Predictive Project

Rohan_Srivastava, Noel Abraham, Ramesh, Vaibhav, Saurav

2022-05-12

Importing packages

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

library(sjPlot)

## Registered S3 method overwritten by 'parameters':
##   method                         from
##   format.parameters_distribution datawizard

## #refugeeswelcome

library(sjmisc)

## Learn more about sjmisc with 'browseVignettes("sjmisc")'.
```

```
##
## Attaching package: 'sjmisc'

## The following object is masked from 'package:purrr':
##
##     is_empty

## The following object is masked from 'package:tidyr':
##
##     replace_na

## The following object is masked from 'package:tibble':
##
##     add_case

library(sjlabelled)

##
## Attaching package: 'sjlabelled'

## The following object is masked from 'package:forcats':
##
##     as_factor

## The following object is masked from 'package:dplyr':
##
##     as_label

## The following object is masked from 'package:ggplot2':
##
##     as_label

library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

library(caTools)
library(ROCR)
library(precrec)
library(pROC)

## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'
## The following object is masked from 'package:precrec':
##
##     auc

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var

library(randomForest)

## randomForest 4.7-1

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

##
## Attaching package: 'randomForest'

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

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

library(ggplot2)
library(lessR)

##
## lessR 4.1.9                         feedback: gerbing@pdx.edu
## --------------------------------------------------------------
## > d <- Read("")   Read text, Excel, SPSS, SAS, or R data file
##   d is default data frame, data= in analysis routines optional
##
## Learn about reading, writing, and manipulating data, graphics,
## testing means and proportions, regression, factor analysis,
## customization, and descriptive statistics from pivot tables.
##   Enter:  browseVignettes("lessR")
##
## View changes in this or recent versions of lessR.
##   Enter: help(package=lessR)  Click: Package NEWS
##   Enter: interact()  for access to interactive graphics
##   New function: reshape_long() to move data from wide to long

##
## Attaching package: 'lessR'

## The following objects are masked from 'package:car':
##
##     bc, recode, sp
## The following objects are masked from 'package:dplyr':
##
##     recode, rename
```

```
library(dplyr)
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:lessR':
##
##     reflect, rescale, scree, skew

## The following object is masked from 'package:randomForest':
##
##     outlier

## The following object is masked from 'package:car':
##
##     logit

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

Importing the data

```
df <- read.csv("C:/Users/Rohan.000/Desktop/Predictive/WA_Fn-UseC_-Telco-
Customer-Churn.csv")
head(df)
```

```
##   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
```

```
nrow(df)
```

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

Drop customer ID

```
df <- df[, !(colnames(df) %in% c("customerID"))]
```

count and remove null values

```
names(which(colSums(is.na(df))>0))
```

```
## [1] "TotalCharges"
```

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

```
## [1] 11
```

```
df <- na.omit(df)
```

Count number of unique values in each column

```
ulst <- lapply(df, unique)
k <- lengths(ulst)
k
```

```
##           gender    SeniorCitizen          Partner       Dependents 
##                2                2                2                2 
##           tenure     PhoneService    MultipleLines  InternetService 
##               72                2                3                3 
##   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport 
##                3                3                3                3 
##      StreamingTV  StreamingMovies         Contract PaperlessBilling 
##                3                3                3                2 
##    PaymentMethod   MonthlyCharges     TotalCharges            Churn 
##                4             1584             6530                2 
```

Find unique values for each variable

```
unique(df$gender)
```

```
## [1] "Female" "Male"
```

```
unique(df$SeniorCitizen)
```

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

```
unique(df$Partner)
```

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

```
unique(df$Dependents)
```

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

```
unique(df$PhoneService)
```

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

```
unique(df$MultipleLines)
```

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

```
unique(df$PhoneService)
```

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

```
unique(df$InternetService)
```

```
## [1] "DSL"         "Fiber optic" "No"
```

```
unique(df$OnlineSecurity)
```

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

```
unique(df$OnlineBackup)
```

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

```
unique(df$DeviceProtection)
```

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

```
unique(df$OnlineBackup)
```

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

```
unique(df$TechSupport)
```

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

```
unique(df$StreamingTV)
```

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

```
unique(df$StreamingMovies)

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

unique(df$Contract)

## [1] "Month-to-month" "One year"       "Two year"

unique(df$PaperlessBilling)

## [1] "Yes" "No"

unique(df$PaymentMethod)

## [1] "Electronic check"          "Mailed check"             
## [3] "Bank transfer (automatic)" "Credit card (automatic)"

unique(df$Churn)

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

Summary of Data

```
summary(df)

##     gender          SeniorCitizen      Partner           Dependents       
##  Length:7032        Min.   :0.0000   Length:7032        Length:7032       
##  Class :character   1st Qu.:0.0000   Class :character   Class :character  
##  Mode  :character   Median :0.0000   Mode  :character   Mode  :character  
##                     Mean   :0.1624                                        
##                     3rd Qu.:0.0000                                        
##                     Max.   :1.0000                                        
##      tenure      PhoneService       MultipleLines      InternetService   
##  Min.   : 1.00   Length:7032        Length:7032        Length:7032       
##  1st Qu.: 9.00   Class :character   Class :character   Class :character  
##  Median :29.00   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :32.42                                                           
##  3rd Qu.:55.00                                                           
##  Max.   :72.00                                                           
##  OnlineSecurity     OnlineBackup       DeviceProtection   TechSupport       
##  Length:7032        Length:7032        Length:7032        Length:7032       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##  StreamingTV        StreamingMovies      Contract         PaperlessBilling  
##  Length:7032        Length:7032        Length:7032        Length:7032       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
```

```
##  PaymentMethod      MonthlyCharges    TotalCharges       Churn
##  Length:7032        Min.   : 18.25   Min.   :  18.8   Length:7032
##  Class :character   1st Qu.: 35.59   1st Qu.: 401.4   Class :character
##  Mode  :character   Median : 70.35   Median :1397.5   Mode  :character
##                     Mean   : 64.80   Mean   :2283.3
##                     3rd Qu.: 89.86   3rd Qu.:3794.7
##                     Max.   :118.75   Max.   :8684.8
```

Convert all No internet service to No

```
df <- data.frame(lapply(df, function(x) {
  gsub("No internet service", "No", x)}))
```

Split data into two categories: categorical and continuous

```
int <- c("tenure", "MonthlyCharges", "TotalCharges")
df[int] <- sapply(df[int], as.numeric)
df_int <- df[,c("tenure", "MonthlyCharges", "TotalCharges")]
df_int <- data.frame(scale(df_int))                            ## Scaling the
numeric data

df_cat <- df[,-c(5,7,8,15,17,18,19)]
df_dummy<- data.frame(sapply(df_cat,function(x) data.frame(model.matrix(~x-
1,data =df_cat))[,-1]))
```

#create dummy variables for for than 2 categories

```
df_cat2 <- df[,c(7,8,15,17)]

df_cat2$MultipleLines[df_cat2$MultipleLines == "Yes"] <- 1      # Replace
"Yes" by 1
df_cat2$MultipleLines[df_cat2$MultipleLines == "No"] <- 0      # Replace "No"
by 0
df_cat2$MultipleLines[df_cat2$MultipleLines == "No phone service"] <- 0
# Replace "No phone service" by 0
df_cat2$MultipleLines <- as.factor(df_cat2$MultipleLines)

df_cat2$InternetService[df_cat2$InternetService == "DSL"] <- 1      # Replace
"DSL" by 1
df_cat2$InternetService[df_cat2$InternetService == "No"] <- 0      # Replace
"No" by 0
df_cat2$InternetService[df_cat2$InternetService == "Fiber optic"] <- 2      #
Replace "Fiber optic" by 2
df_cat2$InternetService <- as.factor(df_cat2$InternetService)

df_cat2$Contract[df_cat2$Contract == "Month-to-month"] <- 1      # Replace
"Month-to-month" by 1
df_cat2$Contract[df_cat2$Contract == "One year"] <- 0      # Replace "One
year" by 0
df_cat2$Contract[df_cat2$Contract == "Two year"] <- 2      # Replace "Two
year" by 2
df_cat2$Contract <- as.factor(df_cat2$Contract)
```

```
df_cat2$PaymentMethod[df_cat2$PaymentMethod == "Electronic check"] <- 1
# Replace "Electronic check" by 1
df_cat2$PaymentMethod[df_cat2$PaymentMethod == "Mailed check"] <- 0      #
Replace "Mailed check" by 0
df_cat2$PaymentMethod[df_cat2$PaymentMethod == "Bank transfer (automatic)"]
<- 2      # Replace "Bank transfer (automatic)" by 2
df_cat2$PaymentMethod[df_cat2$PaymentMethod == "Credit card (automatic)"] <-
3      # Replace "Credit card (automatic)" by 3
df_cat2$PaymentMethod <- as.factor(df_cat2$PaymentMethod)

nrow(df_cat2)

## [1] 7032

nrow(df_int)

## [1] 7032

nrow(df_dummy)

## [1] 7032
```

recombining final data set : data

```
data <- cbind(df_int,df_dummy,df_cat2)
sapply(data, class)

##           tenure   MonthlyCharges     TotalCharges           gender
##        "numeric"        "numeric"        "numeric"        "numeric"
##    SeniorCitizen          Partner       Dependents     PhoneService
##        "numeric"        "numeric"        "numeric"        "numeric"
##   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport
##        "numeric"        "numeric"        "numeric"        "numeric"
##      StreamingTV  StreamingMovies PaperlessBilling            Churn
##        "numeric"        "numeric"        "numeric"        "numeric"
##    MultipleLines  InternetService         Contract    PaymentMethod
##         "factor"         "factor"         "factor"         "factor"

head(data)

##        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 OnlineSecurity OnlineBackup DeviceProtection
## 1          0            0              0            1                0
## 2          0            1              1            0                1
## 3          0            1              1            1                0
## 4          0            0              1            0                1
## 5          0            1              0            0                0
## 6          0            1              0            0                1
##   TechSupport StreamingTV StreamingMovies PaperlessBilling Churn
MultipleLines
## 1           0           0               0                1     0
0
## 2           0           0               0                0     0
0
## 3           0           0               0                1     1
0
## 4           1           0               0                0     0
0
## 5           0           0               0                1     1
0
## 6           0           1               1                1     1
1
##   InternetService Contract PaymentMethod
## 1               1        1             1
## 2               1        0             0
## 3               1        1             0
## 4               1        0             2
## 5               2        1             1
## 6               2        1             1
```

