# EDA - Dataset 01

Yutika Rege

17/06/2020

## Exploratory Data Analysis for Dataset 1

Installing all the necessary libraries :

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.3
```

```
## corrplot 0.84 loaded
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages -------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

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

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

Downloading the dataset and encoding all the categorical variables :

```
ds1 <- read.csv(file.choose(), header = T)
num <-data.matrix(ds1)
df1 <- data.frame(num)
head(df1)
```

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1       5376      1             0       2          1      1            1
## 2       3963      2             0       1          1     34            2
## 3       2565      2             0       1          1      2            2
## 4       5536      2             0       1          1     45            1
## 5       6512      1             0       1          1      2            2
## 6       6552      1             0       1          1      8            2
##   MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
```

```
## 1             2              1             1           3               1
## 2             1              1             3           1               3
## 3             1              1             3           3               1
## 4             2              1             3           1               3
## 5             1              2             1           1               1
## 6             3              2             1           1               3
##   TechSupport StreamingTV StreamingMovies Contract PaperlessBilling
## 1           1           1               1        1                2
## 2           1           1               1        2                1
## 3           1           1               1        1                2
## 4           3           1               1        2                1
## 5           1           1               1        1                2
## 6           1           3               3        1                2
##   PaymentMethod MonthlyCharges TotalCharges Churn
## 1             3          29.85        29.85     1
## 2             4          56.95      1889.50     1
## 3             4          53.85       108.15     2
## 4             1          42.30      1840.75     1
## 5             3          70.70       151.65     2
## 6             3          99.65       820.50     2
```

Viewing the structure of of the newly formed dataframe :

```
structure(head(df1))
```

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1       5376      1             0       2          1      1            1
## 2       3963      2             0       1          1     34            2
## 3       2565      2             0       1          1      2            2
## 4       5536      2             0       1          1     45            1
## 5       6512      1             0       1          1      2            2
## 6       6552      1             0       1          1      8            2
##   MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
## 1             2               1              1            3                1
## 2             1               1              3            1                3
## 3             1               1              3            3                1
## 4             2               1              3            1                3
## 5             1               2              1            1                1
## 6             3               2              1            1                3
##   TechSupport StreamingTV StreamingMovies Contract PaperlessBilling
## 1           1           1               1        1                2
## 2           1           1               1        2                1
## 3           1           1               1        1                2
## 4           3           1               1        2                1
## 5           1           1               1        1                2
## 6           1           3               3        1                2
##   PaymentMethod MonthlyCharges TotalCharges Churn
## 1             3          29.85        29.85     1
## 2             4          56.95      1889.50     1
## 3             4          53.85       108.15     2
## 4             1          42.30      1840.75     1
## 5             3          70.70       151.65     2
## 6             3          99.65       820.50     2
```

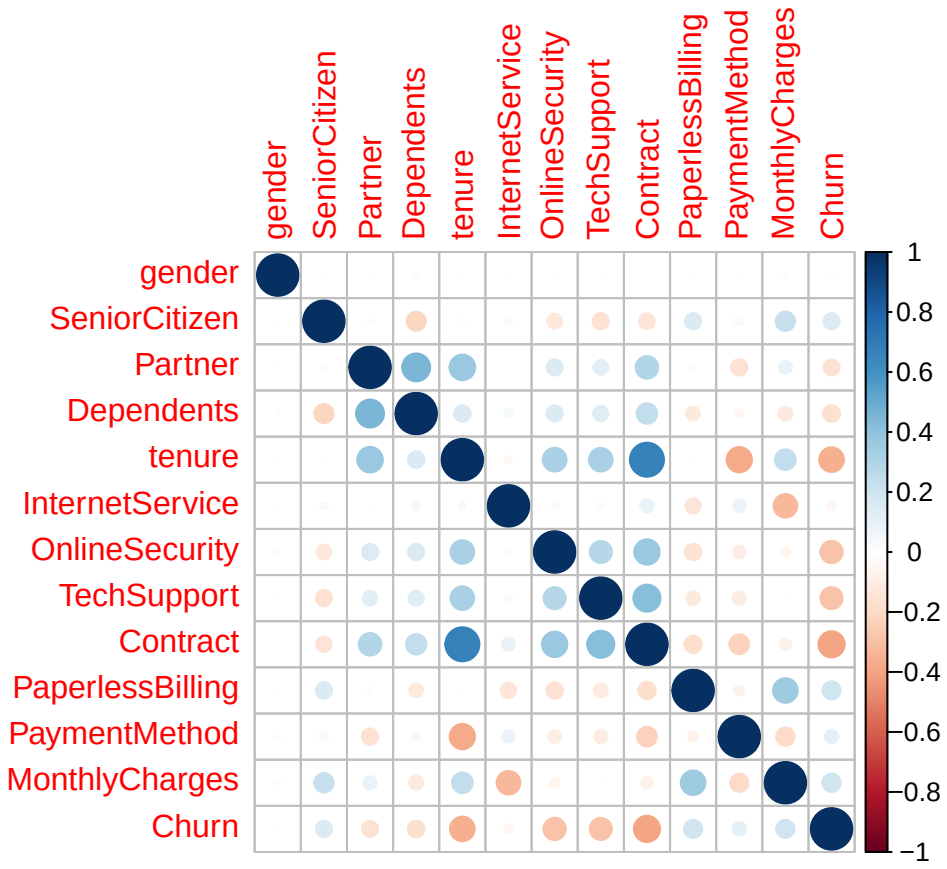
Initiation of EDA :

```r
summary(df1)
```

```
##    customerID       gender      SeniorCitizen       Partner        Dependents
##  Min.   :   1   Min.   :1.000   Min.   :0.0000   Min.   :1.000   Min.   :1.0
##  1st Qu.:1762   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:1.0
##  Median :3522   Median :2.000   Median :0.0000   Median :1.000   Median :1.0
##  Mean   :3522   Mean   :1.505   Mean   :0.1621   Mean   :1.483   Mean   :1.3
##  3rd Qu.:5282   3rd Qu.:2.000   3rd Qu.:0.0000   3rd Qu.:2.000   3rd Qu.:2.0
##  Max.   :7043   Max.   :2.000   Max.   :1.0000   Max.   :2.000   Max.   :2.0
##
##      tenure       PhoneService   MultipleLines   InternetService OnlineSecurity
##  Min.   : 0.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.00
##  1st Qu.: 9.00   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.00
##  Median :29.00   Median :2.000   Median :2.000   Median :2.000   Median :2.00
##  Mean   :32.37   Mean   :1.903   Mean   :1.941   Mean   :1.873   Mean   :1.79
##  3rd Qu.:55.00   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.00
##  Max.   :72.00   Max.   :2.000   Max.   :3.000   Max.   :3.000   Max.   :3.00
##
##   OnlineBackup   DeviceProtection  TechSupport     StreamingTV
##  Min.   :1.000   Min.   :1.000    Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000    1st Qu.:1.000   1st Qu.:1.000
##  Median :2.000   Median :2.000    Median :2.000   Median :2.000
##  Mean   :1.906   Mean   :1.904    Mean   :1.797   Mean   :1.985
##  3rd Qu.:3.000   3rd Qu.:3.000    3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :3.000   Max.   :3.000    Max.   :3.000   Max.   :3.000
##
##  StreamingMovies    Contract    PaperlessBilling PaymentMethod
##  Min.   :1.000   Min.   :1.00   Min.   :1.000    Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.:1.000    1st Qu.:2.000
##  Median :2.000   Median :1.00   Median :2.000    Median :3.000
##  Mean   :1.992   Mean   :1.69   Mean   :1.592    Mean   :2.574
##  3rd Qu.:3.000   3rd Qu.:2.00   3rd Qu.:2.000    3rd Qu.:3.000
##  Max.   :3.000   Max.   :3.00   Max.   :2.000    Max.   :4.000
##
##  MonthlyCharges    TotalCharges        Churn
##  Min.   : 18.25   Min.   :  18.8   Min.   :1.000
##  1st Qu.: 35.50   1st Qu.: 401.4   1st Qu.:1.000
##  Median : 70.35   Median :1397.5   Median :1.000
##  Mean   : 64.76   Mean   :2283.3   Mean   :1.265
##  3rd Qu.: 89.85   3rd Qu.:3794.7   3rd Qu.:2.000
##  Max.   :118.75   Max.   :8684.8   Max.   :2.000
##                   NA's   :11
```

