Unsupervised Learning with R

Brian Michira

9/3/2021

1.Problem Definition

Kira Plastinina is a Russian brand that is sold through a defunct chain of retail stores in Russia, Ukraine, Kazakhstan, Belarus, China, Philippines, and Armenia. The brands Sales and Marketing team would like to understand their customers behavior from data that they have collected over the past year. More specifically, they would like to learn the characteristics of customer groups.

2.Data Sourcing

Loading the Dataset

```
data=read.csv("http://bit.ly/EcommerceCustomersDataset")
head(data,n=10)
```

```
##    Administrative Administrative_Duration Informational Informational_Duration
## 1               0                        0             0                       0
## 2               0                        0             0                       0
## 3               0                       -1             0                      -1
## 4               0                        0             0                       0
## 5               0                        0             0                       0
## 6               0                        0             0                       0
## 7               0                       -1             0                      -1
## 8               1                       -1             0                      -1
## 9               0                        0             0                       0
## 10              0                        0             0                       0
##    ProductRelated ProductRelated_Duration BounceRates  ExitRates PageValues
## 1               1                0.000000  0.20000000 0.20000000          0
## 2               2               64.000000  0.00000000 0.10000000          0
## 3               1               -1.000000  0.20000000 0.20000000          0
## 4               2                2.666667  0.05000000 0.14000000          0
## 5              10              627.500000  0.02000000 0.05000000          0
## 6              19              154.216667  0.01578947 0.02456140          0
## 7               1               -1.000000  0.20000000 0.20000000          0
## 8               1               -1.000000  0.20000000 0.20000000          0
## 9               2               37.000000  0.00000000 0.10000000          0
## 10              3              738.000000  0.00000000 0.02222222          0
##    SpecialDay Month OperatingSystems Browser Region TrafficType
```

```
## 1         0.0   Feb                1       1      1           1
## 2         0.0   Feb                2       2      1           2
## 3         0.0   Feb                4       1      9           3
## 4         0.0   Feb                3       2      2           4
## 5         0.0   Feb                3       3      1           4
## 6         0.0   Feb                2       2      1           3
## 7         0.4   Feb                2       4      3           3
## 8         0.0   Feb                1       2      1           5
## 9         0.8   Feb                2       2      2           3
## 10        0.4   Feb                2       4      1           2
##           VisitorType Weekend Revenue
## 1  Returning_Visitor   FALSE   FALSE
## 2  Returning_Visitor   FALSE   FALSE
## 3  Returning_Visitor   FALSE   FALSE
## 4  Returning_Visitor   FALSE   FALSE
## 5  Returning_Visitor    TRUE   FALSE
## 6  Returning_Visitor   FALSE   FALSE
## 7  Returning_Visitor   FALSE   FALSE
## 8  Returning_Visitor    TRUE   FALSE
## 9  Returning_Visitor   FALSE   FALSE
## 10 Returning_Visitor   FALSE   FALSE
```

3.Cheking the Data

```
# view the bottom of our dataset
tail(data)
```

```
##       Administrative Administrative_Duration Informational
## 12325              0                      0             1
## 12326              3                    145             0
## 12327              0                      0             0
## 12328              0                      0             0
## 12329              4                     75             0
## 12330              0                      0             0
##       Informational_Duration ProductRelated ProductRelated_Duration BounceRates
## 12325                      0             16                503.000 0.000000000
## 12326                      0             53               1783.792 0.007142857
## 12327                      0              5                465.750 0.000000000
## 12328                      0              6                184.250 0.083333333
## 12329                      0             15                346.000 0.000000000
## 12330                      0              3                 21.250 0.000000000
##        ExitRates PageValues SpecialDay Month OperatingSystems Browser Region
## 12325 0.03764706    0.00000          0   Nov                2       2      1
## 12326 0.02903061   12.24172          0   Dec                4       6      1
## 12327 0.02133333    0.00000          0   Nov                3       2      1
## 12328 0.08666667    0.00000          0   Nov                3       2      1
## 12329 0.02105263    0.00000          0   Nov                2       2      3
## 12330 0.06666667    0.00000          0   Nov                3       2      1
##       TrafficType       VisitorType Weekend Revenue
## 12325           1 Returning_Visitor   FALSE   FALSE
## 12326           1 Returning_Visitor    TRUE   FALSE
```

```
## 12327           8 Returning_Visitor    TRUE   FALSE
## 12328          13 Returning_Visitor    TRUE   FALSE
## 12329          11 Returning_Visitor   FALSE   FALSE
## 12330           2       New_Visitor    TRUE   FALSE
```

```
# view the dimensions of our dataset
dim(data)
```

```
## [1] 12330    18
```

Our dataset has 18 columns and 12330 rows.

```
# view the structure of our dataset
str(data)
```

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 1 2 3 ...
##  $ ProductRelated_Duration: num  0 64 -1 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : chr  "Feb" "Feb" "Feb" "Feb" ...
##  $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : chr  "Returning_Visitor" "Returning_Visitor" "Returning_Visitor" "Returni
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

4.Data Cleaninig

```
#viewing the number of missing values
sum(is.na(data))
```

```
## [1] 112
```

The sum of missing values is 112.We went ahead and checked for the missing value in each column.

```
# check for total null values in each column
colSums(is.na(data))
```

```
##          Administrative Administrative_Duration           Informational
##                      14                      14                      14
```

```
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                      14                      14                      14
##             BounceRates               ExitRates             PageValues
##                      14                      14                       0
##              SpecialDay                   Month        OperatingSystems
##                       0                       0                       0
##                 Browser                  Region             TrafficType
##                       0                       0                       0
##             VisitorType                 Weekend                 Revenue
##                       0                       0                       0
```

We noticed there are 14 missing values in most of the columns and we decided to drop and observe if the are originating from the same rows.

```
# dropping the rows with the missing values
df<-na.omit(data)
head(df)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                      -1             0                     -1
## 4              0                       0             0                      0
## 5              0                       0             0                      0
## 6              0                       0             0                      0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000  0.20000000 0.2000000          0
## 2              2               64.000000  0.00000000 0.1000000          0
## 3              1               -1.000000  0.20000000 0.2000000          0
## 4              2                2.666667  0.05000000 0.1400000          0
## 5             10              627.500000  0.02000000 0.0500000          0
## 6             19              154.216667  0.01578947 0.0245614          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType
## 1          0   Feb                1       1      1           1
## 2          0   Feb                2       2      1           2
## 3          0   Feb                4       1      9           3
## 4          0   Feb                3       2      2           4
## 5          0   Feb                3       3      1           4
## 6          0   Feb                2       2      1           3
##         VisitorType Weekend Revenue
## 1 Returning_Visitor   FALSE   FALSE
## 2 Returning_Visitor   FALSE   FALSE
## 3 Returning_Visitor   FALSE   FALSE
## 4 Returning_Visitor   FALSE   FALSE
## 5 Returning_Visitor    TRUE   FALSE
## 6 Returning_Visitor   FALSE   FALSE
```

It was safe to drop the missing values since it was a small percentage of the whole data.

```
#Checking the number of records
dim(df)
```

```
## [1] 12316    18
```

14 rows with majority of the missing values have been dropped.

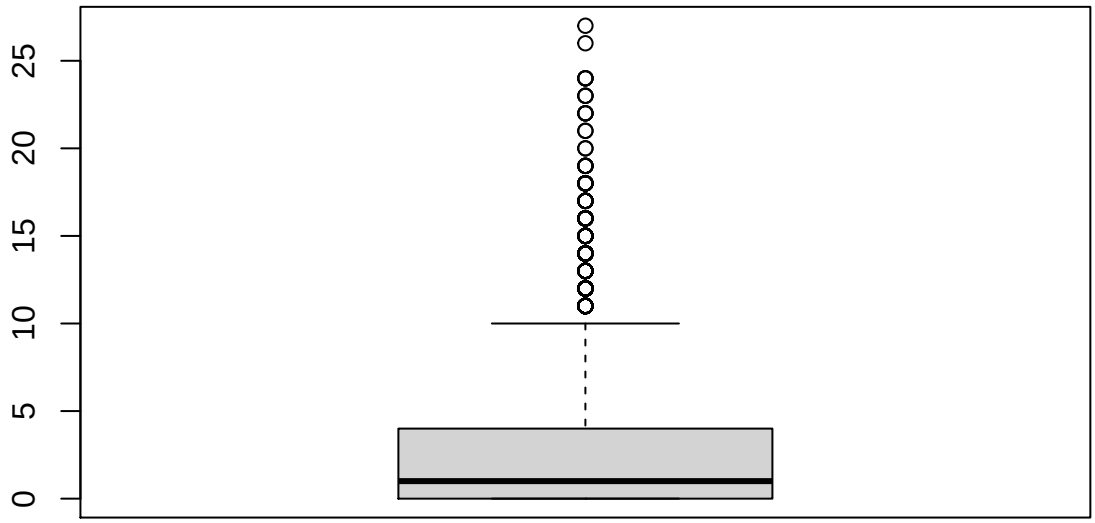
```
# find any duplicated rows in our dataset
duplicated_rows <- df[duplicated(df),]
```

117 rows are duplicated

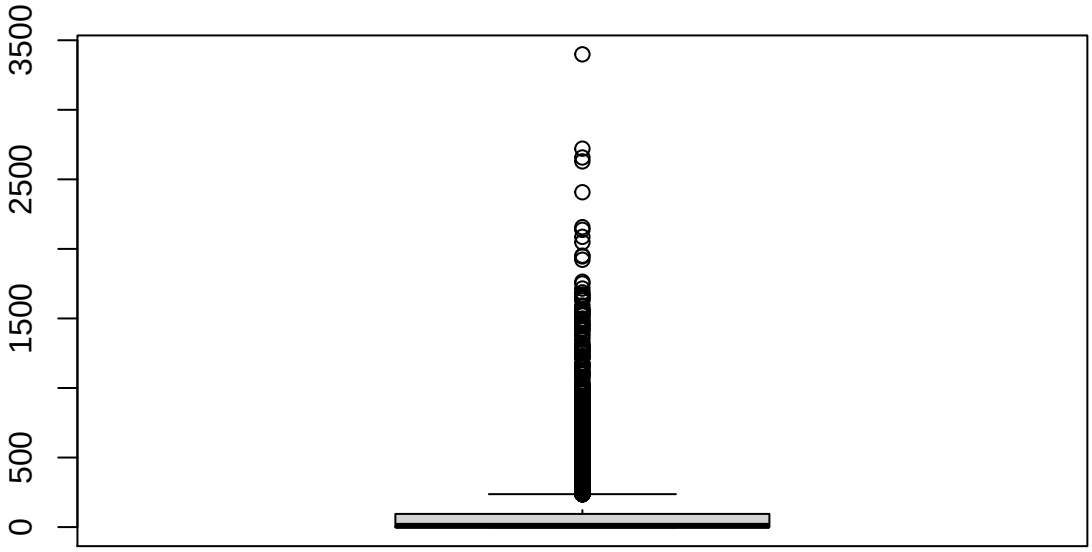
```
# removing the duplicated rows
df_new <- unique(df)
```

Checking for outliers

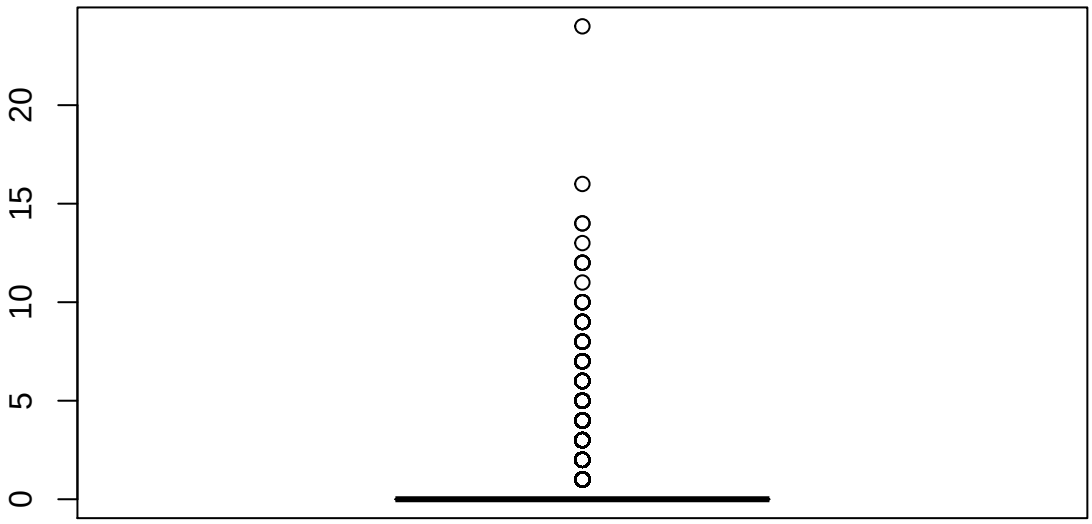
```
# checking for outliers on the Administrative column
boxplot(df_new$Administrative)
```



