Telco-Customer-Churn.R

sshah

Fri Feb 22 16:40:20 2019

```
#Telco Customer Churn
#install.packages("Hotelling")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

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

```
## v ggplot2 3.1.0       v purrr   0.2.5
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.2       v stringr 1.3.1
## v readr   1.3.0       v forcats 0.3.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

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

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

```
library(DMwR)
```

```
## Warning: package 'DMwR' was built under R version 3.5.2
```

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

```
## Warning: package 'lattice' was built under R version 3.5.2
```

```
## Loading required package: grid
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.5.2
```

```
##
## Attaching package: 'car'
```

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

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

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.5.2
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.2
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 3.5.2
```

```
##
## Attaching package: 'cowplot'
```

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

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 3.5.2
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 3.5.2
```

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

```
##
## Attaching package: 'pROC'
```

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

```
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 3.5.2
```

```
library(lattice)
library(sm)
```

```
## Warning: package 'sm' was built under R version 3.5.2
```

```
## Package 'sm', version 2.2-5.6: type help(sm) for summary information
```

```
##
## Attaching package: 'sm'
```

```
## The following object is masked from 'package:MASS':
##
##     muscle
```

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.5.2
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:caret':
##
##     cluster
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 3.5.2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following object is masked from 'package:e1071':
##
##     impute
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(asbio)
```

```
## Warning: package 'asbio' was built under R version 3.5.2
```

```
## Loading required package: tcltk
```

```
##
## Attaching package: 'asbio'
```

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

```
## The following object is masked from 'package:DMwR':
##
##     bootstrap
```

```
library(MVA)
```

```
## Warning: package 'MVA' was built under R version 3.5.2
```

```
## Loading required package: HSAUR2
```

```
## Warning: package 'HSAUR2' was built under R version 3.5.2
```

```
## Loading required package: tools
```

```
library(Hotelling)
```

```
## Warning: package 'Hotelling' was built under R version 3.5.2
```

```
## Loading required package: corpcor
```

```
## Warning: package 'corpcor' was built under R version 3.5.2
```

```
#Reading the dataset
telco_churn <- read.csv("C:\\Users\\sshah\\Desktop\\MVA\\project\\Telco-Customer-Churn.csv")
class(telco_churn)
```

```
## [1] "data.frame"
```

```
# Showing the structure of the data frame.
str(telco_churn)
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 10
03 4771 5605 4535 ...
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
##  $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
##  $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...
##  $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...
##  $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...
##  $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...
##  $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...
##  $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
##  $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
##  $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

```
#size of dataset - output is first # of rows aka points, then columns aka variables
dim(telco_churn)
```

```
## [1] 7043   21
```

```
#list variables in dataset
names(telco_churn)
```

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

```
#print first 10 rows
head(telco_churn, n=10)
```

```
##    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
## 7  1452-KIOVK   Male             0      No        Yes     22          Yes
## 8  6713-OKOMC Female             0      No         No     10           No
## 9  7892-POOKP Female             0     Yes         No     28          Yes
## 10 6388-TABGU   Male             0      No        Yes     62          Yes
##       MultipleLines InternetService OnlineSecurity OnlineBackup
## 1  No phone service             DSL             No          Yes
## 2                No             DSL            Yes           No
## 3                No             DSL            Yes          Yes
## 4  No phone service             DSL            Yes           No
## 5                No     Fiber optic             No           No
## 6               Yes     Fiber optic             No           No
## 7               Yes     Fiber optic             No          Yes
## 8  No phone service             DSL            Yes           No
## 9               Yes     Fiber optic             No           No
## 10               No             DSL            Yes          Yes
##    DeviceProtection TechSupport StreamingTV StreamingMovies       Contract
## 1                No          No          No              No Month-to-month
## 2               Yes          No          No              No       One year
## 3                No          No          No              No Month-to-month
## 4               Yes         Yes          No              No       One year
## 5                No          No          No              No Month-to-month
## 6               Yes          No         Yes             Yes Month-to-month
## 7                No          No         Yes              No Month-to-month
## 8                No          No          No              No Month-to-month
## 9               Yes         Yes         Yes             Yes Month-to-month
## 10               No          No          No              No       One year
##    PaperlessBilling             PaymentMethod MonthlyCharges TotalCharges
## 1               Yes          Electronic check          29.85        29.85
## 2                No              Mailed check          56.95      1889.50
## 3               Yes              Mailed check          53.85       108.15
## 4                No Bank transfer (automatic)          42.30      1840.75
## 5               Yes          Electronic check          70.70       151.65
## 6               Yes          Electronic check          99.65       820.50
## 7               Yes   Credit card (automatic)          89.10      1949.40
## 8                No              Mailed check          29.75       301.90
## 9               Yes          Electronic check         104.80      3046.05
## 10               No Bank transfer (automatic)          56.15      3487.95
##    Churn
## 1     No
## 2     No
## 3    Yes
## 4     No
## 5    Yes
## 6    Yes
## 7     No
## 8     No
## 9    Yes
## 10    No
```

```
#summary of data
summary(telco_churn)
```

```
##       customerID      gender     SeniorCitizen    Partner    Dependents
##  0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933
##  0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110
##  0004-TLHLJ:   1                 Median :0.0000
##  0011-IGKFF:   1                 Mean   :0.1621
##  0013-EXCHZ:   1                 3rd Qu.:0.0000
##  0013-MHZWF:   1                 Max.   :1.0000
##  (Other)   :7037
##      tenure      PhoneService          MultipleLines     InternetService
##  Min.   : 0.00   No : 682     No              :3390   DSL        :2421
##  1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096
##  Median :29.00                Yes             :2971   No         :1526
##  Mean   :32.37
##  3rd Qu.:55.00
##  Max.   :72.00
##
##              OnlineSecurity              OnlineBackup
##  No                 :3498   No                 :3088
##  No internet service:1526   No internet service:1526
##  Yes                :2019   Yes                :2429
##
##
##
##
##             DeviceProtection              TechSupport
##  No                 :3095    No                 :3473
##  No internet service:1526    No internet service:1526
##  Yes                :2422    Yes                :2044
##
##
##
##
##               StreamingTV              StreamingMovies
##  No                 :2810   No                 :2785
##  No internet service:1526   No internet service:1526
##  Yes                :2707   Yes                :2732
##
##
##
##
##           Contract    PaperlessBilling                   PaymentMethod
##  Month-to-month:3875   No :2872         Bank transfer (automatic):1544
##  One year      :1473   Yes:4171         Credit card (automatic)  :1522
##  Two year      :1695                    Electronic check         :2365
##                                         Mailed check             :1612
##
##
##
##  MonthlyCharges    TotalCharges    Churn
##  Min.   : 18.25   Min.   :  18.8   No :5174
##  1st Qu.: 35.50   1st Qu.: 401.4   Yes:1869
##  Median : 70.35   Median :1397.5
##  Mean   : 64.76   Mean   :2283.3
##  3rd Qu.: 89.85   3rd Qu.:3794.7
##  Max.   :118.75   Max.   :8684.8
##                   NA's   :11
```

```
#Data Cleaning
#Converting SeniorCitizen variable into a factor variable
telco_churn$SeniorCitizen <- as.factor(ifelse(telco_churn$SeniorCitizen==0,'Yes','No'))
head(telco_churn, n=10)
```

```
##    customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1  7590-VHVEG Female             Yes     Yes         No      1           No
## 2  5575-GNVDE   Male             Yes      No         No     34          Yes
## 3  3668-QPYBK   Male             Yes      No         No      2          Yes
## 4  7795-CFOCW   Male             Yes      No         No     45           No
## 5  9237-HQITU Female             Yes      No         No      2          Yes
## 6  9305-CDSKC Female             Yes      No         No      8          Yes
## 7  1452-KIOVK   Male             Yes      No        Yes     22          Yes
## 8  6713-OKOMC Female             Yes      No         No     10           No
## 9  7892-POOKP Female             Yes     Yes         No     28          Yes
## 10 6388-TABGU   Male             Yes      No        Yes     62          Yes
##       MultipleLines InternetService OnlineSecurity OnlineBackup
## 1  No phone service             DSL             No          Yes
## 2                No             DSL            Yes           No
## 3                No             DSL            Yes          Yes
## 4  No phone service             DSL            Yes           No
## 5                No     Fiber optic             No           No
## 6               Yes     Fiber optic             No           No
## 7               Yes     Fiber optic             No          Yes
## 8  No phone service             DSL            Yes           No
## 9               Yes     Fiber optic             No           No
## 10               No             DSL            Yes          Yes
##    DeviceProtection TechSupport StreamingTV StreamingMovies       Contract
## 1                No          No          No              No Month-to-month
## 2               Yes          No          No              No       One year
## 3                No          No          No              No Month-to-month
## 4               Yes         Yes          No              No       One year
## 5                No          No          No              No Month-to-month
## 6               Yes          No         Yes             Yes Month-to-month
## 7                No          No         Yes              No Month-to-month
## 8                No          No          No              No Month-to-month
## 9               Yes         Yes         Yes             Yes Month-to-month
## 10               No          No          No              No       One year
##    PaperlessBilling              PaymentMethod MonthlyCharges TotalCharges
## 1               Yes          Electronic check          29.85        29.85
## 2                No              Mailed check          56.95      1889.50
## 3               Yes              Mailed check          53.85       108.15
## 4                No Bank transfer (automatic)          42.30      1840.75
## 5               Yes          Electronic check          70.70       151.65
## 6               Yes          Electronic check          99.65       820.50
## 7               Yes   Credit card (automatic)          89.10      1949.40
## 8                No              Mailed check          29.75       301.90
## 9               Yes          Electronic check         104.80      3046.05
## 10               No Bank transfer (automatic)          56.15      3487.95
##    Churn
## 1     No
## 2     No
## 3    Yes
## 4     No
## 5    Yes
## 6    Yes
## 7     No
## 8     No
## 9    Yes
## 10    No
```