Correlation plot :

```r
c <- cor(df1[,c(2,3,4,5,6,9,10,13,16,17,18,19,21)])
corrplot(c, method = 'circle')
```




**Comments:**

The correlation plot aids us to understand the nature of correlation between various variables of the dataset.The correlation matrix is an intersection of the various variables with each other.The darkness of colour suggests the type of correlation (positive or negative) and size of the circles within the grid determine the strength of the correlation.The blue circles correspond to positive correlations whereas the red ones correspond to negative correlations.

From the plot above we can interpret that pairs of variables such as tenure-online security, tenure-contract, contract-tech support etc. share a positve correlation whereas other pairs such as tenure-churn, contract-churn, tenure-payment method share negative a correlation.

```
library(explore)
```

```
## Warning: package 'explore' was built under R version 3.6.3
```

**Fig-1.1**

```
df1 %>%
  select(OnlineSecurity,Contract,tenure,Churn,TechSupport,MonthlyCharges) %>%
  explore_all()
```

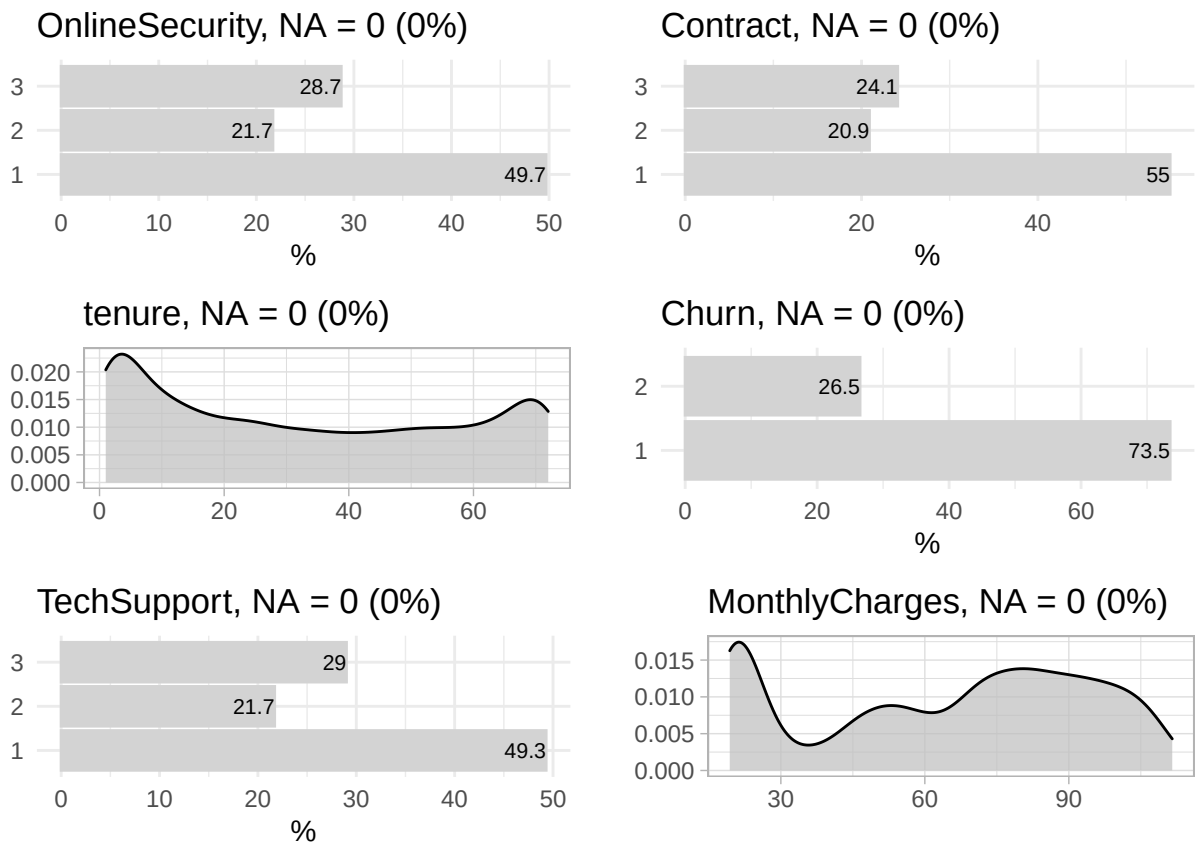
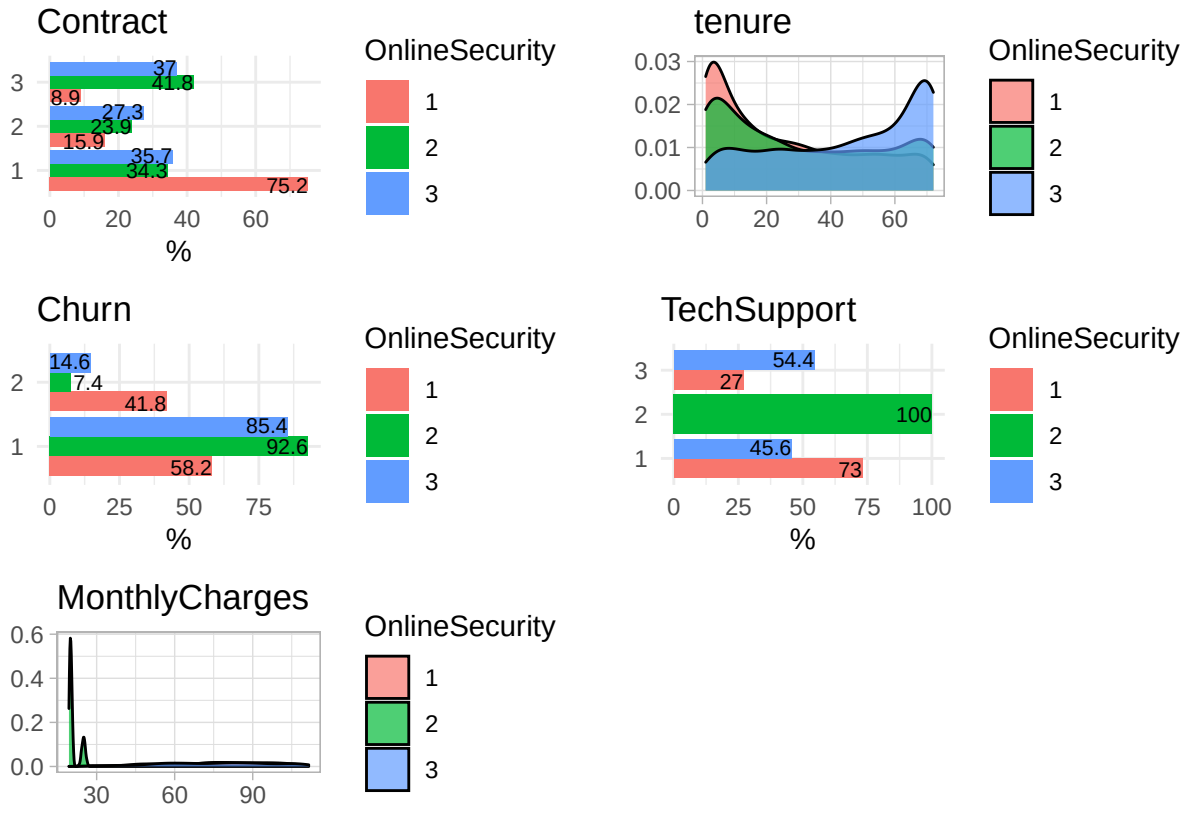



**Fig-1.2**

```
df1 %>%
  select(OnlineSecurity,Contract,tenure,Churn,TechSupport,MonthlyCharges) %>%
  explore_all(target = OnlineSecurity)
```




**Comments:**

1) 55% of the customers have a 'month-to-month' contract out of which 75.2% have no online security, 35.7% have online security and the rest have no internet which is inconsequential for the purpose of analysis.

2) The **no churn** is to **churn** ratio is **7.35**:**2.65** implying that most of the customers are content with the carrier service. Out of the population of customers that don't churn, 92.6% do not have internet at all. 85.4% customers have proper online security which might also factor in as a possible reason of the high percentage of customer retention.

3)49.3% of the customers don't get to avail any technical support out of which 73% do not have online security at all. Only 29% get tech support out of which 54.4% have online security and 27% don't.

