Customer Churn Analysis

2023-05-14

```
#Data Cleaning
#CustID column was removed before importing as it can cause redundancy in data.
```

```
#imported data turned out to be as follows
df <- read.csv('A2Customer.csv')
head(df)
```

```
##      Sex SeniorCard Married HasChildren LengthOfPlan BundledPlan
## 1 Female          1      No          No           28         Yes
## 2   Male          0     Yes          No           12         Yes
## 3   Male          0      No          No            1         Yes
## 4 Female          0     Yes         Yes           30         Yes
## 5   Male          0      No          No           38         Yes
## 6 Female          0      No          No           14         Yes
##   MultipleLinesPlan InternetServicePlan OnlineSecurityEnabled
## 1               Yes         Fiber optic                    No
## 2                No         Fiber optic                   Yes
## 3                No         Fiber optic                    No
## 4               Yes                  No   No internet service
## 5               Yes         Fiber optic                   Yes
## 6                No                 DSL                   Yes
##   OnlineBackupEnabled DeviceProtectionEnabled  TechSupportEnabled
## 1                  No                     Yes                 Yes
## 2                 Yes                     Yes                  No
## 3                  No                      No                  No
## 4 No internet service     No internet service No internet service
## 5                 Yes                      No                  No
## 6                  No                      No                 Yes
##       StreamingTVPlan StreamingMoviesPlan   ContractType ElectronicBilling
## 1                 Yes                 Yes Month-to-month               Yes
## 2                  No                  No Month-to-month               Yes
## 3                 Yes                  No Month-to-month               Yes
## 4 No internet service No internet service       Two year               Yes
## 5                 Yes                 Yes       One year               Yes
## 6                  No                  No       One year                No
##        PaymentType MonthlyFees TotalFees Switched
## 1 Electronic check      103.30   2890.65      Yes
## 2 Electronic check       84.60    959.90       No
## 3 Electronic check       79.95     79.95      Yes
## 4     Mailed check       25.10    789.55       No
## 5 Electronic check      104.85   3887.25       No
## 6     Mailed check       55.70    795.15       No
```

```r
#installing necessary packages
install.packages("caret",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

```r
install.packages("ggthemes",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

```r
install.packages("party",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

```r
install.packages("tidyverse",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

```r
install.packages("randomForest",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

```r
#importing libraries
library(plyr)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```r
library(ggplot2)
library(gridExtra)
library(ggthemes)
library(caret)
```

```
## Loading required package: lattice
```

```r
library(MASS)
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(party)
```

```
## Loading required package: grid

## Loading required package: mvtnorm

## Loading required package: modeltools

## Loading required package: stats4

##
## Attaching package: 'modeltools'

## The following object is masked from 'package:plyr':
##
##     empty

## Loading required package: strucchange

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich
```

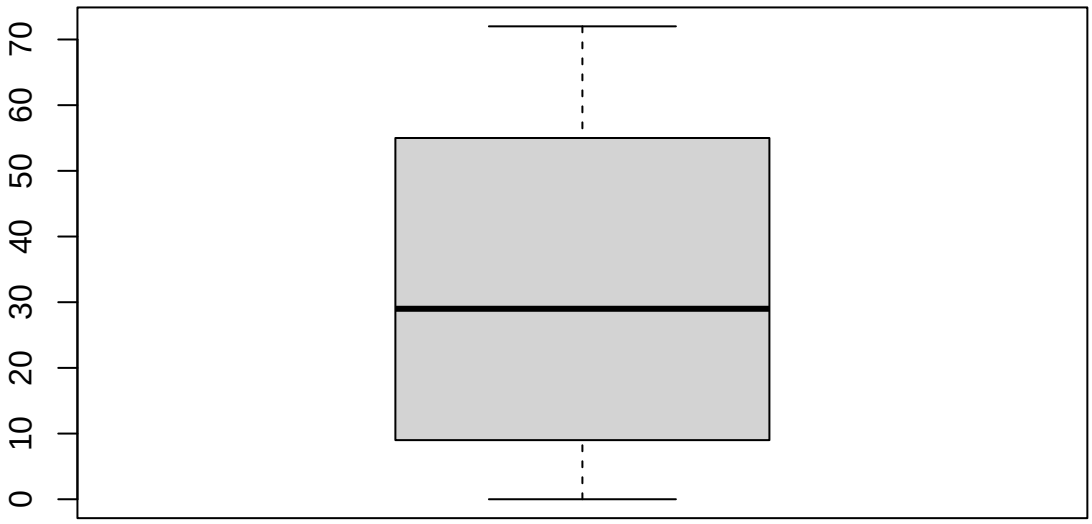
```
str(df)
```

