EDA - Dataset 02

Yutika Rege

18/06/2020

Exploratory Data Analysis for Dataset 2

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)
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
```

```
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
```

Downloading the dataset and encoding all the categorical variables :

```
ds2 <- read.csv(file.choose(), header = T)
num <-data.matrix(ds2)
df2 <- data.frame(num)
head(df2)
```

```
##   state account.length area.code phone.number international.plan
## 1    17            128       415         1927                 1
## 2    36            107       415         1576                 1
## 3    32            137       415         1118                 1
## 4    36             84       408         1708                 2
## 5    37             75       415          111                 2
## 6     2            118       510         2254                 2
##   voice.mail.plan number.vmail.messages total.day.minutes total.day.calls
## 1               2                    25             265.1             110
## 2               2                    26             161.6             123
## 3               1                     0             243.4             114
## 4               1                     0             299.4              71
## 5               1                     0             166.7             113
## 6               1                     0             223.4              98
##   total.day.charge total.eve.minutes total.eve.calls total.eve.charge
## 1            45.07             197.4              99            16.78
## 2            27.47             195.5             103            16.62
## 3            41.38             121.2             110            10.30
## 4            50.90              61.9              88             5.26
## 5            28.34             148.3             122            12.61
## 6            37.98             220.6             101            18.75
##   total.night.minutes total.night.calls total.night.charge total.intl.minutes
## 1               244.7               91              11.01               10.0
## 2               254.4              103              11.45               13.7
## 3               162.6              104               7.32               12.2
## 4               196.9               89               8.86                6.6
## 5               186.9              121               8.41               10.1
## 6               203.9              118               9.18                6.3
##   total.intl.calls total.intl.charge customer.service.calls churn
## 1                3              2.70                      1     1
```

```
## 2                3              3.70                      1     1
## 3                5              3.29                      0     1
## 4                7              1.78                      2     1
## 5                3              2.73                      3     1
## 6                6              1.70                      0     1
```

Viewing the structure of of the newly formed dataframe :

```
structure(head(df2))
```

```
##   state account.length area.code phone.number international.plan
## 1    17            128       415         1927                  1
## 2    36            107       415         1576                  1
## 3    32            137       415         1118                  1
## 4    36             84       408         1708                  2
## 5    37             75       415          111                  2
## 6     2            118       510         2254                  2
##   voice.mail.plan number.vmail.messages total.day.minutes total.day.calls
## 1               2                    25             265.1             110
## 2               2                    26             161.6             123
## 3               1                     0             243.4             114
## 4               1                     0             299.4              71
## 5               1                     0             166.7             113
## 6               1                     0             223.4              98
##   total.day.charge total.eve.minutes total.eve.calls total.eve.charge
## 1            45.07             197.4              99            16.78
## 2            27.47             195.5             103            16.62
## 3            41.38             121.2             110            10.30
## 4            50.90              61.9              88             5.26
## 5            28.34             148.3             122            12.61
## 6            37.98             220.6             101            18.75
##   total.night.minutes total.night.calls total.night.charge total.intl.minutes
## 1               244.7                91              11.01               10.0
## 2               254.4               103              11.45               13.7
## 3               162.6               104               7.32               12.2
## 4               196.9                89               8.86                6.6
## 5               186.9               121               8.41               10.1
## 6               203.9               118               9.18                6.3
##   total.intl.calls total.intl.charge customer.service.calls churn
## 1                3              2.70                      1     1
## 2                3              3.70                      1     1
## 3                5              3.29                      0     1
## 4                7              1.78                      2     1
## 5                3              2.73                      3     1
## 6                6              1.70                      0     1
```

EDA :

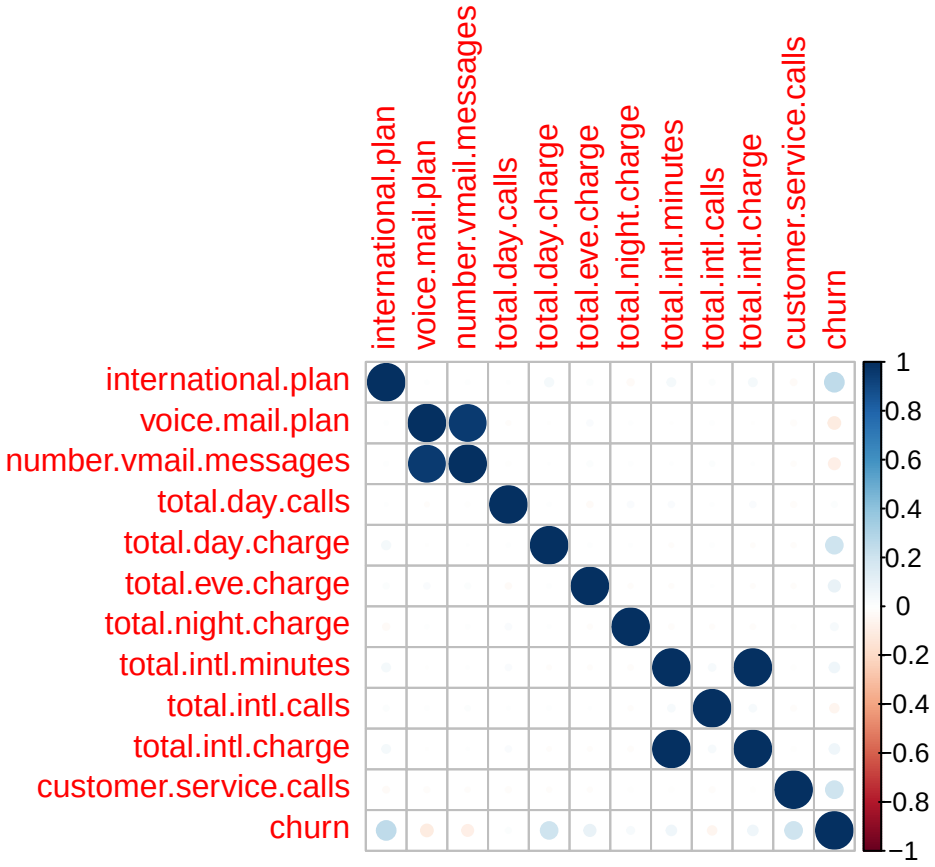
```
summary(df2)
```