**Fig-2.1:**Boxplot for tenure around online security.

```
ggplot(df1,aes(factor(OnlineSecurity),tenure))+geom_boxplot(aes(colour = factor(OnlineSecurity)))+ggtitl
```

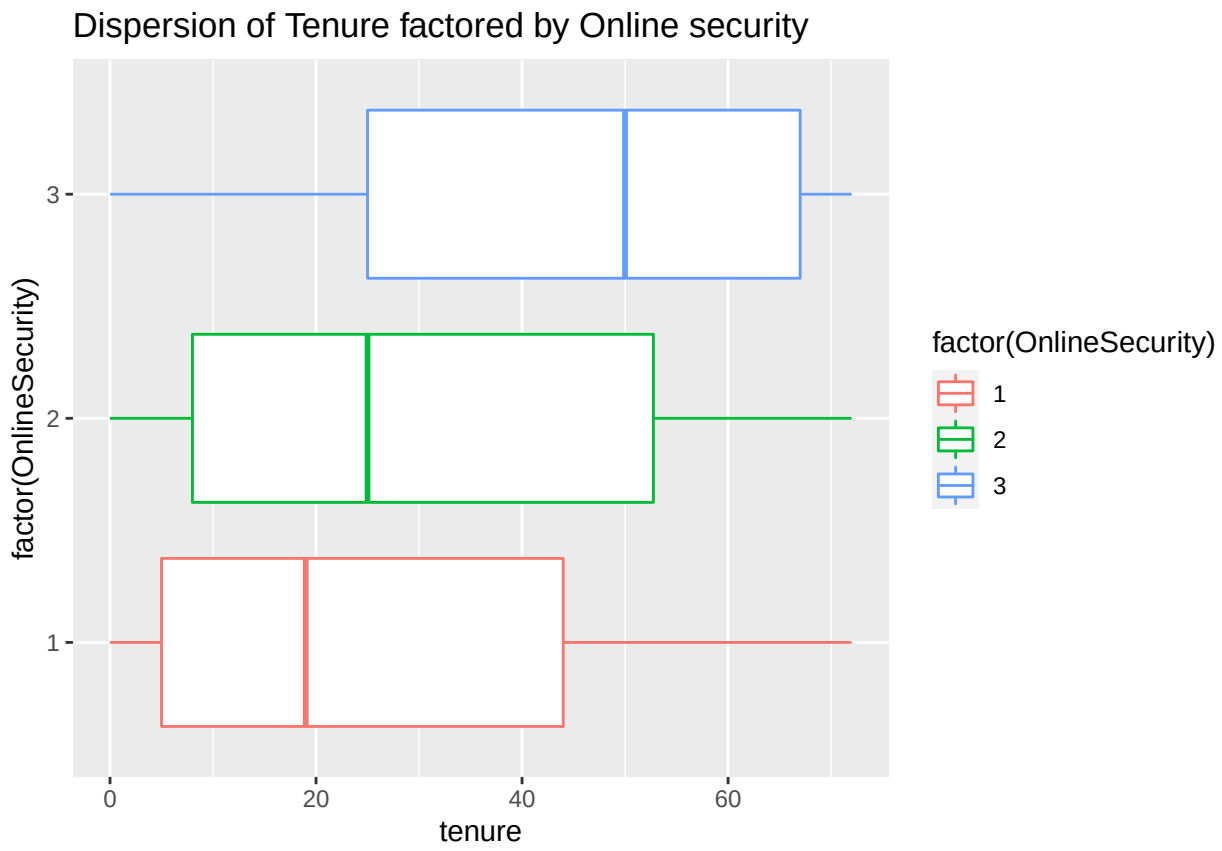



**Fig-2.2:**Comparative density plot for online security

```
library(sm)
```

```
## Warning: package 'sm' was built under R version 3.6.3
```

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

```
sm.density.compare(df1$tenure,df1$OnlineSecurity,xlab = 'Tenure')
title(main = 'Tenure as per Online security')
```

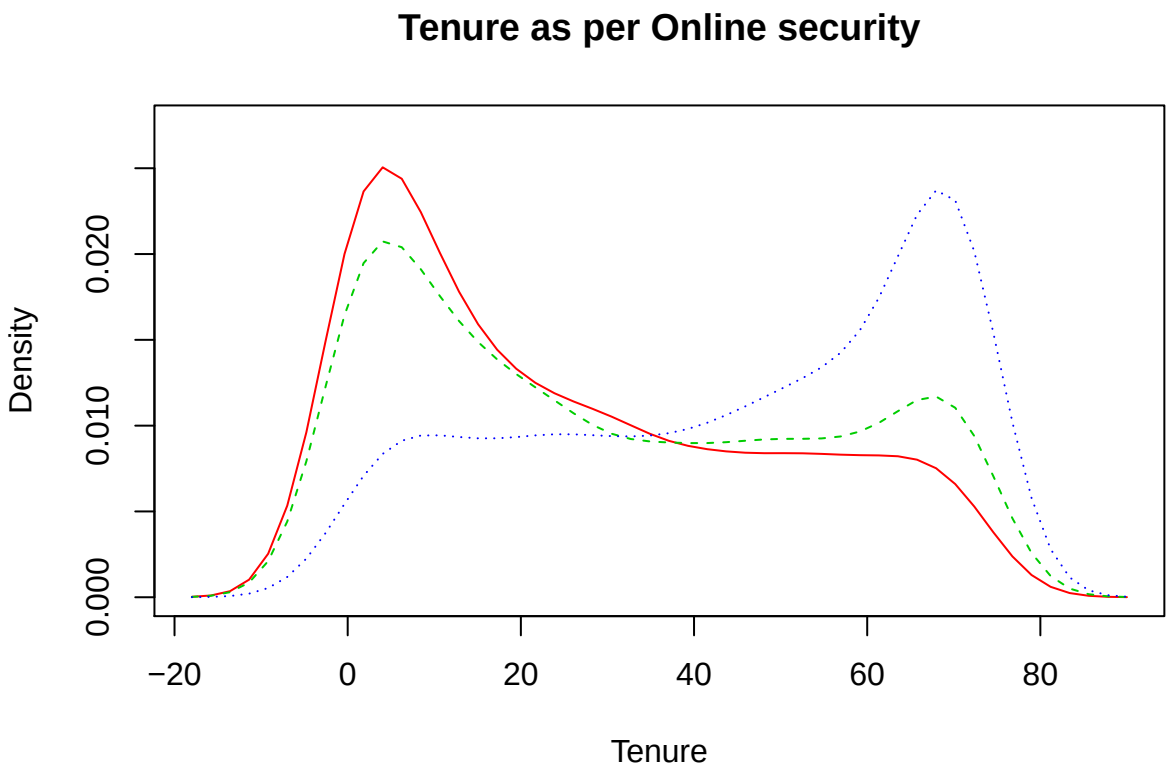



**Fig-3.1:**Boxplot for tenure around contract.

```
ggplot(df1,aes(factor(Contract),tenure))+geom_boxplot(aes(colour = factor(Contract)))+ggtitle('Dispersi
```