```
## 'data.frame':    7045 obs. of  20 variables:
##  $ Sex                     : chr  "Female" "Male" "Male" "Female" ...
##  $ SeniorCard              : int  1 0 0 0 0 0 1 0 0 0 ...
##  $ Married                 : chr  "No" "Yes" "No" "Yes" ...
##  $ HasChildren             : chr  "No" "No" "No" "Yes" ...
##  $ LengthOfPlan            : int  28 12 1 30 38 14 65 68 13 47 ...
##  $ BundledPlan             : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ MultipleLinesPlan       : chr  "Yes" "No" "No" "Yes" ...
##  $ InternetServicePlan     : chr  "Fiber optic" "Fiber optic" "Fiber optic" "No" ...
##  $ OnlineSecurityEnabled   : chr  "No" "Yes" "No" "No internet service" ...
##  $ OnlineBackupEnabled     : chr  "No" "Yes" "No" "No internet service" ...
##  $ DeviceProtectionEnabled: chr  "Yes" "Yes" "No" "No internet service" ...
##  $ TechSupportEnabled      : chr  "Yes" "No" "No" "No internet service" ...
##  $ StreamingTVPlan         : chr  "Yes" "No" "Yes" "No internet service" ...
##  $ StreamingMoviesPlan     : chr  "Yes" "No" "No" "No internet service" ...
##  $ ContractType            : chr  "Month-to-month" "Month-to-month" "Month-to-month" "Two year" ...
##  $ ElectronicBilling       : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ PaymentType             : chr  "Electronic check" "Electronic check" "Electronic check" "Mailed ch
##  $ MonthlyFees             : num  103.3 84.6 80 25.1 104.8 ...
##  $ TotalFees               : num  2891 960 80 790 3887 ...
##  $ Switched                : chr  "Yes" "No" "Yes" "No" ...
```

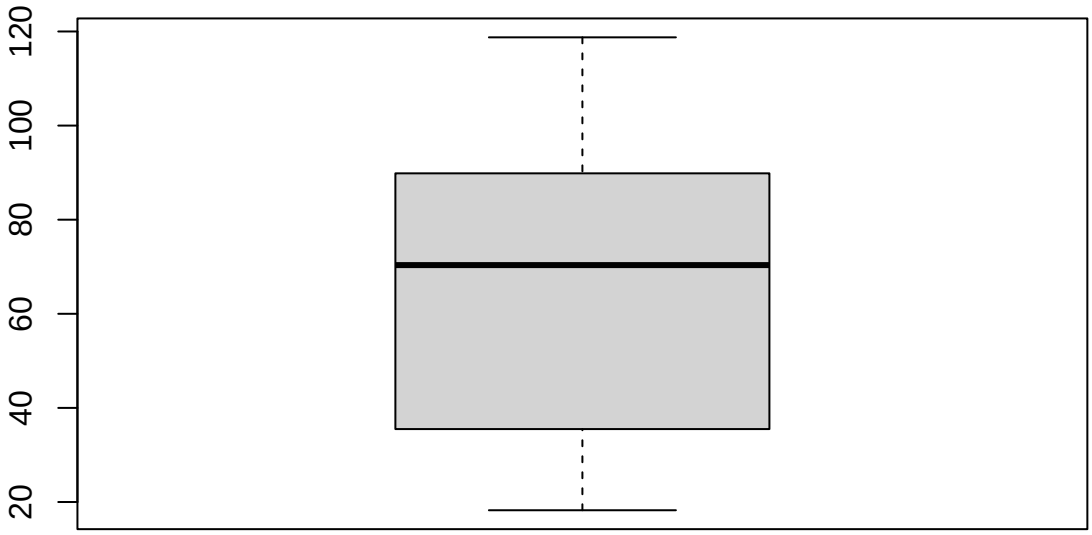
```
sum(is.na(df))
```

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

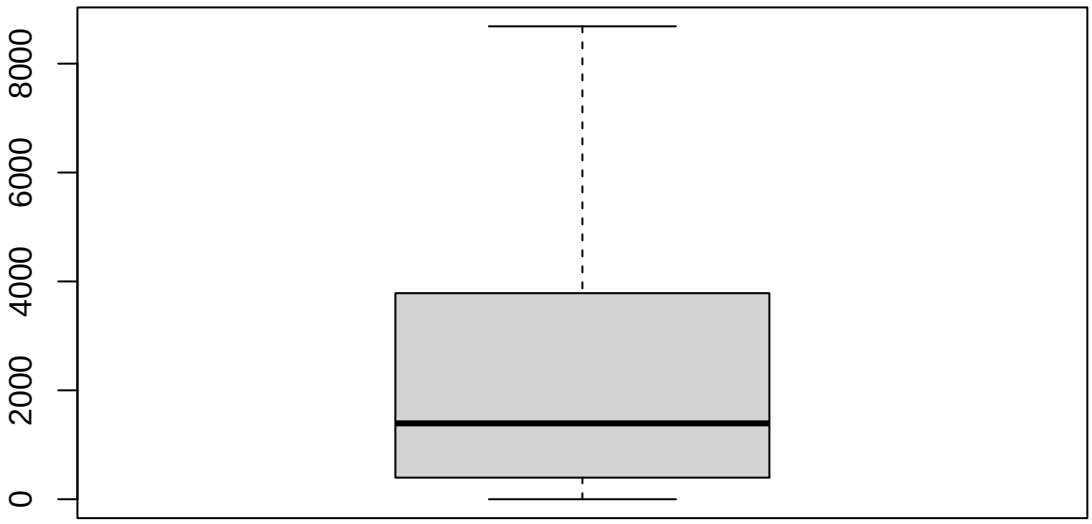
```
# Create a boxplot to check whether their exists any outliers in the dataset
boxplot(df$LengthOfPlan)
```



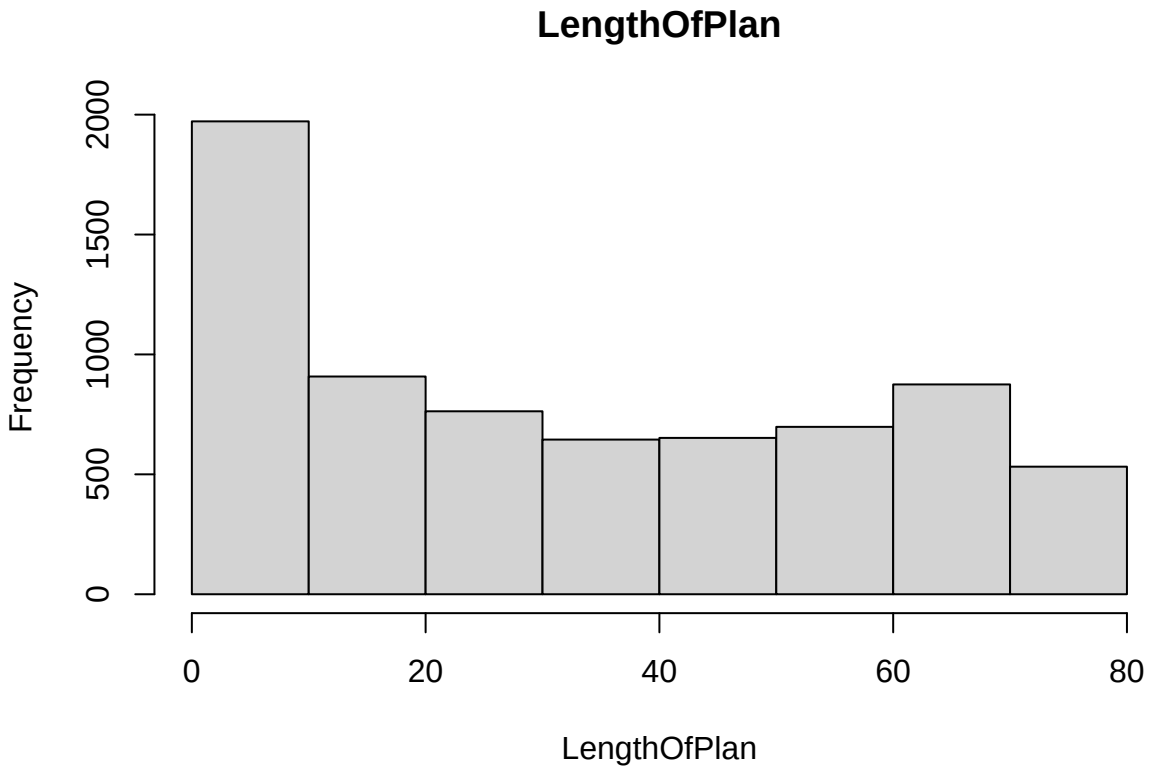