```
describe(data)
```

```
##                  vars    n mean   sd median trimmed  mad   min  max range
skew
## tenure              1 7032 0.00 1.00  -0.14   -0.04 1.33 -1.28 1.61  2.89
0.24
## MonthlyCharges      2 7032 0.00 1.00   0.18    0.01 1.19 -1.55 1.79  3.34
-0.22
## TotalCharges        3 7032 0.00 1.00  -0.39   -0.14 0.80 -1.00 2.82  3.82
0.96
## gender              4 7032 0.50 0.50   1.00    0.51 0.00  0.00 1.00  1.00
-0.02
## SeniorCitizen       5 7032 0.16 0.37   0.00    0.08 0.00  0.00 1.00  1.00
1.83
## Partner             6 7032 0.48 0.50   0.00    0.48 0.00  0.00 1.00  1.00
0.07
## Dependents          7 7032 0.30 0.46   0.00    0.25 0.00  0.00 1.00  1.00
0.88
## PhoneService        8 7032 0.90 0.30   1.00    1.00 0.00  0.00 1.00  1.00
-2.73
## OnlineSecurity      9 7032 0.29 0.45   0.00    0.23 0.00  0.00 1.00  1.00
0.94
## OnlineBackup       10 7032 0.34 0.48   0.00    0.31 0.00  0.00 1.00  1.00
0.65
## DeviceProtection   11 7032 0.34 0.48   0.00    0.30 0.00  0.00 1.00  1.00
```

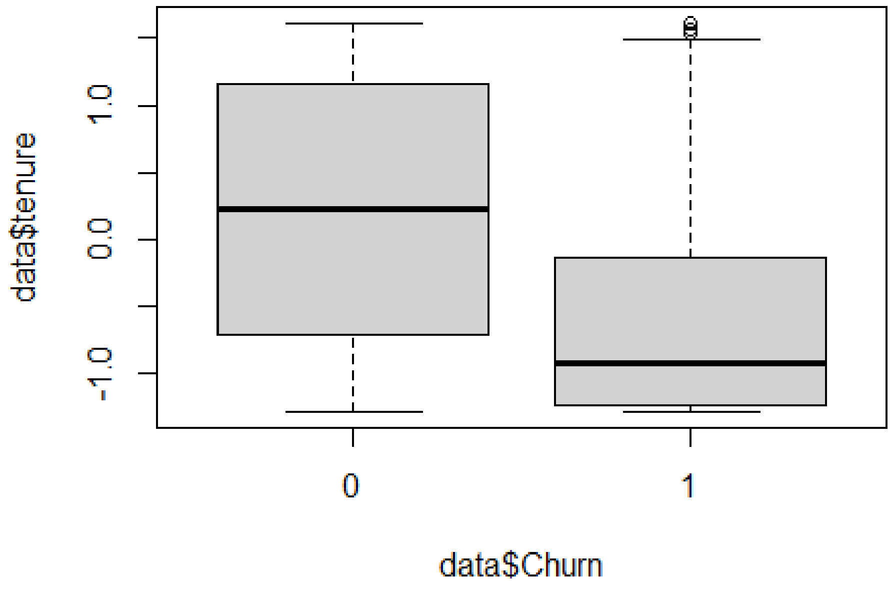
```
0.66
## TechSupport        12 7032 0.29 0.45   0.00    0.24 0.00  0.00 1.00  1.00
0.92
## StreamingTV        13 7032 0.38 0.49   0.00    0.36 0.00  0.00 1.00  1.00
0.48
## StreamingMovies    14 7032 0.39 0.49   0.00    0.36 0.00  0.00 1.00  1.00
0.46
## PaperlessBilling   15 7032 0.59 0.49   1.00    0.62 0.00  0.00 1.00  1.00
-0.38
## Churn              16 7032 0.27 0.44   0.00    0.21 0.00  0.00 1.00  1.00
1.06
## MultipleLines*     17 7032 1.42 0.49   1.00    1.40 0.00  1.00 2.00  1.00
0.32
## InternetService*   18 7032 2.22 0.78   2.00    2.28 1.48  1.00 3.00  2.00
-0.41
## Contract*          19 7032 2.03 0.67   2.00    2.04 0.00  1.00 3.00  2.00
-0.03
## PaymentMethod*     20 7032 2.42 1.06   2.00    2.40 1.48  1.00 4.00  3.00
0.17
##                  kurtosis   se
## tenure              -1.39 0.01
## MonthlyCharges      -1.26 0.01
## TotalCharges        -0.23 0.01
## gender              -2.00 0.01
## SeniorCitizen        1.35 0.00
## Partner             -2.00 0.01
## Dependents          -1.22 0.01
## PhoneService         5.45 0.00
## OnlineSecurity      -1.11 0.01
## OnlineBackup        -1.57 0.01
## DeviceProtection    -1.57 0.01
## TechSupport         -1.14 0.01
## StreamingTV         -1.77 0.01
## StreamingMovies     -1.79 0.01
## PaperlessBilling    -1.86 0.01
## Churn               -0.88 0.01
## MultipleLines*      -1.90 0.01
## InternetService*    -1.24 0.01
## Contract*           -0.77 0.01
## PaymentMethod*      -1.20 0.01
```

```
unique(data$MultipleLines)
```

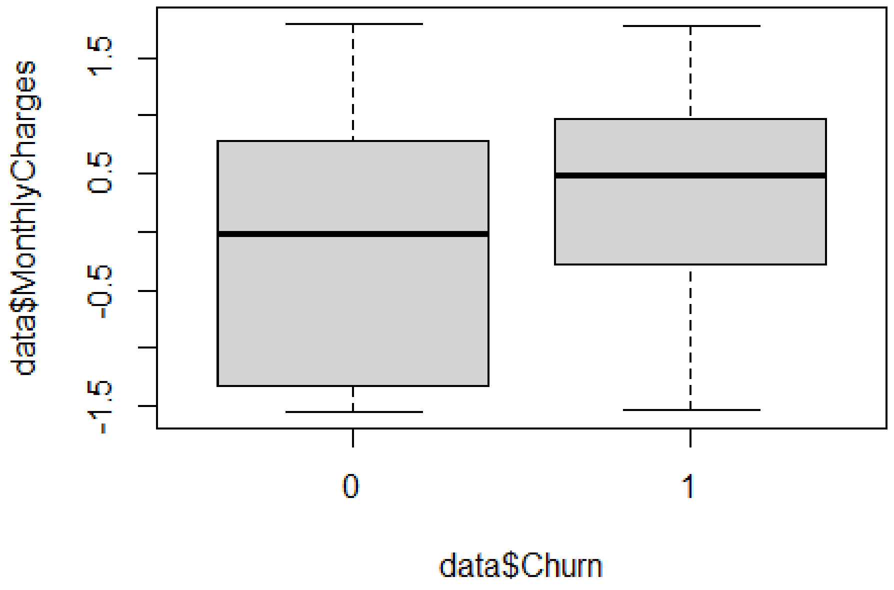
```
## [1] 0 1
## Levels: 0 1
```

Considering Out-liers

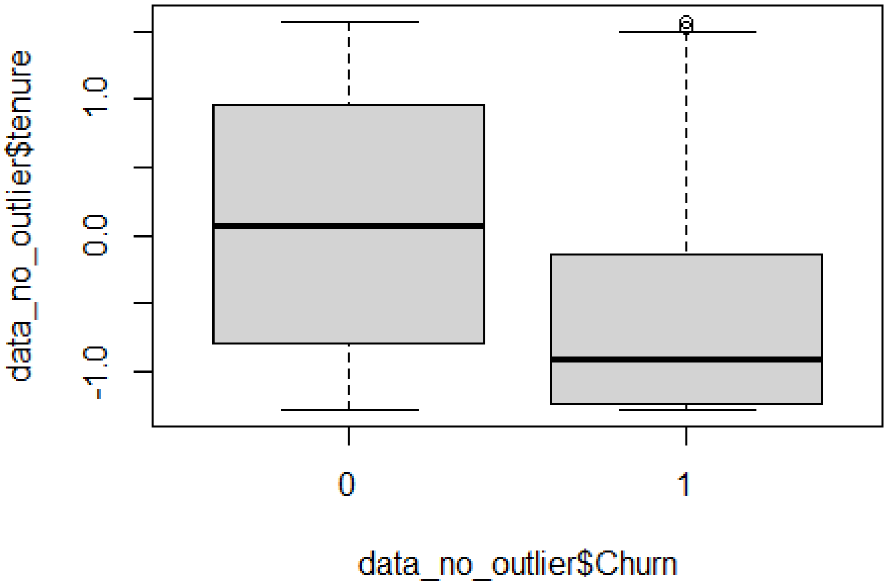
```
boxplot(data$tenure~data$Churn)
```



```
boxplot(data$MonthlyCharges~data$Churn)
```



```
quartiles <- quantile(data$tenure, probs=c(.25, .75), na.rm = FALSE)
IQR <- IQR(data$tenure)

Upper <- quartiles[2] + 0.369*IQR

data_no_outlier <- subset(data,data$tenure < Upper)
nrow(data_no_outlier)

## [1] 6670

boxplot(data_no_outlier$tenure~data_no_outlier$Churn)
```



## 3% data lost for outliers which are not too far out, therefore we keep original data with outliers.

Variable Selection

```
library(reshape2)

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

creating correlation matrix

```
corr_mat <- round(cor(df_int),2)
```

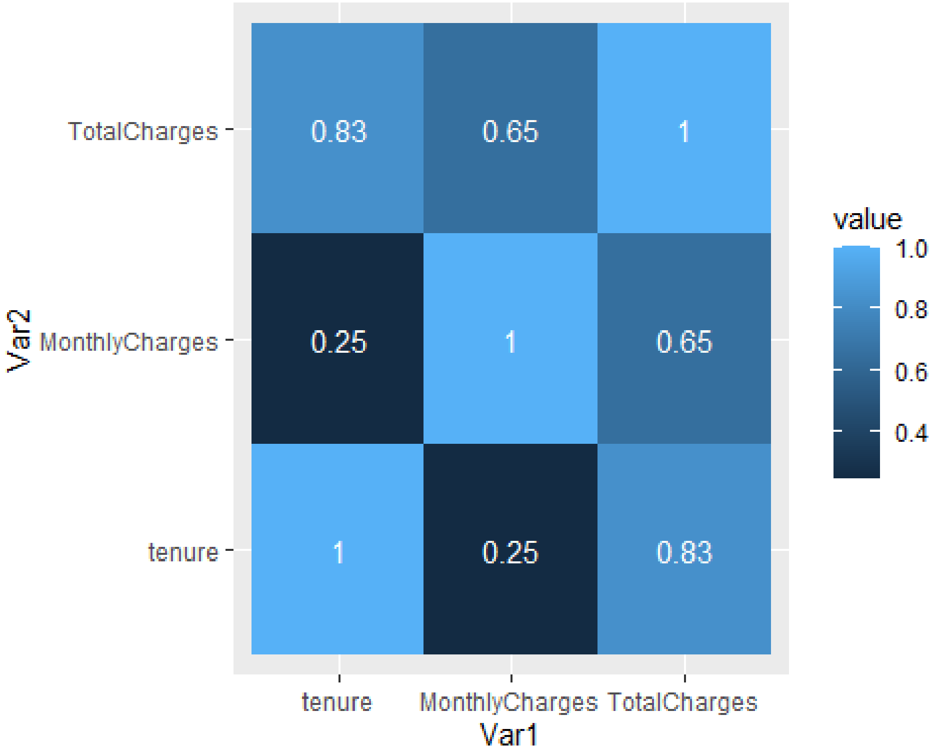
reduce the size of correlation matrix

```
melted_corr_mat <- melt(corr_mat)
head(melted_corr_mat)
```

```
##             Var1           Var2 value
## 1         tenure         tenure  1.00
## 2 MonthlyCharges         tenure  0.25
## 3   TotalCharges         tenure  0.83
## 4         tenure MonthlyCharges  0.25
## 5 MonthlyCharges MonthlyCharges  1.00
## 6   TotalCharges MonthlyCharges  0.65
```

plotting the correlation heatmap

```
library(ggplot2)
ggplot(data = melted_corr_mat, aes(x=Var1, y=Var2,
                                   fill=value)) +
  geom_tile() +
  geom_text(aes(Var2, Var1, label = value),
            color = "white", size = 4)
```