```
##      state       account.length    area.code      phone.number
##  Min.   : 1.00   Min.   :  1.0   Min.   :408.0   Min.   :   1
##  1st Qu.:15.00   1st Qu.: 74.0   1st Qu.:408.0   1st Qu.: 834
##  Median :27.00   Median :101.0   Median :415.0   Median :1667
##  Mean   :27.06   Mean   :101.1   Mean   :437.2   Mean   :1667
##  3rd Qu.:40.00   3rd Qu.:127.0   3rd Qu.:510.0   3rd Qu.:2500
```

```
##  Max.   :51.00   Max.   :243.0   Max.   :510.0   Max.   :3333
##  international.plan voice.mail.plan number.vmail.messages total.day.minutes
##  Min.   :1.000      Min.   :1.000   Min.   : 0.000        Min.   :  0.0
##  1st Qu.:1.000      1st Qu.:1.000   1st Qu.: 0.000        1st Qu.:143.7
##  Median :1.000      Median :1.000   Median : 0.000        Median :179.4
##  Mean   :1.097      Mean   :1.277   Mean   : 8.099        Mean   :179.8
##  3rd Qu.:1.000      3rd Qu.:2.000   3rd Qu.:20.000        3rd Qu.:216.4
##  Max.   :2.000      Max.   :2.000   Max.   :51.000        Max.   :350.8
##  total.day.calls total.day.charge total.eve.minutes total.eve.calls
##  Min.   :  0.0   Min.   : 0.00    Min.   :  0.0     Min.   :  0.0
##  1st Qu.: 87.0   1st Qu.:24.43    1st Qu.:166.6     1st Qu.: 87.0
##  Median :101.0   Median :30.50    Median :201.4     Median :100.0
##  Mean   :100.4   Mean   :30.56    Mean   :201.0     Mean   :100.1
##  3rd Qu.:114.0   3rd Qu.:36.79    3rd Qu.:235.3     3rd Qu.:114.0
##  Max.   :165.0   Max.   :59.64    Max.   :363.7     Max.   :170.0
##  total.eve.charge total.night.minutes total.night.calls total.night.charge
##  Min.   : 0.00    Min.   : 23.2       Min.   : 33.0     Min.   : 1.040
##  1st Qu.:14.16    1st Qu.:167.0       1st Qu.: 87.0     1st Qu.: 7.520
##  Median :17.12    Median :201.2       Median :100.0     Median : 9.050
##  Mean   :17.08    Mean   :200.9       Mean   :100.1     Mean   : 9.039
##  3rd Qu.:20.00    3rd Qu.:235.3       3rd Qu.:113.0     3rd Qu.:10.590
##  Max.   :30.91    Max.   :395.0       Max.   :175.0     Max.   :17.770
##  total.intl.minutes total.intl.calls total.intl.charge customer.service.calls
##  Min.   : 0.00      Min.   : 0.000   Min.   :0.000     Min.   :0.000
##  1st Qu.: 8.50      1st Qu.: 3.000   1st Qu.:2.300     1st Qu.:1.000
##  Median :10.30      Median : 4.000   Median :2.780     Median :1.000
##  Mean   :10.24      Mean   : 4.479   Mean   :2.765     Mean   :1.563
##  3rd Qu.:12.10      3rd Qu.: 6.000   3rd Qu.:3.270     3rd Qu.:2.000
##  Max.   :20.00      Max.   :20.000   Max.   :5.400     Max.   :9.000
##      churn
##  Min.   :1.000
##  1st Qu.:1.000
##  Median :1.000
##  Mean   :1.145
##  3rd Qu.:1.000
##  Max.   :2.000
```

Correlation plot :

```
c <- cor(df2[,c(5,6,7,9,10,13,16,17,18,19,20,21)])
corrplot(c, method = 'circle')
```



Comments:

From the plot above, it can be can interpretted that pairs of variables such as number.vmail.messages - voice.mail.plan, total.day.charge - churn, total.intl.charge - churn share and total.intl.minutes - total.intl.charge share a positve correlation whereas other pairs such as voicemail.plan - churn share a negative a correlation.