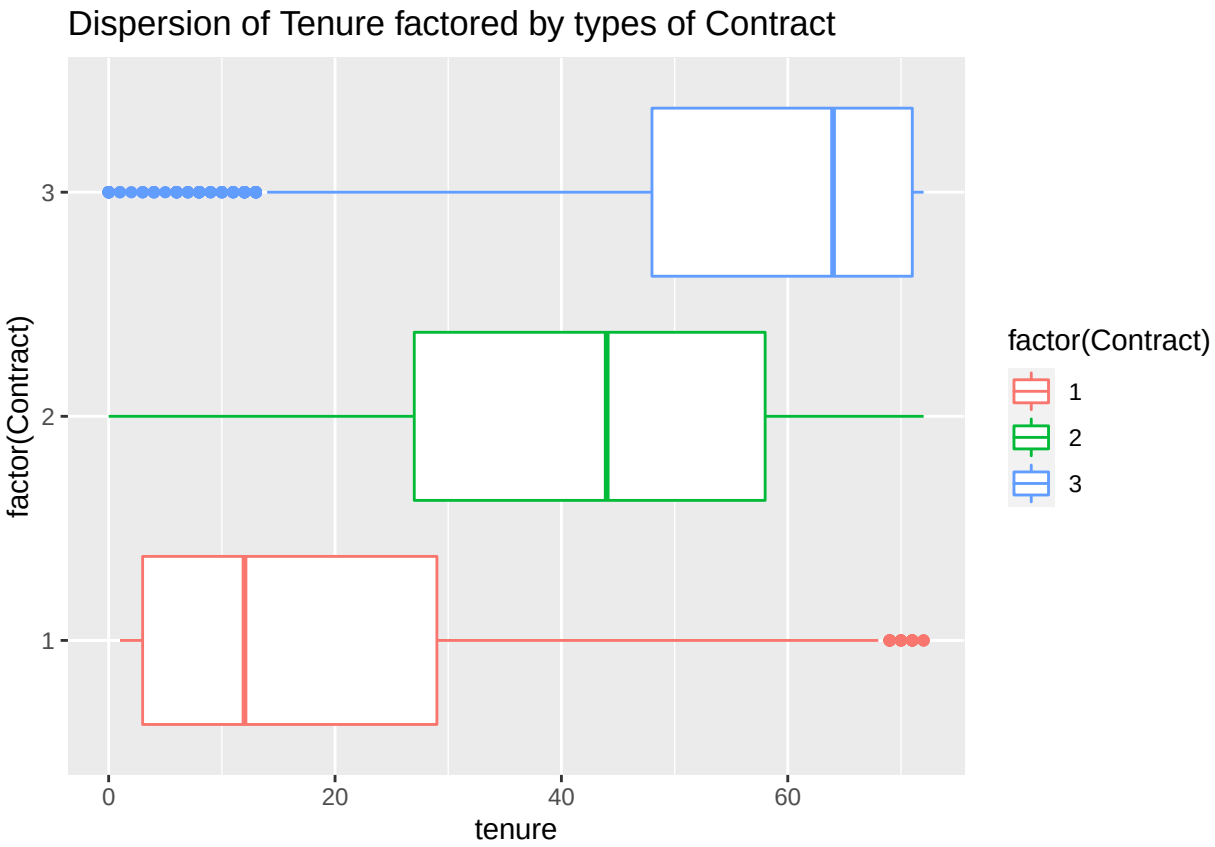



**Fig-3.2:**Comparative density plot for contract

```
sm.density.compare(df1$tenure,df1$Contract,xlab = 'Tenure')
title(main = 'Tenure as per Contract')
```