```
boxplot(df$MonthlyFees)
```




```
boxplot(df$TotalFees)
```

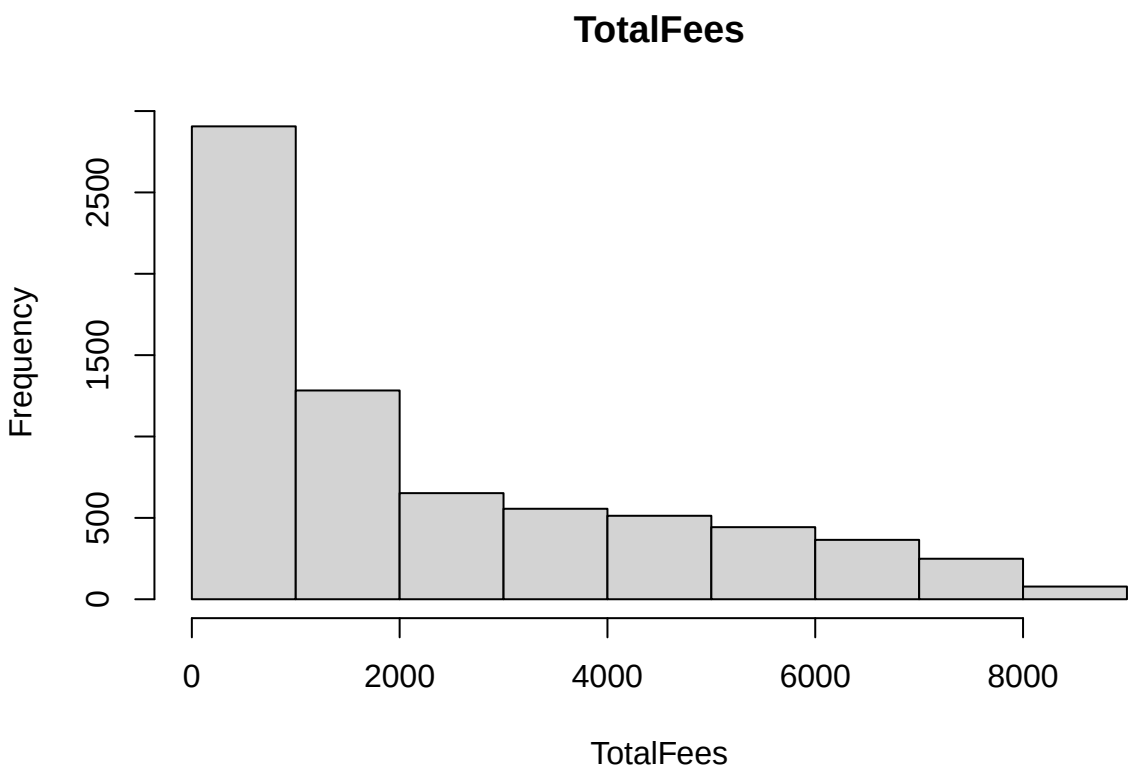



```r
# Visualize the distribution of numerical variables using histograms
hist(df$LengthOfPlan, breaks = 10, main = "LengthOfPlan", xlab = "LengthOfPlan")
```




```
hist(df$MonthlyFees, breaks = 10, main = "MonthlyFees", xlab = "MonthlyFees")
```


```r
hist(df$TotalFees, breaks = 10, main = "TotalFees", xlab = "TotalFees")
```




```
#Following the Central Limit Theoreom, it can be assumed that numerical variables are normally distribu
#more than 7000 values

#no outliers detected from the boxplots
```

```
#after examining the dataset, we do chi-square test on 4 categorial variables to check if there is sign
#If their association is proved then we do cramerV test to check their correlation and decide on which
```

```
cont_table <- table(df$OnlineSecurityEnabled, df$OnlineBackupEnabled)
chi_square_result <- chisq.test(cont_table)
chi_square_value <- chi_square_result$statistic
p_value <- chi_square_result$p.value
n <- sum(chi_square_result$observed)
k <- min(dim(chi_square_result$observed))


cramers_v <- sqrt(chi_square_value/ (n * (k - 1)))


print(chi_square_result)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 7272.1, df = 4, p-value < 2.2e-16
```

```r
print(p_value)
```

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

```r
print(cramers_v)
```

```
## X-squared
## 0.7184112
```

```r
#Value of CramerV coefficent close to one suggests that the two variables are highly correlated.
#Thus keeping one of them is the best choice
```

```r
cont_table1 <- table(df$OnlineSecurityEnabled, df$TechSupportEnabled)
chi_square_result1 <- chisq.test(cont_table1)
chi_square_value1 <- chi_square_result1$statistic
p_value1 <- chi_square_result1$p.value
n <- sum(chi_square_result1$observed)
k <- min(dim(chi_square_result1$observed))
cramers_v1 <- sqrt(chi_square_value1 / (n * (k - 1)))


print(chi_square_result1)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont_table1
## X-squared = 7571, df = 4, p-value < 2.2e-16
```

```r
print(p_value1)
```

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

```r
print(cramers_v1)
```

```
## X-squared
## 0.7330272
```

```r
#Value of CramerV coefficent close to one suggests that the two variables are highly correlated. Thus k
```

```r
cont_table2 <- table(df$OnlineSecurityEnabled, df$DeviceProtectionEnabled)
chi_square_result2 <- chisq.test(cont_table2)
chi_square_value2 <- chi_square_result2$statistic
p_value2 <- chi_square_result2$p.value
n <- sum(chi_square_result2$observed)
k <- min(dim(chi_square_result2$observed))
cramers_v2 <- sqrt(chi_square_value2 / (n * (k - 1)))


print(chi_square_result2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  cont_table2
## X-squared = 7249, df = 4, p-value < 2.2e-16
```

```
print(p_value2)
```

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

```
print(cramers_v2)
```

```
## X-squared
## 0.7172696
```

```
#Value of CramerV coefficent close to one suggests that the two variables are highly correlated. Thus k
```

```
#columns OnlineSecurityEnabled, OnlineBackupEnabled, DeviceProtectionEnabled, TechSupportEnabled, Strea
cols_recode1 <- c(9:14)
for(i in 1:ncol(df[,cols_recode1])) {
        df[,cols_recode1][,i] <- as.factor(mapvalues
                                           (df[,cols_recode1][,i], from =c("No internet service"),to=
}
```