## We can observe a high correlation for Total charge, therefore we can drop it.

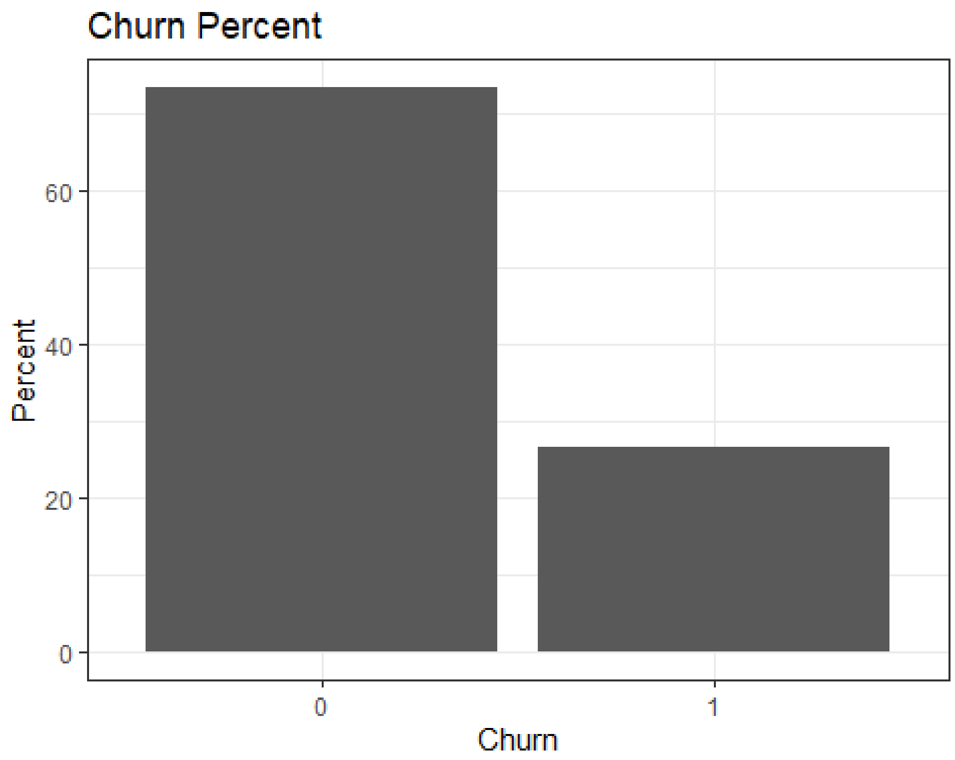
```
data <- data[, !(colnames(data) %in% c("TotalCharges"))]
head(data)
```

```
##        tenure MonthlyCharges gender SeniorCitizen Partner Dependents
## 1 -1.28015700     -1.1616113      0             0       1          0
## 2  0.06429811     -0.2608594      1             0       0           0
## 3 -1.23941594     -0.3638974      1             0       0           0
## 4  0.51244982     -0.7477972      1             0       0           0
## 5 -1.23941594      0.1961642      0             0       0           0
## 6 -0.99496955      1.1584066      0             0       0           0
##   PhoneService OnlineSecurity OnlineBackup DeviceProtection TechSupport
## 1            0              0            1                0           0
## 2            1              1            0                1           0
## 3            1              1            1                0           0
## 4            0              1            0                1           1
## 5            1              0            0                0           0
## 6            1              0            0                1           0
##   StreamingTV StreamingMovies PaperlessBilling Churn MultipleLines
## 1           0               0                1     0             0
## 2           0               0                0     0             0
## 3           0               0                1     1             0
## 4           0               0                0     0             0
## 5           0               0                1     1             0
## 6           1               1                1     1             1
##   InternetService Contract PaymentMethod
## 1               1        1             1
## 2               1        0             0
## 3               1        1             0
## 4               1        0             2
## 5               2        1             1
## 6               2        1             1
```

EDA

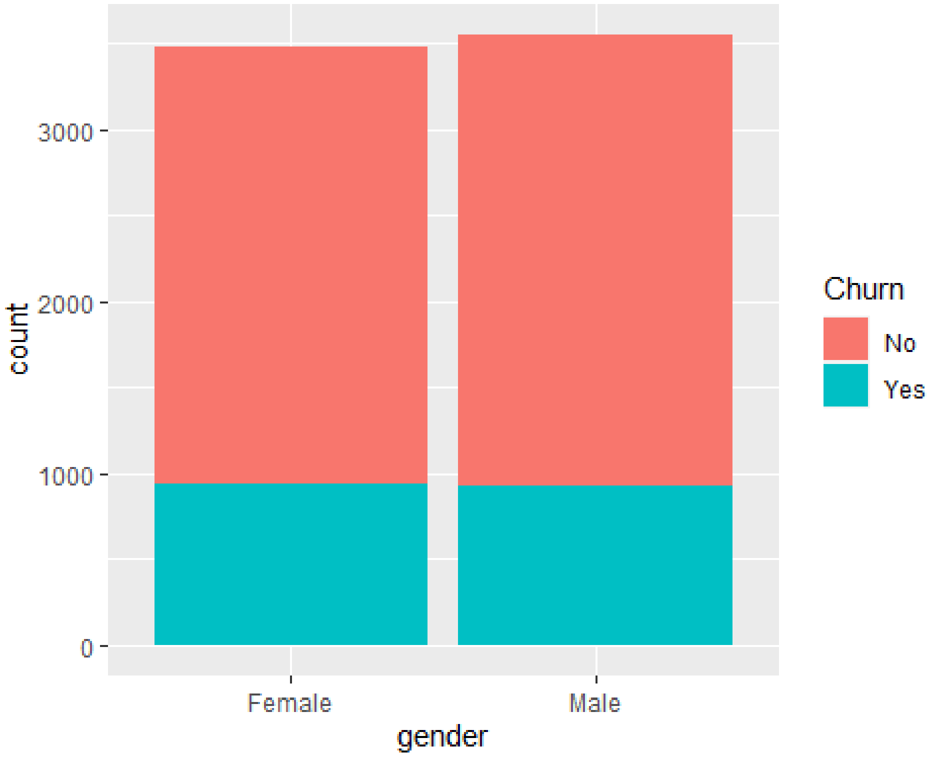
Churn percent

```
data %>%
  group_by(Churn) %>%
  summarise(Count = n())%>%
  mutate(percent = prop.table(Count)*100)%>%
  ggplot(aes(reorder(Churn, -percent), percent), fill = Churn)+
  geom_col()+
  theme_bw()+
  xlab("Churn") +
  ylab("Percent")+
  ggtitle("Churn Percent")
```

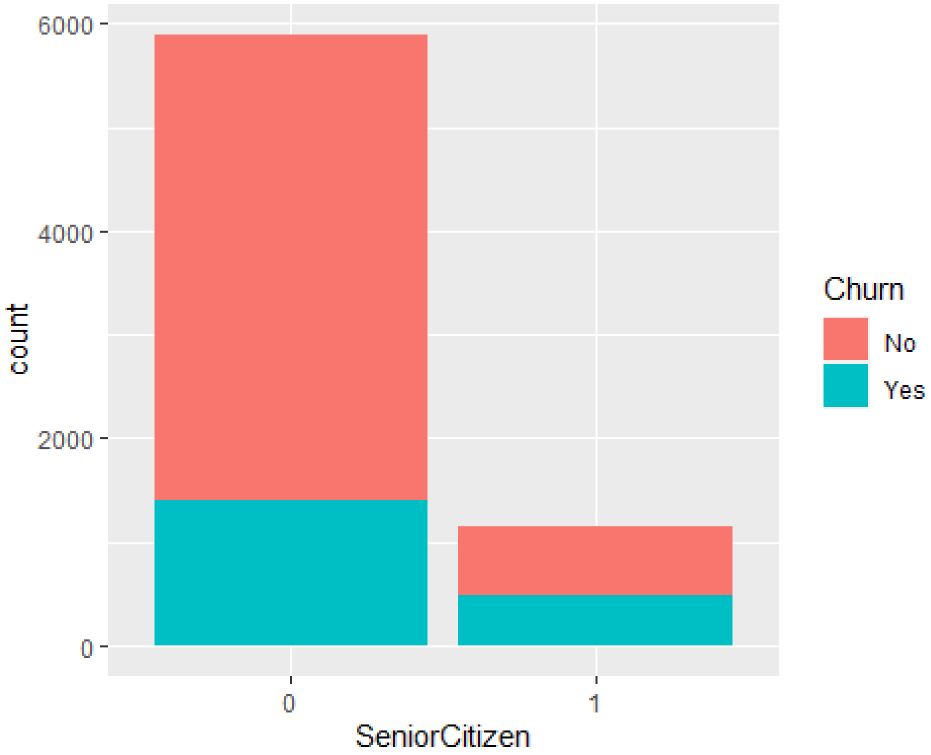


## Chart for demographic data

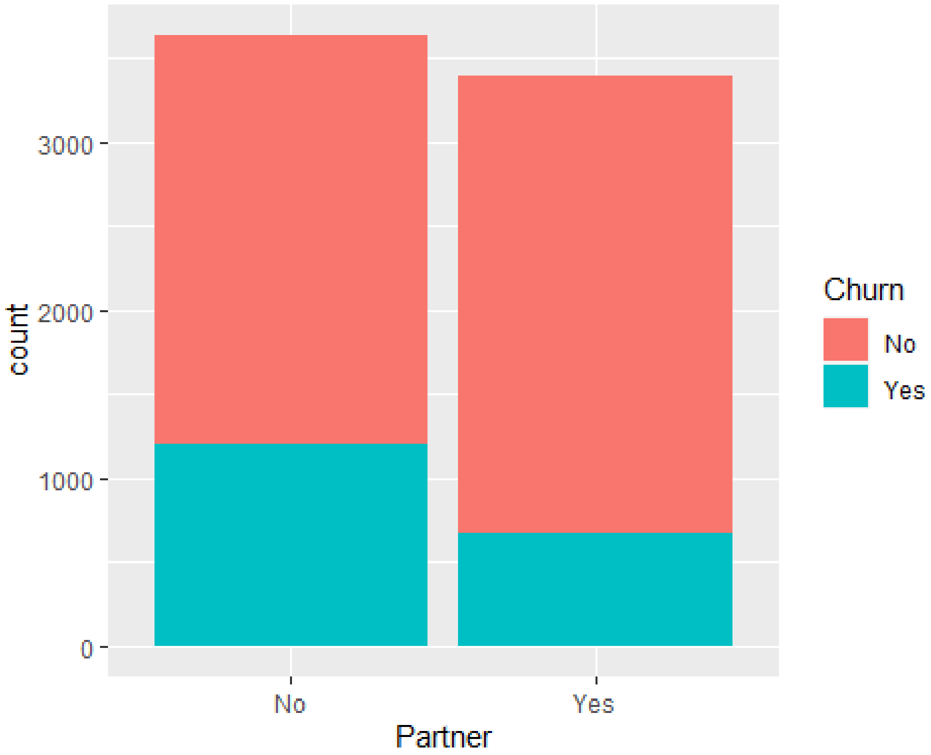
```
ggplot(df, aes(x=gender,fill=Churn))+ geom_bar()
```



```
ggplot(df, aes(x=SeniorCitizen,fill=Churn))+ geom_bar()
```



```
ggplot(df, aes(x=Partner,fill=Churn))+ geom_bar()
```



```
ggplot(df, aes(x=Dependents,fill=Churn))+ geom_bar()
```

