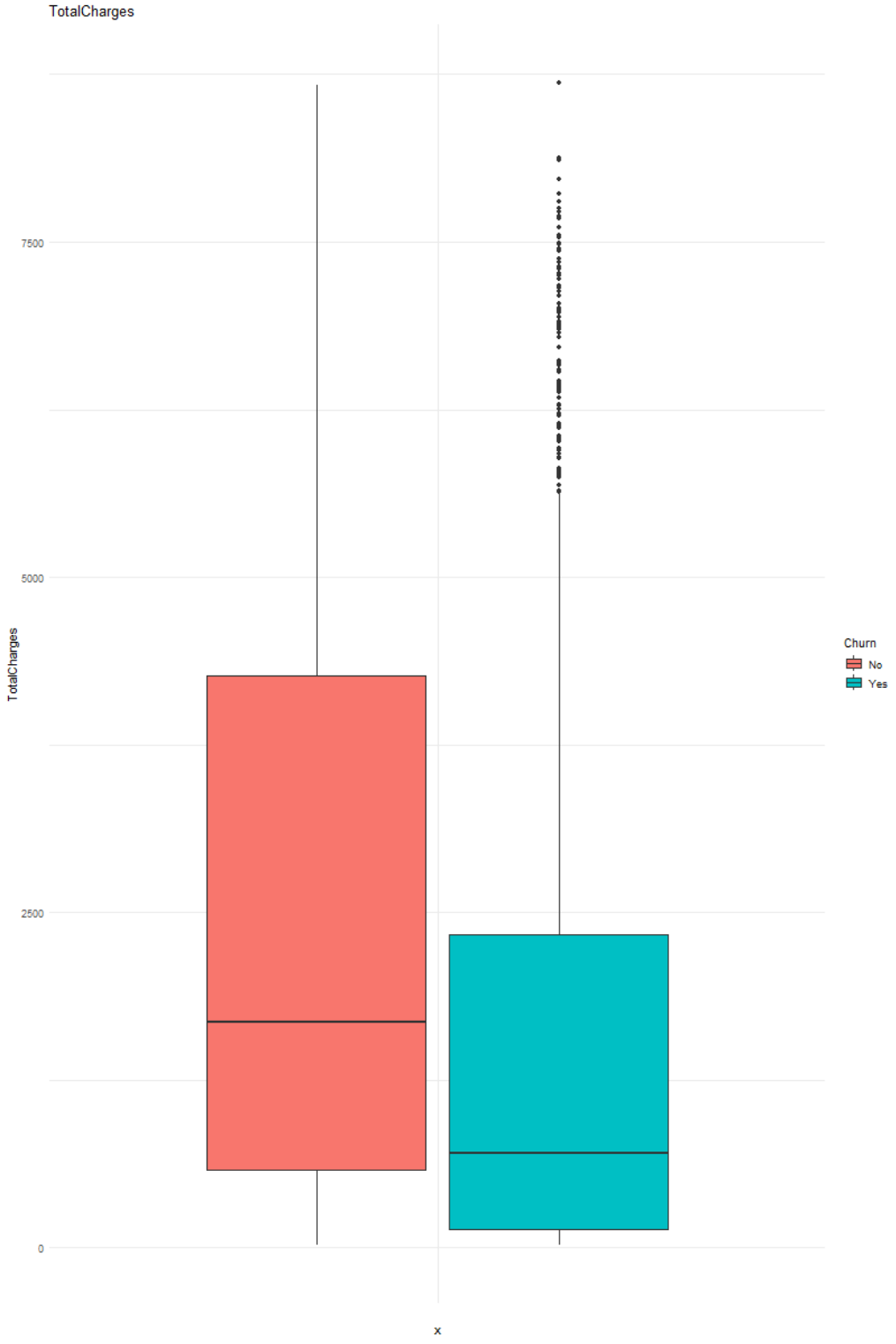


```
# create data visual for explore this data when data is numerical
plot18 <- ggplot(telco_clean_na, aes(y = tenure, x = " ", fill = Churn)) +
  geom_boxplot() +
  labs(title = "tenure") +
  theme_minimal()

plot19 <- ggplot(telco_clean_na, aes(y = MonthlyCharges, x = " ", fill = Churn)) +
  geom_boxplot() +
  labs(title = "MonthlyCharges") +
  theme_minimal()

plot20 <- ggplot(telco_clean_na, aes(y = TotalCharges, x = " ", fill = Churn)) +
  geom_boxplot() +
  labs(title = "TotalCharges") +
  theme_minimal()

# create graph plot in boxplot when data is stats
grid_5 <- plot_grid(plot18, ncol = 1)
grid_6 <- plot_grid(plot19, ncol = 1)
grid_7 <- plot_grid(plot20, ncol = 1)



# show graph in boxplot
grid_5
grid_6
grid_7
```