```
library(explore)
```

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

Fig-1.1

```
df2 %>%
  select(total.day.charge,international.plan,total.intl.charge,voice.mail.plan,customer.service.calls,c
  explore_all()
```

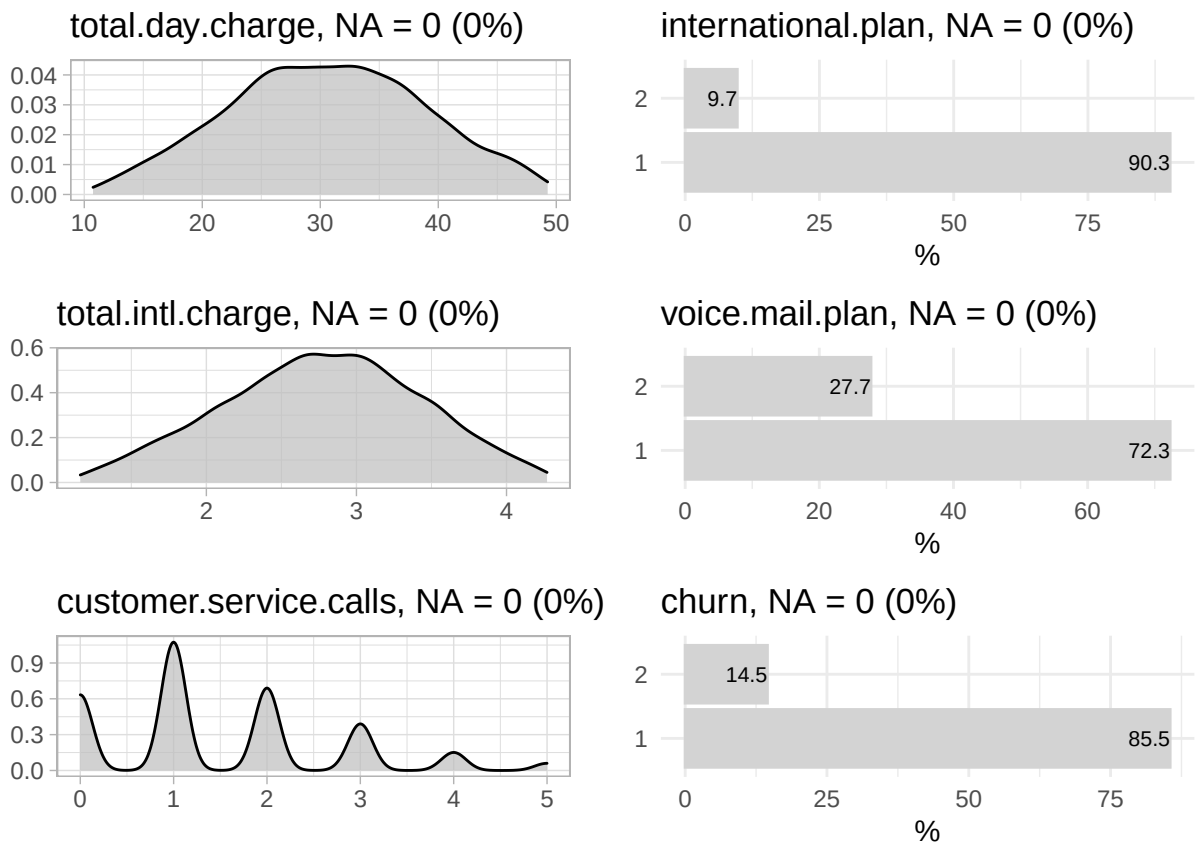
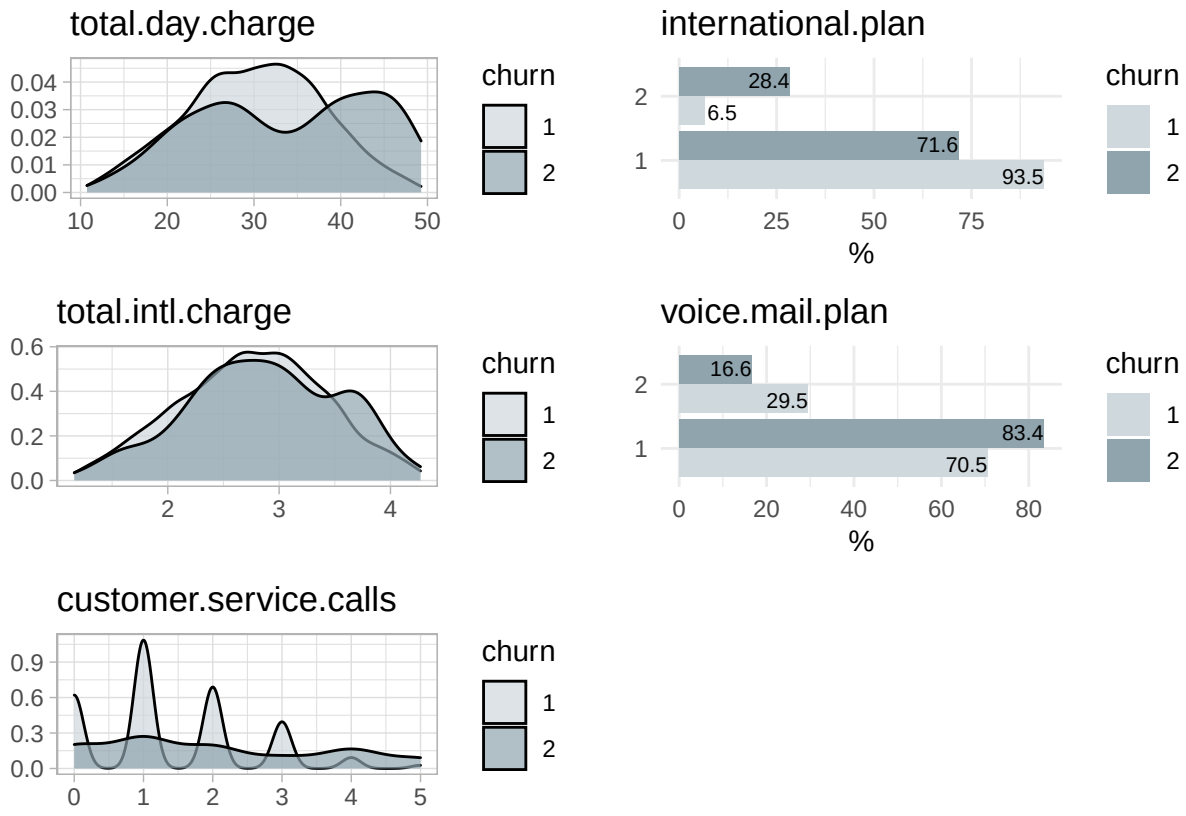


Fig-1.2