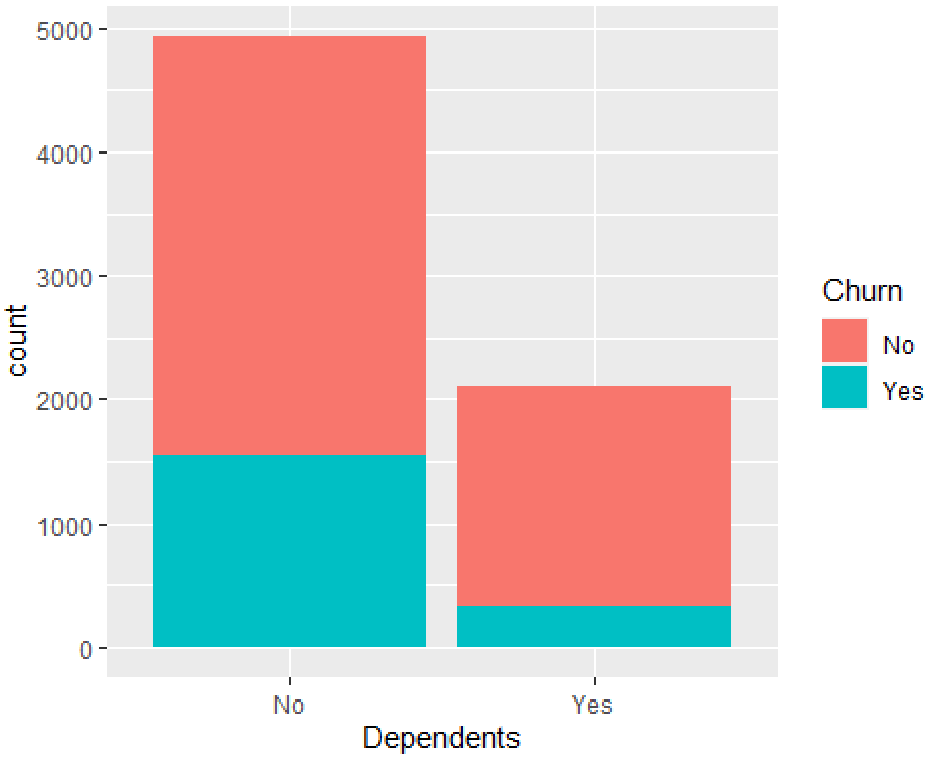
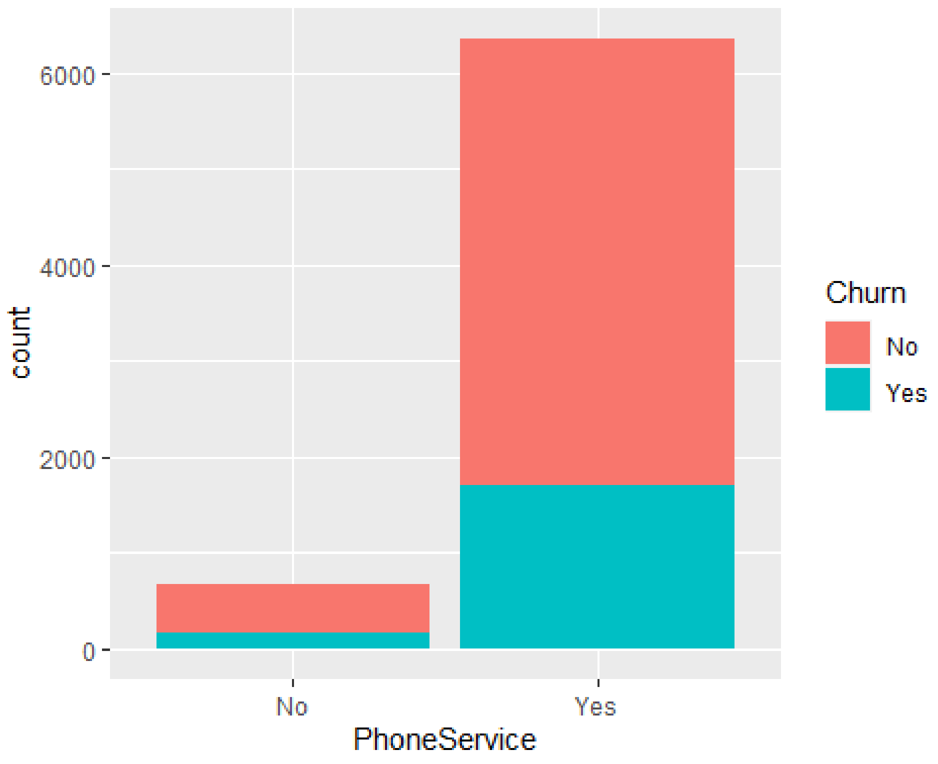
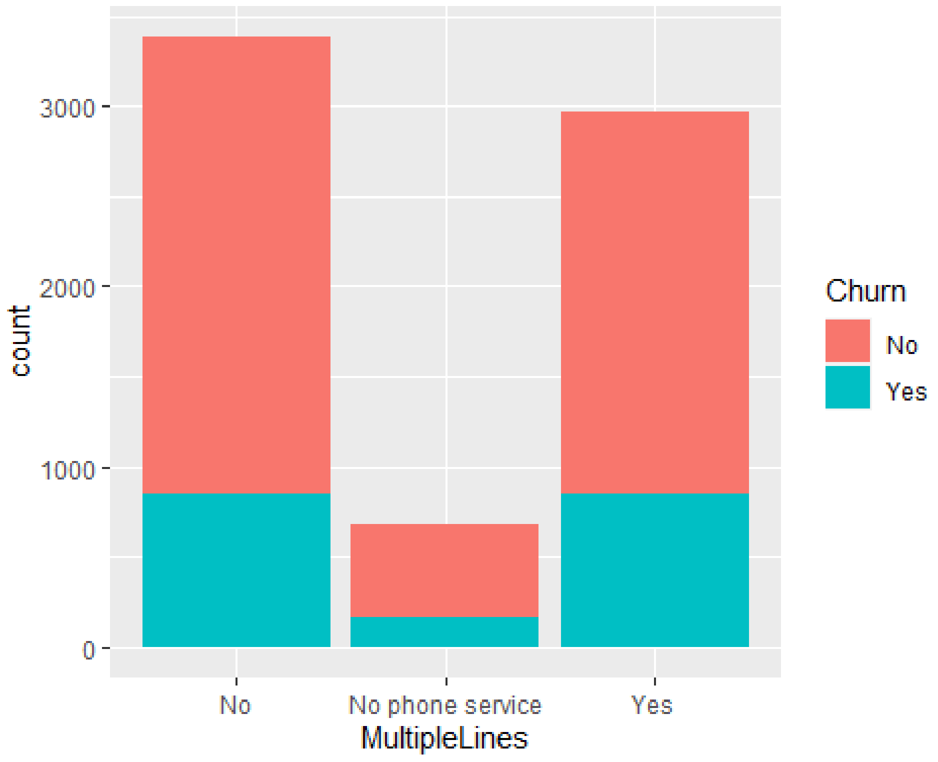


Chart for Service data

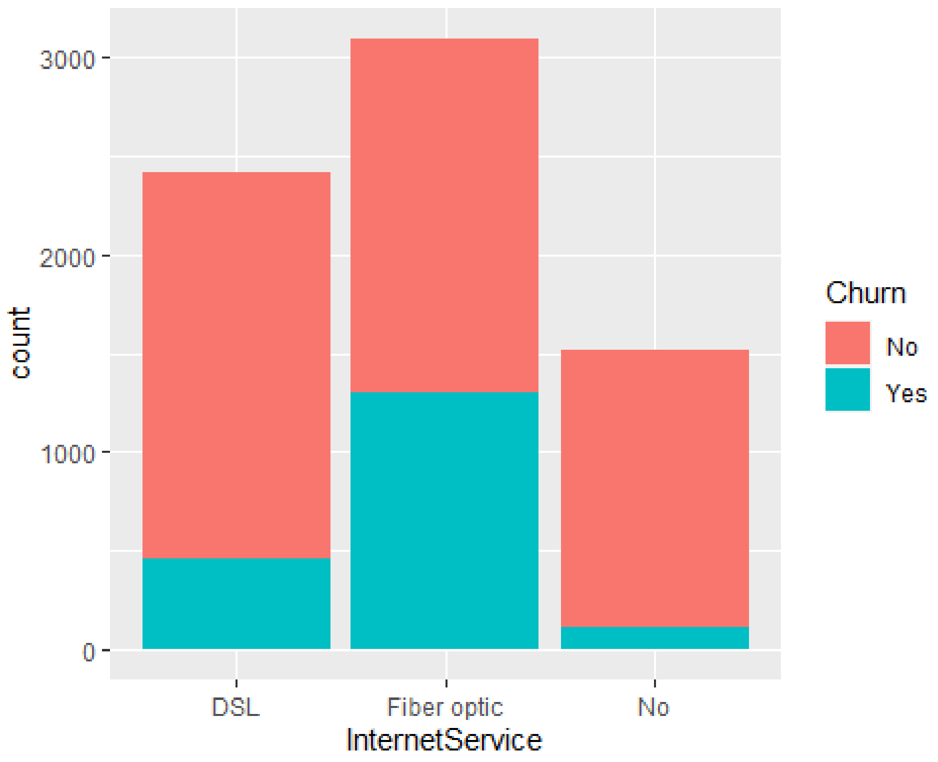
```
ggplot(df, aes(x=PhoneService,fill=Churn))+ geom_bar()
```