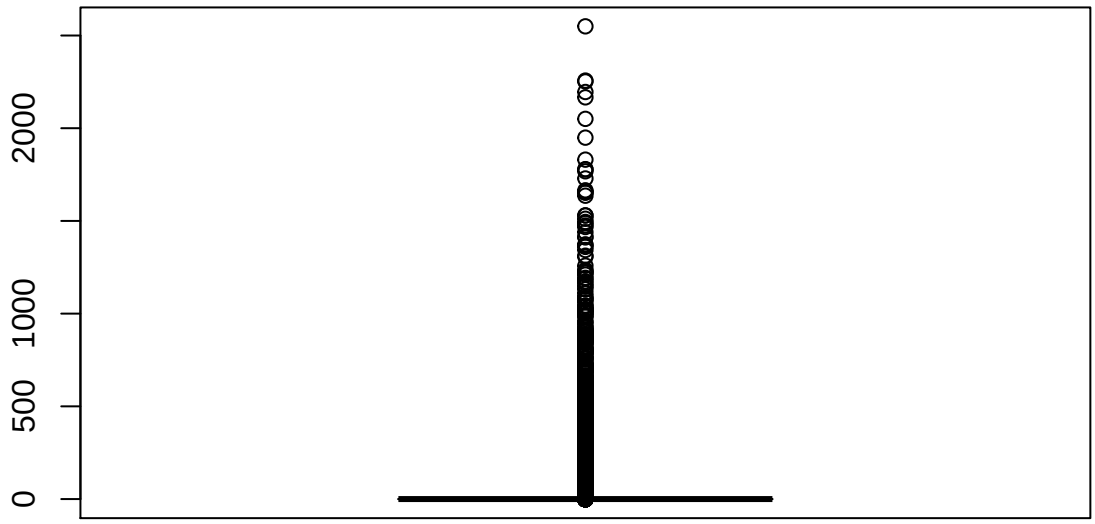

```
# checking for outliers in Administrative_Duration
boxplot(df_new$Administrative_Duration)
```



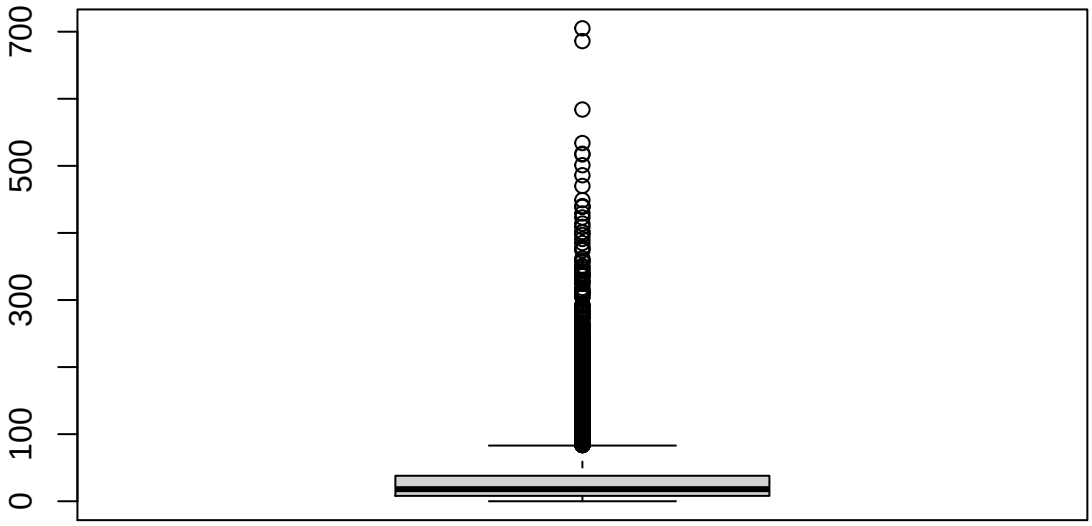

```
# check for outliers in Informational
boxplot(df_new$Informational)
```



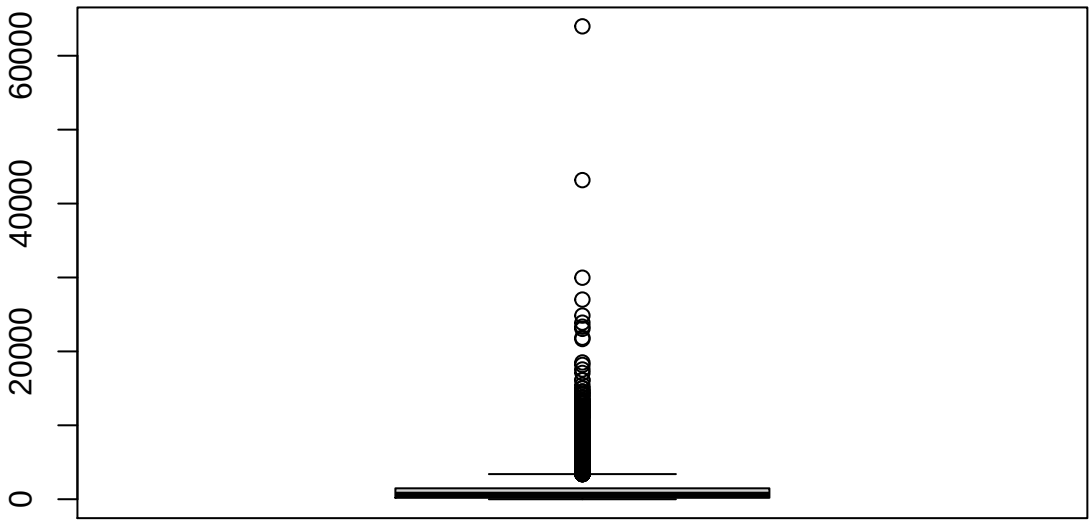

```
# check for outliers in Informational_Duration
boxplot(df_new$Informational_Duration)
```



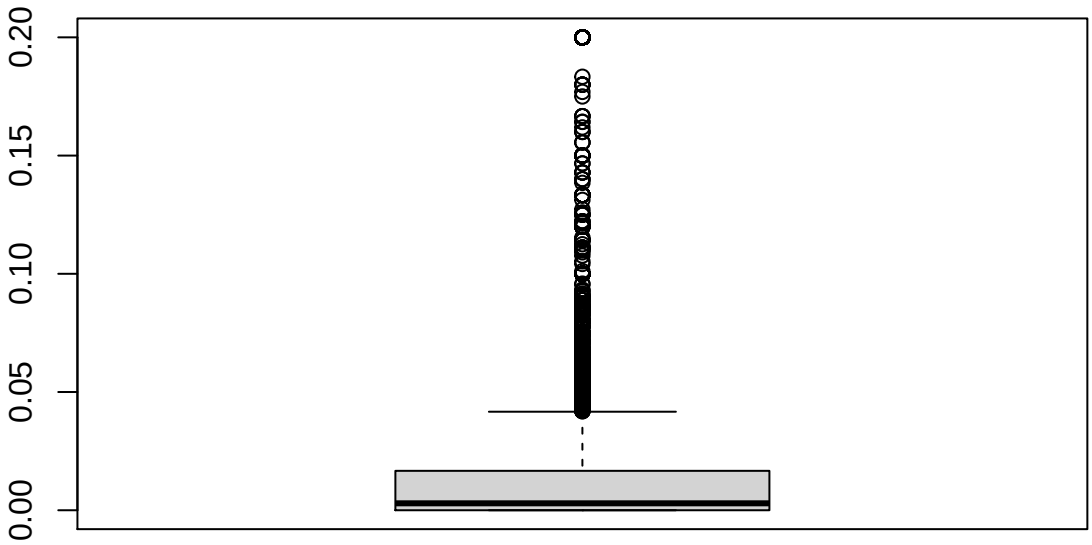

```
# check for outliers in ProductRelated
boxplot(df_new$ProductRelated)
```



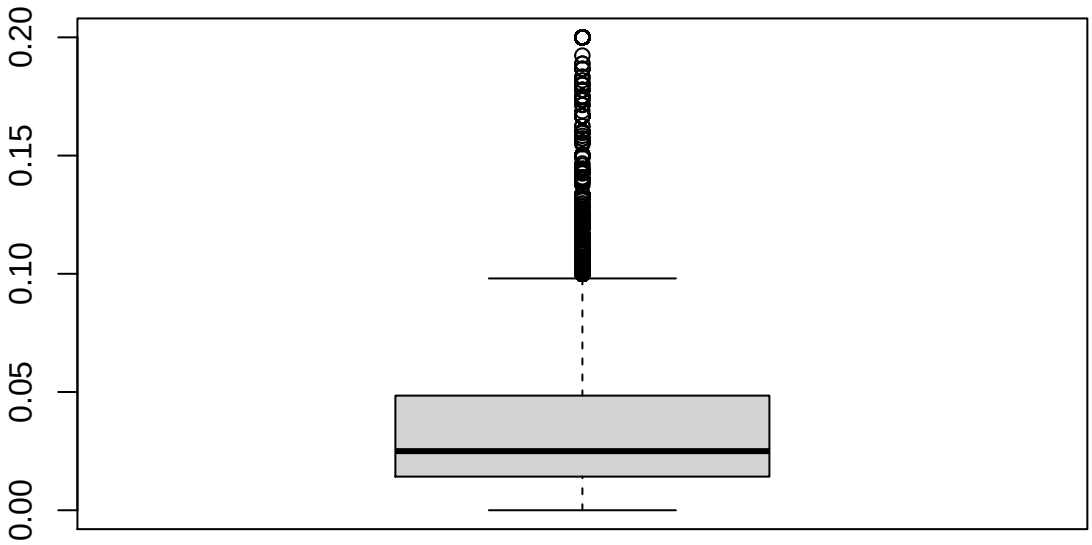

```
# check for outliers in ProductRelated_Duration
boxplot(df_new$ProductRelated_Duration)
```



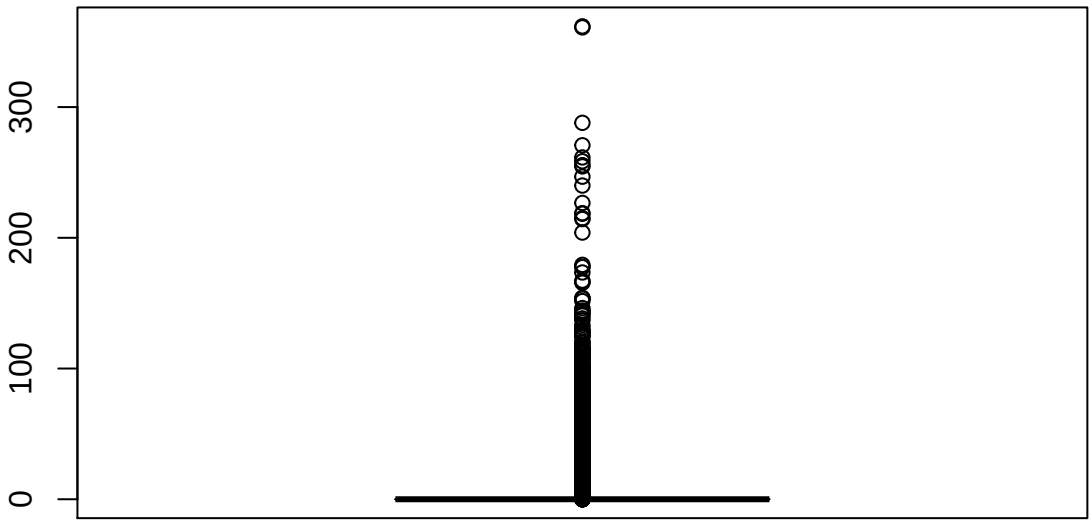

```
# check for outliers in BounceRates
boxplot(df_new$BounceRates)
```