```
df2 %>%
  select(total.day.charge,international.plan,total.intl.charge,churn,voice.mail.plan,customer.service.ca
  explore_all(target = churn)
```



Comments:

1) A total of 14.5% customers churn out of which 28.4% have an international plan and 16.6% have a voicemail plan.

2) 85.5% customers don't churn 6.5% have an international plan and 29.5% have a voicemail plan.

Fig-2: Boxplot for number.vmail.messages-voice.mail.plan

```
ggplot(df2, aes(factor(voice.mail.plan),number.vmail.messages))+geom_boxplot(aes(colour = factor(voice.
```

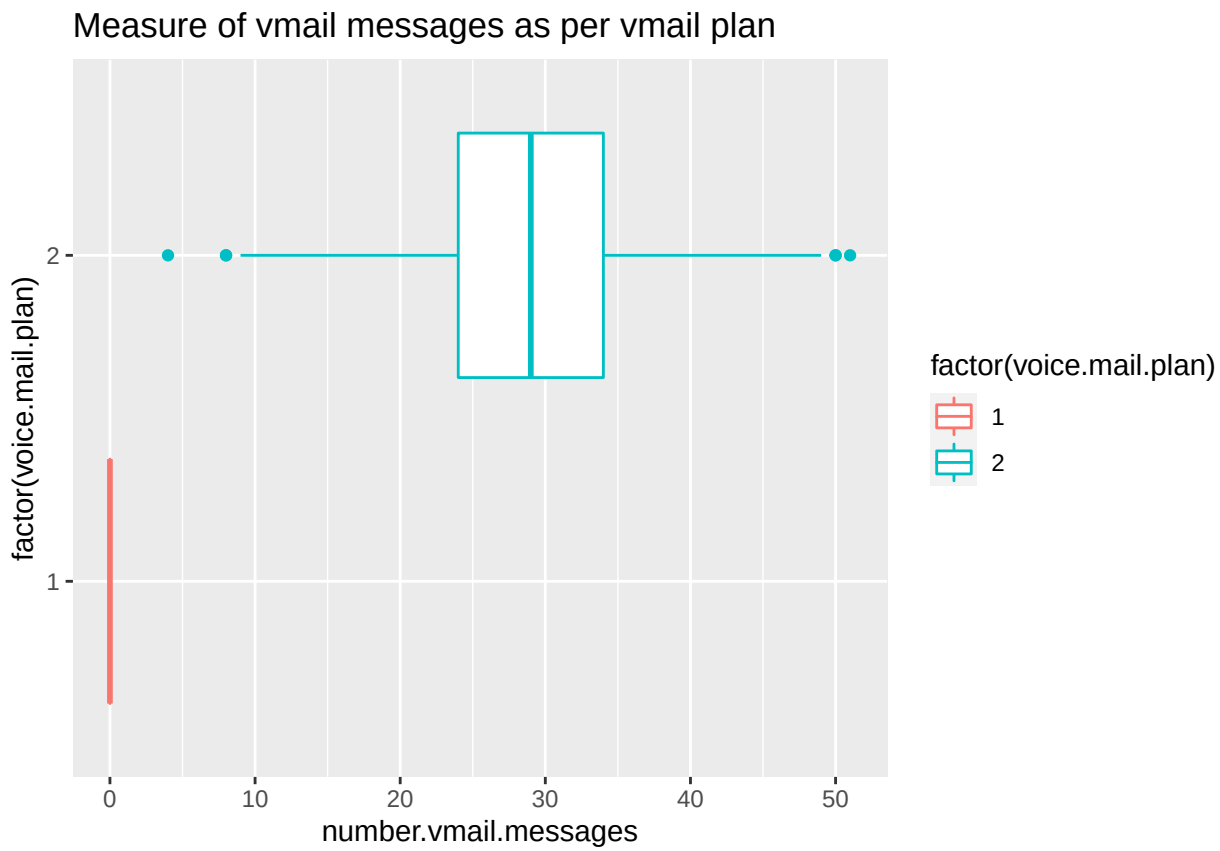


Fig-3.1: Boxplot for total.day.charge and churn