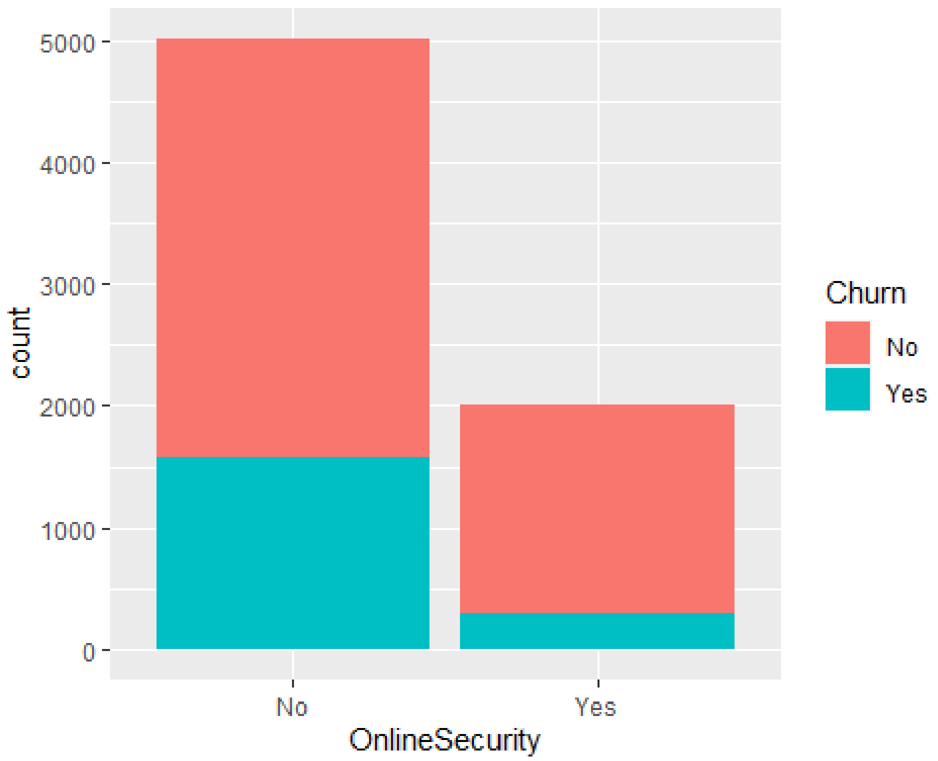
```
ggplot(df, aes(x=MultipleLines,fill=Churn))+ geom_bar()
```



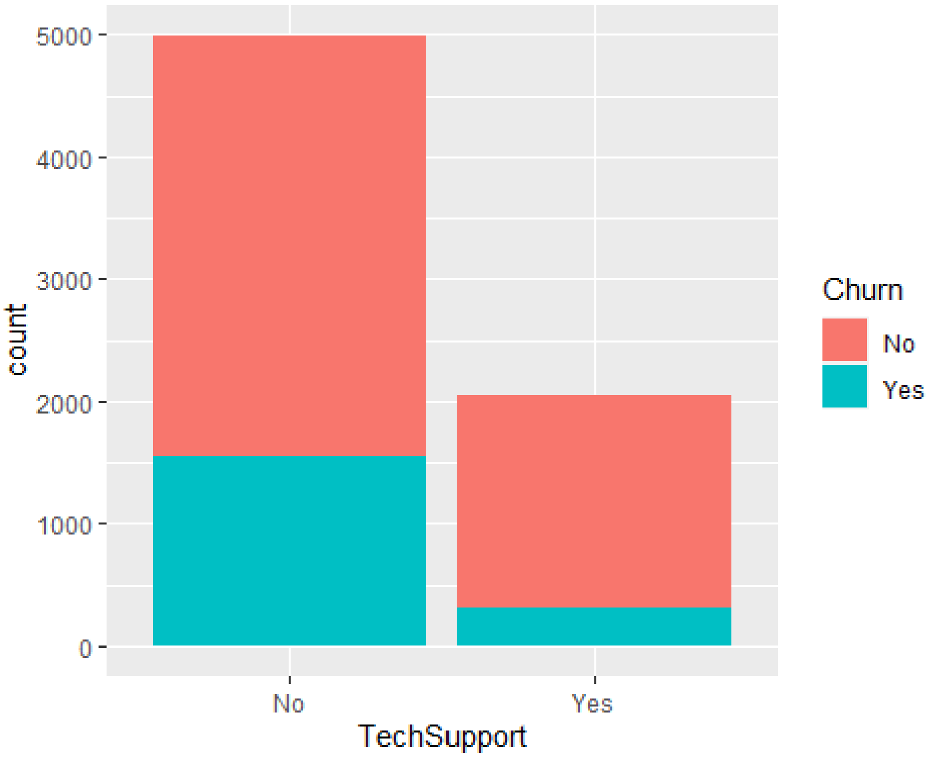
```
ggplot(df, aes(x=InternetService,fill=Churn))+ geom_bar()
```



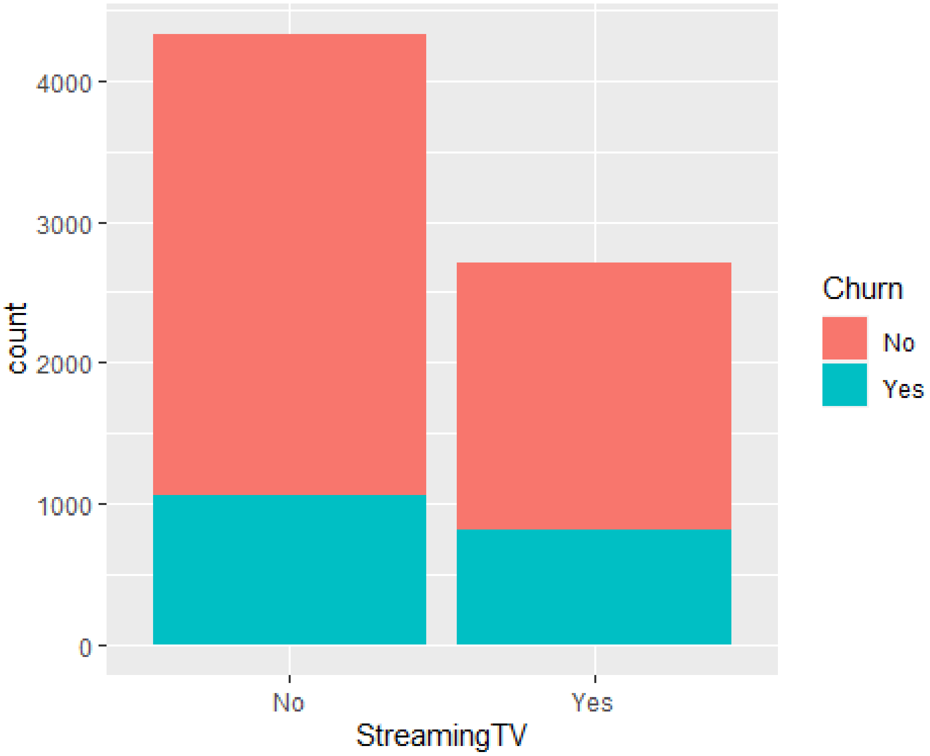
```
ggplot(df, aes(x=OnlineSecurity,fill=Churn))+ geom_bar()
```



```
ggplot(df, aes(x=TechSupport,fill=Churn))+ geom_bar()
```

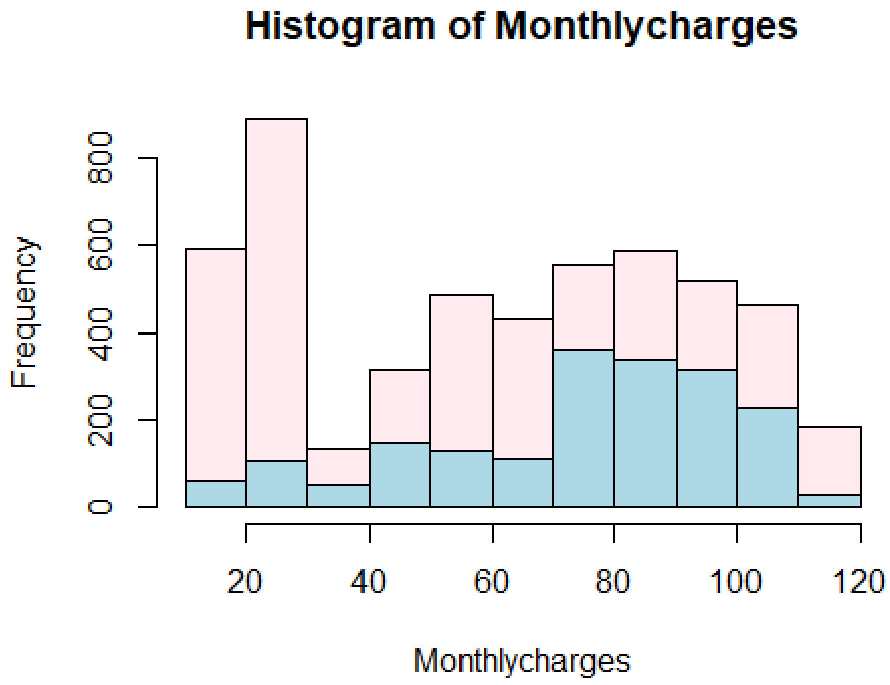


```
ggplot(df, aes(x=StreamingTV,fill=Churn))+ geom_bar()
```



## Comaprong churn rate for continuous variables

```
Churn_by_Tenure <- df$tenure[df$Churn == "Yes"]
tenchn <- df$tenure[df$Churn == "No"]
a <- hist(Churn_by_Tenure, plot = FALSE)
b <- hist(tenchn, plot = FALSE)
c1 <- rgb(173,216,230,max = 255,  names = "lt.blue")
c2 <- rgb(255,192,203, max = 255, alpha = 80, names = "lt.pink")

Churn_by_MonthlyCharges<- df$MonthlyCharges[df$Churn == "Yes"]
Monthlycharges <- df$MonthlyCharges[df$Churn == "No"]
a <- hist(Churn_by_MonthlyCharges, plot = FALSE)
b <- hist(Monthlycharges, plot = FALSE)
c1 <- rgb(173,216,230,max = 255,  names = "lt.blue")
c2 <- rgb(255,192,203, max = 255, alpha = 80, names = "lt.pink")
plot(b, col = c2)
plot(a, col = c1, add = TRUE)
```