```
#column MulitpleLines has No internet service option other than "yes" and "no". That option is also con
df$MultipleLinesPlan <- as.factor(mapvalues(df$MultipleLinesPlan,
                                            from=c("No phone service"),
                                            to=c("No")))
```

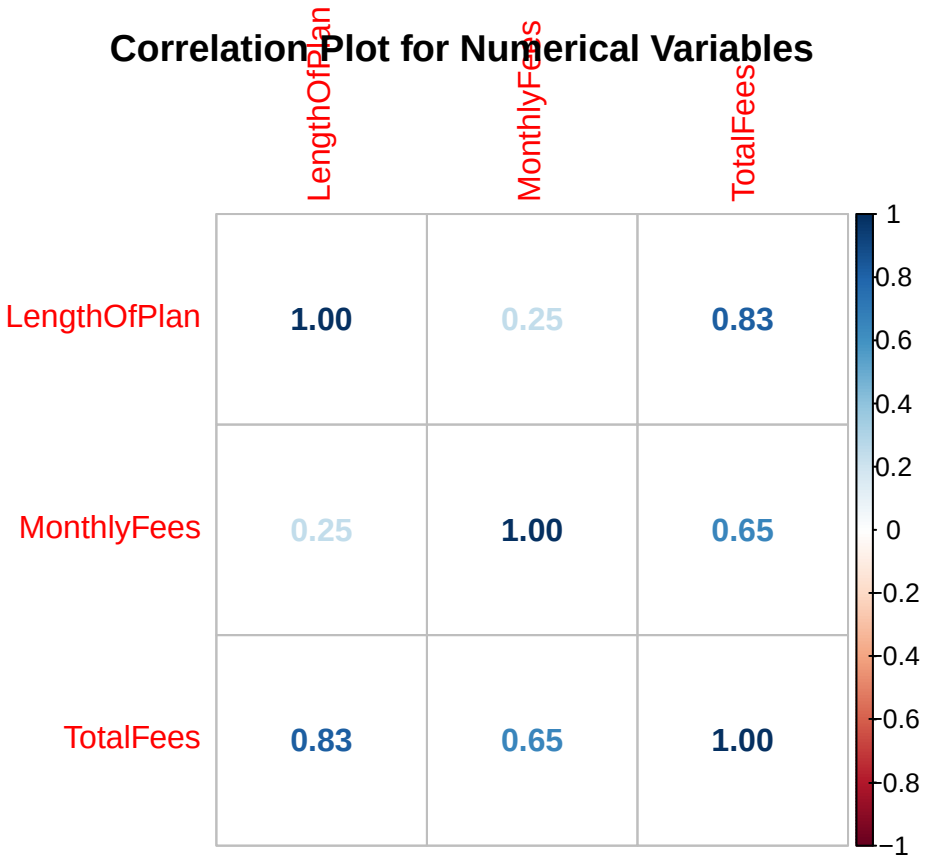
```
#SeniorCard column has 1 and 0  which are changed to yes and no and then factored for model
df$SeniorCard <- as.factor(mapvalues(df$SeniorCard,
                                     from=c("0","1"),
                                     to=c("No", "Yes")))
```

```
#factorizing the remaining categorical variables so that they can be fed onto the main model

remaining_cat <- c('Sex','Married', 'HasChildren','BundledPlan', 'InternetServicePlan', 'ContractType',
for (var in remaining_cat) {
  df[[var]] <- factor(df[[var]])
}
```

```
#correlation of numeric variables

numeric_var <- sapply(df, is.numeric)
corr.matrix <- cor(df[,numeric_var])
corrplot(corr.matrix,
         main = "\n\nCorrelation Plot for Numerical Variables", method ='number')
```




```
group_lop <- function(LengthOfPlan){
    if (LengthOfPlan >= 0 & LengthOfPlan <= 12){
        return('0-12 Month')
    }else if(LengthOfPlan > 12 & LengthOfPlan <= 24){
        return('12-24 Month')
    }else if (LengthOfPlan > 24 & LengthOfPlan <= 48){
        return('24-48 Month')
    }else if (LengthOfPlan > 48 & LengthOfPlan <=60){
        return('48-60 Month')
    }else if (LengthOfPlan > 60){
        return('> 60 Month')
    }
}
df$lop_group <- sapply(df$LengthOfPlan,group_lop)
df$lop_group <- as.factor(df$lop_group)
```

```
str(df)
```

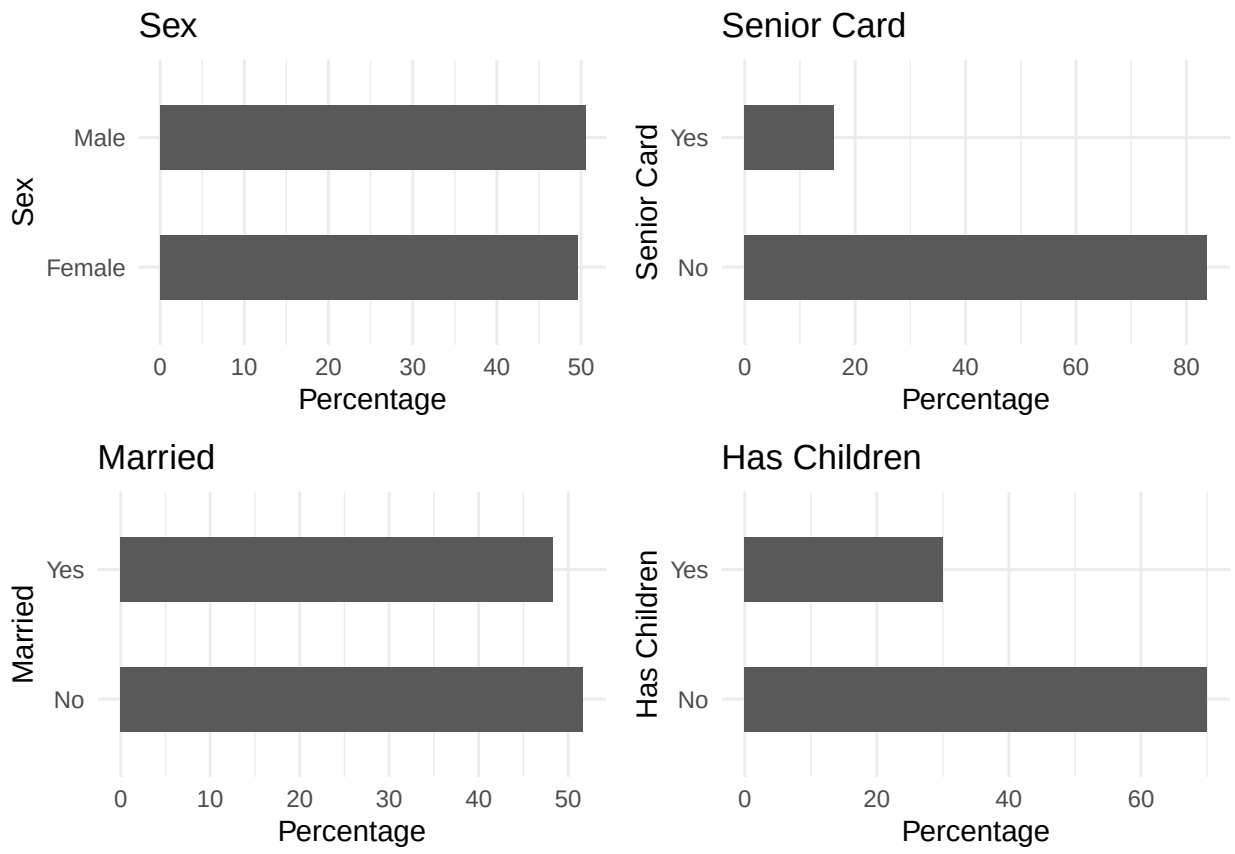
```
## 'data.frame':    7045 obs. of  21 variables:
##  $ Sex                  : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 1 1 1 1 1 ...
##  $ SeniorCard           : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 2 1 1 1 ...
##  $ Married              : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 1 2 2 1 1 ...
##  $ HasChildren          : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 2 1 1 ...
##  $ LengthOfPlan         : int  28 12 1 30 38 14 65 68 13 47 ...
##  $ BundledPlan          : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
```