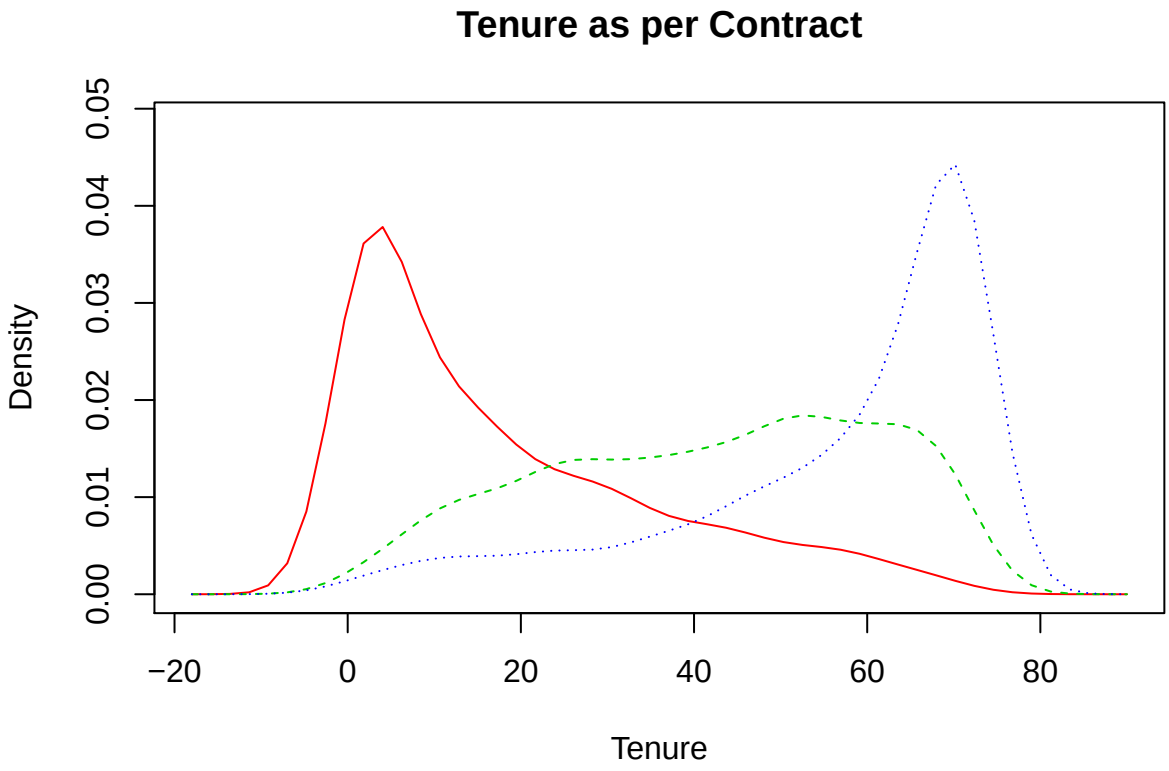



**Fig-4.1:**Boxplot for tenure around churn.

```
ggplot(df1,aes(factor(Churn),tenure))+geom_boxplot(aes(colour = factor(Churn)))+ggtitle('Dispersion of '
```

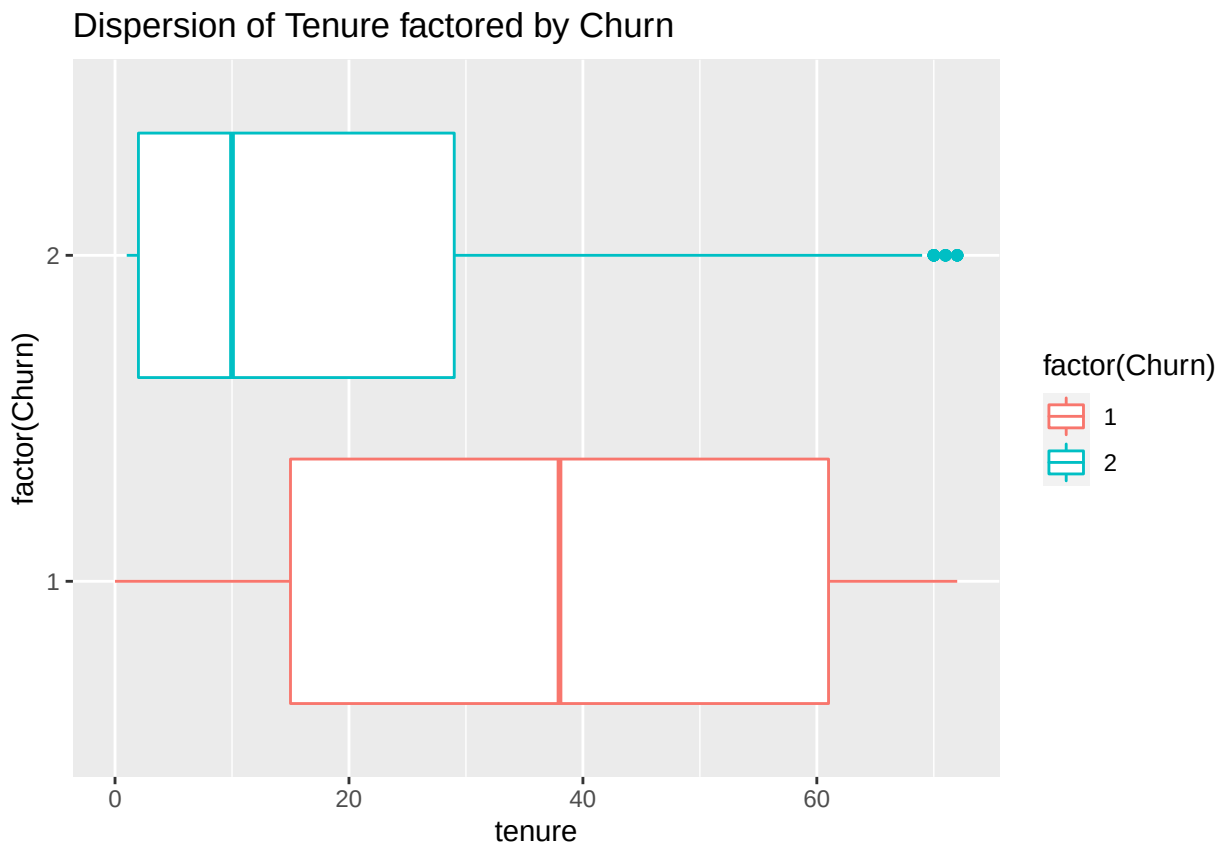
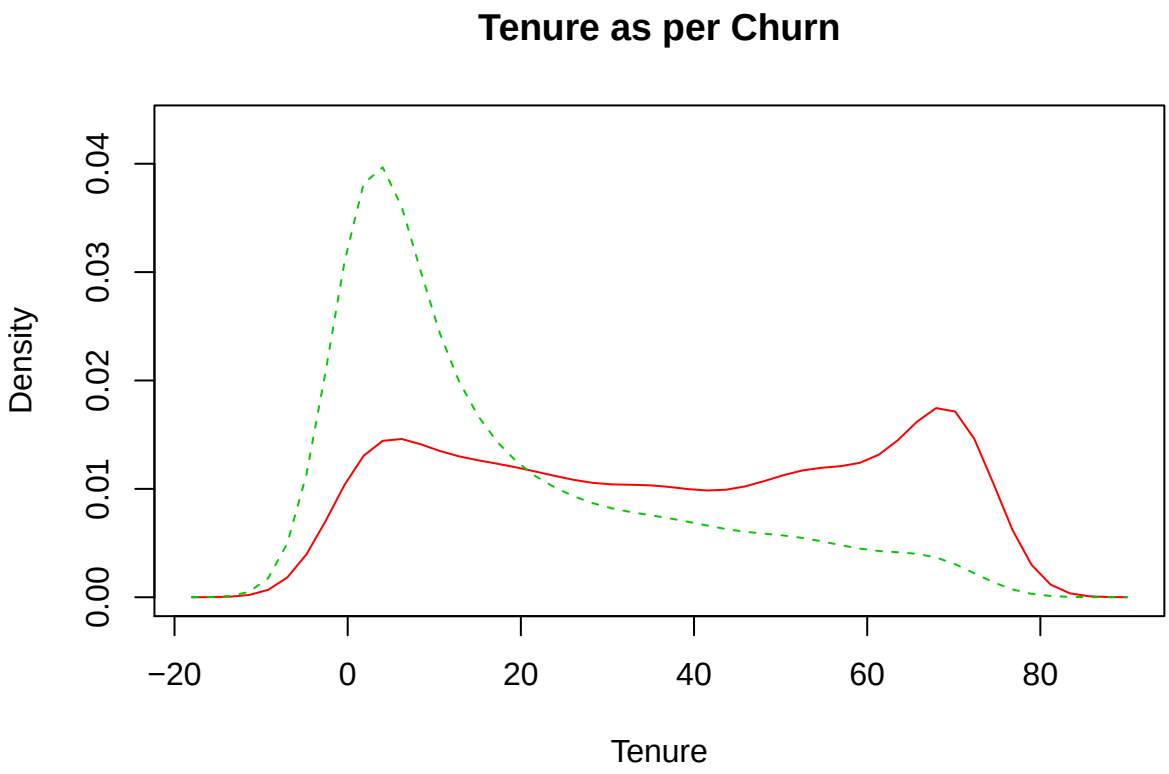