## Logit Model

```
split1<- sample(c(rep(0, 0.7 * nrow(data)), rep(1, 0.3 * nrow(data))))
train <- data[split1 == 0, ]
test <- data[split1== 1, ]
```

with all varibles

```
glm <- glm(Churn ~., data = train)
summary(glm)

##
## Call:
## glm(formula = Churn ~ ., data = train)
```

```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.75244  -0.26355  -0.07115   0.29746   1.11259
##
## Coefficients:
##                    Estimate Std. Error t value             Pr(>|t|)
## (Intercept)      -0.292572   0.338229  -0.865             0.387075
## tenure           -0.115352   0.008833 -13.060 < 0.0000000000000002 ***
## MonthlyCharges   -0.153136   0.156238  -0.980             0.327063
## gender           -0.012809   0.010604  -1.208             0.227118
## SeniorCitizen     0.035855   0.015652   2.291             0.022021 *
## Partner          -0.014109   0.012884  -1.095             0.273545
## Dependents       -0.012581   0.013603  -0.925             0.355077
## PhoneService      0.051533   0.106181   0.485             0.627464
## OnlineSecurity   -0.048511   0.029561  -1.641             0.100853
## OnlineBackup     -0.001023   0.028905  -0.035             0.971774
## DeviceProtection  0.006639   0.029452   0.225             0.821678
## TechSupport      -0.048398   0.029681  -1.631             0.103035
## StreamingTV       0.076892   0.053462   1.438             0.150422
## StreamingMovies   0.080387   0.053574   1.500             0.133558
## PaperlessBilling  0.054380   0.011842   4.592       0.000004501279 ***
## MultipleLines1    0.057788   0.029004   1.992             0.046382 *
## InternetService1  0.267080   0.131970   2.024             0.043046 *
## InternetService2  0.532871   0.260883   2.043             0.041149 *
## Contract1         0.104059   0.016474   6.317       0.000000000291 ***
## Contract2         0.056675   0.016902   3.353             0.000805 ***
## PaymentMethod1    0.078751   0.016409   4.799       0.000001640139 ***
## PaymentMethod2    0.015318   0.017129   0.894             0.371218
## PaymentMethod3    0.006366   0.016983   0.375             0.707811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1379319)
##
##     Null deviance: 942.86  on 4922  degrees of freedom
## Residual deviance: 675.87  on 4900  degrees of freedom
## AIC: 4243.4
##
## Number of Fisher Scoring iterations: 2
```

```
tab_model(glm)
```

```
vif(glm)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## tenure             2.793465  1        1.671366
## MonthlyCharges   870.461945  1       29.503592
## gender             1.003232  1        1.001615
## SeniorCitizen      1.157381  1        1.075817
## Partner            1.479610  1        1.216392
## Dependents         1.382614  1        1.175846
## PhoneService      35.145940  1        5.928401
## OnlineSecurity     6.475409  1        2.544682
## OnlineBackup       6.730225  1        2.594268
## DeviceProtection   7.007018  1        2.647077
## TechSupport        6.509397  1        2.551352
## StreamingTV       24.156834  1        4.914960
## StreamingMovies   24.156538  1        4.914930
## PaperlessBilling   1.209260  1        1.099664
## MultipleLines      7.310568  1        2.703806
## InternetService  629.341349  2        5.008660
## Contract           2.543514  2        1.262870
## PaymentMethod      1.565242  3        1.077532
```

High VIFs, insignificant variables, so we can also use step (naive method)

```
model_2<- stepAIC(glm, direction="both")
```

```
## Start:  AIC=4243.38
## Churn ~ tenure + MonthlyCharges + gender + SeniorCitizen + Partner + 
##     Dependents + PhoneService + OnlineSecurity + OnlineBackup + 
##     DeviceProtection + TechSupport + StreamingTV + StreamingMovies + 
##     PaperlessBilling + MultipleLines + InternetService + Contract + 
##     PaymentMethod
## 
##                    Df Deviance    AIC
## - OnlineBackup      1   675.87 4241.4
## - DeviceProtection  1   675.87 4241.4
## - PhoneService      1   675.90 4241.6
## - Dependents        1   675.98 4242.2
## - MonthlyCharges    1   676.00 4242.3
## - Partner           1   676.03 4242.6
## - gender            1   676.07 4242.8
## <none>                  675.87 4243.4
## - StreamingTV       1   676.15 4243.5
## - InternetService   2   676.44 4243.6
## - StreamingMovies   1   676.18 4243.6
## - TechSupport       1   676.23 4244.0
## - OnlineSecurity    1   676.24 4244.1
## - MultipleLines     1   676.41 4245.4
## - SeniorCitizen     1   676.59 4246.6
## - PaperlessBilling  1   678.77 4262.5
## - PaymentMethod     3   680.30 4269.5
## - Contract          2   681.73 4281.9
## - tenure            1   699.39 4409.8
## 
## Step:  AIC=4241.38
## Churn ~ tenure + MonthlyCharges + gender + SeniorCitizen + Partner + 
##     Dependents + PhoneService + OnlineSecurity + DeviceProtection + 
##     TechSupport + StreamingTV + StreamingMovies + PaperlessBilling +
##     MultipleLines + InternetService + Contract + PaymentMethod
##
##                    Df Deviance    AIC
## - DeviceProtection  1   675.89 4239.6
## - Dependents        1   675.98 4240.2
## - PhoneService      1   676.02 4240.5
## - Partner           1   676.03 4240.6
## - gender            1   676.07 4240.8
## <none>                  675.87 4241.4
## + OnlineBackup      1   675.87 4243.4
## - MonthlyCharges    1   676.56 4244.4
## - SeniorCitizen     1   676.59 4244.6
## - TechSupport       1   676.74 4245.7
## - OnlineSecurity    1   676.77 4245.9
## - StreamingTV       1   677.00 4247.6
## - StreamingMovies   1   677.10 4248.4
## - MultipleLines     1   677.42 4250.7
## - InternetService   2   678.64 4257.5
## - PaperlessBilling  1   678.78 4260.5
## - PaymentMethod     3   680.30 4267.6
## - Contract          2   681.73 4279.9
## - tenure            1   699.63 4409.5
##
## Step:  AIC=4239.55
## Churn ~ tenure + MonthlyCharges + gender + SeniorCitizen + Partner +
##     Dependents + PhoneService + OnlineSecurity + TechSupport +
##     StreamingTV + StreamingMovies + PaperlessBilling + MultipleLines +
##     InternetService + Contract + PaymentMethod
##
##                    Df Deviance    AIC
## - Dependents        1   676.01 4238.4
## - PhoneService      1   676.03 4238.6
## - Partner           1   676.05 4238.7
## - gender            1   676.09 4239.0
## <none>                  675.89 4239.6
## + DeviceProtection  1   675.87 4241.4
## + OnlineBackup      1   675.87 4241.4
## - SeniorCitizen     1   676.61 4242.8
## - MonthlyCharges    1   676.84 4244.4
## - TechSupport       1   677.07 4246.1
## - OnlineSecurity    1   677.13 4246.6
## - StreamingTV       1   677.24 4247.3
## - StreamingMovies   1   677.33 4248.1
## - MultipleLines     1   677.63 4250.2
## - PaperlessBilling  1   678.78 4258.6
## - InternetService   2   680.01 4265.5
## - PaymentMethod     3   680.31 4265.7
## - Contract          2   681.74 4277.9
## - tenure            1   699.73 4408.2
```

```
##
## Step:  AIC=4238.43
## Churn ~ tenure + MonthlyCharges + gender + SeniorCitizen + Partner +
##     PhoneService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + PaperlessBilling + MultipleLines + InternetService +
##     Contract + PaymentMethod
##
##                    Df Deviance    AIC
## - PhoneService      1   676.15 4237.5
## - gender            1   676.21 4237.9
## <none>                  676.01 4238.4
## - Partner           1   676.40 4239.3
## + Dependents        1   675.89 4239.6
## + DeviceProtection  1   675.98 4240.2
## + OnlineBackup      1   675.99 4240.3
## - SeniorCitizen     1   676.88 4242.7
## - MonthlyCharges    1   676.96 4243.3
## - TechSupport       1   677.19 4245.0
## - OnlineSecurity    1   677.27 4245.6
## - StreamingTV       1   677.35 4246.2
## - StreamingMovies   1   677.47 4247.0
## - MultipleLines     1   677.78 4249.3
## - PaperlessBilling  1   678.92 4257.6
## - InternetService   2   680.15 4264.5
## - PaymentMethod     3   680.48 4264.9
## - Contract          2   681.91 4277.2
## - tenure            1   699.77 4406.5
##
## Step:  AIC=4237.48
## Churn ~ tenure + MonthlyCharges + gender + SeniorCitizen + Partner +
##     OnlineSecurity + TechSupport + StreamingTV + StreamingMovies +
##     PaperlessBilling + MultipleLines + InternetService + Contract +
##     PaymentMethod
##
##                    Df Deviance    AIC
## - gender            1   676.36 4237.0
## <none>                  676.15 4237.5
## - Partner           1   676.55 4238.4
## + PhoneService      1   676.01 4238.4
## + Dependents        1   676.03 4238.6
## + OnlineBackup      1   676.04 4238.6
## + DeviceProtection  1   676.14 4239.4
## - SeniorCitizen     1   677.00 4241.6
## - StreamingTV       1   677.71 4246.8
## - MonthlyCharges    1   677.77 4247.2
## - MultipleLines     1   677.80 4247.5
## - StreamingMovies   1   677.93 4248.4
## - TechSupport       1   678.25 4250.7
## - OnlineSecurity    1   678.33 4251.3
## - PaperlessBilling  1   679.06 4256.6
## - PaymentMethod     3   680.69 4264.4
## - Contract          2   682.11 4276.6
## - InternetService   2   685.66 4302.2
## - tenure            1   701.97 4419.9
## 
## Step:  AIC=4236.98
## Churn ~ tenure + MonthlyCharges + SeniorCitizen + Partner + OnlineSecurity 
+ 
##     TechSupport + StreamingTV + StreamingMovies + PaperlessBilling + 
##     MultipleLines + InternetService + Contract + PaymentMethod
## 
##                    Df Deviance    AIC
## <none>                  676.36 4237.0
## + gender            1   676.15 4237.5
## - Partner           1   676.75 4237.8
## + PhoneService      1   676.21 4237.9
## + Dependents        1   676.24 4238.1
## + OnlineBackup      1   676.25 4238.2
## + DeviceProtection  1   676.35 4238.9
## - SeniorCitizen     1   677.21 4241.1
## - StreamingTV       1   677.91 4246.2
## - MonthlyCharges    1   677.97 4246.7
## - MultipleLines     1   678.00 4246.9
## - StreamingMovies   1   678.14 4247.9
## - TechSupport       1   678.45 4250.1
## - OnlineSecurity    1   678.54 4250.8
## - PaperlessBilling  1   679.29 4256.2
## - PaymentMethod     3   680.87 4263.7
## - Contract          2   682.33 4276.2
## - InternetService   2   685.84 4301.5
## - tenure            1   702.22 4419.7
```

```
summary(model_2)
```

```
## 
## Call:
## glm(formula = Churn ~ tenure + MonthlyCharges + SeniorCitizen + 
##     Partner + OnlineSecurity + TechSupport + StreamingTV + StreamingMovies 
+ 
##     PaperlessBilling + MultipleLines + InternetService + Contract + 
##     PaymentMethod, data = train)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -0.75703  -0.26450  -0.07257   0.30015   1.12719  
## 
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)    
## (Intercept)      -0.160523   0.044098  -3.640             0.000275 ***
## tenure           -0.117318   0.008567 -13.694 < 0.0000000000000002 ***
## MonthlyCharges   -0.092002   0.026954  -3.413             0.000647 ***
## SeniorCitizen     0.038094   0.015366   2.479             0.013204 *  
## Partner          -0.019475   0.011532  -1.689             0.091318 .  
## OnlineSecurity   -0.059297   0.014932  -3.971   0.0000725545152108 ***
## TechSupport      -0.059152   0.015203  -3.891             0.000101 ***
## StreamingTV       0.055573   0.016583   3.351             0.000811 ***
## StreamingMovies   0.059590   0.016574   3.595             0.000327 ***
## PaperlessBilling  0.054465   0.011823   4.607   0.0000041957883788 ***
## MultipleLines1    0.049383   0.014338   3.444             0.000577 ***
## InternetService1  0.208342   0.027910   7.465   0.0000000000000983 ***
## InternetService2  0.426856   0.051583   8.275 < 0.0000000000000002 ***
## Contract1         0.104907   0.016433   6.384   0.0000000001884617 ***
## Contract2         0.056708   0.016889   3.358             0.000792 ***
## PaymentMethod1    0.079953   0.016384   4.880   0.0000010946792201 ***
## PaymentMethod2    0.016618   0.017110   0.971             0.331461    
## PaymentMethod3    0.007373   0.016969   0.434             0.663969    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.1378921)
## 
##     Null deviance: 942.86  on 4922  degrees of freedom
## Residual deviance: 676.36  on 4905  degrees of freedom
## AIC: 4237
## 
## Number of Fisher Scoring iterations: 2
```

```
tab_model(model_2)
```

```
vif(model_2)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## tenure            2.628613  1        1.621300
## MonthlyCharges   25.913845  1        5.090564
## SeniorCitizen     1.115766  1        1.056298
## Partner           1.185616  1        1.088860
## OnlineSecurity    1.652659  1        1.285558
## TechSupport       1.708351  1        1.307039
## StreamingTV       2.324971  1        1.524785
## StreamingMovies   2.312540  1        1.520704
## PaperlessBilling  1.205700  1        1.098044
## MultipleLines     1.786937  1        1.336764
## InternetService  20.951003  2        2.139445
## Contract          2.524024  2        1.260443
## PaymentMethod     1.557410  3        1.076632
```