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

MonthlyCharges: Customers who have churned, have high monthly charges. The median is above 75.

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




```
## Prepare data
# Cleaning data

telco_clean_na_1 <- data.frame(lapply(telco_clean_na,
                          function(x) gsub("No internet service", "No", x)))
```

```
telco_clean_na_1 <- data.frame(lapply(telco_clean_na_1,
                           function(x) gsub("No phone service", "No", x)))
```

- cleaning data by change No internet service and No phone service to No

```
## Standardization

num_columns <- c("tenure", "MonthlyCharges", "TotalCharges")
telco_clean_na_1[num_columns] <- sapply(telco_clean_na_1[num_columns], as.numeric)

telco_clean_na_2 <- telco_clean_na_1[,c("tenure", "MonthlyCharges", "TotalCharges")]
telco_std_1 <- data.frame(scale(telco_clean_na_2))
```

- standardization data

```
## create dummy for analyse

telco_cat <- telco_clean_na_1[,-c(1,6,19,20)]

#Creating Dummy Variables
dummy<- data.frame(sapply(telco_cat,function(x) data.frame(model.matrix(~x-1,data =telco_cat))[,-1]))

head(dummy)

#Combining the data
telco_final <- cbind(telco_std_1,dummy)
head(telco_final)
```

- Create Dummy variables.
- Creating the final dataset by combining the numeric and dummy data frames.

```
> head(telco_final)
      tenure MonthlyCharges TotalCharges gender SeniorCitizen Partner
1 -1.28015700     -1.1616113   -0.9941234      0             0       1
2  0.06429811     -0.2608594   -0.1737275      1             0       0
3 -1.23941594     -0.3638974   -0.9595809      1             0       0
4  0.51244982     -0.7477972   -0.1952338      1             0       0
5 -1.23941594      0.1961642   -0.9403906      0             0       0
6 -0.99496955      1.1584066   -0.6453233      0             0       0
  Dependents PhoneService MultipleLines InternetService.xFiber.optic
1          0            0             0                           0
2          0            1             0                           0
3          0            1             0                           0
4          0            0             0                           0
5          0            1             0                           1
6          0            1             1                           1
  InternetService.xNo OnlineSecurity OnlineBackup DeviceProtection TechSupport
1                   0              0            1                 0           0
2                   0              1            0                 1           0
3                   0              1            1                 0           0
4                   0              1            0                 1           1
5                   0              0            0                 0           0
6                   0              0            0                 1           0
  StreamingTV StreamingMovies Contract.xOne.year Contract.xTwo.year
1           0               0                 0                 0
2           0               0                 1                 0
3           0               0                 0                 0
4           0               0                 1                 0
5           0               0                 0                 0
6           1               1                 0                 0
  PaperlessBilling PaymentMethod.xCredit.card..automatic.
1                1                                     0
2                0                                     0
3                1                                     0
4                0                                     0
5                1                                     0
6                1                                     0
  PaymentMethod.xElectronic.check PaymentMethod.xMailed.check Churn
1                               1                           0     0
2                               0                           1     0
3                               0                           1     1
4                               0                           0     0
5                               1                           0     1
6                               1                           0     1
```

```
## SPLIT DATA
set.seed(42)
n <- nrow(telco_final)
id <- sample(1:n, size=n*0.7) ## 70% train 30% test
train_data <- telco_final[id, ]
test_data <- telco_final[-id, ]
```

- Splitting the data into train and validation data.

```
#Build the first model using all variables
model_1 = glm(Churn ~ ., data = train_data, family = "binomial")
```