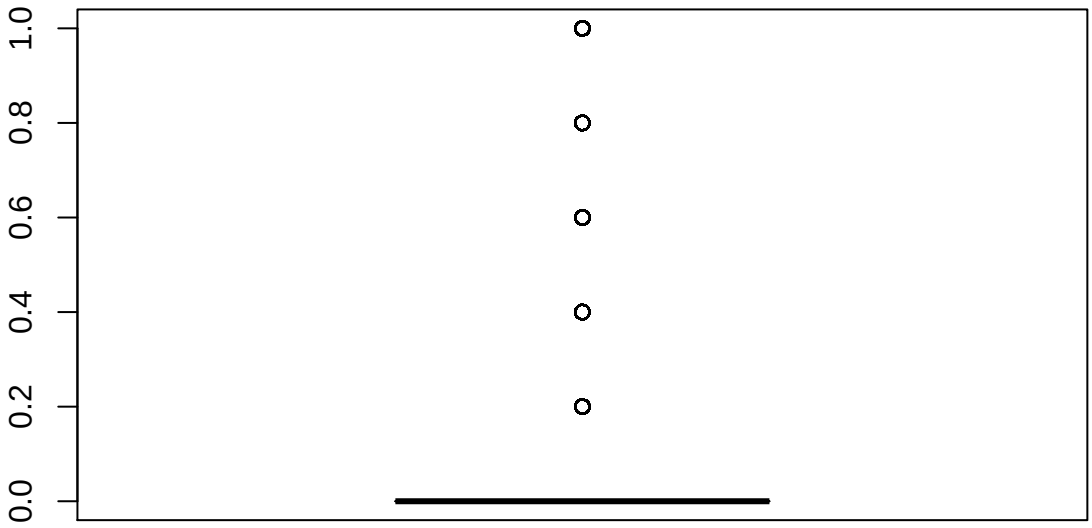

```
# check for outliers in ExitRates
boxplot(df_new$ExitRates)
```



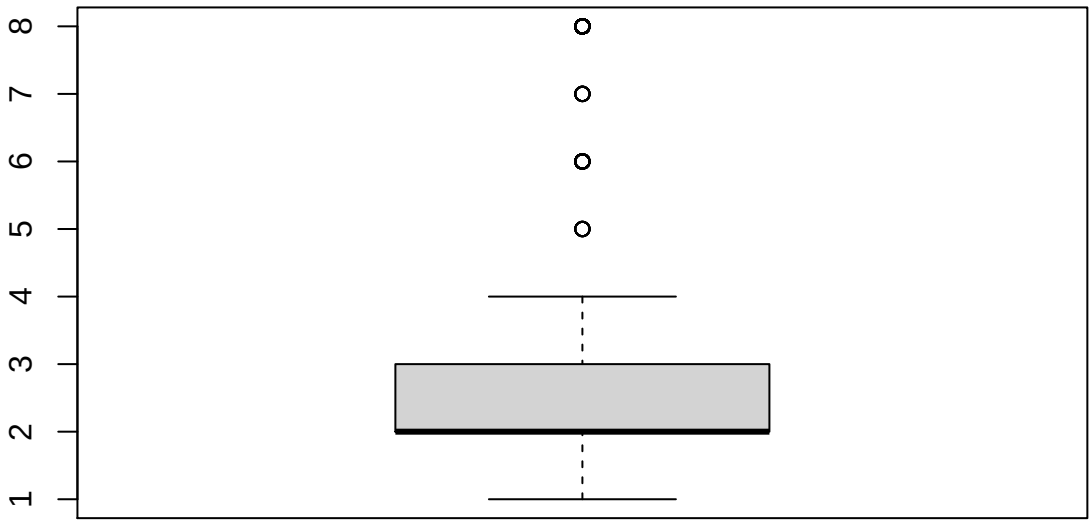

```
# check for outliers in PageValues
boxplot(df_new$PageValues)
```



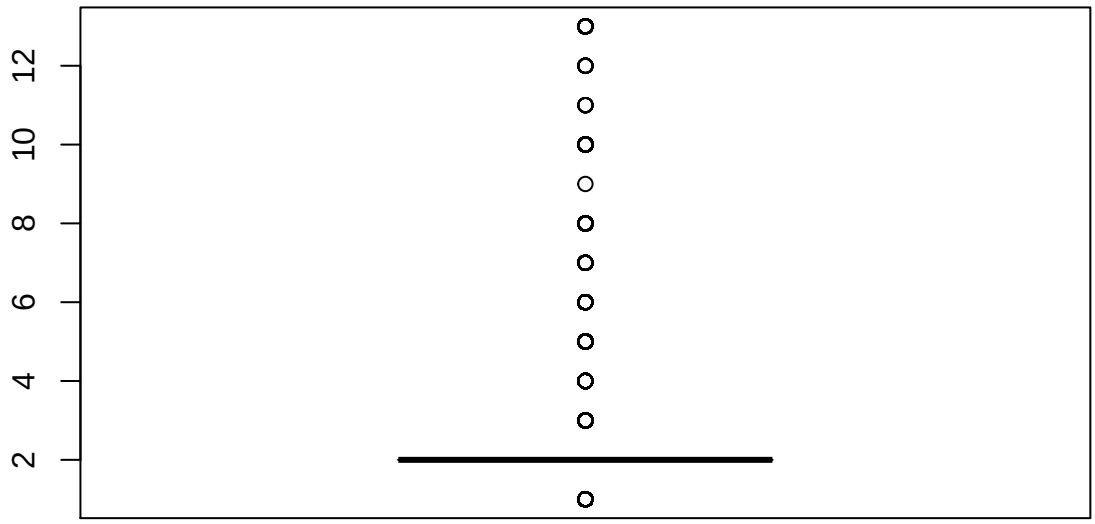

```
# check for outliers in SpecialDay
boxplot(df_new$SpecialDay)
```



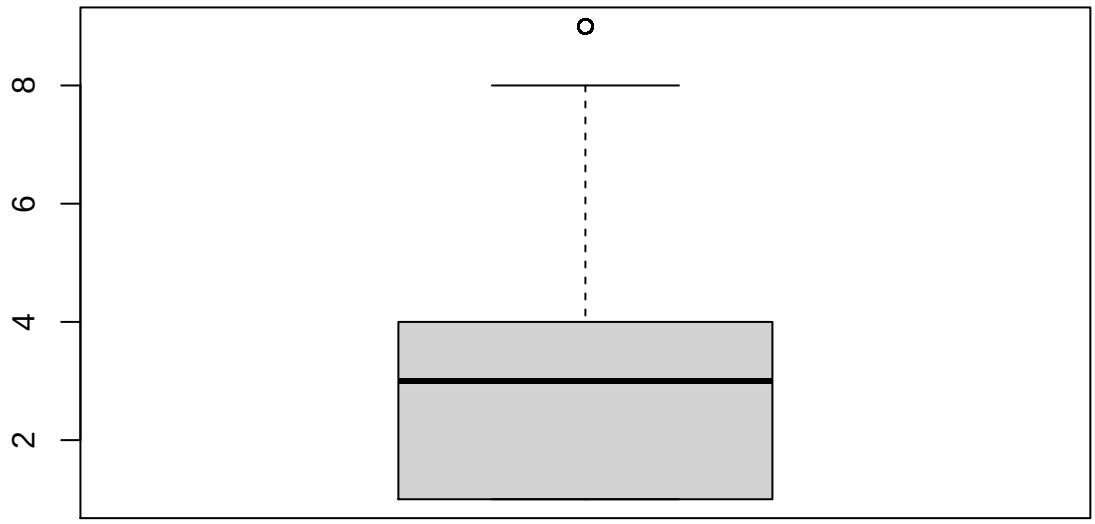

```
# check for outliers in OperatingSystems
boxplot(df_new$OperatingSystems)
```



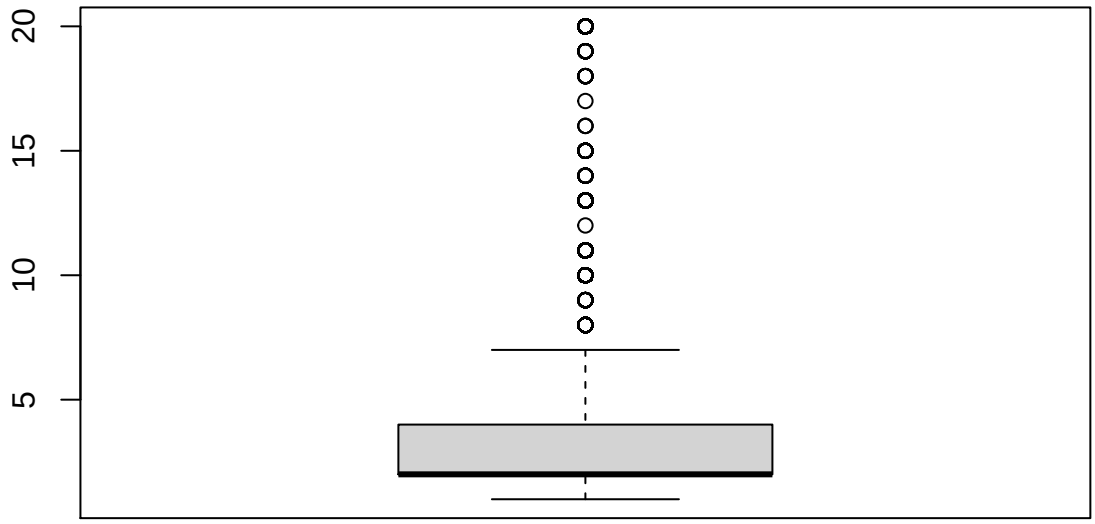

```
# check for outliers in Browser
boxplot(df_new$Browser)
```




```
# check for outliers in Region
boxplot(df_new$Region)
```




```
# check for outliers in TrafficType
boxplot(df_new$TrafficType)
```



5.Exploratory Data Analysis

Univariate Analysis

1.Measures of Central Tendency & Measures of Dispersion

```
# mean, median, Min, Max
summary(df_new)
```

```
##  Administrative  Administrative_Duration Informational
##  Min.   : 0.00   Min.   :  -1.00         Min.   : 0.0000
##  1st Qu.: 0.00   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.00   Median :   9.00         Median : 0.0000
##  Mean   : 2.34   Mean   :  81.68         Mean   : 0.5088
##  3rd Qu.: 4.00   3rd Qu.:  94.75         3rd Qu.: 0.0000
##  Max.   :27.00   Max.   :3398.75         Max.   :24.0000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :  -1.00        Min.   :  0.00   Min.   :   -1.0
##  1st Qu.:   0.00        1st Qu.:  8.00   1st Qu.:  193.6
##  Median :   0.00        Median : 18.00   Median :  609.5
##  Mean   :  34.84        Mean   : 32.06   Mean   : 1207.5
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1477.6
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
```

```
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.00000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.01422   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.00293   Median :0.02500   Median :  0.000   Median :0.00000
##  Mean   :0.02045   Mean   :0.04150   Mean   :  5.952   Mean   :0.06197
##  3rd Qu.:0.01667   3rd Qu.:0.04848   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.20000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##     Month           OperatingSystems    Browser           Region
##  Length:12199       Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mode  :character   Median :2.000    Median : 2.000   Median :3.000
##                     Mean   :2.124    Mean   : 2.358   Mean   :3.153
##                     3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##                     Max.   :8.000    Max.   :13.000   Max.   :9.000
##   TrafficType     VisitorType         Weekend         Revenue
##  Min.   : 1.000   Length:12199       Mode :logical   Mode :logical
##  1st Qu.: 2.000   Class :character   FALSE:9343      FALSE:10291
##  Median : 2.000   Mode  :character   TRUE :2856      TRUE :1908
##  Mean   : 4.075
##  3rd Qu.: 4.000
##  Max.   :20.000
```

```
library(moments)
# mode function
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

```
#mean
mean(df_new$Administrative)
```

```
## [1] 2.340028
```

```
#meadian
median(df_new$Administrative)
```

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

```
#mode
getmode(df_new$Administrative)
```

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

```
#variance
var(df_new$Administrative)
```

```
## [1] 11.09457
```

```
#standard deviation
sd(df_new$Administrative)
```

```
## [1] 3.330851
```

```
#variance
var(df_new$Administrative)
```

```
## [1] 11.09457
```

```
#skewness
skewness(df_new$Administrative)
```

```
## [1] 1.946248
```

Administrative was positively skewed/right skewed showing most of the values were greater than the mean.