```
ggplot(df2,aes(factor(churn),total.day.charge))+ geom_boxplot(aes(colour = factor(churn)))+ ggtitle('Mea
per Churn')+coord_flip()
```

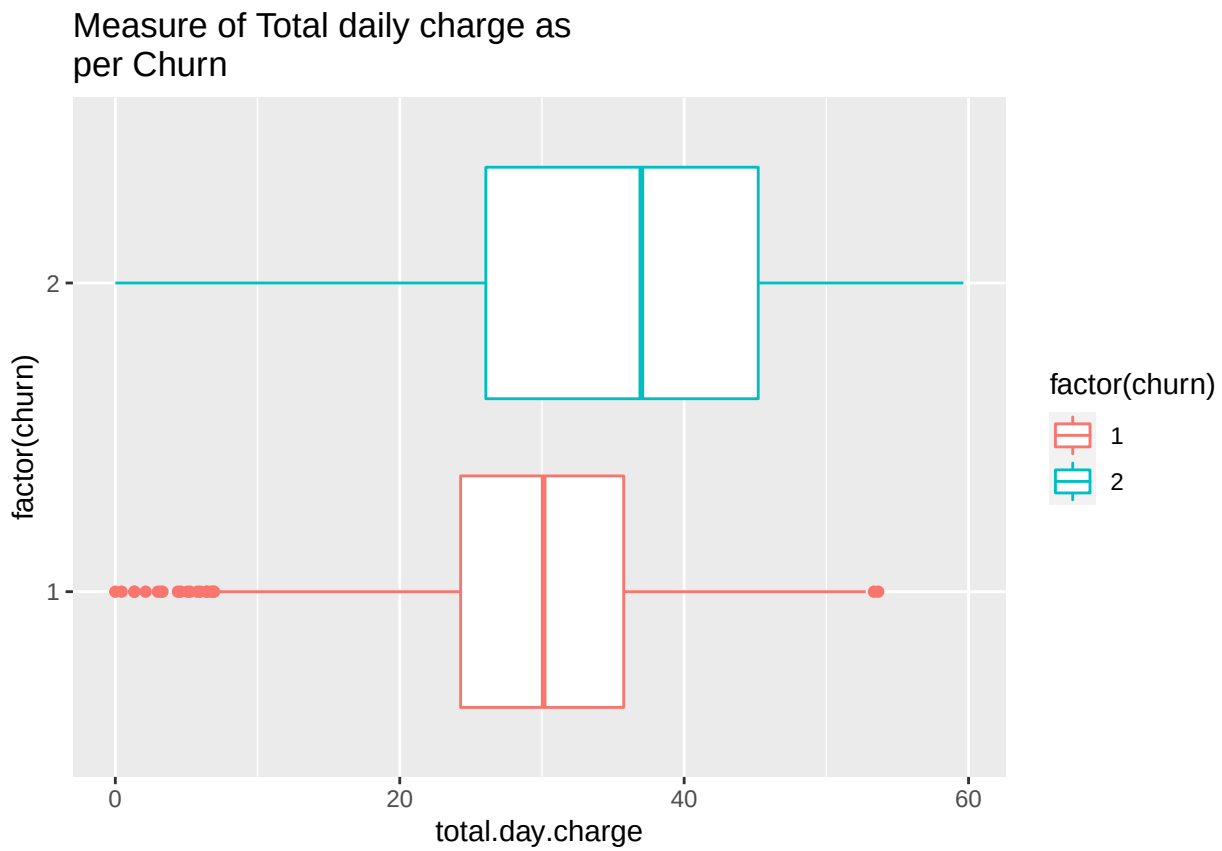


Fig-3.2: Comparative density plot for total.day.charge and churn

```
sm.density.compare(df2$total.day.charge, df2$churn, xlab ='Total daily charge')
title(main = 'Total day charged factored by Churn')
```

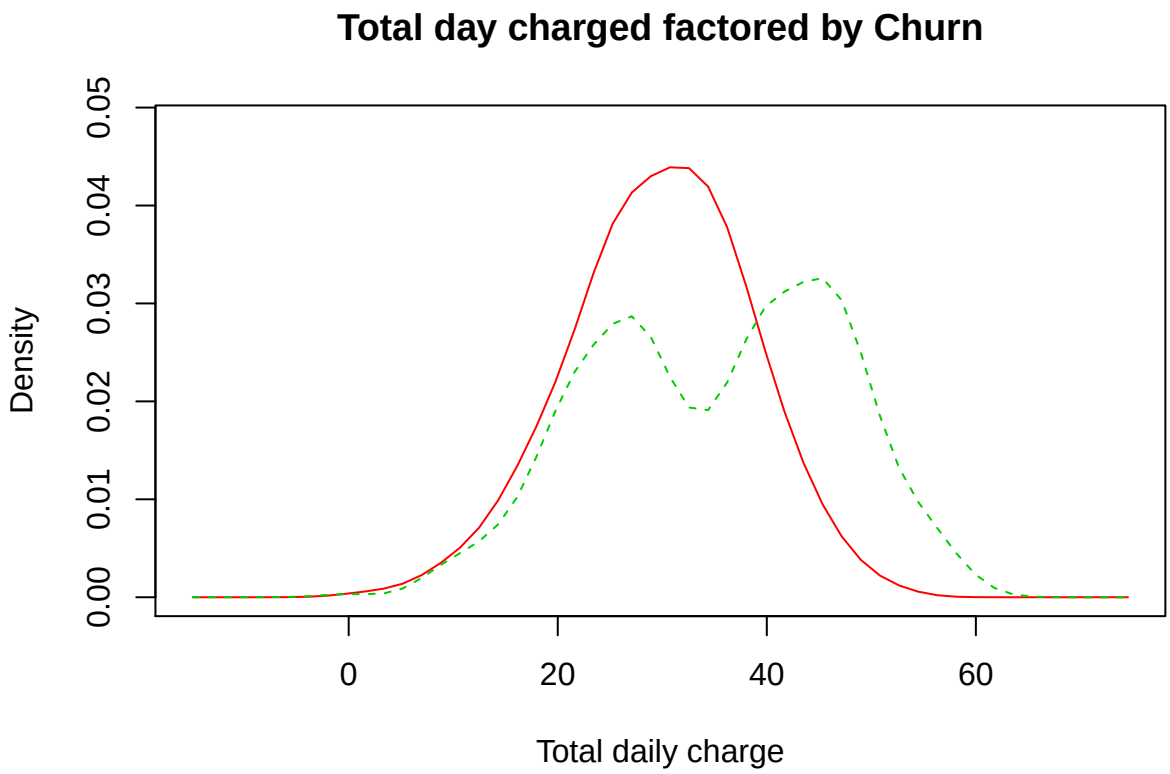


Fig-4.1: Boxplot for total.intl.charge and churn