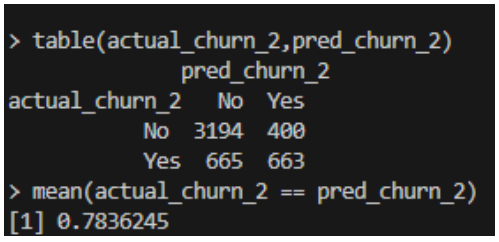
```
summary(model_1)
```

- Trian Model 1 by use logistic regression

```
> summary(model_1)

Call:
glm(formula = Churn ~ ., family = "binomial", data = train_data)

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -1.872233   1.515400  -1.235  0.21666
tenure                                 -1.495009   0.178391  -8.381  < 2e-16
MonthlyCharges                         -0.575986   1.136425  -0.507  0.61227
TotalCharges                            0.744849   0.187608   3.970 7.18e-05
gender                                  0.011314   0.077292   0.146  0.88362
SeniorCitizen                           0.160049   0.101326   1.580  0.11421
Partner                                 0.015005   0.092249   0.163  0.87078
Dependents                             -0.180685   0.107207  -1.685  0.09192
PhoneService                           -0.301209   0.773600  -0.389  0.69701
MultipleLines                           0.366851   0.209762   1.749  0.08031
InternetService.xFiber.optic            1.233766   0.949331   1.300  0.19373
InternetService.xNo                    -1.059757   0.960123  -1.104  0.26969
OnlineSecurity                         -0.343312   0.211754  -1.621  0.10496
OnlineBackup                           -0.005812   0.208924  -0.028  0.97781
DeviceProtection                        0.046124   0.210735   0.219  0.82675
TechSupport                            -0.304830   0.216223  -1.410  0.15860
StreamingTV                             0.476659   0.388079   1.228  0.21935
StreamingMovies                         0.351589   0.388081   0.906  0.36495
Contract.xOne.year                     -0.603351   0.126028  -4.787 1.69e-06
Contract.xTwo.year                     -1.402453   0.211502  -6.631 3.34e-11
PaperlessBilling                        0.372299   0.088430   4.210 2.55e-05
PaymentMethod.xCredit.card..automatic. -0.111953   0.135607  -0.826  0.40905
PaymentMethod.xElectronic.check         0.339038   0.112948   3.002  0.00268
PaymentMethod.xMailed.check            -0.063824   0.136783  -0.467  0.64078

(Intercept)
tenure                                 ***
MonthlyCharges
TotalCharges                           ***
gender
SeniorCitizen
Partner
Dependents                             .
PhoneService
MultipleLines                          .
InternetService.xFiber.optic
InternetService.xNo
OnlineSecurity
OnlineBackup
DeviceProtection
TechSupport
StreamingTV
StreamingMovies
Contract.xOne.year                     ***
Contract.xTwo.year                     ***
PaperlessBilling                       ***
PaymentMethod.xCredit.card..automatic.
PaymentMethod.xElectronic.check        **
PaymentMethod.xMailed.check
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

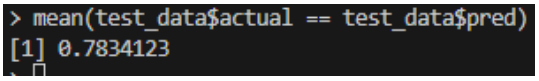
```
# Train Model 2 for P value <= 0.5
model_2 = glm(Churn ~ tenure + TotalCharges + Contract.xOne.year + Contract.xTwo.year
                    + PaperlessBilling + PaymentMethod.xElectronic.check
                    , data = train_data, family = "binomial" )
p_train_2 <- predict(model_2, type = "response")

pred_churn_2 <- factor(ifelse(p_train_2 >= 0.50, "Yes", "No"))
actual_churn_2 <- factor(ifelse(train_data$Churn == 1, "Yes", "No"))
table(actual_churn_2,pred_churn_2)
mean(actual_churn_2 == pred_churn_2)
```

- Train Model 2 by use logistic regression and pick variables that p value < 0.05




```
## test model 2
p_test_2 <- predict(model_2, newdata = test_data, type = "response")

test_data$pred <- if_else(p_test_2 >= 0.5, "Yes", "No")
test_data$actual <- if_else(test_data$Churn == 1, "Yes", "No")
mean(test_data$actual == test_data$pred)
```

- Test Model 2 by use logistic regression