```
#kurtosis
kurtosis(df_new$Administrative)
```

```
## [1] 7.636106
```

Administrative had a positive kurtosis.Showing the presence of outliers.

Administrative Duration

```
#mean
mean(df_new$Administrative_Duration)
```

```
## [1] 81.68214
```

```
#median
median(df_new$Administrative_Duration)
```

```
## [1] 9
```

```
#mode
getmode(df_new$Administrative_Duration)
```

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

```
#variance
var(df_new$Administrative_Duration)
```

```
## [1] 31516.25
```

```
#standard deviation
sd(df_new$Administrative_Duration)
```

```
## [1] 177.5282
```

```
#variance
var(df_new$Administrative_Duration)
```

```
## [1] 31516.25
```

```
#skewness
skewness(df_new$Administrative_Duration)
```

```
## [1] 5.59021
```

```
#kurtosis
kurtosis(df_new$Administrative_Duration)
```

```
## [1] 53.09389
```

Informational

```
#mean
mean(df_new$Informational)
```

```
## [1] 0.5088122
```

```
#median
median(df_new$Informational)
```

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

```
#mode
getmode(df_new$Informational)
```

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

```
#variance
var(df_new$Informational)
```

```
## [1] 1.62771
```

```
#standard deviation
sd(df_new$Informational)
```

```
## [1] 1.275817
```

```
#variance
var(df_new$Informational)
```

```
## [1] 1.62771
```

```
#skewness
skewness(df_new$Informational)
```

```
## [1] 4.013451
```

```
#kurtosis
kurtosis(df_new$Informational)
```

```
## [1] 29.64254
```

Informational Duration

```
#mean
mean(df_new$Informational_Duration)
```

```
## [1] 34.83734
```

```
#median
median(df_new$Informational_Duration)
```

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

```
#mode
getmode(df_new$Informational_Duration)
```

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

```
#variance
var(df_new$Informational_Duration)
```

```
## [1] 20010.51
```

```
#standard deviation
sd(df_new$Informational_Duration)
```

```
## [1] 141.4585
```

```
#variance
var(df_new$Informational_Duration)
```

```
## [1] 20010.51
```

```
#skewness
skewness(df_new$Informational_Duration)
```

```
## [1] 7.537435
```

```
#kurtosis
kurtosis(df_new$Informational_Duration)
```

```
## [1] 78.46409
```

Product Related

```
#mean
mean(df_new$ProductRelated)
```

```
## [1] 32.05845
```

```
#median
median(df_new$ProductRelated)
```

```
## [1] 18
```

```
#mode
getmode(df_new$ProductRelated)
```

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

```
#variance
var(df_new$ProductRelated)
```

```
## [1] 1989.241
```

```
#standard deviation
sd(df_new$ProductRelated)
```

```
## [1] 44.60091
```

```
#variance
var(df_new$ProductRelated)
```

```
## [1] 1989.241
```

```
#skewness
skewness(df_new$ProductRelated)
```

```
## [1] 4.332134
```

```
#kurtosis
kurtosis(df_new$ProductRelated)
```

```
## [1] 34.04903
```

Product Related Duration

```
#mean
mean(df_new$ProductRelated_Duration)
```

```
## [1] 1207.508
```

```
#median
median(df_new$ProductRelated_Duration)
```

```
## [1] 609.5417
```

```
#mode
getmode(df_new$ProductRelated_Duration)
```

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

```
#variance
var(df_new$ProductRelated_Duration)
```

```
## [1] 3686121
```

```
#standard deviation
sd(df_new$ProductRelated_Duration)
```

```
## [1] 1919.927
```

```
#variance
var(df_new$ProductRelated_Duration)
```

```
## [1] 3686121
```

```
#skewness
#skewness(df_new$ProductRelated_Duration)
```

```
#kurtosis
#kurtosis(df_new$ProductRelated_Duration)
```

Bonus Rates

```
#mean
mean(df_new$BounceRates)
```

```
## [1] 0.02044674
```

```
#median
median(df_new$BounceRates)
```

```
## [1] 0.002930403
```

```
#mode
getmode(df_new$BounceRates)
```

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

```
#variance
var(df_new$BounceRates)
```

```
## [1] 0.002061387
```

```
#standard deviation
sd(df_new$BounceRates)
```

```
## [1] 0.0454025
```

```
#variance
var(df_new$BounceRates)
```

```
## [1] 0.002061387
```

```
#skewness
#skewness(df_new$BounceRates)
```

```
#kurtosis
#kurtosis(df_new$BounceRates)
```

Exit Rates

```
#mean
mean(df_new$ExitRates)
```

```
## [1] 0.04149678
```

```
#median
median(df_new$ExitRates)
```

```
## [1] 0.025
```

```
#mode
getmode(df_new$ExitRates)
```

```
## [1] 0.2
```

```
#variance
var(df_new$ExitRates)
```

```
## [1] 0.0021388
```

```
#standard deviation
sd(df_new$ExitRates)
```

```
## [1] 0.04624716
```

```
#variance
var(df_new$ExitRates)
```

```
## [1] 0.0021388
```

```
#skewness
skewness(df_new$ExitRates)
```

```
## [1] 2.233125
```

```
#kurtosis
kurtosis(df_new$ExitRates)
```

```
## [1] 7.624252
```

Page Values

```
#mean
mean(df_new$PageValues)
```

```
## [1] 5.9525
```

```
#median
median(df_new$PageValues)
```

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

```
#mode
getmode(df_new$PageValues)
```

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

```
#variance
var(df_new$PageValues)
```

```
## [1] 348.1132
```

```
#standard deviation
sd(df_new$PageValues)
```

```
## [1] 18.65779
```

```
#variance
var(df_new$PageValues)
```

```
## [1] 348.1132
```

```
#skewness
#skewness(df_new$PageValues)
```

```
#kurtosis
kurtosis(df_new$PageValues)
```

```
## [1] 67.94031
```

Special Day

```
#mean
mean(df_new$SpecialDay)
```

```
## [1] 0.06197229
```

```
#meadian
median(df_new$SpecialDay)
```

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

```
#mode
getmode(df_new$SpecialDay)
```

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

```
#variance
var(df_new$SpecialDay)
```

```
## [1] 0.03988432
```

```
#standard deviation
sd(df_new$SpecialDay)
```

```
## [1] 0.1997106
```

```
#variance
var(df_new$SpecialDay)
```

```
## [1] 0.03988432
```

```
#skewness
skewness(df_new$SpecialDay)
```

```
## [1] 3.284481
```

```
#kurtosis
kurtosis(df_new$SpecialDay)
```

```
## [1] 12.78605
```

Operating Systems

```
#mean
mean(df_new$OperatingSystems)
```

```
## [1] 2.124354
```

```
#meadian
median(df_new$OperatingSystems)
```

```
## [1] 2
```

```
#mode
getmode(df_new$OperatingSystems)
```

```
## [1] 2
```

```
#variance
var(df_new$OperatingSystems)
```

```
## [1] 0.8226229
```

```
#standard deviation
sd(df_new$OperatingSystems)
```

```
## [1] 0.9069856
```

```
#variance
var(df_new$OperatingSystems)
```

```
## [1] 0.8226229
```

```
#skewness
skewness(df_new$OperatingSystems)
```

```
## [1] 2.031955
```

```
#kurtosis
kurtosis(df_new$OperatingSystems)
```

```
## [1] 13.26887
```

Browser

```
#mean
mean(df_new$Browser)
```

```
## [1] 2.358144
```

```
#median
median(df_new$Browser)
```

```
## [1] 2
```

```
#mode
getmode(df_new$Browser)
```

```
## [1] 2
```

```
#variance
var(df_new$Browser)
```

```
## [1] 2.926075
```

```
#standard deviation
sd(df_new$Browser)
```

```
## [1] 1.710578
```

```
#variance
var(df_new$Browser)
```

```
## [1] 2.926075
```

```
#skewness
skewness(df_new$Browser)
```

```
## [1] 3.215653
```

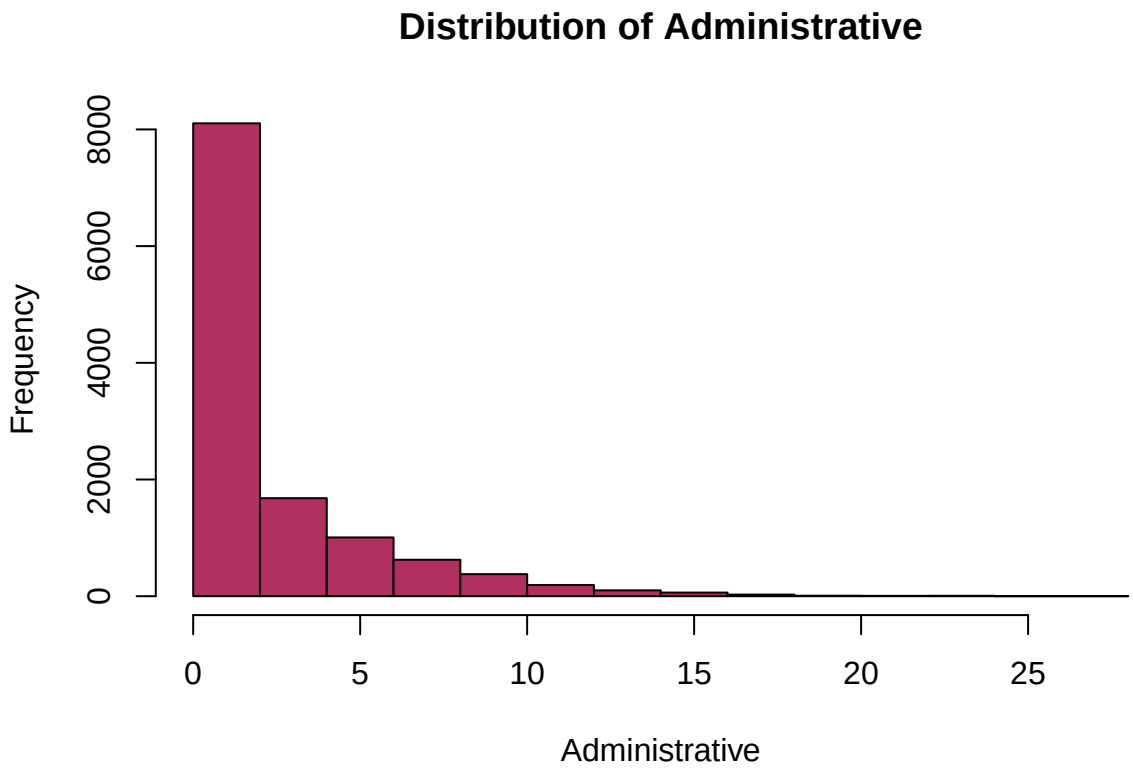
```
#kurtosis
kurtosis(df_new$Browser)
```

```
## [1] 15.53659
```

All the variables were positively skewed and had a positive kurtosis indicating the presence of outliers.

Univariate Graphical