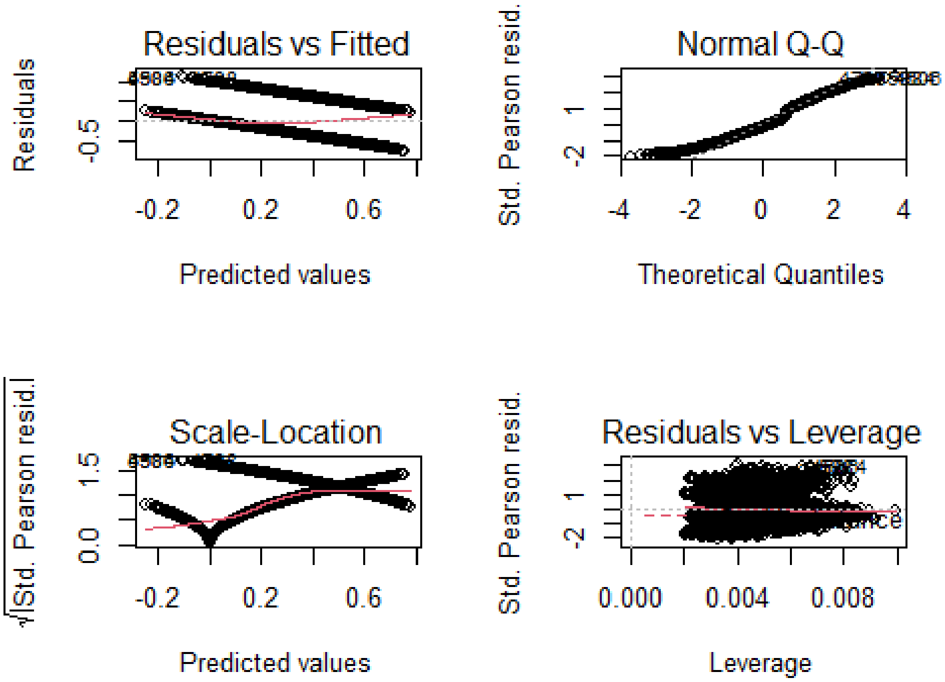
high VIF for monthly charge and Internet Service, we try two models by removing both, one at a time and choose least AIC

```
glm2 <- glm(Churn ~.-InternetService, data = train)
summary(glm2)

## 
## Call:
## glm(formula = Churn ~ . - InternetService, data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.7515  -0.2629  -0.0714   0.2996   1.1305  
## 
## Coefficients:
##                   Estimate Std. Error t value             Pr(>|t|)    
## (Intercept)       0.394062   0.036568  10.776 < 0.0000000000000002 ***
## tenure           -0.115316   0.008789 -13.120 < 0.0000000000000002 ***
## MonthlyCharges    0.165126   0.012133  13.609 < 0.0000000000000002 ***
## gender           -0.013086   0.010605  -1.234             0.217273    
## SeniorCitizen     0.035989   0.015649   2.300             0.021501 *  
## Partner          -0.013685   0.012883  -1.062             0.288166    
## Dependents       -0.013124   0.013603  -0.965             0.334709    
## PhoneService     -0.160252   0.021701  -7.385    0.000000000000179 ***
## OnlineSecurity   -0.101320   0.013706  -7.392    0.000000000000169 ***
## OnlineBackup     -0.053207   0.013348  -3.986    0.000068119660961 ***
## DeviceProtection -0.046303   0.013896  -3.332             0.000869 ***
## TechSupport      -0.101292   0.013958  -7.257    0.000000000000458 ***
## StreamingTV      -0.027907   0.015072  -1.852             0.064146 .  
## StreamingMovies  -0.024648   0.015085  -1.634             0.102342    
## PaperlessBilling  0.054744   0.011841   4.623    0.000003876980549 ***
## MultipleLines1    0.005128   0.013303   0.386             0.699876    
## Contract1         0.104805   0.016442   6.374    0.000000000200696 ***
## Contract2         0.056245   0.016814   3.345             0.000829 ***
## PaymentMethod1    0.079506   0.016380   4.854    0.000001248652427 ***
## PaymentMethod2    0.015567   0.017109   0.910             0.362940    
## PaymentMethod3    0.006943   0.016961   0.409             0.682315    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for gaussian family taken to be 0.1379931)
## 
##     Null deviance: 942.86  on 4922  degrees of freedom
## Residual deviance: 676.44  on 4902  degrees of freedom
## AIC: 4243.6
## 
## Number of Fisher Scoring iterations: 2

tab_model(glm2)
```

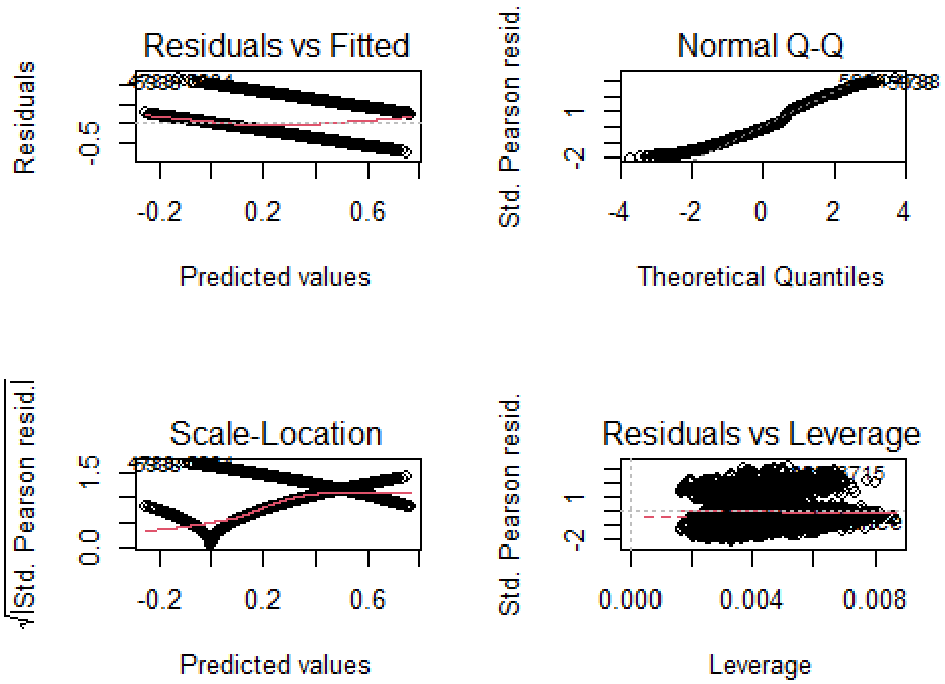
```
vif(glm2)

##                      GVIF Df GVIF^(1/(2*Df))
## tenure           2.764945  1        1.662812
## MonthlyCharges   5.247490  1        2.290740
## gender           1.003051  1        1.001524
## SeniorCitizen    1.156373  1        1.075348
## Partner          1.478739  1        1.216034
## Dependents       1.382026  1        1.175596
## PhoneService     1.467397  1        1.211361
## OnlineSecurity   1.391516  1        1.179625
## OnlineBackup     1.434536  1        1.197721
## DeviceProtection 1.559187  1        1.248674
## TechSupport      1.438998  1        1.199583
## StreamingTV      1.919183  1        1.385346
## StreamingMovies  1.914348  1        1.383600
## PaperlessBilling 1.208510  1        1.099323
## MultipleLines    1.537228  1        1.239850
## Contract         2.491099  2        1.256313
## PaymentMethod    1.557574  3        1.076651
```

heteroskedasticity check for best model from above

```
par(mfrow = c(2, 2))
plot(glm)
```



```
par(mfrow = c(2, 2))
plot(glm2)
```