```
#Confusion Matrix.
table(test_data$pred, test_data$actual)

test_pred_churn_2 <- factor(ifelse(p_test_2 >= 0.50, 1, 0))
test_actual_churn_2 <- factor(ifelse(test_data$Churn == 1, 1, 0))
conf_matrix_2 <- confusionMatrix(test_actual_churn_2, test_pred_churn_2)


# Accuracy calculation
accuracy_2 <- sum(test_pred_churn_2 == test_actual_churn_2) / length(test_actual_churn_2)


# Accuray, Recall and F1 score
cat("Accuracy:", accuracy_2, "\n")
```

```
cat("Recall:", conf_matrix_2$byClass["Recall"], "\n")
cat("F1 Score:", conf_matrix_2$byClass["F1"], "\n")
```

- **Model Evaluation using the validation data**
  - Accuracy = 0.7834123
  - Recall = 0.8398533
  - F1 Score = 0.8574103

```
> table(test_data$pred, test_data$actual)

       No  Yes
  No  1374  262
  Yes  195  279
> cat("Accuracy:", accuracy_2, "\n")
Accuracy: 0.7834123
> cat("Recall:", conf_matrix_2$byClass["Recall"], "\n")
Recall: 0.8398533
> cat("F1 Score:", conf_matrix_2$byClass["F1"], "\n")
F1 Score: 0.8574103
```

```
### Train Model 3 Randomforrest

model.rf <- randomForest(Churn ~ ., data = train_data, proximity=FALSE,importance = FALSE,
                         ntree=500,mtry=4, do.trace=FALSE)

p_train_3 <- predict(model.rf, type = "response")

pred_churn_3 <- factor(ifelse(p_train_3 >= 0.50, "Yes", "No"))
actual_churn_3 <- factor(ifelse(train_data$Churn == 1, "Yes", "No"))
table(actual_churn_3,pred_churn_3)
mean(actual_churn_3 == pred_churn_3)
```

- Train Model 3 by use Randomforrest

```
> table(actual_churn_3,pred_churn_3)
              pred_churn_3
actual_churn_3   No  Yes
           No  3237  357
           Yes  654  674
> mean(actual_churn_3 == pred_churn_3)
[1] 0.7945957
```

```
### Test model 3 Randomforrest
```

```
p_test_3 <- predict(model.rf, newdata = test_data, type = "response")

test_data$pred_rf <- if_else(p_test_3 >= 0.5, "Yes", "No")
test_data$actual <- if_else(test_data$Churn == 1, "Yes", "No")
mean(test_data$actual == test_data$pred_rf)
```

- Test Model 3 by use Randomforrest

```
> mean(test_data$actual == test_data$pred_rf)
[1] 0.8080569
```

```
#Confusion Matrix.
table(test_data$pred_rf, test_data$actual)

test_pred_churn_3 <- factor(ifelse(p_test_3 >= 0.50, 1, 0))
test_actual_churn_3 <- factor(ifelse(test_data$Churn == 1, 1, 0))
conf_matrix_3 <- confusionMatrix(test_actual_churn_3, test_pred_churn_3)