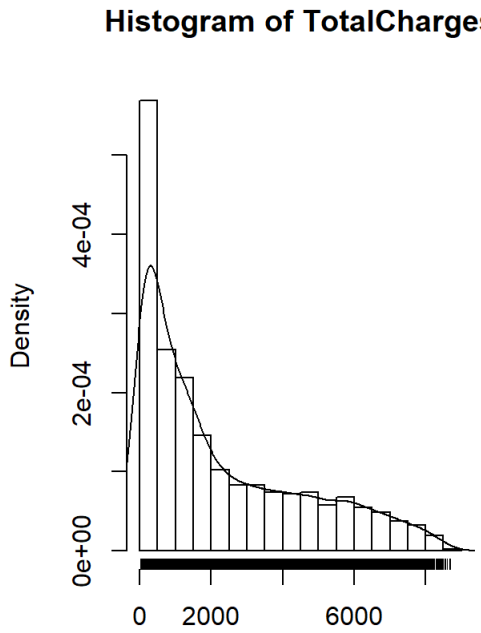
```
#Converting tenure values into ranges of 12 months

telco_churn <- mutate(telco_churn,tenure_range = tenure)
telco_churn_tenure <- cut(telco_churn$tenure_range,6,labels = c('0-1 Years','1-2 Years','2-3 Years','4-5 Yea
rs','5-6 Years','6-7 Years'))
head(telco_churn_tenure)
```

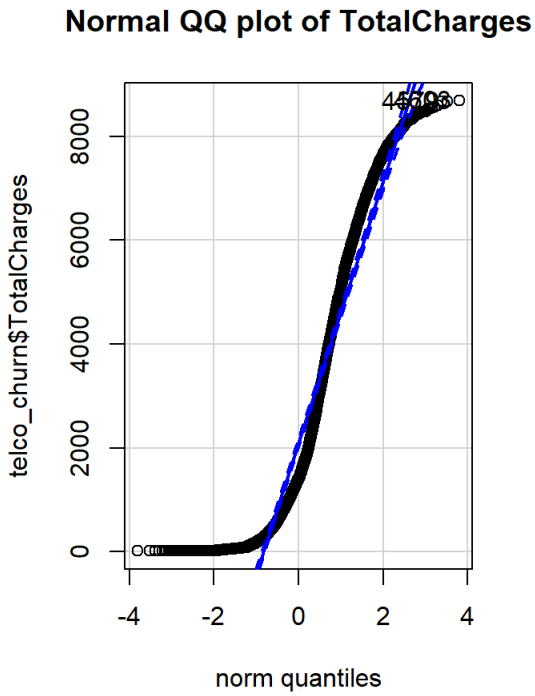
```
## [1] 0-1 Years 2-3 Years 0-1 Years 4-5 Years 0-1 Years 0-1 Years
## Levels: 0-1 Years 1-2 Years 2-3 Years 4-5 Years 5-6 Years 6-7 Years
```

```
#Replacing 'No Internet Service' value in Streaming Movies, Online Security, Device Prtection,
#Tech Support and Streaming TV with No'
telco_churn$StreamingTV[telco_churn$StreamingTV=='No internet service'] <- 'No'
telco_churn$StreamingMovies[telco_churn$StreamingMovies=='No internet service'] <- 'No'
telco_churn$OnlineSecurity[telco_churn$OnlineSecurity=='No internet service'] <- 'No'
telco_churn$OnlineBackup[telco_churn$OnlineBackup=='No internet service'] <- 'No'
telco_churn$DeviceProtection[telco_churn$DeviceProtection=='No internet service'] <- 'No'
telco_churn$TechSupport[telco_churn$TechSupport=='No internet service'] <- 'No'

#Deleting the unused levels from the factor variables
telco_churn$StreamingMovies <- factor(telco_churn$StreamingMovies)
telco_churn$StreamingTV <- factor(telco_churn$StreamingTV)
telco_churn$OnlineSecurity <- factor(telco_churn$OnlineSecurity)
telco_churn$OnlineBackup <- factor(telco_churn$OnlineBackup)
telco_churn$DeviceProtection <- factor(telco_churn$DeviceProtection)
telco_churn$TechSupport <- factor(telco_churn$TechSupport)

#Calculating the number of null values in each of the columns
nullvalues <- colSums(is.na(telco_churn))
nullvalues <- (nullvalues/nrow(telco_churn))*100
nullvalues
```

```
##       customerID           gender    SeniorCitizen          Partner
##        0.0000000        0.0000000        0.0000000        0.0000000
##       Dependents           tenure     PhoneService    MultipleLines
##        0.0000000        0.0000000        0.0000000        0.0000000
##  InternetService   OnlineSecurity     OnlineBackup DeviceProtection
##        0.0000000        0.0000000        0.0000000        0.0000000
##      TechSupport      StreamingTV  StreamingMovies         Contract
##        0.0000000        0.0000000        0.0000000        0.0000000
## PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
##        0.0000000        0.0000000        0.0000000        0.1561834
##            Churn     tenure_range
##        0.0000000        0.0000000
```

```
#Removing the rows containing null values as there are just 11 rows out of 7043 rows
#in total which is 0.15% and hence we can afford dropping those
telco_churn <- telco_churn[complete.cases(telco_churn), ]

#Exploratory Data Analysis

#Checking for distributions in numerical columns
#The qqplotd show a few extreme outliers which break the assumption of 95% confidence
#normal distribution
par(mfrow = c(1,2))
hist(telco_churn$TotalCharges,xlab='',main = 'Histogram of TotalCharges',freq = FALSE)
lines(density(telco_churn$TotalCharges,na.rm = T))
rug(jitter(telco_churn$TotalCharges))
qqPlot(telco_churn$TotalCharges,main='Normal QQ plot of TotalCharges')
```