```
hist(df_new$Administrative,main = "Distribution of Administrative",col="maroon",
     xlab="Administrative")
```

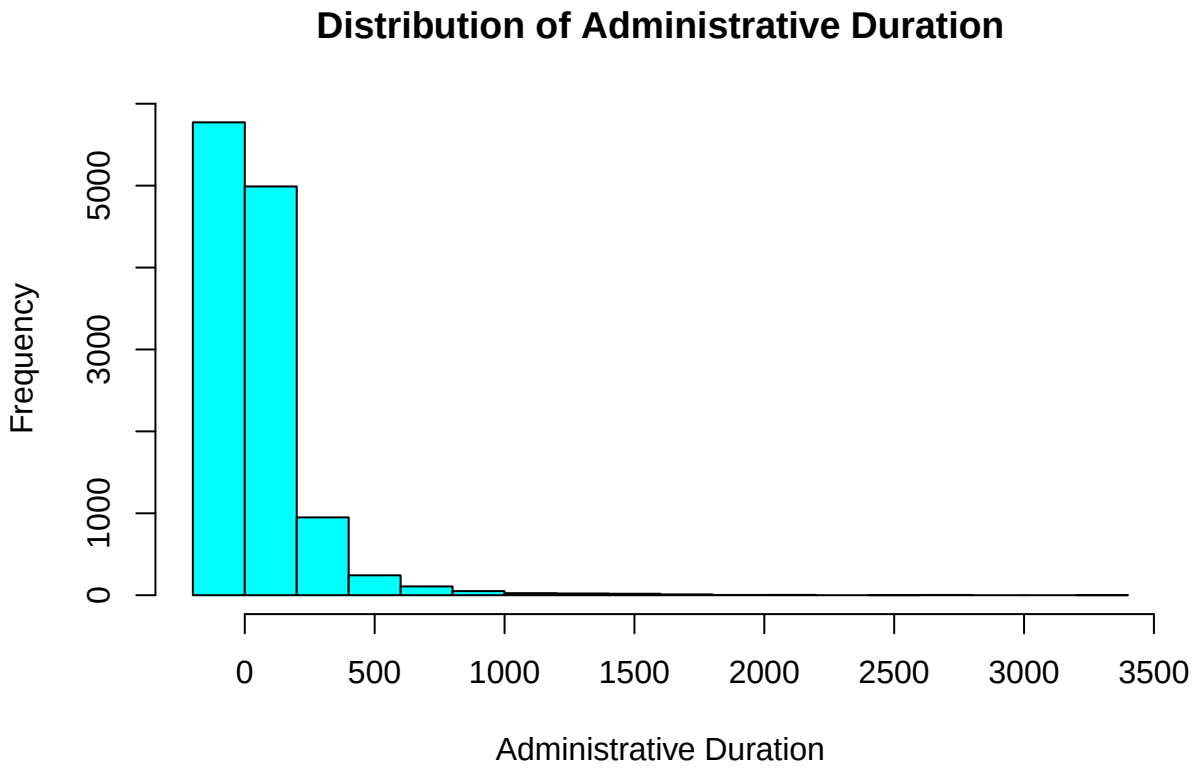


Histogram

Administrative is positively skewed.

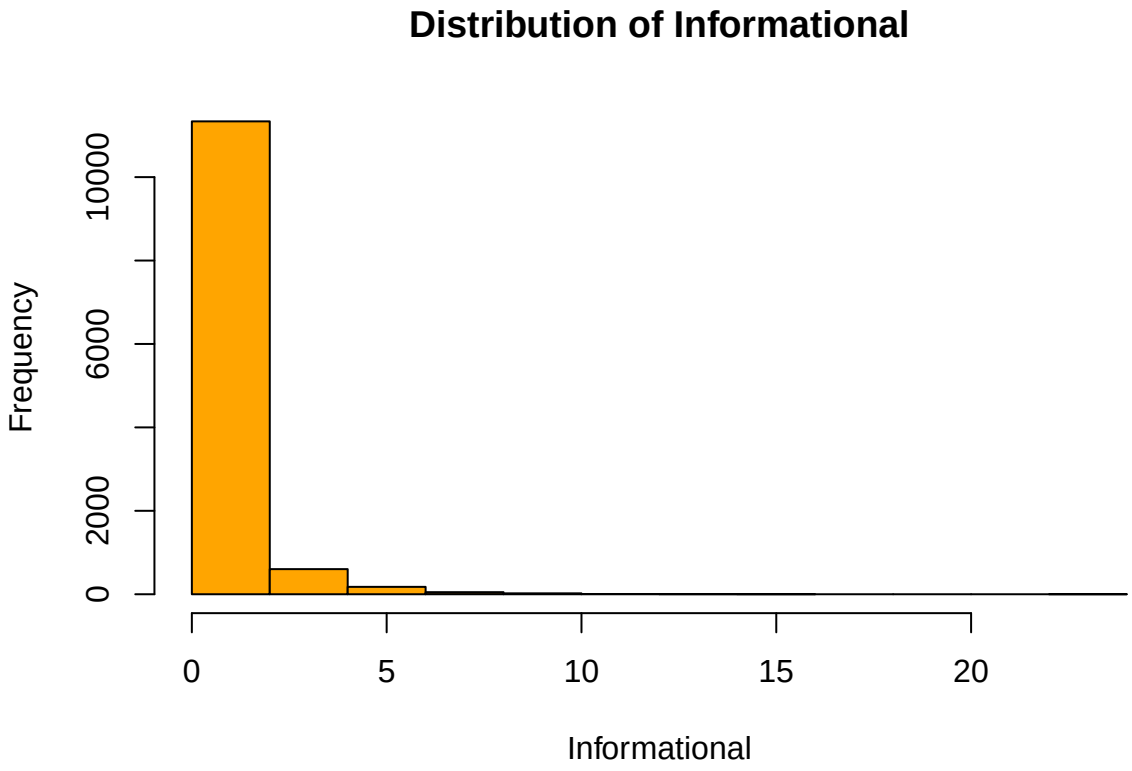
Most of the values were 0.

```
hist(df_new$Administrative_Duration,main = "Distribution of Administrative Duration",col="cyan",
     xlab="Administrative Duration")
```



Administrative Duration is positively skewed.

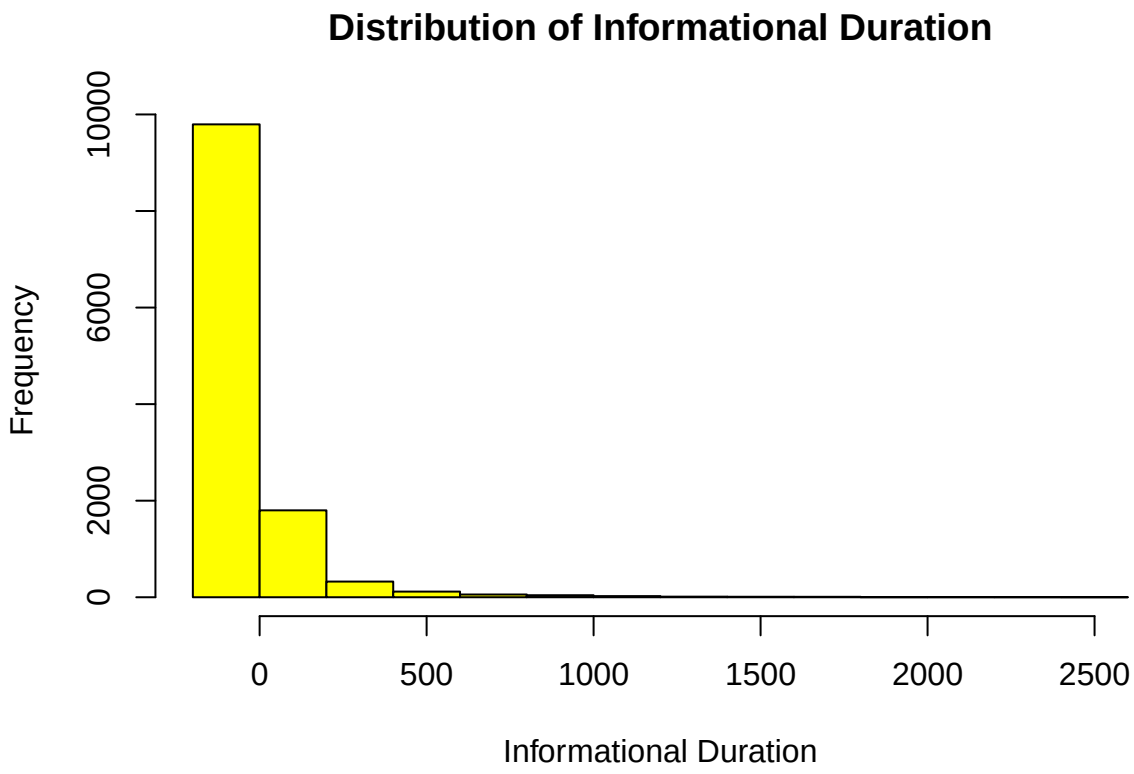
```
hist(df_new$Informational,main = "Distribution of Informational",col="orange",
     xlab="Informational")
```



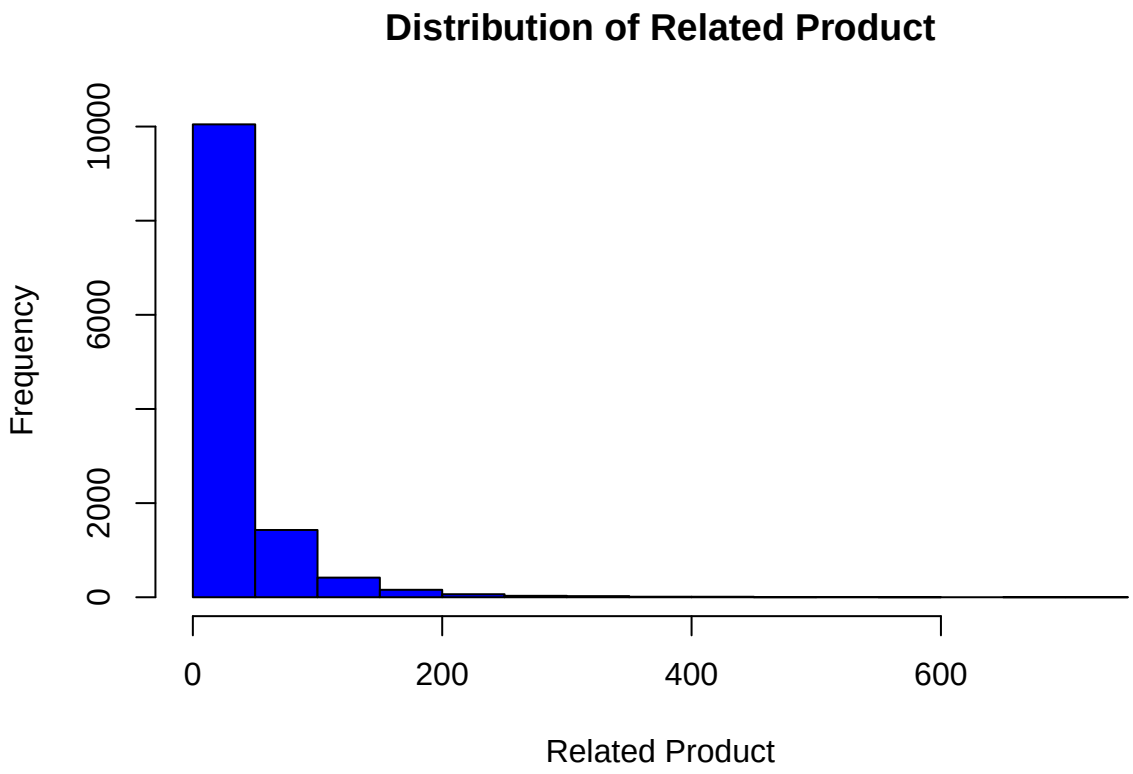
Informational is positively skewed.

Most of the values were 0.

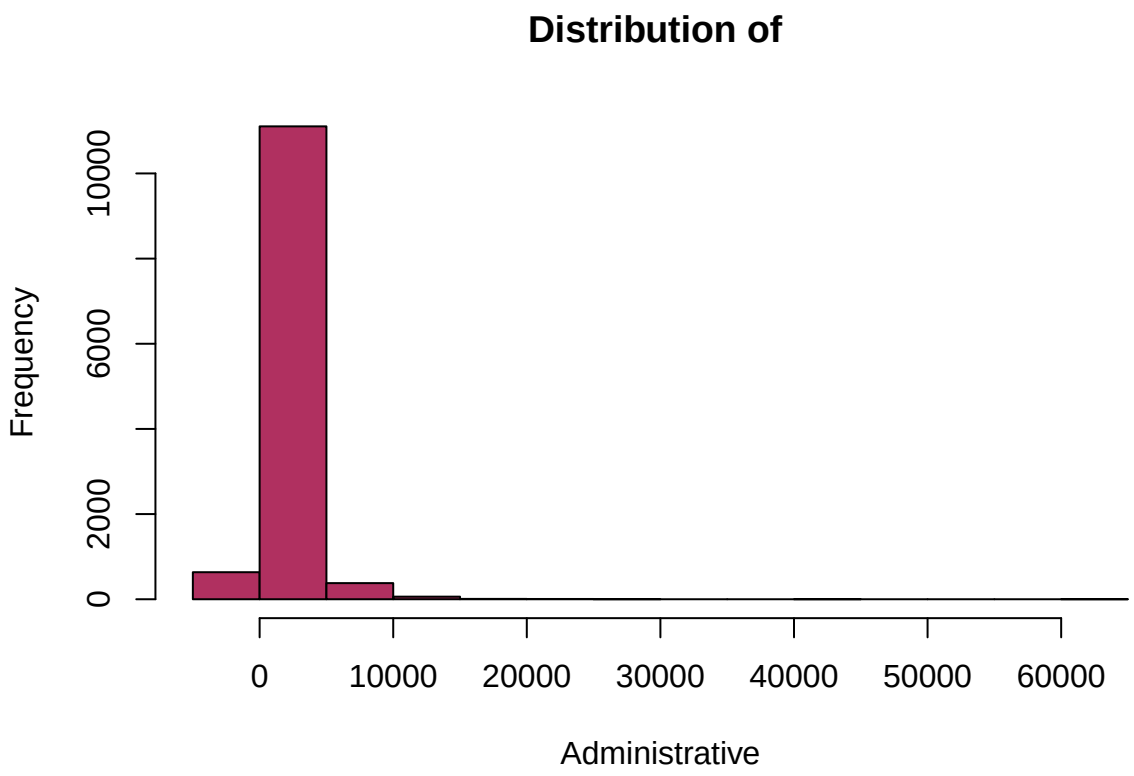
```
hist(df_new$Informational_Duration,main = "Distribution of Informational Duration",col="yellow",
     xlab="Informational Duration")
```



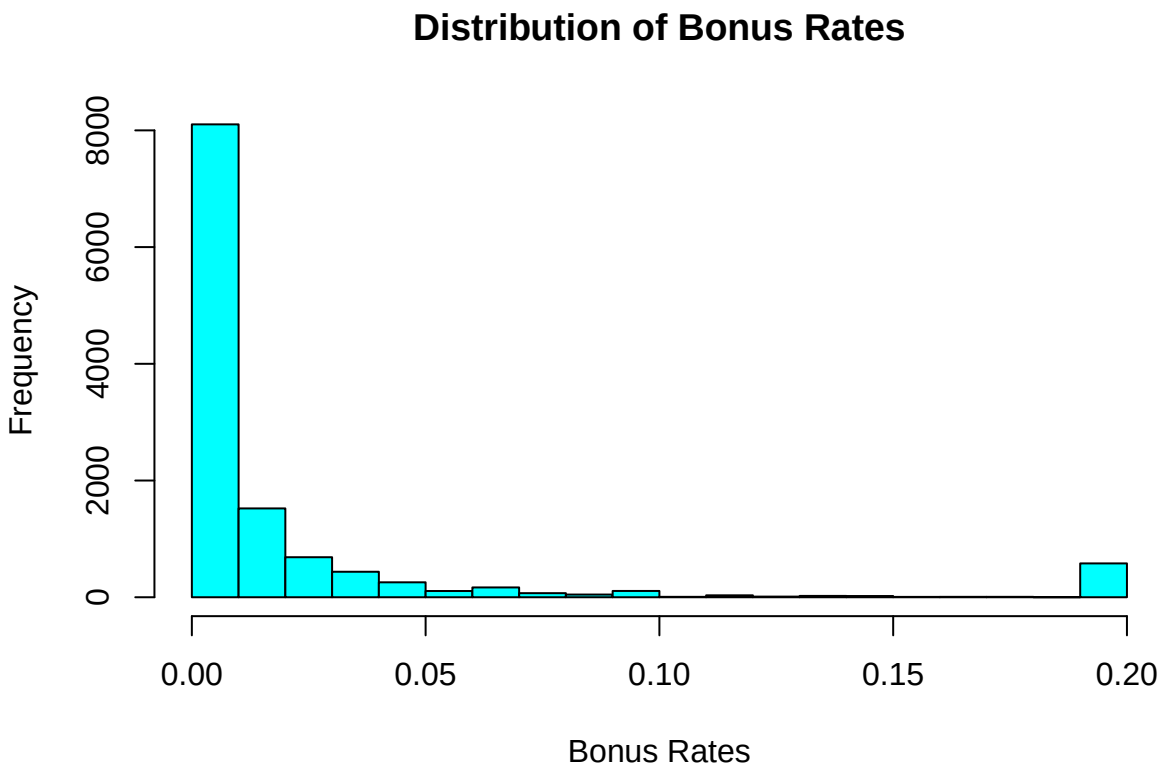