```
ggplot(df2, aes(factor(churn),total.intl.charge))+geom_boxplot(aes(colour =
factor(churn)))+ggtitle('Measure of Total international charge as per
Churn')+coord_flip()
```

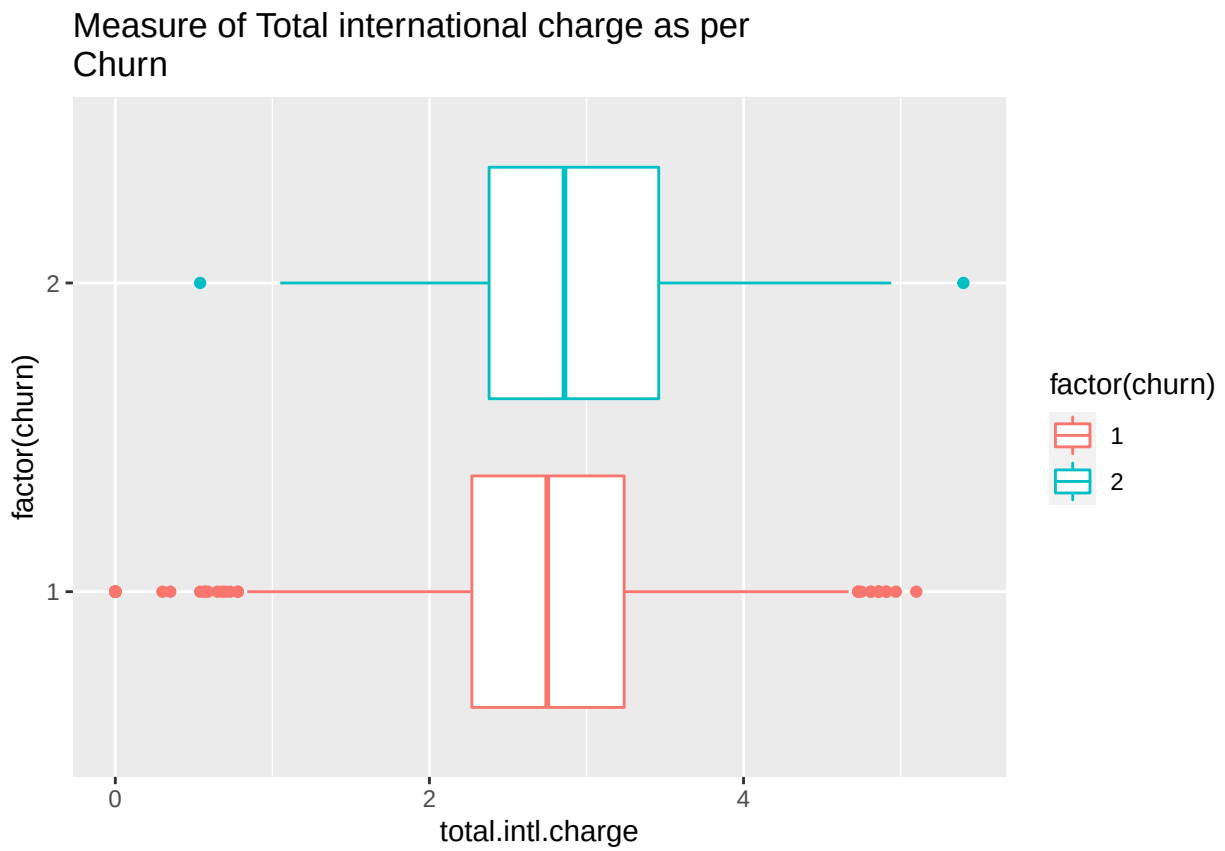
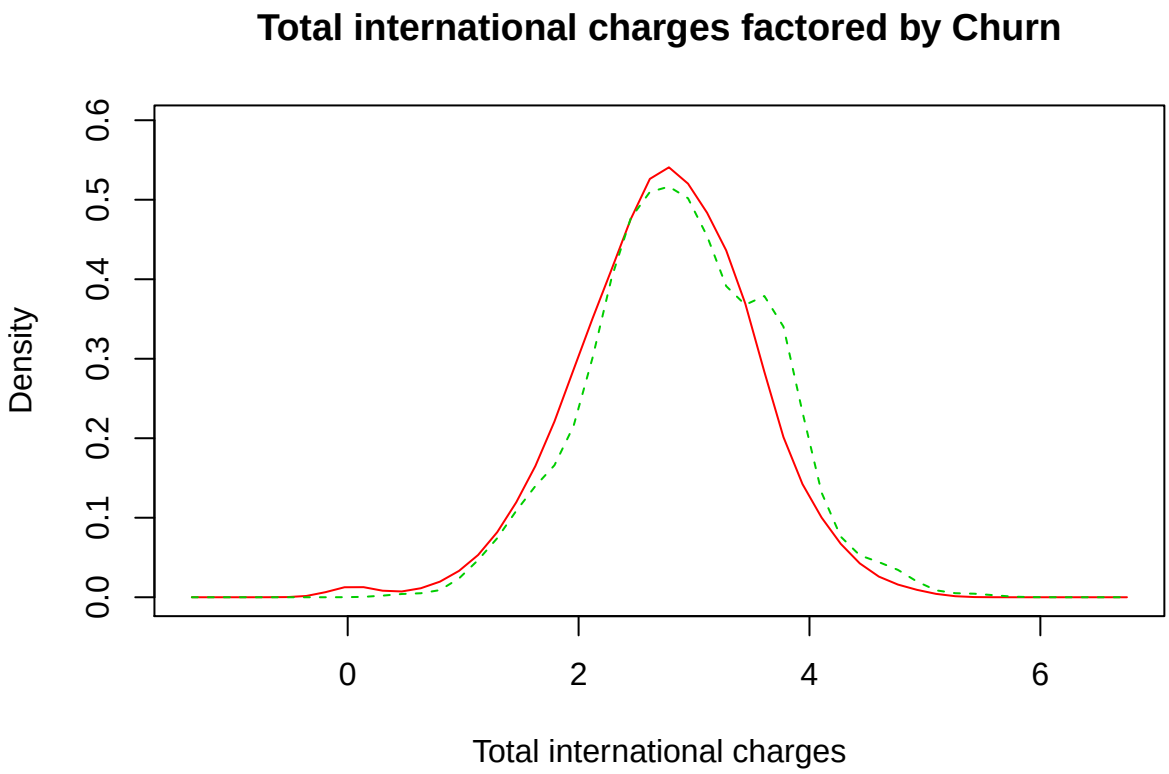


Fig-4.2 Comparative density plot for total.intl.charge and churn