**Fig-4.2:**Comparative density plot for churn

```
library(sm)
sm.density.compare(df1$tenure,df1$Churn,xlab = 'Tenure')
title(main = 'Tenure as per Churn')
```




**Comments:**

1) From the boxplots and comparative density plots it can be concluded that the distribution for online security = 'yes' is negatively skewed. For online security = 'no', the distribution is positively skewed and as per the boxplot, the distribution is somewhat positively skewed for 'No-internet'.

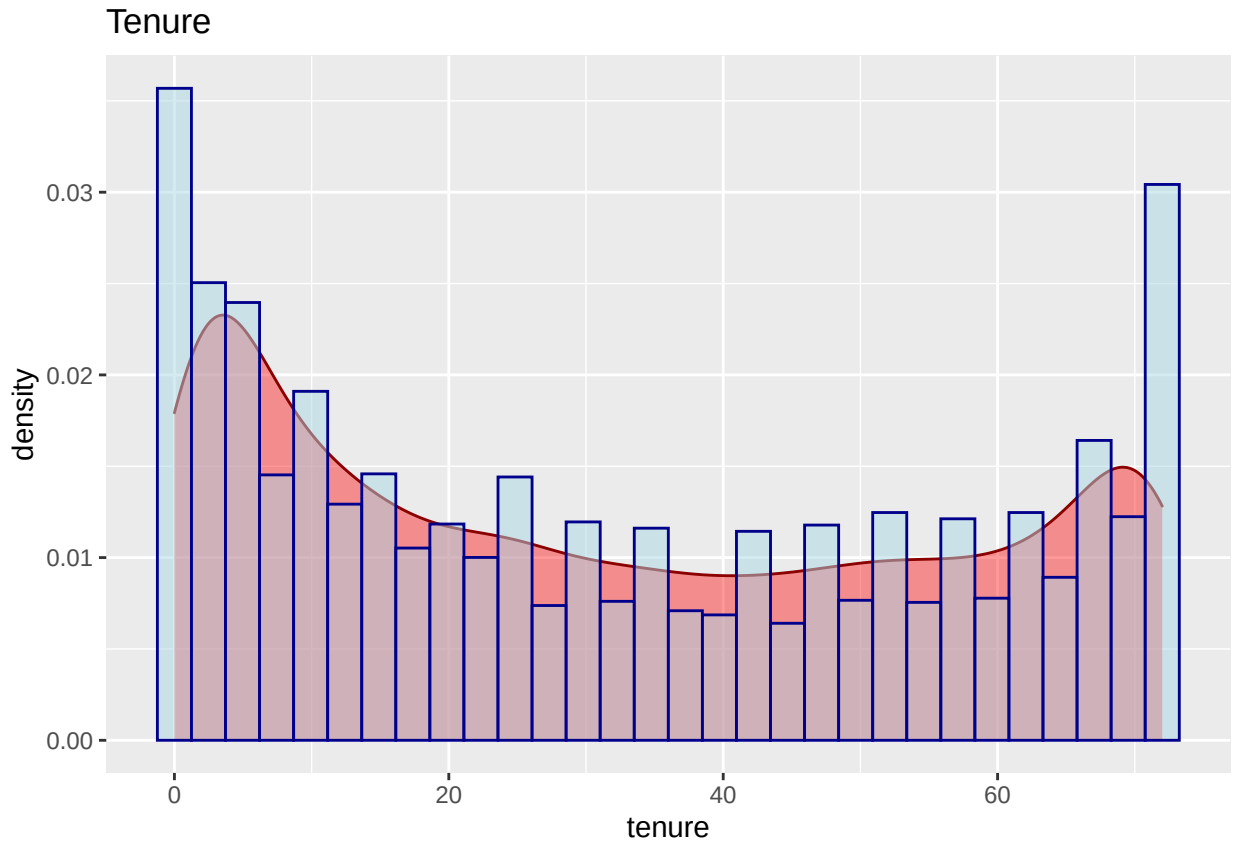
2) The distribution for 'two-year' contract is highly negatively skewed. The 'one-year' contract shows a slight negative skew whereas the 'month-to-month' contract is positively skewed.