# Probit Model

```
glmp <- glm(Churn ~.-InternetService, family=binomial(link="probit"), data =
train)
summary(glmp)

##
## Call:
## glm(formula = Churn ~ . - InternetService, family = binomial(link =
"probit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9272  -0.6719  -0.2782   0.6975   3.3487
##
## Coefficients:
##                  Estimate Std. Error z value             Pr(>|z|)
## (Intercept)      -0.34888    0.15545  -2.244              0.02482 *
## tenure           -0.49058    0.03949 -12.422 < 0.0000000000000002 ***
## MonthlyCharges    0.61959    0.05130  12.079 < 0.0000000000000002 ***
## gender           -0.04847    0.04536  -1.068              0.28530
## SeniorCitizen     0.10647    0.06100   1.745              0.08091 .
## Partner          -0.06990    0.05450  -1.283              0.19962
## Dependents       -0.07215    0.06200  -1.164              0.24455
## PhoneService     -0.64963    0.09330  -6.963     0.00000000000333 ***
## OnlineSecurity   -0.34056    0.05821  -5.851     0.00000000489990 ***
## OnlineBackup     -0.13207    0.05464  -2.417              0.01565 *
## DeviceProtection -0.11022    0.05657  -1.948              0.05138 .
## TechSupport      -0.32468    0.05888  -5.514     0.00000003497902 ***
## StreamingTV      -0.06561    0.06077  -1.080              0.28030    
## StreamingMovies  -0.04226    0.06092  -0.694              0.48792    
## PaperlessBilling  0.23139    0.05179   4.468     0.00000789287324 ***
## MultipleLines1    0.04871    0.05739   0.849              0.39604    
## Contract1         0.38298    0.07113   5.384     0.00000007272784 ***
## Contract2        -0.24383    0.10196  -2.391              0.01678 *  
## PaymentMethod1    0.18891    0.06736   2.805              0.00504 ** 
## PaymentMethod2    0.02120    0.07805   0.272              0.78593    
## PaymentMethod3   -0.02426    0.07872  -0.308              0.75794    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 5623.4  on 4922  degrees of freedom
## Residual deviance: 4003.2  on 4902  degrees of freedom
## AIC: 4045.2
## 
## Number of Fisher Scoring iterations: 6

tab_model(glmp)
```

```
vif(glmp)

##                      GVIF Df GVIF^(1/(2*Df))
## tenure           2.295186  1        1.514987
## MonthlyCharges   4.297743  1        2.073100
## gender           1.004354  1        1.002175
## SeniorCitizen    1.144157  1        1.069653
## Partner          1.390103  1        1.179026
## Dependents       1.310397  1        1.144726
## PhoneService     1.539639  1        1.240822
## OnlineSecurity   1.164142  1        1.078954
## OnlineBackup     1.278170  1        1.130562
## DeviceProtection 1.369812  1        1.170390
## TechSupport      1.204596  1        1.097541
## StreamingTV      1.741392  1        1.319618
## StreamingMovies  1.739153  1        1.318770
## PaperlessBilling 1.141358  1        1.068344
## MultipleLines    1.566709  1        1.251682
## Contract         1.679596  2        1.138417
## PaymentMethod    1.408391  3        1.058735
```

#Random Forest Classifier

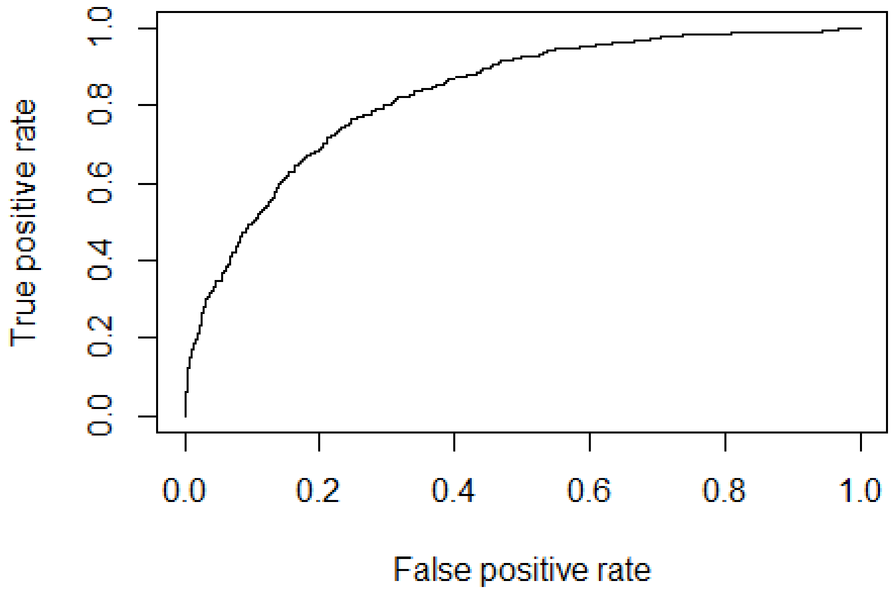
```
data_rf <- data
data_rf$Churn <- as.factor(data$Churn)
indices = sample.split(data_rf$Churn, SplitRatio = 0.7)
```

```
train1 = data_rf[indices,]
test1 = data_rf[!(indices),]

model.rf <- randomForest(Churn ~ ., data=train1, proximity=FALSE,importance =
FALSE,
                         ntree=500,mtry=4, do.trace=FALSE)
model.rf

##
## Call:
##  randomForest(formula = Churn ~ ., data = train1, proximity = FALSE,
importance = FALSE, ntree = 500, mtry = 4, do.trace = FALSE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 21.03%
## Confusion matrix:
##      0   1 class.error
## 0 3223 391   0.1081904
## 1  644 664   0.4923547

accuracy = (3254+644)/(3254+360+664+664)
accuracy

## [1] 0.7887495

RFPred <- predict(model.rf, newdata=test[,-24])
```

ROC Curves and accuracy

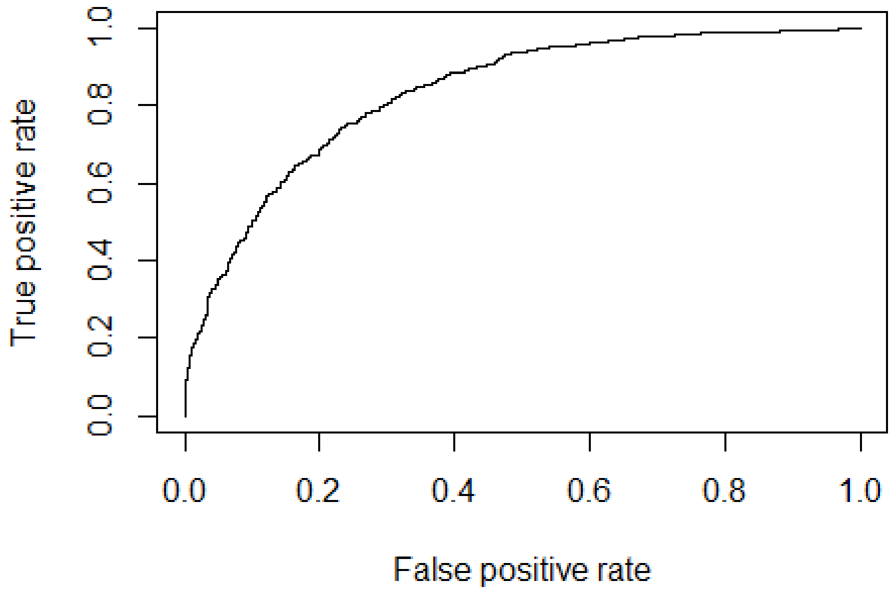
logit

```
glm_result <- predict(glm2, newdata = test, type = "response")

pred_log <- prediction(glm_result, test$Churn)
table(test$Churn, glm_result>0.5)

##
##     FALSE TRUE
##   0  1357  154
##   1   300  298

accuracy = (1418+147)/(1418+147+270+275)
accuracy

## [1] 0.7417062

glmpred <- predict(glm2, type = "response", newdata = test[,-24])

glm_roc <- performance(pred_log, "tpr", "fpr")
plot(glm_roc)
```



## Probit

```
prb_result <- predict(glmp, newdata = test, type = "response")

pred_log_prb <- prediction(prb_result, test$Churn)
table(test$Churn, glm_result>0.6)

##
##      FALSE TRUE
##   0  1468   43
##   1   433  165

accuracy = (1505+145)/(1505+145+60+400)
accuracy

## [1] 0.7819905

glmppred <- predict(glmp, type = "response", newdata = test[,-24])

glmp_roc <- performance(pred_log_prb, "tpr", "fpr")
plot(glmp_roc)
```

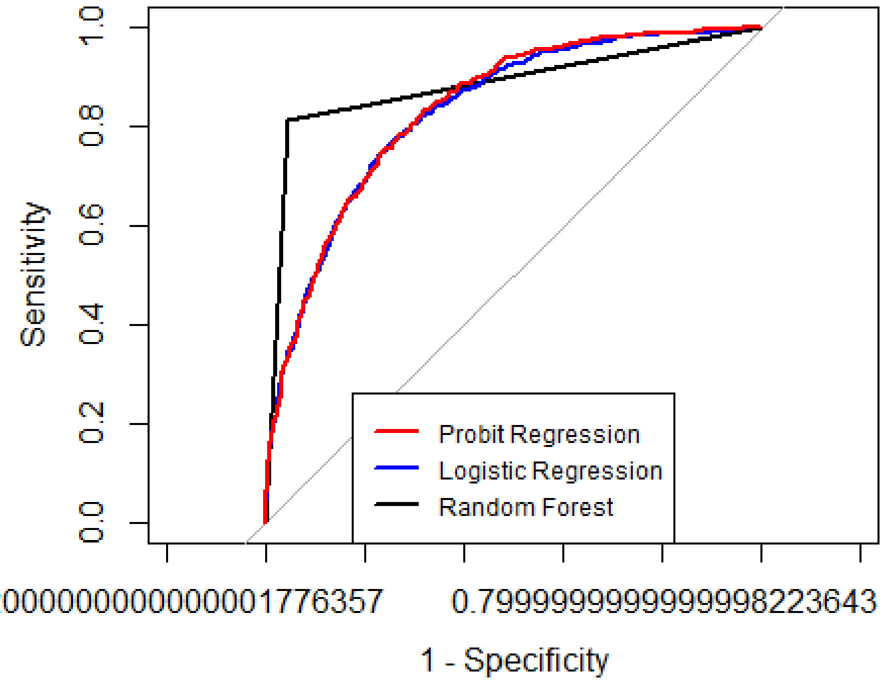


## ROC and accuracy comparison

```
roc1 <- roc(response = test$Churn, predictor = as.numeric(RFPred))

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

roc2 <- roc(response = test$Churn, predictor = as.numeric(glmpred))

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

roc3 <- roc(response = test$Churn, predictor = as.numeric(glmppred))

## Setting levels: control = 0, case = 1
## Setting direction: controls < cases

roc.test( roc3, roc2)

## 
##  DeLong's test for two correlated ROC curves
## 
## data:  roc3 and roc2
## Z = 2.573, p-value = 0.01008
## alternative hypothesis: true difference in AUC is not equal to 0
## 95 percent confidence interval:
##  0.0009094624 0.0067246323
## sample estimates:
```

```
## AUC of roc1 AUC of roc2
##   0.8332313   0.8294143
```

```
plot(roc1, legacy.axes = TRUE)
plot(roc2, col = "blue", add = TRUE)
plot(roc3, col = "red" , add = TRUE)
legend("bottom", c("Probit Regression", "Logistic Regression", "Random
Forest"),
       lty = c(1,1), lwd = c(2, 2), col = c("red", "blue", "black"), cex =
0.75)
```


Sensitivity
Probit Regression
Logistic Regression
Random Forest
1 - Specificity