```
sm.density.compare(df2$total.intl.charge, df2$churn, xlab ='Total international charges')
title(main = 'Total international charges factored by Churn')
```



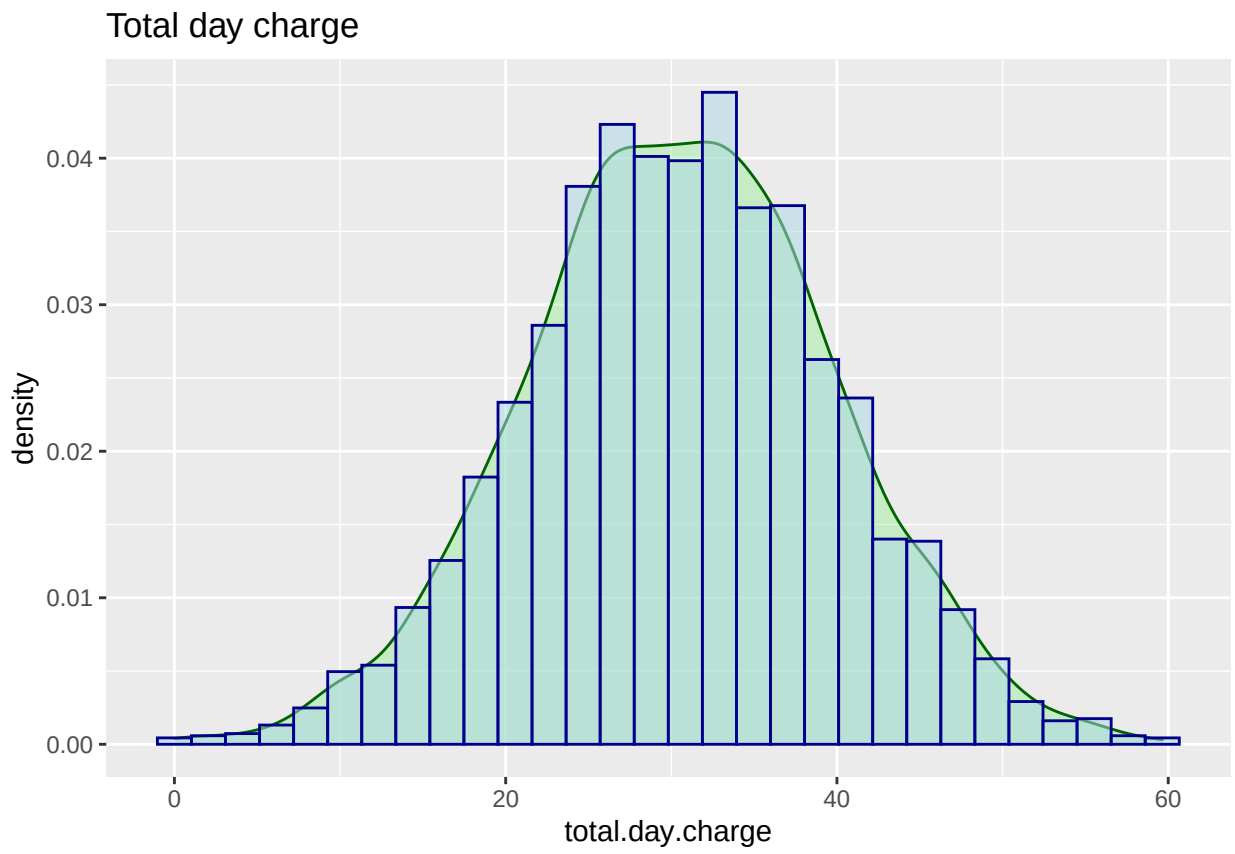
Comments:

1) The distribution for voicemail plan = 'No' is almost negligible whereas the distribution for voicemail plan = 'yes' is roughly symmetrically skewed.
2) The curve for churn = 'False' is symmetrically and normally distributed. Churn = 'True' shows a slight negative skew.
3) The curve for churn = 'False' is symmetrically and normally distributed. Churn = 'True' tends to show a positively skewed distribution.

Fig-5.1: Integrated histogram and density plot for Total day charge