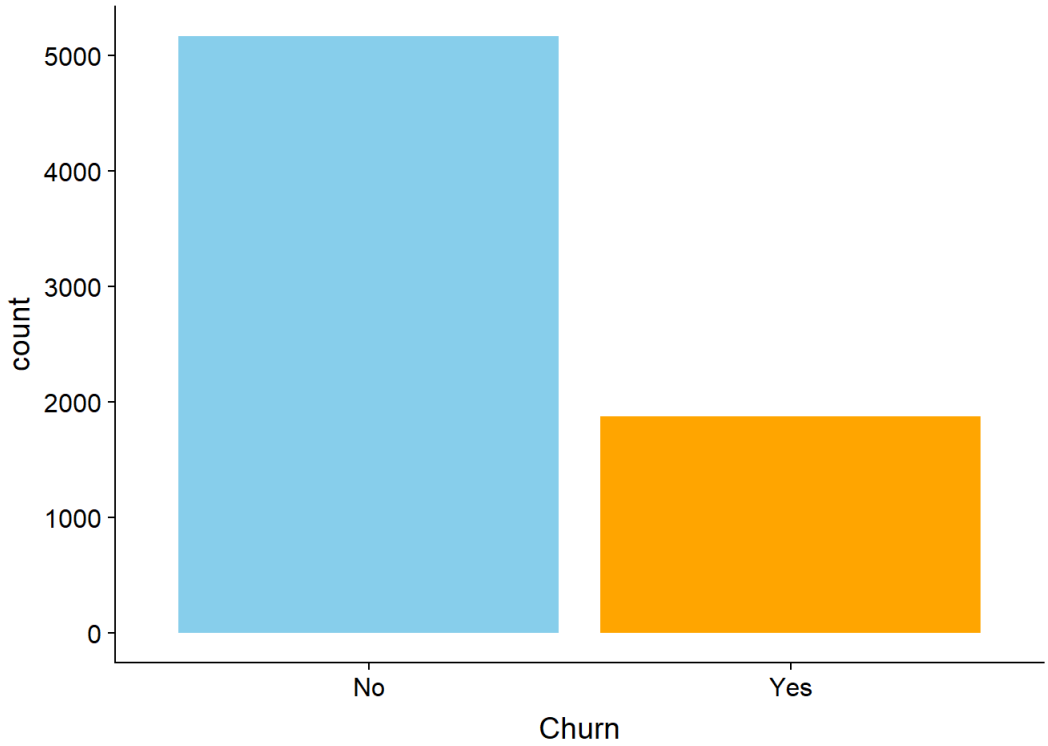




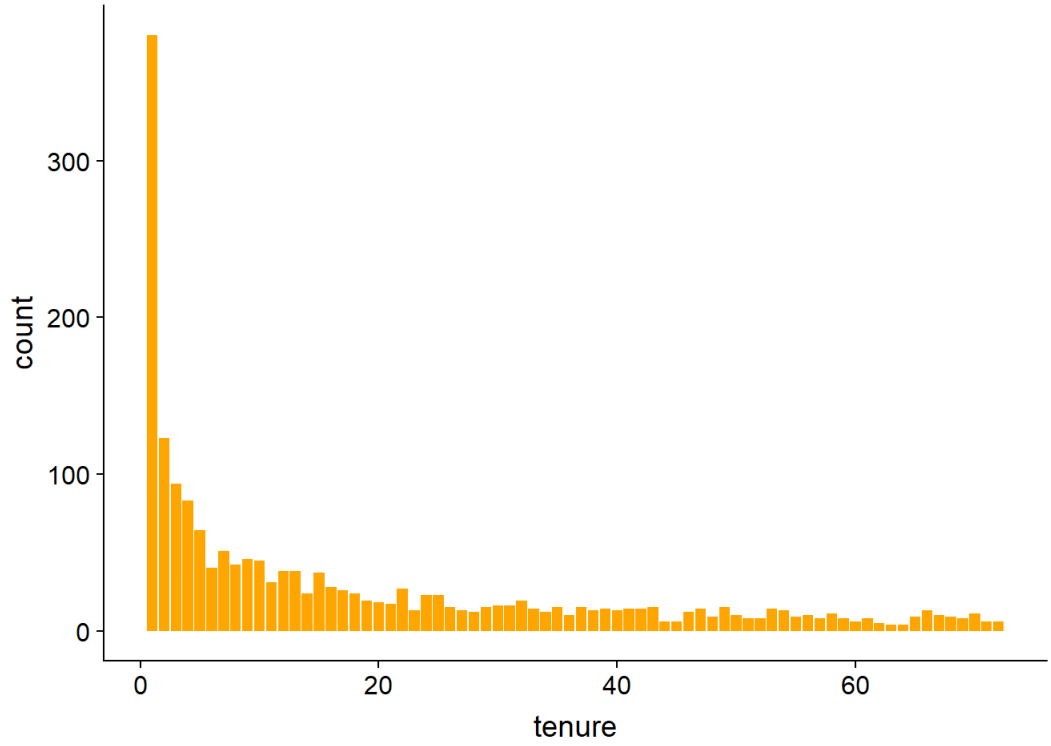

```
## [1] 4603 4579
```

```
par(mfrow=c(1,1))

ggplot(telco_churn, aes(x = Churn))+geom_histogram(stat = "count", fill = c("sky blue", "orange"))
```

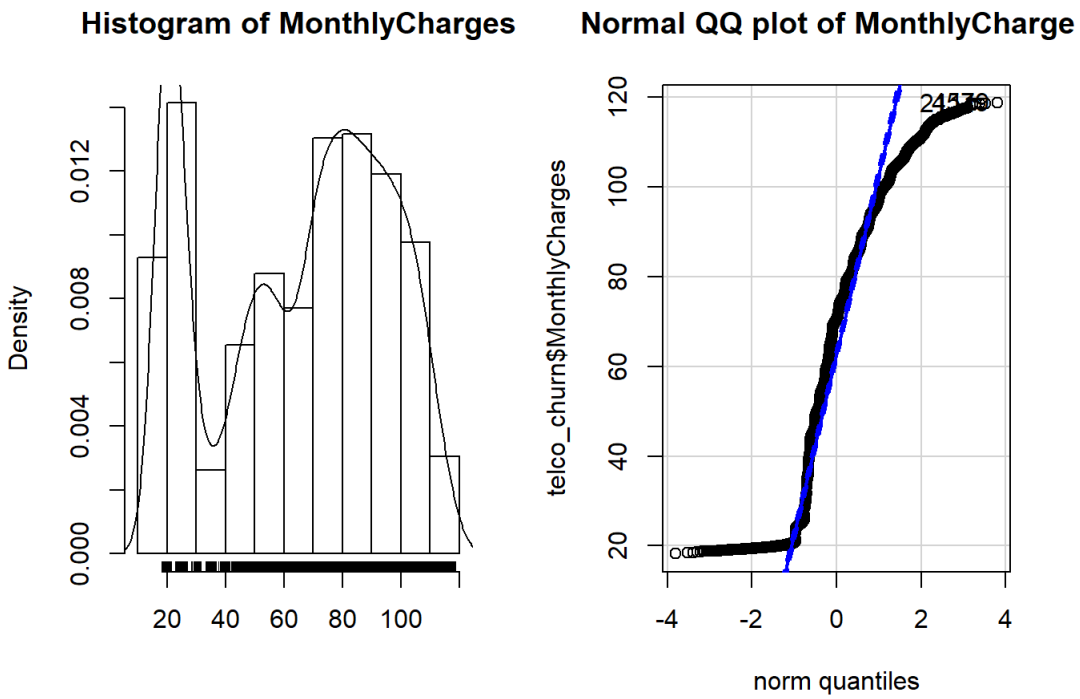
```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




```
telco_churn %>% filter(telco_churn$Churn == "Yes") %>%  ggplot( aes(x=  tenure))+geom_bar(fill = "orange" )
```

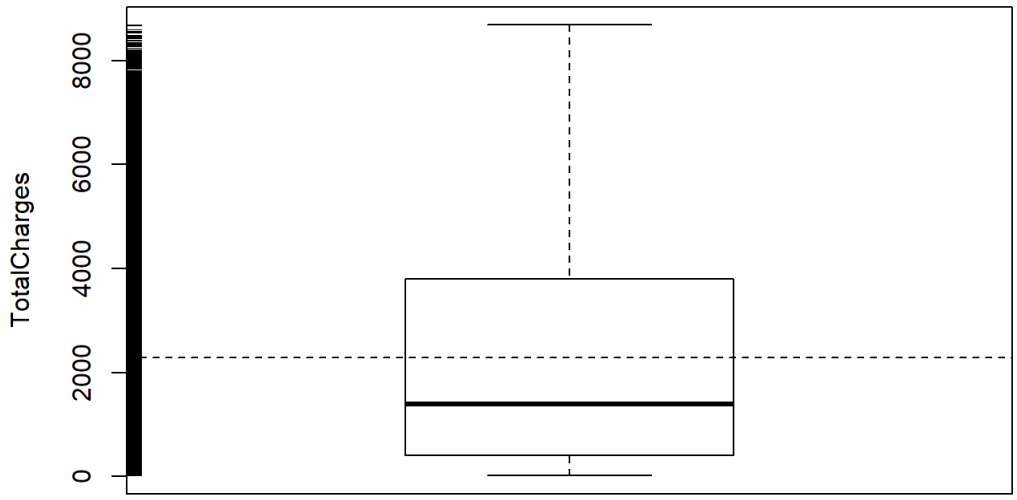



```
par(mfrow = c(1,2))
hist(telco_churn$MonthlyCharges,xlab='',main = 'Histogram of MonthlyCharges',freq = FALSE)
lines(density(telco_churn$MonthlyCharges,na.rm = T))
rug(jitter(telco_churn$MonthlyCharges))
qqPlot(telco_churn$MonthlyCharges,main='Normal QQ plot of MonthlyCharges')
```

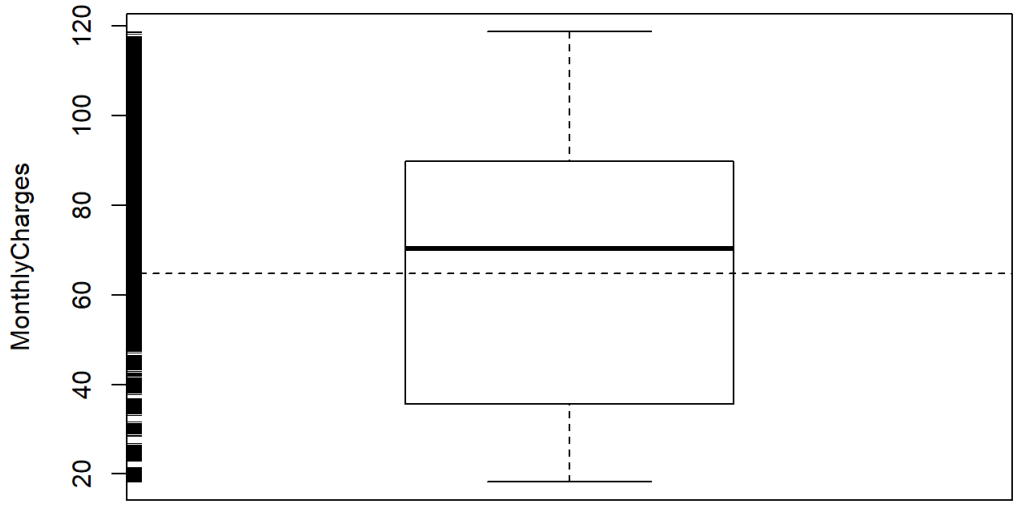