```
hist(df_new$ProductRelated,main = "Distribution of Related Product",col="blue",
     xlab="Related Product")
```



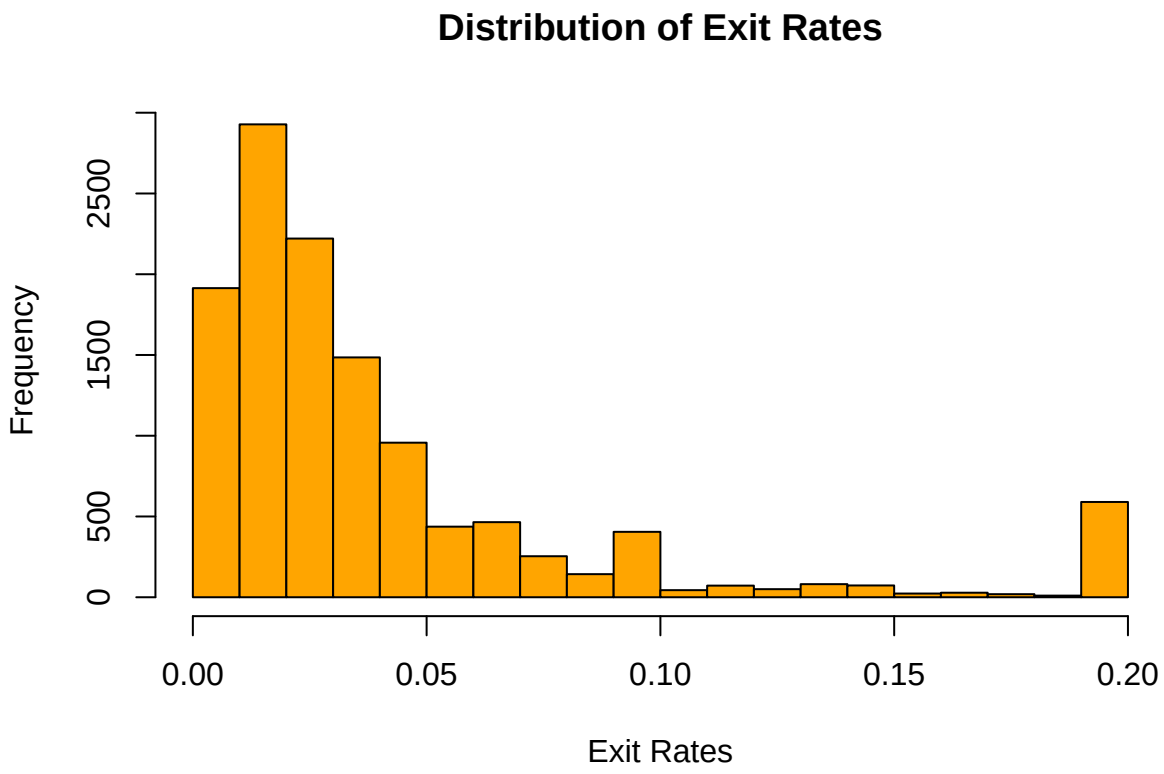

```
hist(df_new$ProductRelated_Duration,main = "Distribution of ",col="maroon",
     xlab="Administrative")
```



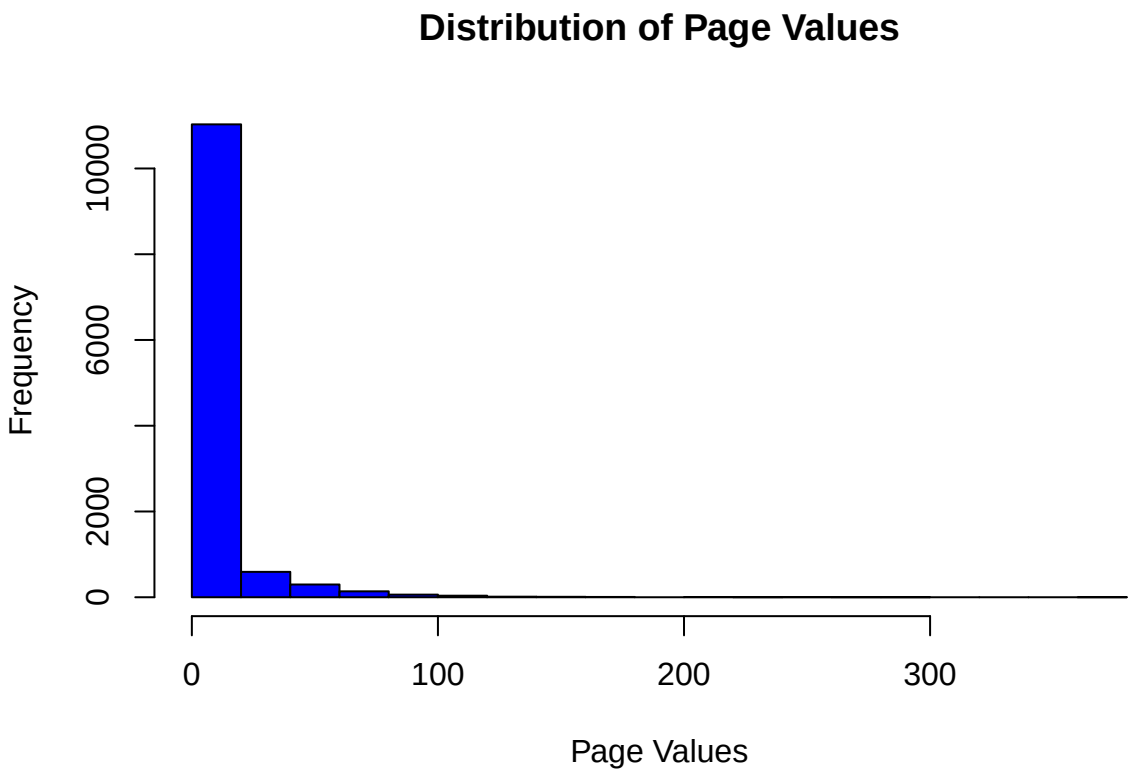