```
##  $ MultipleLinesPlan      : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 2 2 1 2 ...
##  $ InternetServicePlan    : Factor w/ 3 levels "DSL","Fiber optic",..: 2 2 2 3 2 1 2 2 2 2 ...
##  $ OnlineSecurityEnabled  : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 2 2 1 1 2 ...
##  $ OnlineBackupEnabled    : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 1 2 1 1 ...
##  $ DeviceProtectionEnabled: Factor w/ 2 levels "No","Yes": 2 2 1 1 1 1 2 1 1 2 ...
##  $ TechSupportEnabled     : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 1 2 1 1 ...
##  $ StreamingTVPlan        : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 1 2 2 1 2 ...
##  $ StreamingMoviesPlan    : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 1 2 2 1 1 ...
##  $ ContractType           : Factor w/ 3 levels "Month-to-month",..: 1 1 1 3 2 2 1 3 1 2 ...
##  $ ElectronicBilling      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 2 2 1 ...
##  $ PaymentType            : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 3 3 4 3 4 1 2 4 2 .
##  $ MonthlyFees            : num  103.3 84.6 80 25.1 104.8 ...
##  $ TotalFees              : num  2891 960 80 790 3887 ...
##  $ Switched               : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 1 ...
##  $ lop_group              : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 4 2 2 4 4 3 1 1 3 4 ...
```

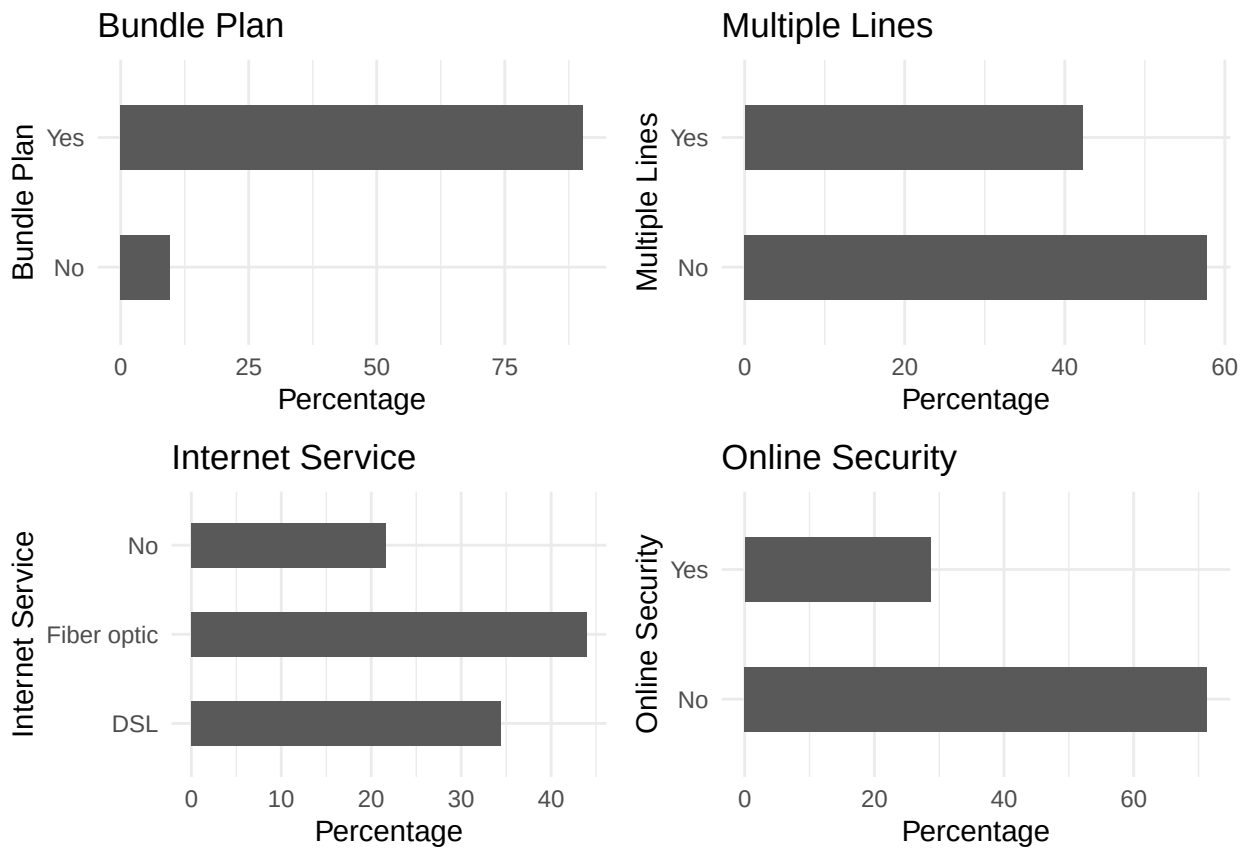
```r
#Plotting distributions to check each column


p1 <- ggplot(df, aes(x=Sex)) + ggtitle("Sex") + xlab("Sex") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
p2 <- ggplot(df, aes(x=SeniorCard)) + ggtitle("Senior Card") + xlab("Senior Card") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
p3 <- ggplot(df, aes(x=Married)) + ggtitle("Married") + xlab("Married") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
p4 <- ggplot(df, aes(x=HasChildren)) + ggtitle("Has Children") + xlab("Has Children") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
grid.arrange(p1, p2, p3, p4, ncol=2)
```

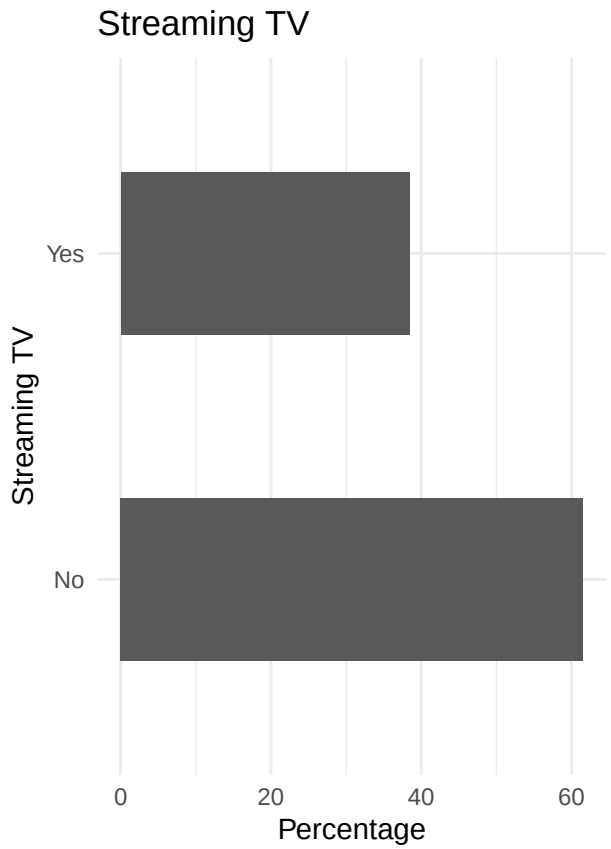
```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(count)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