```
## [1] 4579 2111
```

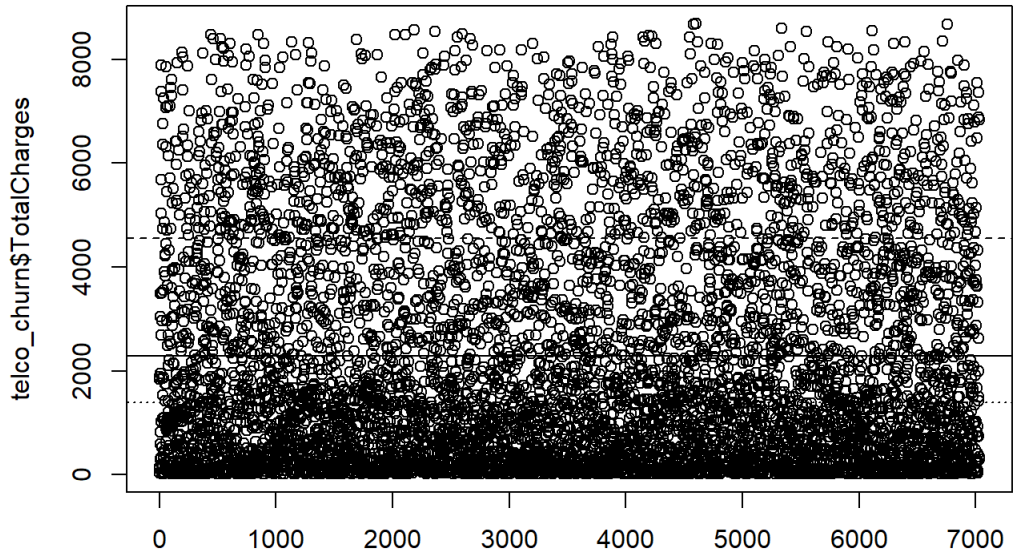
```
par(mfrow=c(1,1))

#Boxplot distributions for our numeric columns
#The dashed line shows the mean and the dark center line shows the median
#Difference between these two lines depict the deviation from the central limit theorem
boxplot(telco_churn$TotalCharges, ylab = "TotalCharges")
rug(jitter(telco_churn$TotalCharges), side = 2)
abline(h = mean(telco_churn$TotalCharges, na.rm = T), lty = 2)
```




```
boxplot(telco_churn$MonthlyCharges, ylab = "MonthlyCharges",outline = TRUE)
rug(jitter(telco_churn$MonthlyCharges), side = 2)
abline(h = mean(telco_churn$MonthlyCharges, na.rm = T), lty = 2)
```



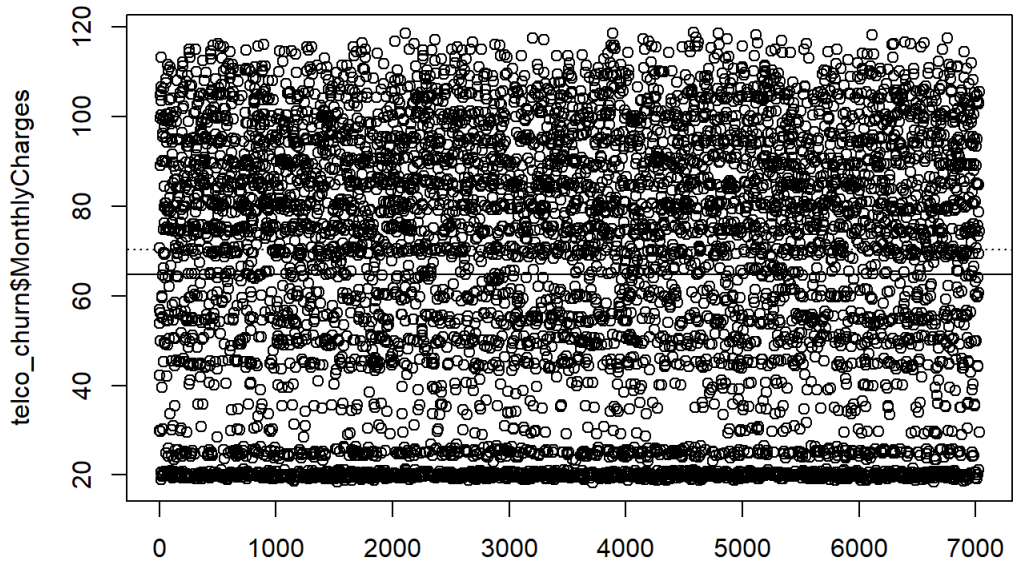
```
#Plotting the TotalCharges and Monthl Charges with 3 lines for mean, median and mean+std
plot(telco_churn$TotalCharges, xlab = "")
abline(h = mean(telco_churn$TotalCharges, na.rm = T), lty = 1)
abline(h = mean(telco_churn$TotalCharges, na.rm = T) + sd(telco_churn$TotalCharges, na.rm = T),lty = 2)
abline(h = median(telco_churn$TotalCharges, na.rm = T), lty = 3)
```



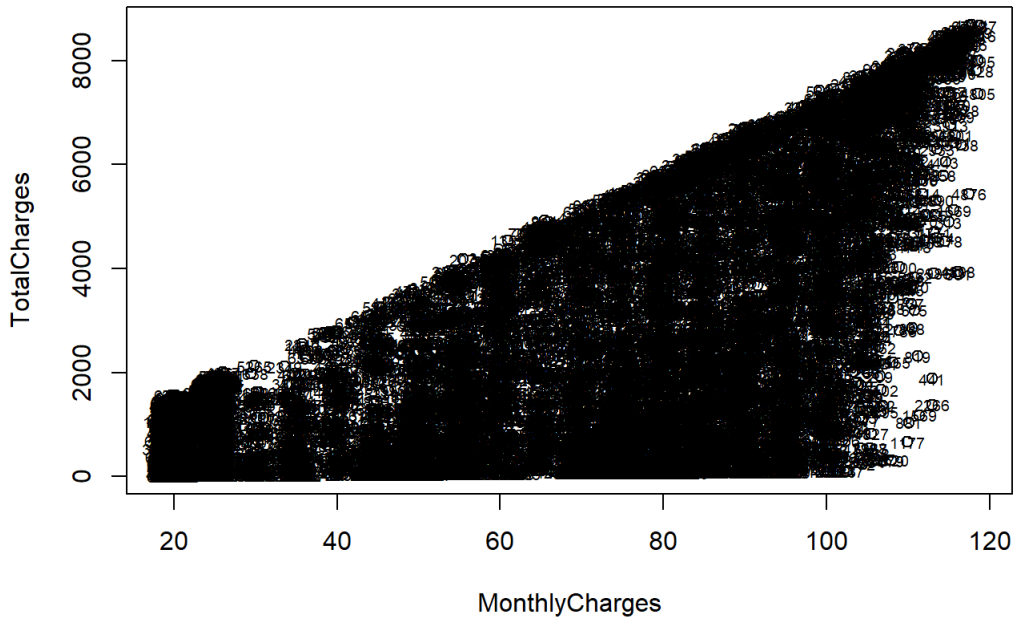

```
#Plotting the TotalCharges and Monthl Charges with 3 lines for mean, median and mean+std
plot(telco_churn$MonthlyCharges, xlab = "")
abline(h = mean(telco_churn$MonthlyCharges, na.rm = T), lty = 1)
abline(h = mean(telco_churn$MonthCharges, na.rm = T) + sd(telco_churn$MonthlyCharges, na.rm = T),lty = 2)
```