3) The distribution for churn = 'no' is symmetrically skewed (acc. to the boxplot) whereas the distribution for churn = 'yes' is highly positively skewed.

**Fig-5.1:**Integrated histogram and density plot for Tenure

```
ggplot(df1, aes(x=tenure))+geom_density(alpha =0.4, fill ='red', colour = 'darkred')+geom_histogram(alp
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
library(PerformanceAnalytics)
```

```
## Warning: package 'PerformanceAnalytics' was built under R version 3.6.3
## Loading required package: xts
## Warning: package 'xts' was built under R version 3.6.3
## Loading required package: zoo
## Warning: package 'zoo' was built under R version 3.6.3
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Attaching package: 'xts'
## The following objects are masked from 'package:dplyr':
##
##     first, last
##
## Attaching package: 'PerformanceAnalytics'
## The following object is masked from 'package:graphics':
##
```

```
##     legend
```

```
kurtosis(df1$tenure)
```

```
## [1] -1.387239
```

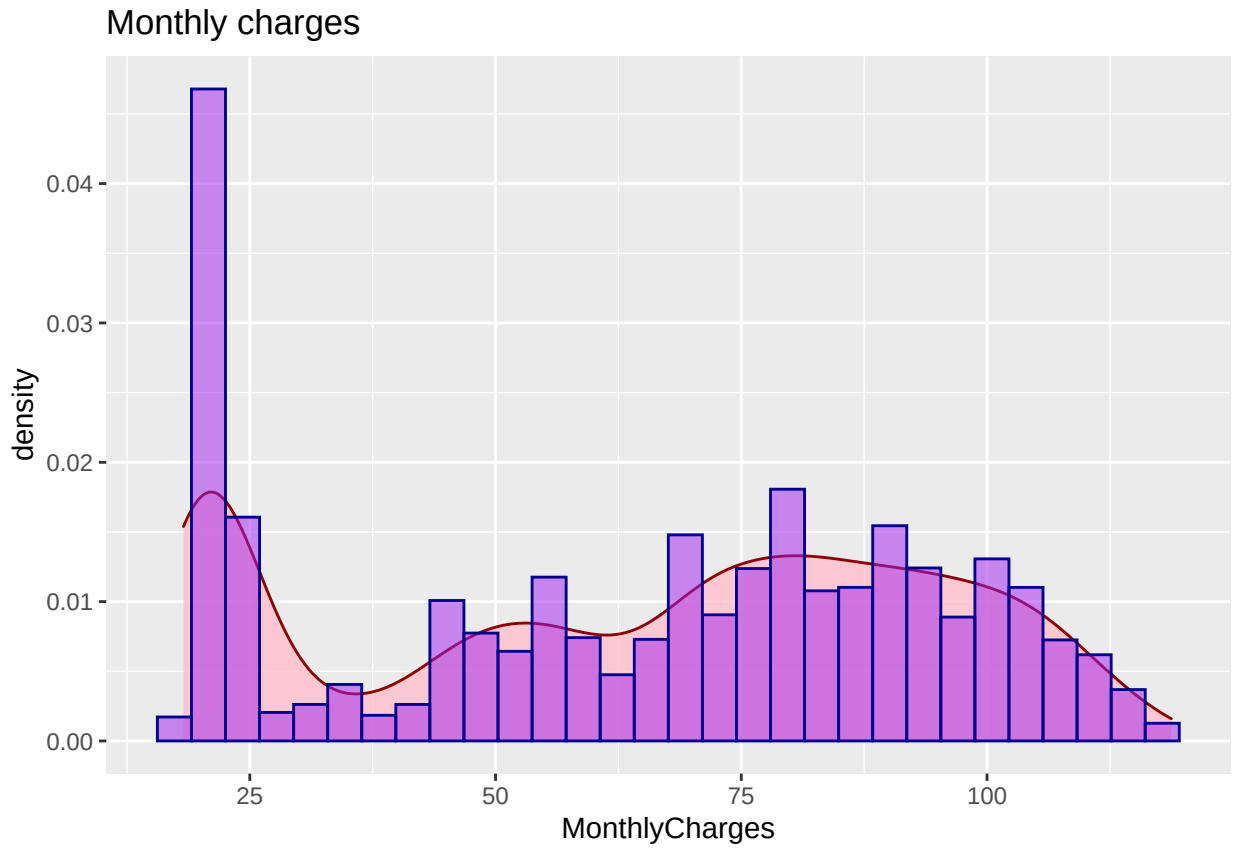
```
skewness(df1$tenure)
```

```
## [1] 0.2394887
```

**Fig-5.2:**Integrated histogram and density plot for Monthly Charges

```
ggplot(df1, aes(x=MonthlyCharges))+geom_density(alpha =0.8, fill ='pink', colour = 'darkred')+geom_hist
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
kurtosis(df1$MonthlyCharges)
```

```
## [1] -1.257219
```

```
skewness(df1$MonthlyCharges)
```

```
## [1] -0.2204775
```

**Comments:**

The integrated plots for both the variables - 'tenure' and 'monthly charges' show that their respective distributions are platykurtic. This distribution of tenure is positively skewed and for monthly charges, the distribution is negatively skewed .