```
ggplot(df2, aes(x=total.day.charge))+geom_density(alpha =0.4, fill ='lightgreen', colour = 'darkgreen')+
```

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




```
kurtosis(df2$total.day.charge)
```

```
## [1] -0.02158172
```

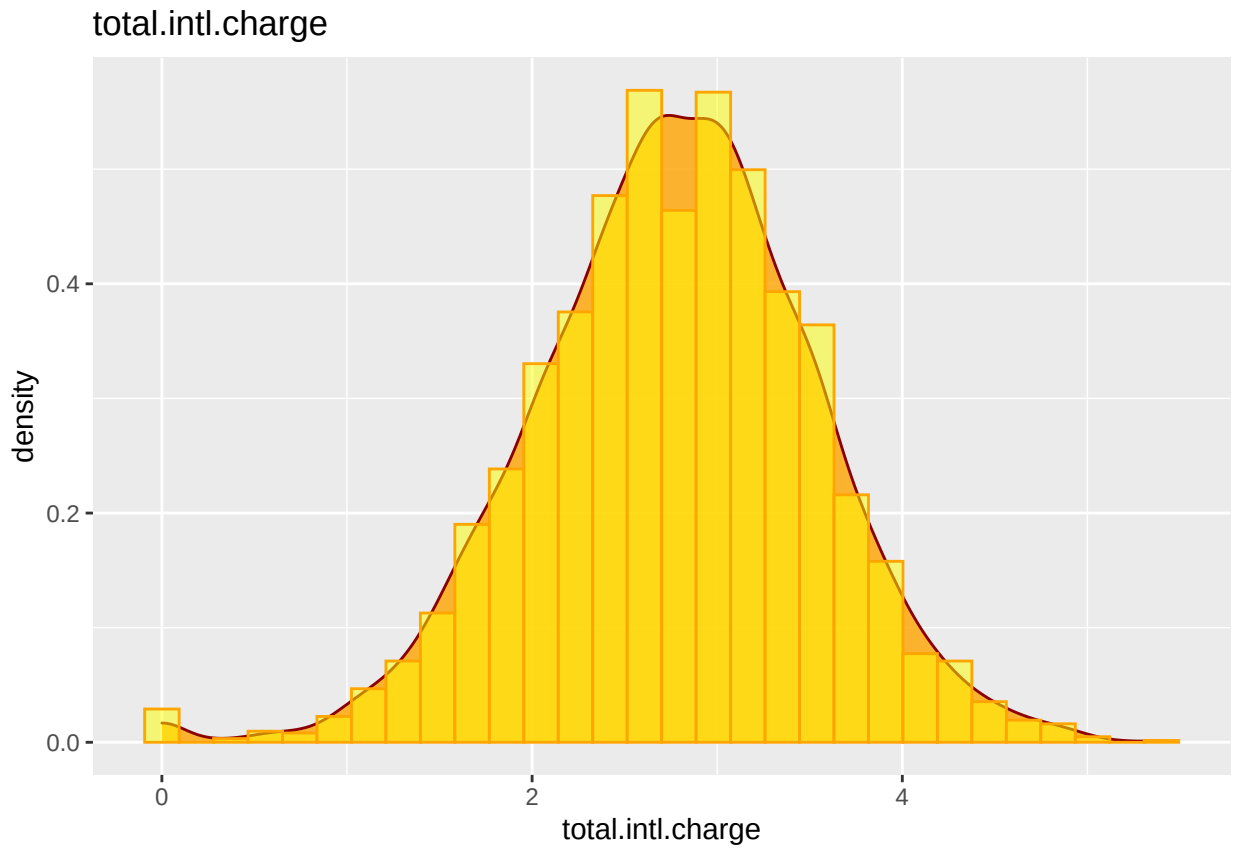
```
skewness(df2$total.day.charge)
```

```
## [1] -0.02907018
```

Fig-5.2: Integrated histogram and density plot for Total international charge

```
ggplot(df2, aes(x=total.intl.charge))+geom_density(alpha =0.8, fill ='orange', colour = 'darkred')+geom_
```

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




```
kurtosis(df2$total.intl.charge)
```

```
## [1] 0.6068967
```

```
skewness(df2$total.intl.charge)
```

```
## [1] -0.2451761
```

Comments:

The distribution for both variables - 'total.day.charge' and total.intl.charge' is negatively skewed. The curve for total day charge is platykurtic and that of total international charge is leptokurtic.