```
## Warning in mean.default(telco_churn$MonthCharges, na.rm = T): argument is
## not numeric or logical: returning NA
```

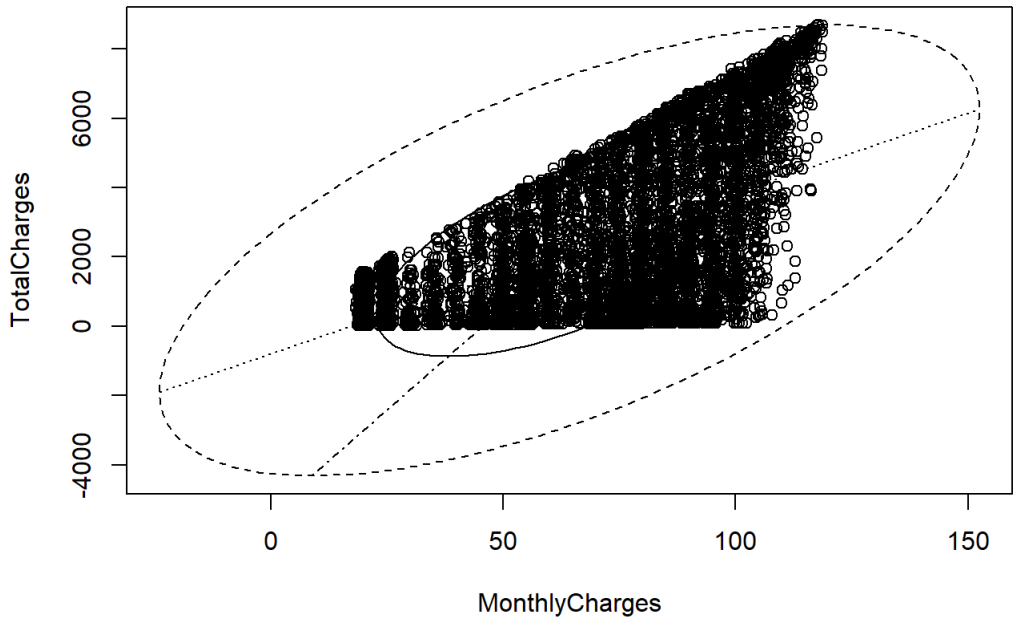
```
abline(h = median(telco_churn$MonthlyCharges, na.rm = T), lty = 3)
```



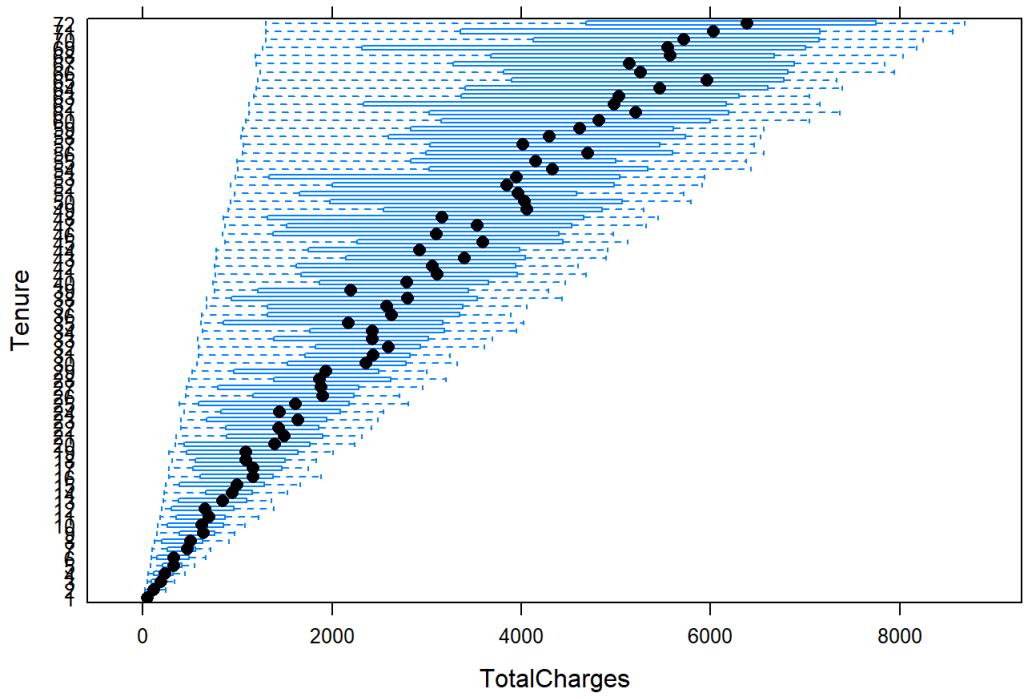
```
#Scatterplot to understand the relationship between monthly and yearly charges
#This shows us that both of them are highly corelated which is kind of obvious
#For other categorical varibles, we compare only on one numeric column which is TotalCharges
#Checking for outliers using bvplot
plot(telco_churn$TotalCharges~telco_churn$MonthlyCharges,data=telco_churn,xlab="MonthlyCharges",ylab="TotalCharges")
with(telco_churn,text(telco_churn$MonthlyCharges,telco_churn$TotalCharges,cex=0.6,labels=abbreviate(row.names(telco_churn))))
```



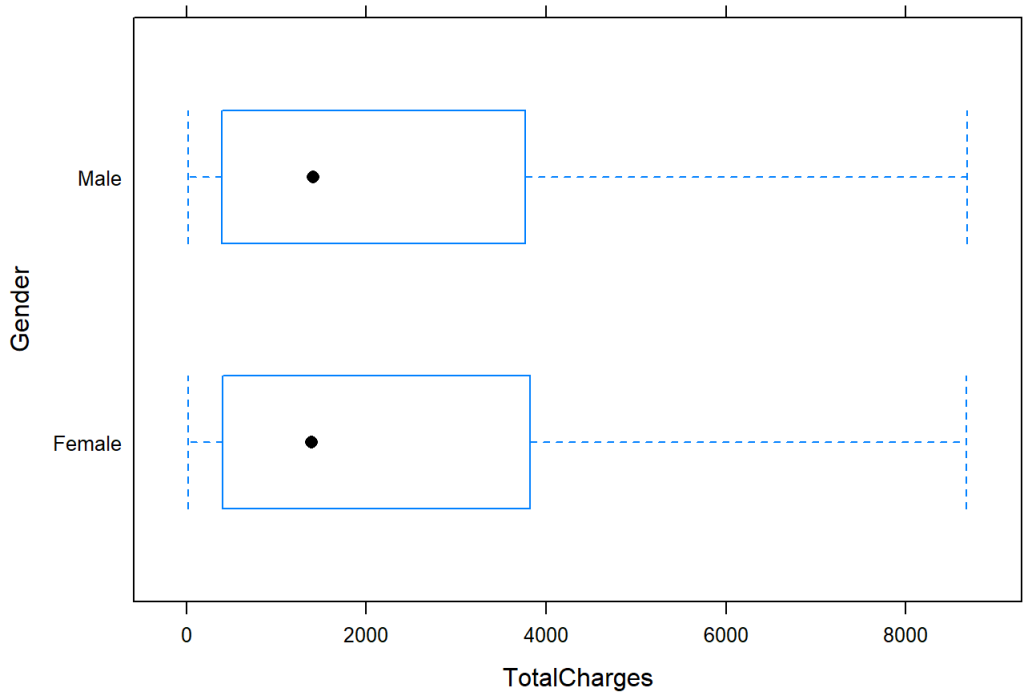

```
x<-telco_churn[,c(19,20)]
bvbox(x,ylab = "TotalCharges",xlab="MonthlyCharges")
```



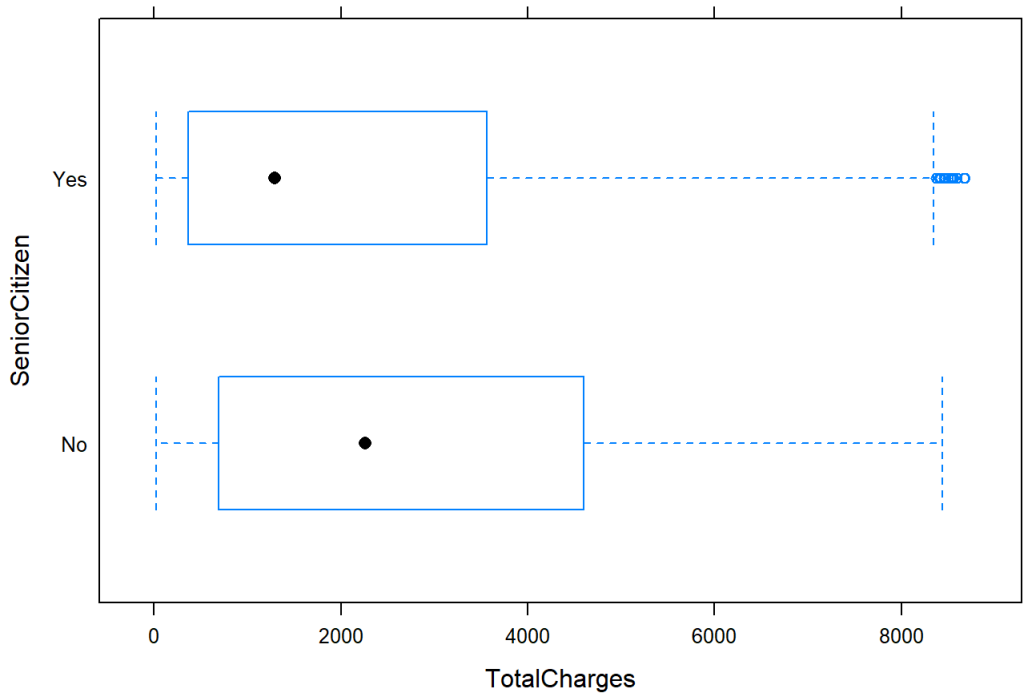

```
#Plotting joint boxplots for various categories wrt numerical column TotalCharges
bwplot(telco_churn$tenure_range ~ telco_churn$TotalCharges, data=telco_churn, ylab='Tenure',xlab='TotalCharges')
```



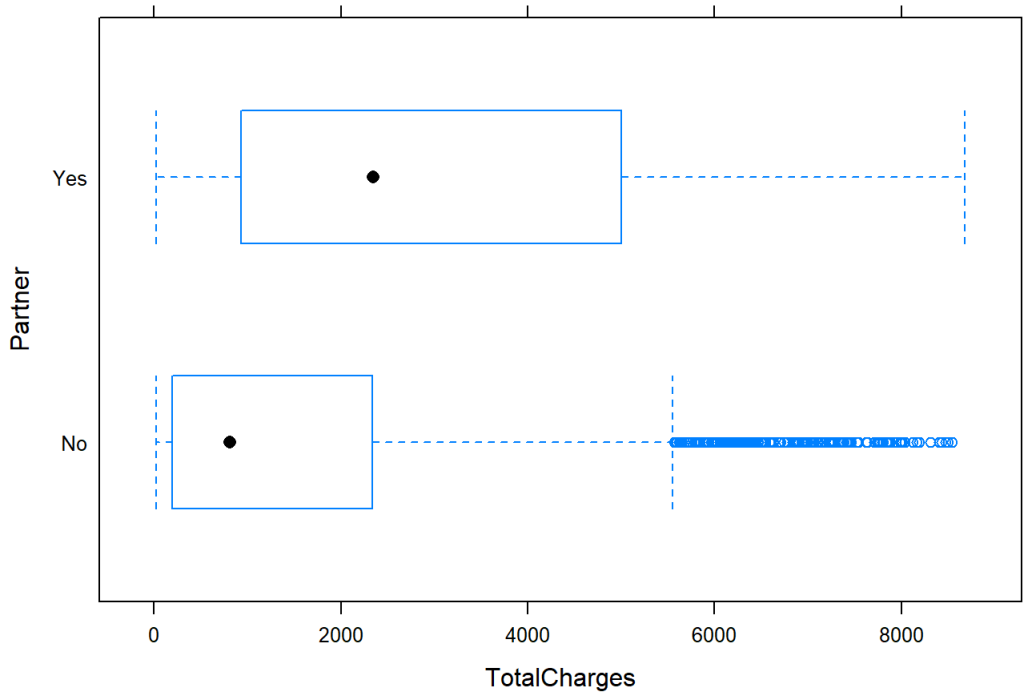