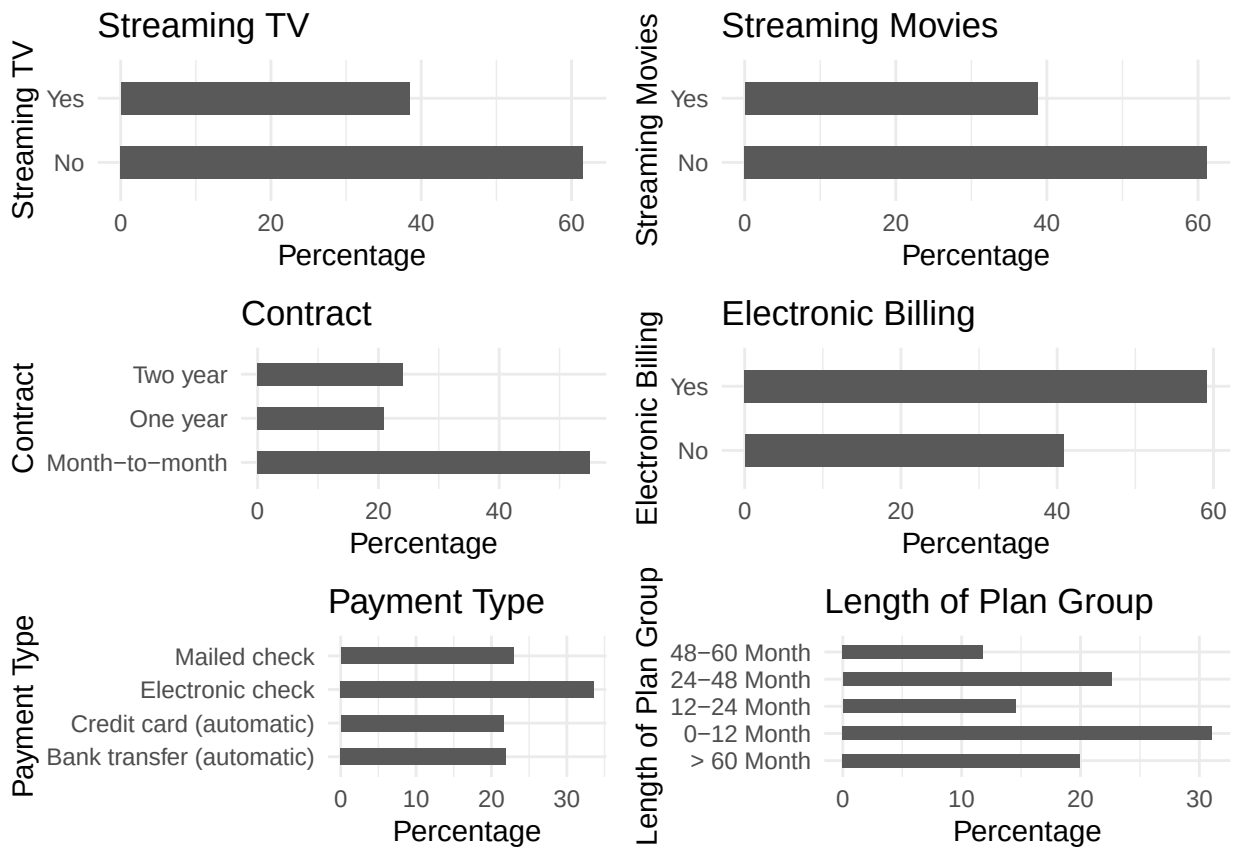

```
p5 <- ggplot(df, aes(x=BundledPlan)) + ggtitle("Bundle Plan") + xlab("Bundle Plan") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() +
p6 <- ggplot(df, aes(x=MultipleLinesPlan)) + ggtitle("Multiple Lines") + xlab("Multiple Lines") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() +
p7 <- ggplot(df, aes(x=InternetServicePlan)) + ggtitle("Internet Service") + xlab("Internet Service") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() +
p8 <- ggplot(df, aes(x=OnlineSecurityEnabled)) + ggtitle("Online Security") + xlab("Online Security") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() +
grid.arrange(p5, p6, p7, p8, ncol=2)
```




```
p12 <- ggplot(df, aes(x=StreamingTVPlan)) + ggtitle("Streaming TV") + xlab("Streaming TV") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
grid.arrange(p12, ncol=2)
```




```
p13 <- ggplot(df, aes(x=StreamingMoviesPlan)) + ggtitle("Streaming Movies") + xlab("Streaming Movies") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
p14 <- ggplot(df, aes(x=ContractType)) + ggtitle("Contract") + xlab("Contract") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
p15 <- ggplot(df, aes(x=ElectronicBilling)) + ggtitle("Electronic Billing") + xlab("Electronic Billing") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
p16 <- ggplot(df, aes(x=PaymentType)) + ggtitle("Payment Type") + xlab("Payment Type") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
p17 <- ggplot(df, aes(x=lop_group)) + ggtitle("Length of Plan Group") + xlab("Length of Plan Group") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + t
grid.arrange(p12, p13, p14, p15, p16, p17, ncol=2)
```




```
#all columns are significant values of each option so keeping them all in the main analysis
```

```
#lengthofPlan and TotalFees are highly correlated to MonthlyFees so keeping only that in the analysis a
#OnlineBackupEnabled , DeviceProtectionEnabled and TechSupportEnabled were found to be highly correalte
```

```
df$LengthOfPlan <- NULL
df$TotalFees <- NULL
df$OnlineBackupEnabled <- NULL
df$DeviceProtectionEnabled <- NULL
df$TechSupportEnabled <- NULL
```

```
str(df)
```

```
## 'data.frame':    7045 obs. of  16 variables:
##  $ Sex                 : Factor w/ 2 levels "Female","Male": 1 2 2 1 2 1 1 1 1 1 ...
##  $ SeniorCard          : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 2 1 1 1 ...
##  $ Married             : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 1 2 2 1 1 ...
##  $ HasChildren         : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 2 1 1 ...
##  $ BundledPlan         : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ MultipleLinesPlan   : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 1 2 2 1 2 ...
##  $ InternetServicePlan : Factor w/ 3 levels "DSL","Fiber optic",..: 2 2 2 3 2 1 2 2 2 2 ...
##  $ OnlineSecurityEnabled: Factor w/ 2 levels "No","Yes": 1 2 1 1 2 2 2 1 1 2 ...
##  $ StreamingTVPlan     : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 1 2 2 1 2 ...
##  $ StreamingMoviesPlan : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 1 2 2 1 1 ...
##  $ ContractType        : Factor w/ 3 levels "Month-to-month",..: 1 1 1 3 2 2 1 3 1 2 ...
```

```
##  $ ElectronicBilling    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 1 2 2 1 ...
##  $ PaymentType          : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 3 3 4 3 4 1 2 4 2 ...
##  $ MonthlyFees          : num  103.3 84.6 80 25.1 104.8 ...
##  $ Switched             : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 1 ...
##  $ lop_group            : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 4 2 2 4 4 3 1 1 3 4 ...
```

```
#dividing the data into train and test sets
intrain<- createDataPartition(df$Switched,p=0.8,list=FALSE)
set.seed(2017)
training<- df[intrain,]
testing<- df[-intrain,]
```

```
#70/30 ratio ensured
dim(training); dim(testing)
```

```
## [1] 5636   16
```

```
## [1] 1409   16
```

```
#activating Logistic regression model
LogModel <- glm(Switched ~ .,family=binomial(),data=training)
print(summary(LogModel))
```

```
##
## Call:
## glm(formula = Switched ~ ., family = binomial(), data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9848  -0.6746  -0.2894   0.6836   3.1822
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.097672   0.383232  -2.864 0.004180 **
## SexMale                         -0.087836   0.072382  -1.213 0.224939
## SeniorCardYes                    0.238616   0.093423   2.554 0.010645 *
## MarriedYes                      -0.061644   0.086041  -0.716 0.473714
## HasChildrenYes                  -0.116410   0.099640  -1.168 0.242685
## BundledPlanYes                   0.055210   0.234663   0.235 0.813997
## MultipleLinesPlanYes             0.440968   0.099728   4.422 9.79e-06 ***
## InternetServicePlanFiber optic   1.616089   0.251360   6.429 1.28e-10 ***
## InternetServicePlanNo           -1.624126   0.306932  -5.291 1.21e-07 ***
## OnlineSecurityEnabledYes        -0.298396   0.106866  -2.792 0.005234 **
## StreamingTVPlanYes               0.567056   0.135406   4.188 2.82e-05 ***
## StreamingMoviesPlanYes           0.514211   0.134641   3.819 0.000134 ***
## ContractTypeOne year            -0.672046   0.118142  -5.688 1.28e-08 ***
## ContractTypeTwo year            -1.770787   0.204154  -8.674  < 2e-16 ***
## ElectronicBillingYes             0.363224   0.083011   4.376 1.21e-05 ***
## PaymentTypeCredit card (automatic) -0.136266   0.127121  -1.072 0.283746
## PaymentTypeElectronic check      0.259979   0.105652   2.461 0.013866 *
## PaymentTypeMailed check         -0.030373   0.127747  -0.238 0.812068
## MonthlyFees                     -0.028554   0.009336  -3.058 0.002225 **
```