```
hist(df_new$BounceRates,main = "Distribution of Bonus Rates",col="cyan",
     xlab="Bonus Rates")
```



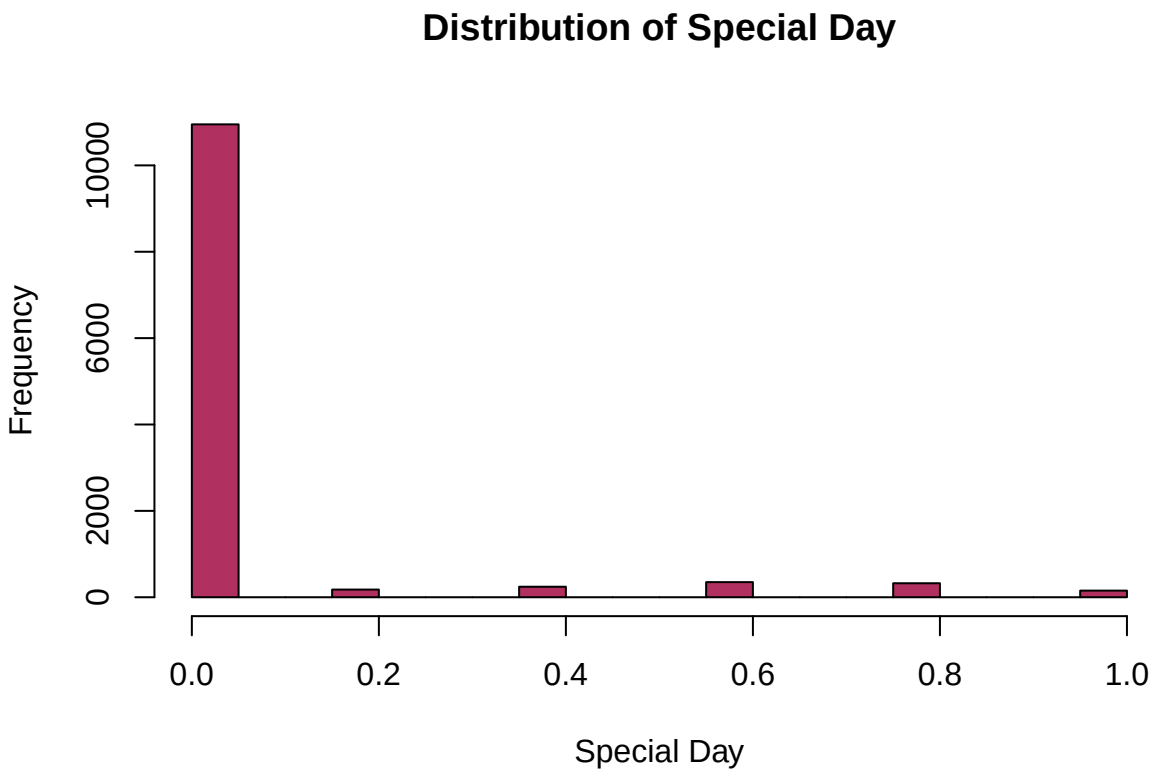

```
hist(df_new$ExitRates,main = "Distribution of Exit Rates",col="orange",
     xlab="Exit Rates")
```




```
hist(df_new$PageValues,main = "Distribution of Page Values",col="blue",
     xlab="Page Values")
```

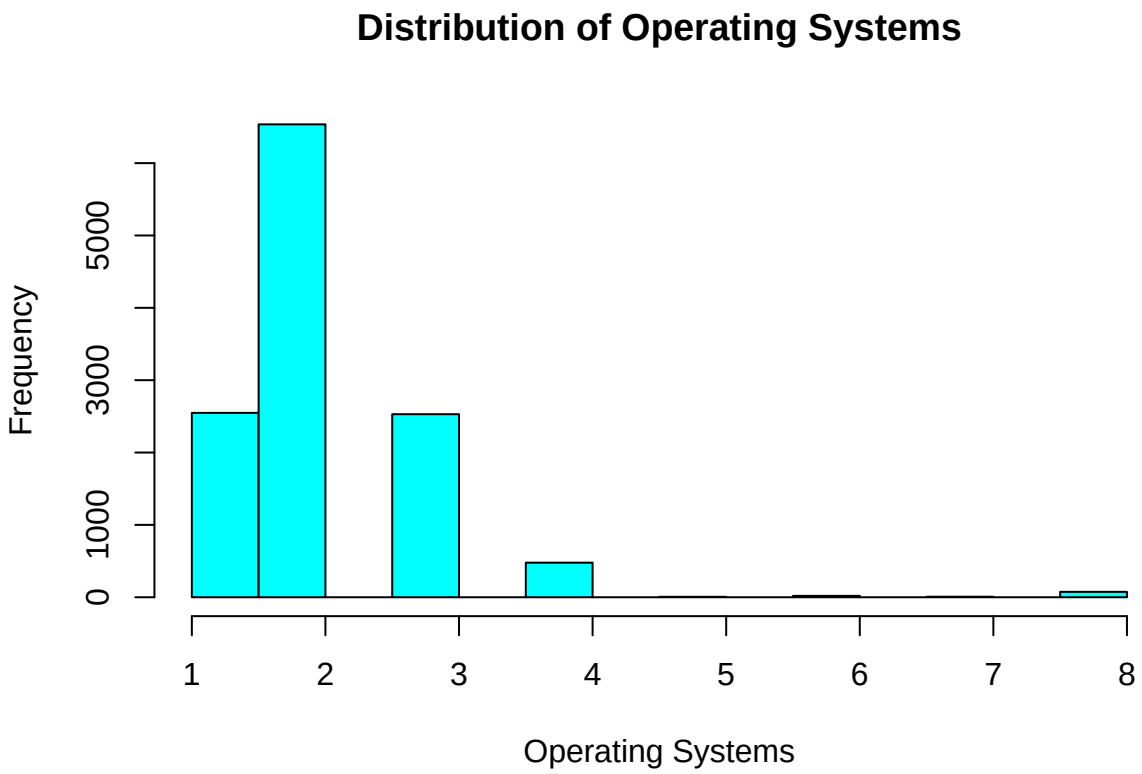



```
hist(df_new$SpecialDay,main = "Distribution of Special Day",col="maroon",
     xlab="Special Day")
```

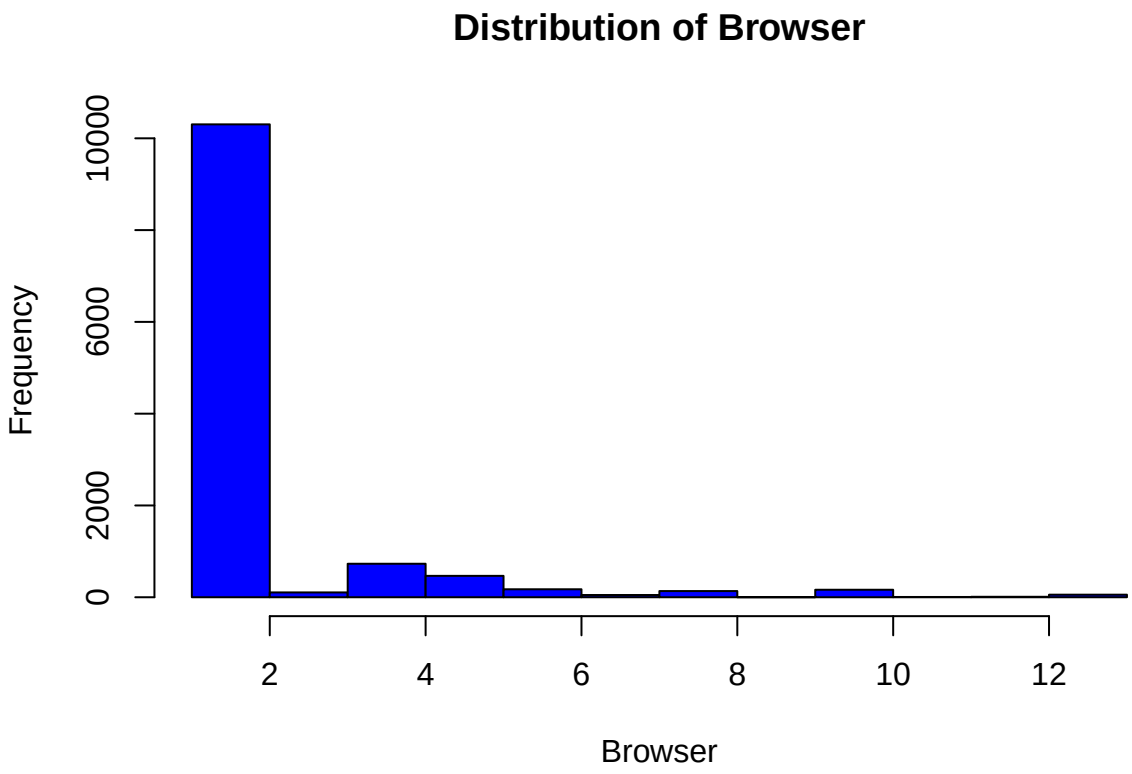


MOst of the values oin the Special Day column were 0.

```
hist(df_new$OperatingSystems,main = "Distribution of Operating Systems",col="cyan",
     xlab="Operating Systems")
```

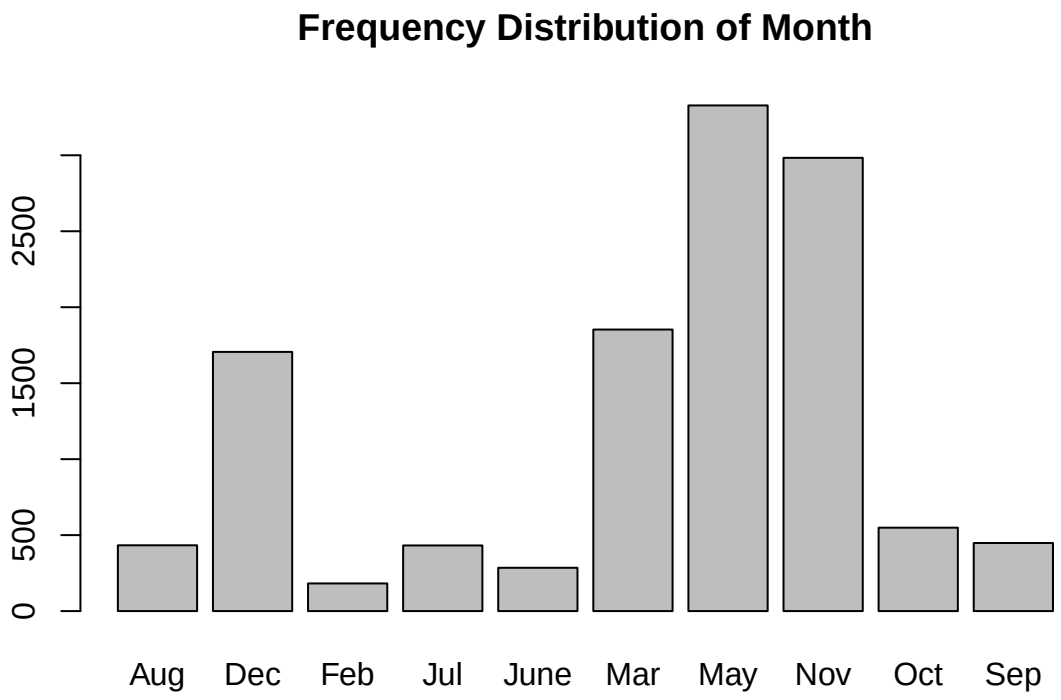