```
bwplot(telco_churn$gender ~ telco_churn$TotalCharges, data=telco_churn, ylab='Gender',xlab='TotalCharges')
```



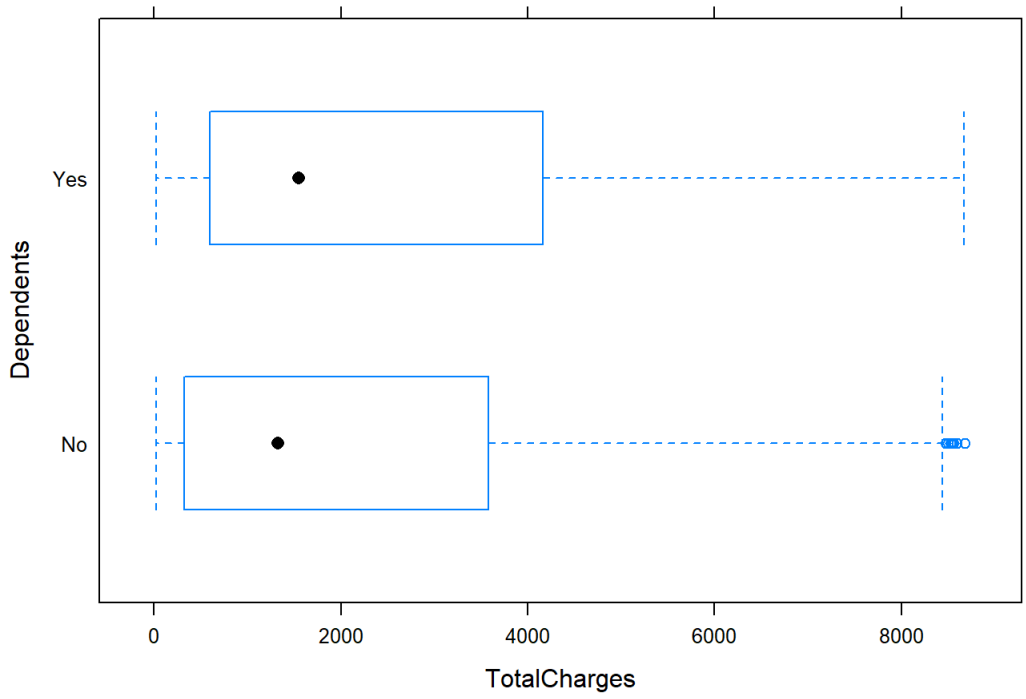

```
bwplot(telco_churn$SeniorCitizen ~ telco_churn$TotalCharges, data=telco_churn, ylab='SeniorCitizen',xlab='TotalCharges')
```



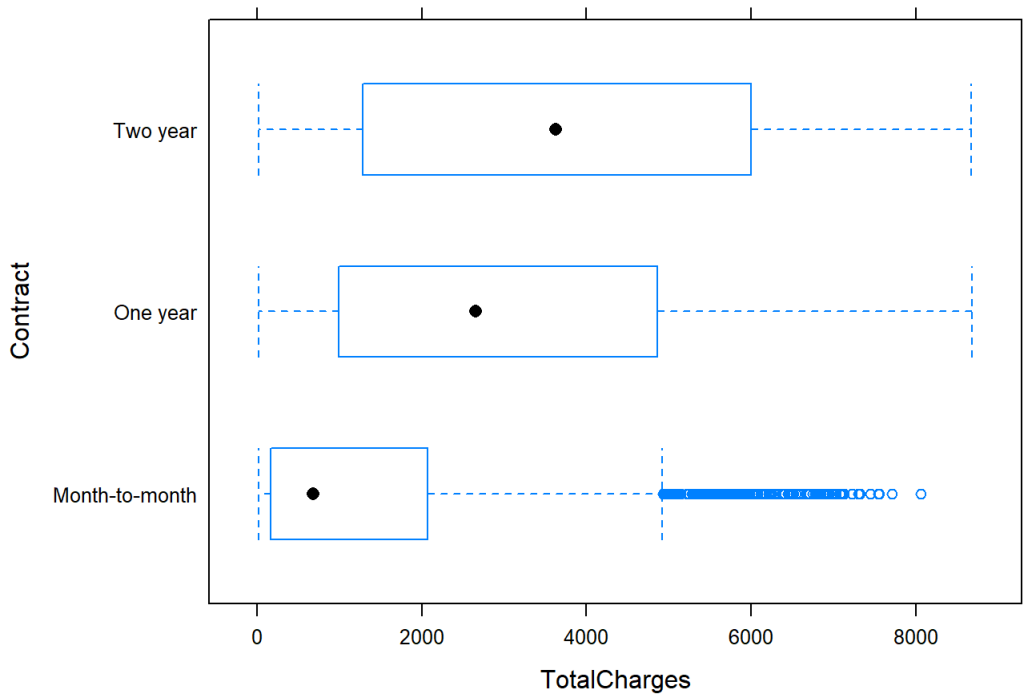

```
bwplot(telco_churn$Partner ~ telco_churn$TotalCharges, data=telco_churn, ylab='Partner',xlab='TotalCharges')
```



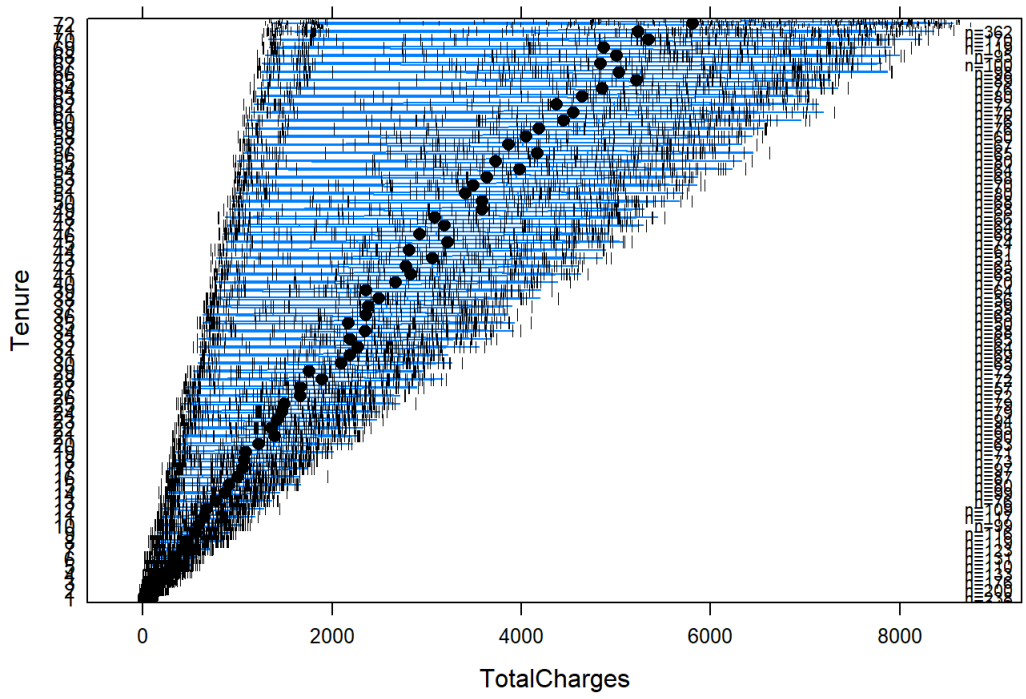

```
bwplot(telco_churn$Dependents ~ telco_churn$TotalCharges, data=telco_churn, ylab='Dependents',xlab='TotalCharges')
```



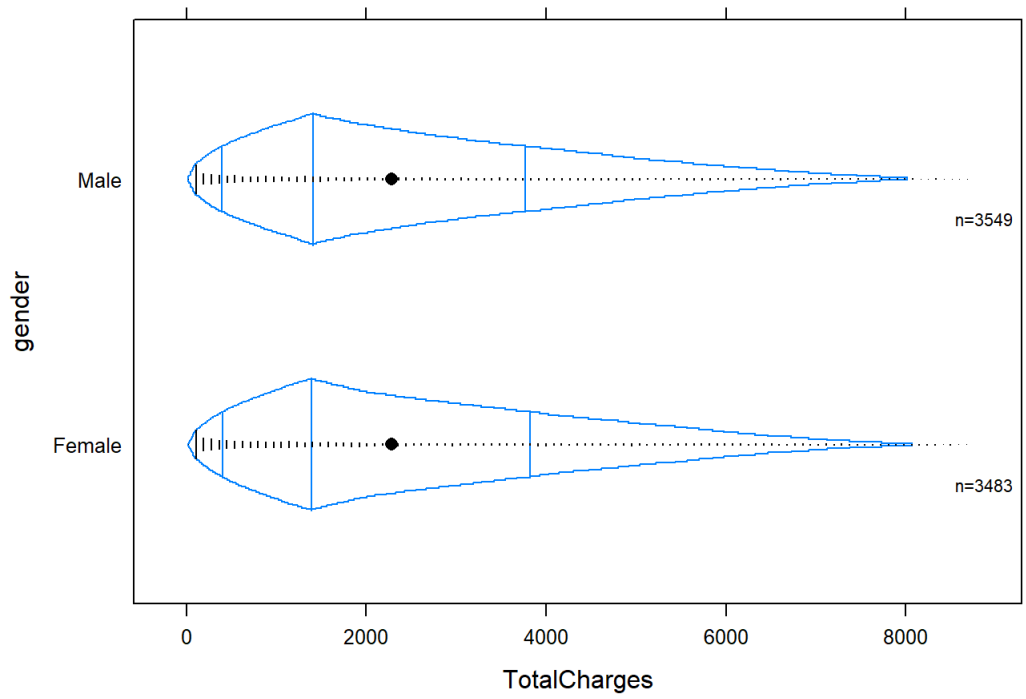