# Accuracy calculation
accuracy_3 <- sum(test_pred_churn_3 == test_actual_churn_3) / length(test_actual_churn_3)


# Accuray, Recall and F1 score
cat("Accuracy:", accuracy_3, "\n")
cat("Recall:", conf_matrix_3$byClass["Recall"], "\n")
cat("F1 Score:", conf_matrix_3$byClass["F1"], "\n")
```

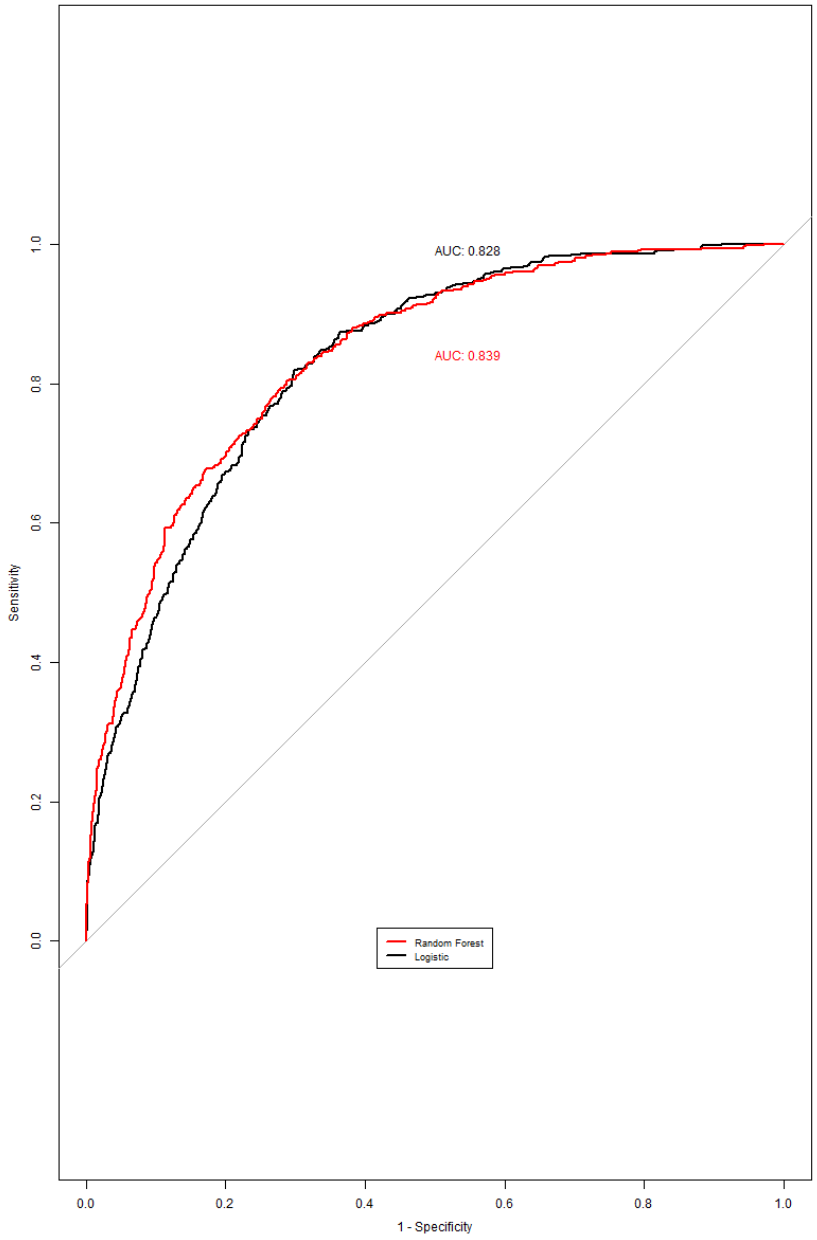
- **Model Evaluation using the validation data**
  - Accuracy = 0.8080569
  - Recall = 0.8501805
  - F1 Score = 0.8746518

```
> table(test_data$pred_rf, test_data$actual)

       No  Yes
  No  1413  249
  Yes  156  292
> cat("Accuracy:", accuracy_3, "\n")
Accuracy: 0.8080569
> cat("Recall:", conf_matrix_3$byClass["Recall"], "\n")
Recall: 0.8501805
> cat("F1 Score:", conf_matrix_3$byClass["F1"], "\n")
F1 Score: 0.8746518
```

```
## Checking the AUC for all two models:
options(repr.plot.width =10, repr.plot.height = 8)

glm.roc <- roc(response = test_data$Churn, predictor = as.numeric(p_test_2))
```

```
rf.roc <- roc(response = test_data$Churn, predictor = as.numeric(p_test_3))

plot(glm.roc,      legacy.axes = TRUE, print.auc.y = 1.0, print.auc = TRUE)
plot(rf.roc, col = "red" , add = TRUE, print.auc.y = 0.85, print.auc = TRUE)
legend("bottom", c("Random Forest", "Logistic"),
       lty = c(1,1), lwd = c(2, 2), col = c("red", "black"), cex = 0.75)
```

- **Checking the AUC for all models:**



A brief Summary of all the models:

Logistic Regression:

- Accuracy = 78.34 %
- Recall = 83.98 %
- F1 Score = 85.74 %

RandomForest:

- Accuracy = 80.80 %
- Recall = 85.02 %
- F1 Score = 87.46 %

Random Forest model performs better in churn predicting compared to Logistic Regression, and I have a plan to further enhance the prediction efficiency.