```
## lop_group0-12 Month                 1.658035   0.188866   8.779  < 2e-16 ***
## lop_group12-24 Month                0.854968   0.185073   4.620 3.84e-06 ***
## lop_group24-48 Month                0.417259   0.169170   2.466 0.013644 *
## lop_group48-60 Month                0.148618   0.185554   0.801 0.423165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6522.7  on 5635  degrees of freedom
## Residual deviance: 4692.3  on 5613  degrees of freedom
## AIC: 4738.3
##
## Number of Fisher Scoring iterations: 6
```

```
anova(LogModel, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Switched
##
## Terms added sequentially (first to last)
##
##
##                       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                  5635     6522.7
## Sex                    1     1.21      5634     6521.5  0.270936
## SeniorCard             1   124.05      5633     6397.5 < 2.2e-16 ***
## Married                1   138.73      5632     6258.7 < 2.2e-16 ***
## HasChildren            1    36.62      5631     6222.1 1.439e-09 ***
## BundledPlan            1     0.12      5630     6222.0  0.731066
## MultipleLinesPlan      1     7.29      5629     6214.7  0.006952 **
## InternetServicePlan    2   549.30      5627     5665.4 < 2.2e-16 ***
## OnlineSecurityEnabled  1   219.81      5626     5445.6 < 2.2e-16 ***
## StreamingTVPlan        1     4.94      5625     5440.7  0.026168 *
## StreamingMoviesPlan    1     7.35      5624     5433.3  0.006697 **
## ContractType           2   471.54      5622     4961.8 < 2.2e-16 ***
## ElectronicBilling      1    19.52      5621     4942.2 9.937e-06 ***
## PaymentType            3    42.51      5618     4899.7 3.122e-09 ***
## MonthlyFees            1    37.48      5617     4862.2 9.215e-10 ***
## lop_group              4   169.99      5613     4692.3 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#printing testing results and accuracy
testing$Switched <- as.character(testing$Switched)
testing$Switched[testing$Switched=="No"] <- 0
testing$Switched[testing$Switched=="Yes"] <- 1
fitted.results <- predict(LogModel,newdata=testing)
fitted.results <- ifelse(fitted.results > 0.5,"1","0")
misClasificError <- mean(fitted.results != testing$Switched)
print(paste('Logistic Regression Accuracy',1-misClasificError))
```

```
## [1] "Logistic Regression Accuracy 0.787792760823279"
```

```
fres = as.factor(fitted.results)
```

```
table(fitted.results)
```

```
## fitted.results
##    0    1
## 1236  173
```

```
cm1 <- confusionMatrix(table(testing$Switched,fres))
print(cm1)
```

```
## Confusion Matrix and Statistics
##
##    fres
##       0   1
##   0 986  49
##   1 250 124
##
##                Accuracy : 0.7878
##                  95% CI : (0.7655, 0.8089)
##     No Information Rate : 0.8772
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3431
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7977
##             Specificity : 0.7168
##          Pos Pred Value : 0.9527
##          Neg Pred Value : 0.3316
##              Prevalence : 0.8772
##          Detection Rate : 0.6998
##    Detection Prevalence : 0.7346
##       Balanced Accuracy : 0.7572
##
##        'Positive' Class : 0
##
```

```
#precision
truepositive <- 997
falsepositive <- 38
truenegative <- 115
falsenegative <- 259
precision <- (truepositive)/(truepositive+falsepositive)
recall <- (truepositive)/(truepositive+falsenegative)
f1score <- 2 * (precision * recall) / (precision + recall)
print(precision)
```

```
## [1] 0.963285
```

```
print(recall)
```

```
## [1] 0.7937898
```

```
print(f1score)
```

```
## [1] 0.8703623
```

```
library(MASS)
exp(cbind(OR=coef(LogModel), confint(LogModel)))
```

```
## Waiting for profiling to be done...
```

```
##                                         OR     2.5 %    97.5 %
## (Intercept)                      0.3336468 0.1569508 0.7053275
## SexMale                          0.9159113 0.7947113 1.0554909
## SeniorCardYes                    1.2694905 1.0569156 1.5244769
## MarriedYes                       0.9402172 0.7943190 1.1130238
## HasChildrenYes                   0.8901103 0.7317335 1.0815232
## BundledPlanYes                   1.0567623 0.6670504 1.6742889
## MultipleLinesPlanYes             1.5542117 1.2785504 1.8903460
## InternetServicePlanFiber optic   5.0333683 3.0769106 8.2454184
## InternetServicePlanNo            0.1970838 0.1080232 0.3600009
## OnlineSecurityEnabledYes         0.7420071 0.6014091 0.9144514
## StreamingTVPlanYes               1.7630690 1.3522750 2.2997402
## StreamingMoviesPlanYes           1.6723188 1.2849230 2.1785403
## ContractTypeOne year             0.5106626 0.4041878 0.6424203
## ContractTypeTwo year             0.1701990 0.1126137 0.2510967
## ElectronicBillingYes             1.4379582 1.2224367 1.6926978
## PaymentTypeCredit card (automatic) 0.8726103 0.6798870 1.1192965
## PaymentTypeElectronic check      1.2969028 1.0549511 1.5964774
## PaymentTypeMailed check          0.9700838 0.7554315 1.2466477
## MonthlyFees                      0.9718498 0.9542365 0.9898262
## lop_group0-12 Month              5.2489842 3.6365847 7.6280076
## lop_group12-24 Month             2.3512988 1.6400314 3.3894896
## lop_group24-48 Month             1.5177953 1.0923715 2.1214113
## lop_group48-60 Month             1.1602300 0.8065274 1.6704440
```

```
#installing packages for decision trees model
install.packages("survival",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