```
bwplot(telco_churn$Contract ~ telco_churn$TotalCharges, data=telco_churn, ylab='Contract',xlab='TotalCharges
')
```



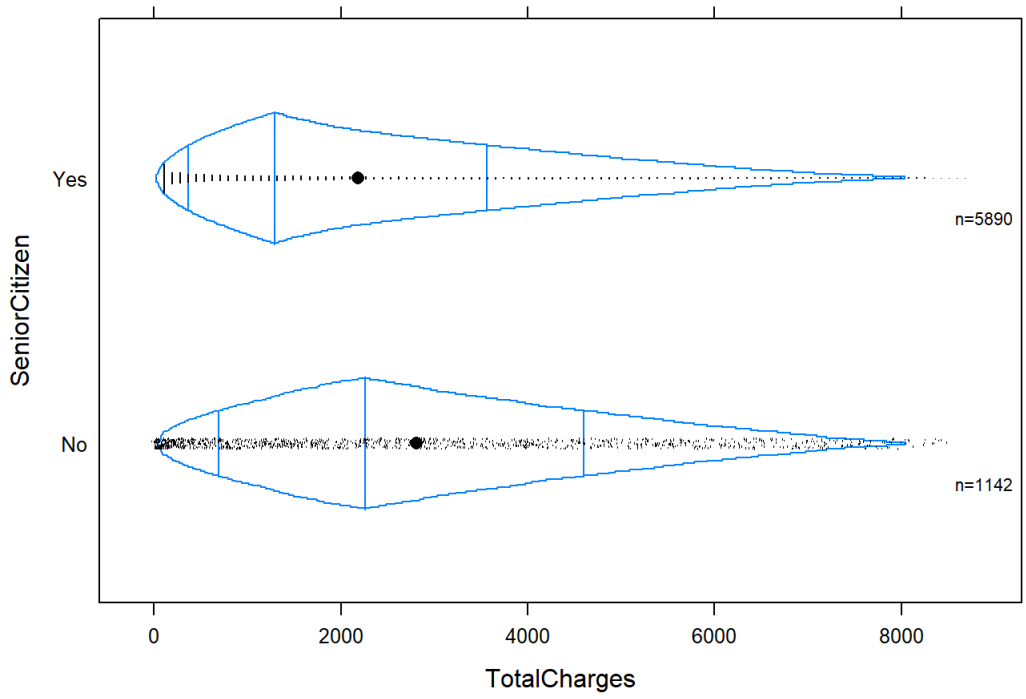

```
#Plotting stripplots for various categories wrt numerical column TotalCharges
bwplot(telco_churn$tenure_range ~ telco_churn$TotalCharges, data=telco_churn,panel=panel.bpplot,
       probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='Tenure',xlab='TotalCharges')
```



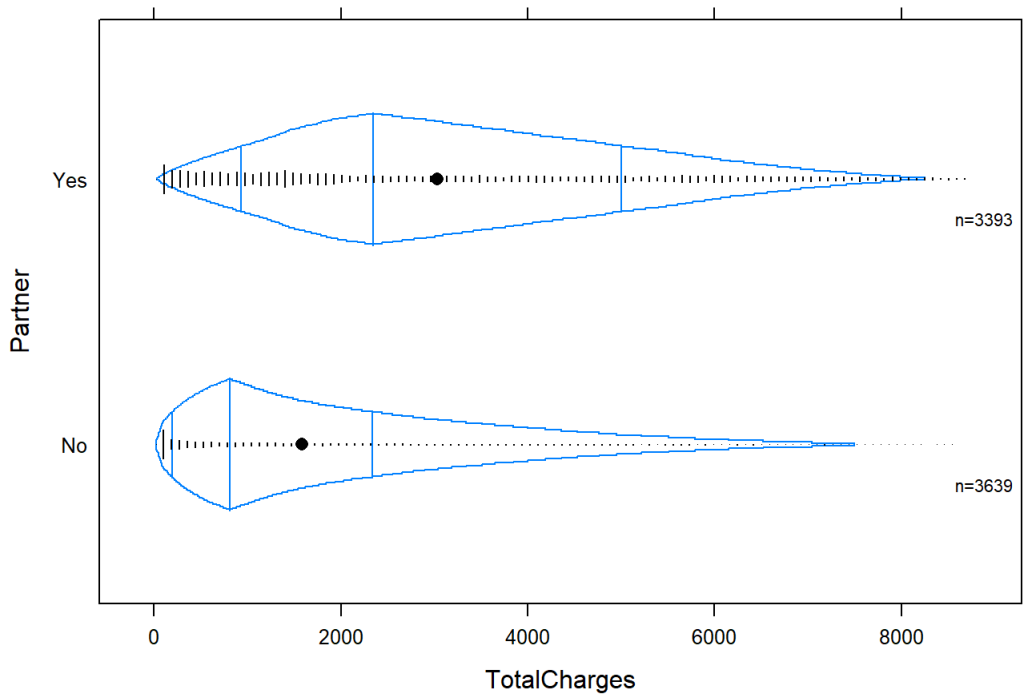

```
bwplot(telco_churn$gender ~ telco_churn$TotalCharges, data=telco_churn,panel=panel.bpplot,
       probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='gender',xlab='TotalCharges')
```



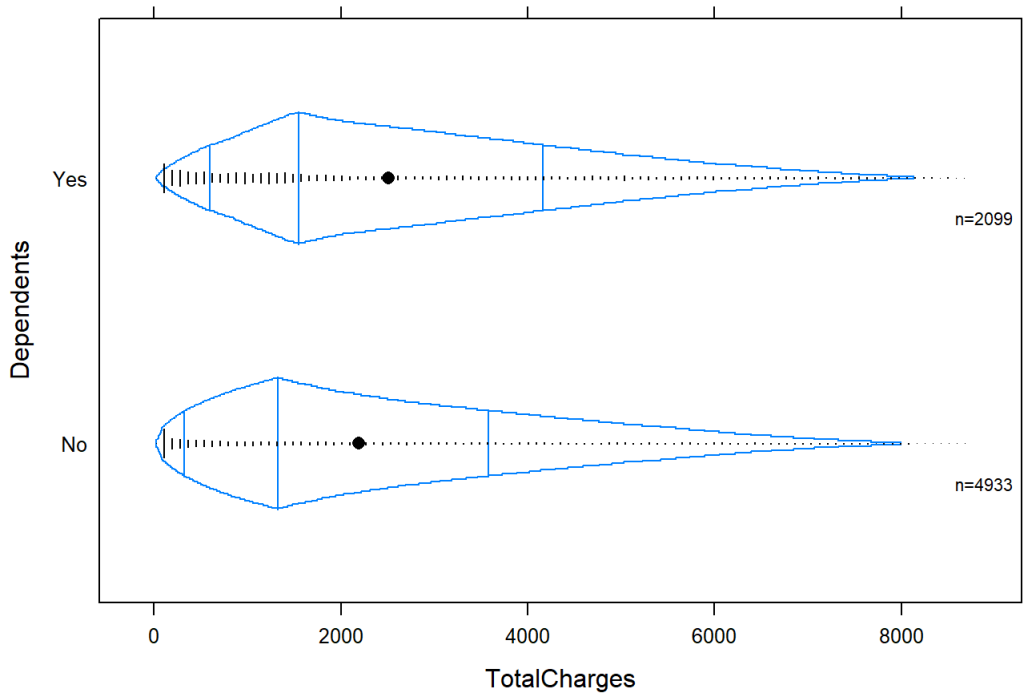

```
bwplot(telco_churn$SeniorCitizen ~ telco_churn$TotalCharges, data=telco_churn,panel=panel.bpplot,
       probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='SeniorCitizen',xlab='TotalCharges')
```



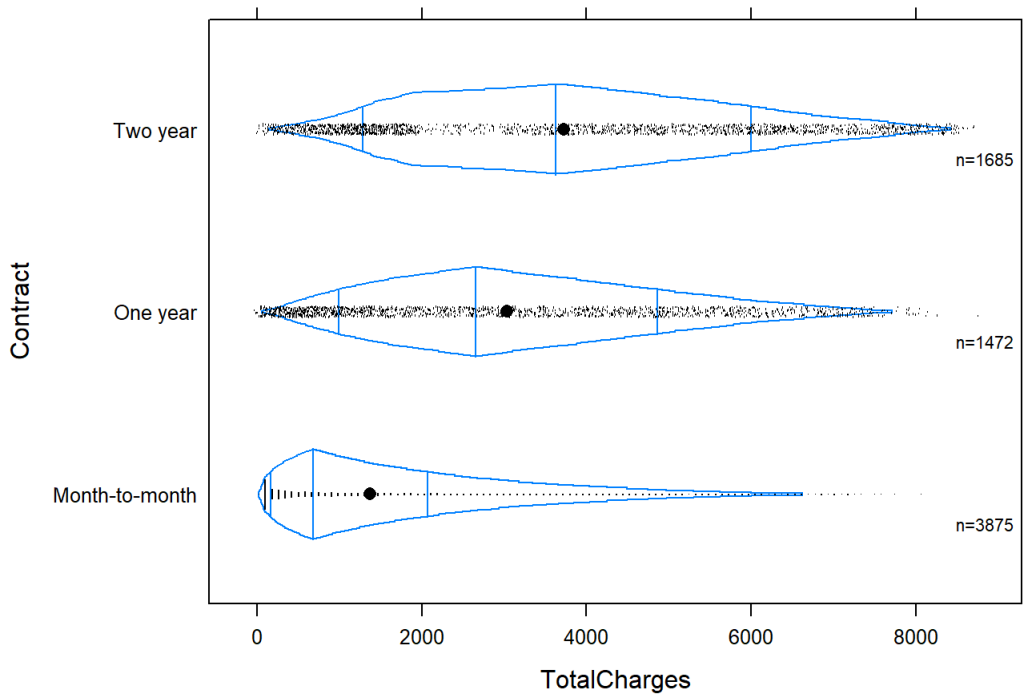