```
install.packages("rpart",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

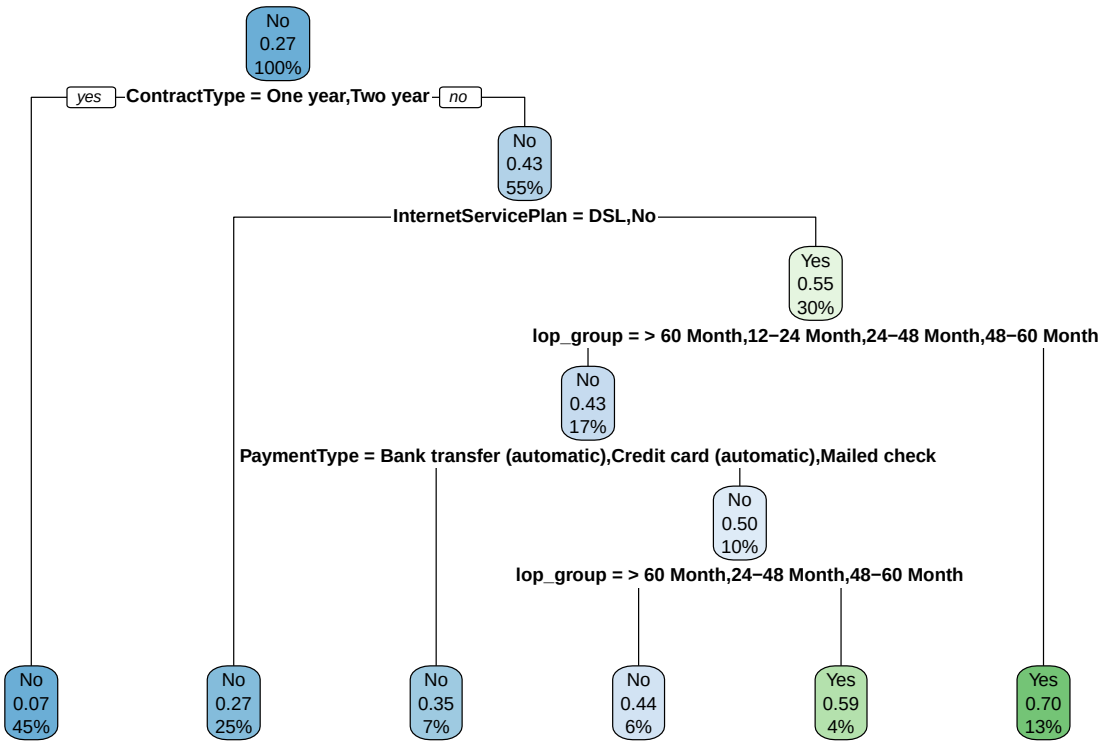
```
install.packages("rpart.plot",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

```
library(rpart)  # For building decision trees
library(rpart.plot)  # For visualizing decision trees
```

```
#running decision trees model
modeldt <- rpart(Switched ~ ., data = training, method = "class")
```

```
rpart.plot(modeldt)
```




```
predictions <- predict(modeldt, newdata = testing, type = "class")
```

```
#pritnting test results
confusion_matrix <- table(actual = testing$Switched, predicted = predictions)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(confusion_matrix)
```

```
##        predicted
## actual  No Yes
```

```
##      0 962  73
##      1 222 152
```

```
print(accuracy)
```

```
## [1] 0.7906317
```

```
truepositive1 <- 934
falsepositive1 <- 101
truenegative1 <- 188
falsenegative1 <- 186
precision1 <- (truepositive1)/(truepositive1+falsepositive1)
recall1 <- (truepositive1)/(truepositive1+falsenegative1)
f1score1 <- 2 * (precision1 * recall1) / (precision1 + recall1)
print(precision1)
```

```
## [1] 0.9024155
```

```
print(recall1)
```

```
## [1] 0.8339286
```

```
print(f1score1)
```

```
## [1] 0.8668213
```

```
#running random forest model for list of important features
modelrf <- randomForest(Switched ~ ., data = training, ntree = 100)
```

```
#printing important features in accordance with their Decrease in Gini Value
importance <- importance(modelrf)
print(importance)
```

```
##                       MeanDecreaseGini
## Sex                           43.53489
## SeniorCard                    38.04954
## Married                       38.66781
## HasChildren                   35.93538
## BundledPlan                   15.54539
## MultipleLinesPlan             33.62662
## InternetServicePlan          105.70280
## OnlineSecurityEnabled         50.09489
## StreamingTVPlan               31.82972
## StreamingMoviesPlan           32.83465
## ContractType                 201.14315
## ElectronicBilling             47.89589
## PaymentType                  136.27676
## MonthlyFees                  326.24497
## lop_group                    202.57052
```

```r
#10 of the most important features are as follows
sorted_importance <- importance[order(importance[, 1], decreasing = TRUE), ]
print(sorted_importance[1:10])
```

```
##           MonthlyFees             lop_group          ContractType
##             326.24497             202.57052             201.14315
##           PaymentType   InternetServicePlan OnlineSecurityEnabled
##             136.27676             105.70280              50.09489
##     ElectronicBilling                   Sex               Married
##              47.89589              43.53489              38.66781
##            SeniorCard
##              38.04954
```

```r
predictions1 <- predict(modelrf, newdata = testing)
```

```r
confusion_matrix1 <- table(actual = testing$Switched, predicted = predictions1)
accuracy1 <- sum(diag(confusion_matrix1)) / sum(confusion_matrix1)
print(confusion_matrix1)
```

```
##       predicted
## actual  No Yes
##      0 931 104
##      1 194 180
```

```r
print(accuracy1)
```

```
## [1] 0.7885025
```

```r
install.packages("tinytex",repos="http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/kv/q8v8kt9n5dg8h7tfdqqxl0v00000gn/T//Rtmp84sydv/downloaded_packages
```

```r
tinytex::install_tinytex(force=TRUE)
```

```
## The directory /usr/local/bin is not writable. I recommend that you make it writable. See https://gitl
```