```
bwplot(telco_churn$Partner ~ telco_churn$TotalCharges, data=telco_churn,panel=panel.bpplot,
       probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='Partner',xlab='TotalCharges')
```




```
bwplot(telco_churn$Dependents ~ telco_churn$TotalCharges, data=telco_churn,panel=panel.bpplot,
       probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='Dependents',xlab='TotalCharges')
```




```
bwplot(telco_churn$Contract ~ telco_churn$TotalCharges, data=telco_churn,panel=panel.bpplot,
       probs=seq(.01,.49,by=.01), datadensity=TRUE, ylab='Contract',xlab='TotalCharges')
```



```r
##Creating Dummy Variables

#Converting double/int columns to numeric
numeric_col <- c("tenure","MonthlyCharges", "TotalCharges")
telco_churn[numeric_col] <- sapply(telco_churn[numeric_col], as.numeric)

#Segregating the numeric columns from categorical columns and storing them as a seperate dataframe
telco_churn_int <- telco_churn[,c("tenure","MonthlyCharges", "TotalCharges")]
telco_churn_int <- data.frame(scale(telco_churn_int))

#Creating dummy variables for the categorical data
telco_churn_cat <- telco_churn[,-c(1,6,19,20)]
dummy<- data.frame(sapply(telco_churn_cat,function(x) data.frame(model.matrix(~x-1,data =telco_churn_cat))[,
-1]))
head(dummy)
```

```
##   gender SeniorCitizen Partner Dependents PhoneService
## 1      0             1       1          0            0
## 2      1             1       0          0            1
## 3      1             1       0          0            1
## 4      1             1       0          0            0
## 5      0             1       0          0            1
## 6      0             1       0          0            1
##   MultipleLines.xNo.phone.service MultipleLines.xYes
## 1                               1                  0
## 2                               0                  0
## 3                               0                  0
## 4                               1                  0
## 5                               0                  0
## 6                               0                  1
##   InternetService.xFiber.optic InternetService.xNo OnlineSecurity
## 1                            0                   0              0
## 2                            0                   0              1
## 3                            0                   0              1
## 4                            0                   0              1
## 5                            1                   0              0
## 6                            1                   0              0
##   OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
## 1            1                0           0           0               0
## 2            0                1           0           0               0
## 3            1                0           0           0               0
## 4            0                1           1           0               0
## 5            0                0           0           0               0
## 6            0                1           0           1               1
##   Contract.xOne.year Contract.xTwo.year PaperlessBilling
## 1                  0                  0                1
## 2                  1                  0                0
## 3                  0                  0                1
## 4                  1                  0                0
## 5                  0                  0                1
## 6                  0                  0                1
##   PaymentMethod.xCredit.card..automatic. PaymentMethod.xElectronic.check
## 1                                      0                               1
## 2                                      0                               0
## 3                                      0                               0
## 4                                      0                               0
## 5                                      0                               1
## 6                                      0                               1
##   PaymentMethod.xMailed.check Churn
## 1                           0     0
## 2                           1     0
## 3                           1     1
## 4                           0     0
## 5                           0     1
## 6                           0     1
```

```r
#Combining the dummy and the numeric columns to form the final dataset
telco_churn_final <- cbind(telco_churn_int,dummy)
head(telco_churn_final)
```

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

```
##Matrix Plots, Covariance and Corelations Plots

#Next 4 lines were used to solve the error "Figure margins too large"
par("mar")
```

```
## [1] 5.1 4.1 4.1 2.1
```

```
par(mar=c(1,1,1,1))
graphics.off()

#ScatterPlot matrix
pairs(telco_churn_final[,1:3],pch=".",cex=1.5)

#CorrelationMatrix
cormatrix <- round(cor(telco_churn_final),4)
#Heatmap for correlation matrix
#Negative correlations are shown in blue and positive in red
col<- colorRampPalette(c("blue", "white", "red"))(20)
heatmap(cormatrix, col=col, symm=TRUE)

#Covariance Matrix
covmatrix <- round(cov(telco_churn_final),4)
#Heatmap for covariance matrix
#Negative correlations are shown in blue and positive in red
col<- colorRampPalette(c("blue", "white", "red"))(20)
heatmap(covmatrix, col=col, symm=TRUE)
```

```
##Test of Significance
```

```
#T-Test
#Null Hypothesis - The two means are equal
#Alternate Hypothesis - Difference in the two means is not zero
#pvalue >= 0.05, accept null hypothesis
#Or else accept the alternate hypothesis

#Univariate mean comparison using t test

#Totalcharges and Churn
with(data=telco_churn,t.test(telco_churn$TotalCharges[telco_churn$Churn=="Yes"],telco_churn$TotalCharges[tel
co_churn$Churn=="No"],var.equal=TRUE))
```

```
##
##  Two Sample t-test
##
## data:  telco_churn$TotalCharges[telco_churn$Churn == "Yes"] and telco_churn$TotalCharges[telco_churn$Chur
n == "No"]
## t = -17.069, df = 7030, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1141.0993  -905.9968
## sample estimates:
## mean of x mean of y
##  1531.796  2555.344
```

```
#MonthlyCharges and Churn
with(data=telco_churn,t.test(telco_churn$MonthlyCharges[telco_churn$Churn=="Yes"],telco_churn$MonthlyCharges
[telco_churn$Churn=="No"],var.equal=TRUE))
```

```
##
##  Two Sample t-test
##
## data:  telco_churn$MonthlyCharges[telco_churn$Churn == "Yes"] and telco_churn$MonthlyCharges[telco_churn$
Churn == "No"]
## t = 16.48, df = 7030, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.57160 14.69625
## sample estimates:
## mean of x mean of y
##  74.44133  61.30741
```

```
#Totalcharges and Churn
with(data=telco_churn,t.test(telco_churn$TotalCharges[telco_churn$gender=="Male"],telco_churn$TotalCharges[t
elco_churn$gender=="Female"],var.equal=TRUE))
```

```
##
##  Two Sample t-test
##
## data:  telco_churn$TotalCharges[telco_churn$gender == "Male"] and telco_churn$TotalCharges[telco_churn$ge
nder == "Female"]
## t = 0.0040111, df = 7030, p-value = 0.9968
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -105.7747  106.2084
## sample estimates:
## mean of x mean of y
##  2283.408  2283.191
```

```
#MonthlyCharges and Churn
with(data=telco_churn,t.test(telco_churn$MonthlyCharges[telco_churn$gender=="Male"],telco_churn$MonthlyCharg
es[telco_churn$gender=="Female"],var.equal=TRUE))
```

```
##
##  Two Sample t-test
##
## data:  telco_churn$MonthlyCharges[telco_churn$gender == "Male"] and telco_churn$MonthlyCharges[telco_chur
n$gender == "Female"]
## t = -1.1554, df = 7030, p-value = 0.248
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.2357573  0.5775443
## sample estimates:
## mean of x mean of y
##  64.38755  65.21665
```

```
#Multivariate mean comparison using Hotelling test

#Charges and gender
t2testgender <- hotelling.test(telco_churn$TotalCharges + telco_churn$MonthlyCharges ~ telco_churn$gender, d
ata=telco_churn)
cat("T2 statistic =",t2testgender$stat[[1]],"\n")
```

```
## T2 statistic = 2.328147
```

```
print(t2testgender)
```

```
## Test stat:  1.1639
## Numerator df:  2
## Denominator df:  7029
## P-value:  0.3123
```

```
#Charges and Churn
t2testelco_churn <- hotelling.test(telco_churn$TotalCharges + telco_churn$MonthlyCharges ~ telco_churn$Churn
, data=telco_churn)
cat("T2 statistic =",t2testelco_churn$stat[[1]],"\n")
```

```
## T2 statistic = 1989.619
```

```
print(t2testelco_churn)
```

```
## Test stat:  994.67
## Numerator df:  2
## Denominator df:  7029
## P-value:  0
```

```
#F Test
#Null Hypothesis - The two samples have same variance
#Alternate Hypothesis - Difference in the variance of two samples
#pvalue >= 0.05, accept null hypothesis
#Or else accept the alternate hypothesis

#The numerical columns we have do not have a normal distribution. Therefore we skip the F test
```