```
hist(df_new$Browser,main = "Distribution of Browser",col="blue",
     xlab="Browser")
```



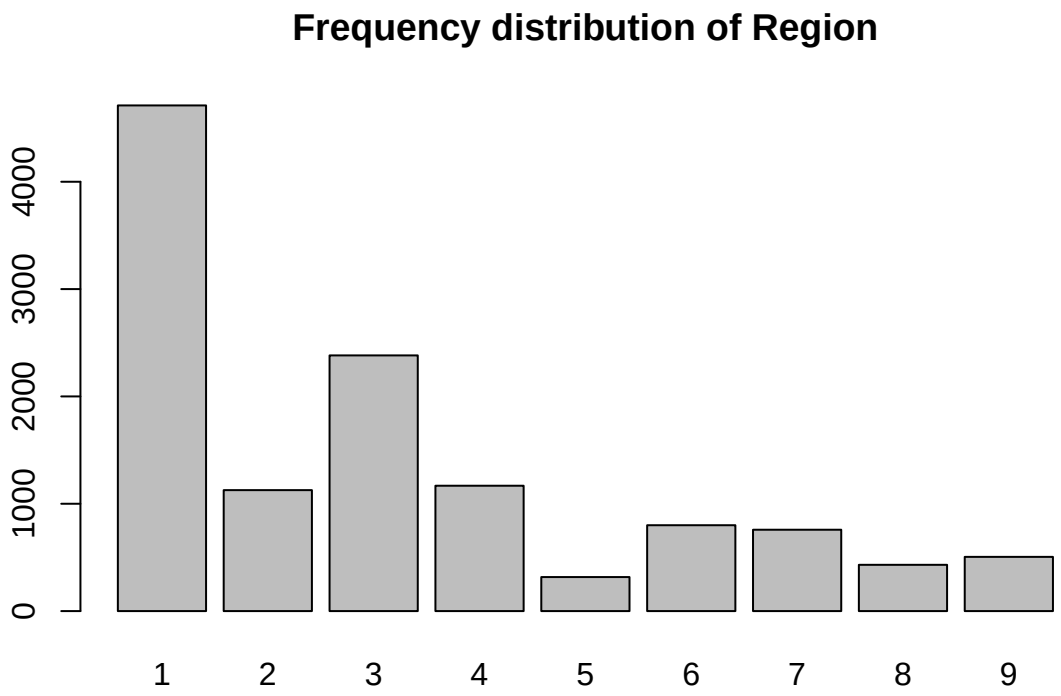
Bar Chart

```
#Month
frequency<-table(df_new$Month)
barplot(frequency,main = "Frequency Distribution of Month")
```



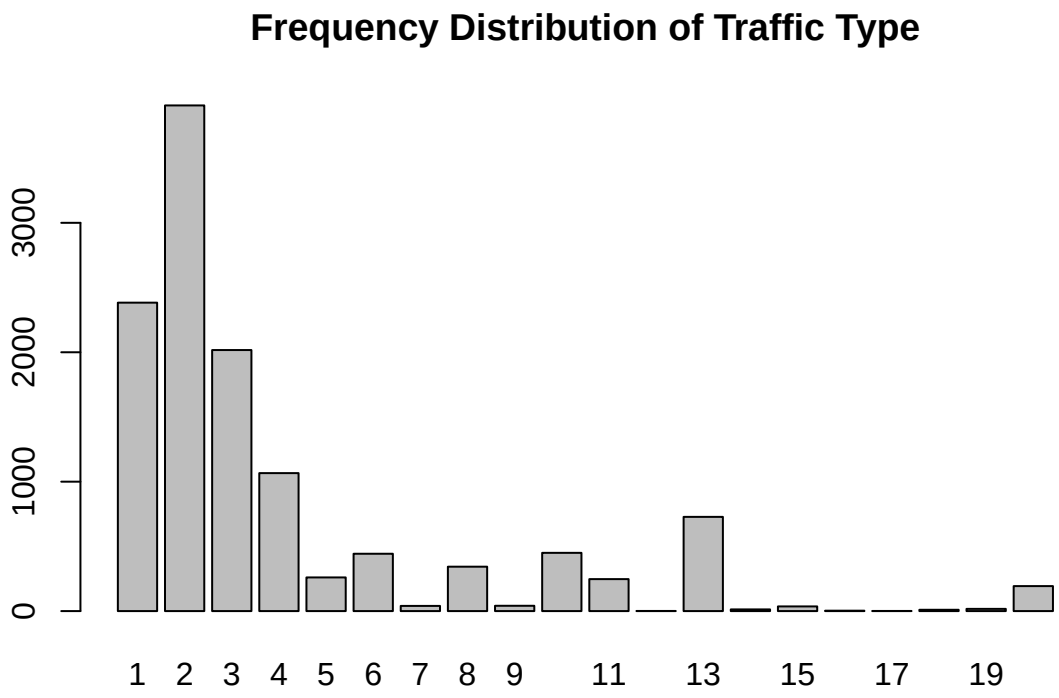
The Month of May had the highest number of records.

```
frequency_region<-table(df_new$Region)
barplot(frequency_region,main="Frequency distribution of Region")
```

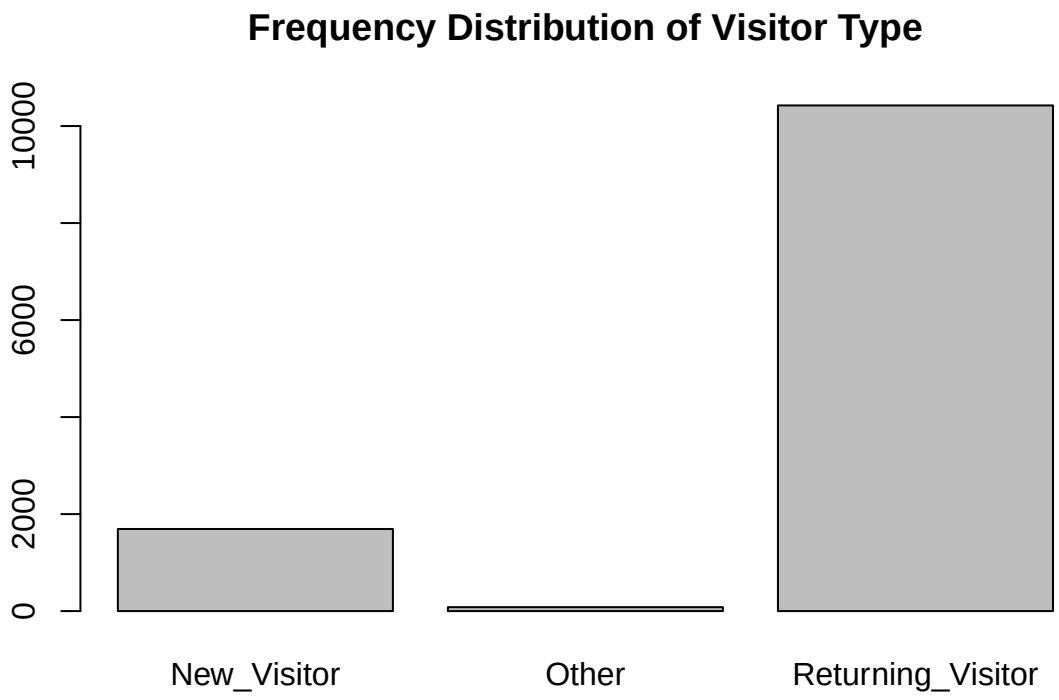


Region 1 was visited mostly followed by region 3.

```
frequency_traffic<-table(df_new$TrafficType)
barplot(frequency_traffic,main = "Frequency Distribution of Traffic Type")
```

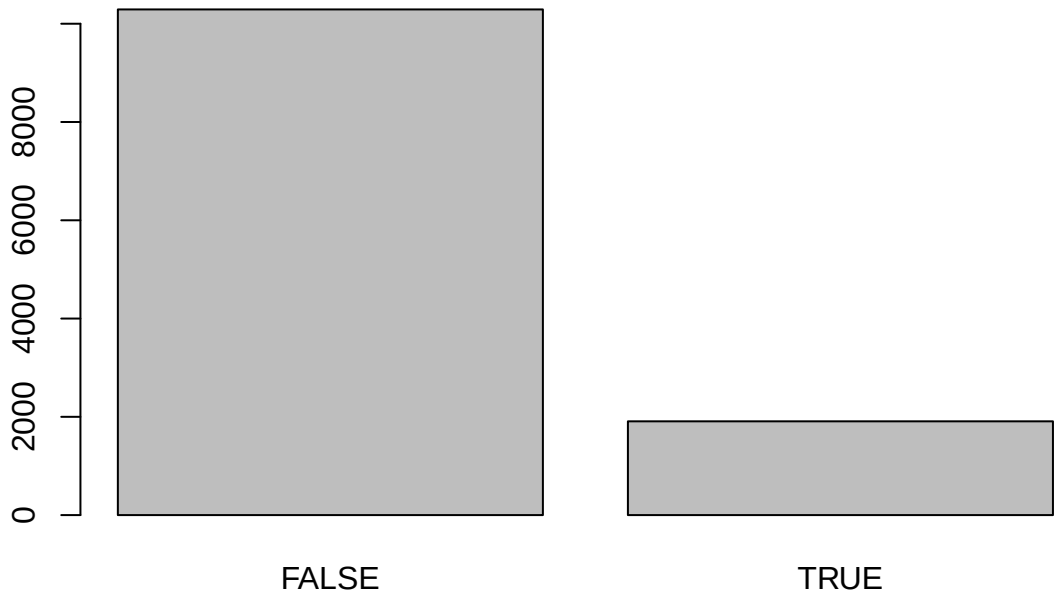



```
frequency_visitor<-table(df_new$VisitorType)
barplot(frequency_visitor,main="Frequency Distribution of Visitor Type")
```

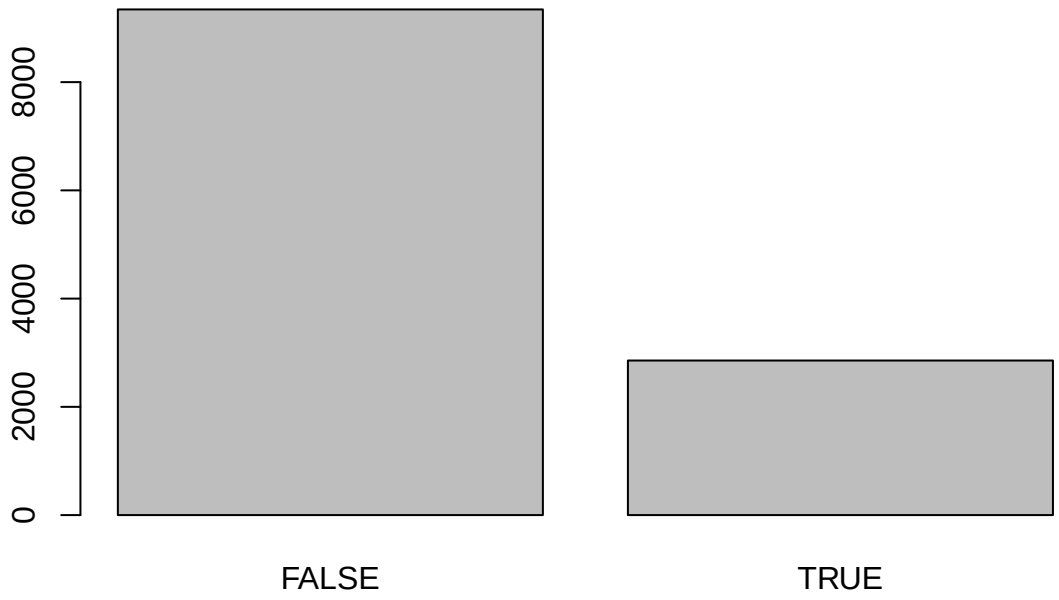


Most of the visitors were Returning Visitors.

```
frequency_rev<-table(df_new$Revenue)
barplot(frequency_rev,main="Frequency Distribution of Revenue")
```




```
frequency_wkd<-table(df_new$Weekend)
barplot(frequency_wkd)
```



Majority of the respondents visited the site on weekdays.

Bivariate Analysis

```
#correlation matrix
cor(df_new[,unlist(lapply(df_new, is.numeric))])
```

```
##                          Administrative Administrative_Duration Informational
## Administrative             1.000000000             0.600409653    0.37528761
## Administrative_Duration    0.600409653             1.000000000    0.30143630
## Informational              0.375287611             0.301436296    1.00000000
## Informational_Duration     0.254786021             0.237189860    0.61867795
## ProductRelated             0.428191515             0.286783914    0.37260472
## ProductRelated_Duration    0.371027224             0.353513793    0.38608372
## BounceRates               -0.213666635            -0.137333397   -0.10950530
## ExitRates                 -0.311274132            -0.202024452   -0.15956681
## PageValues                 0.096920968             0.066168365    0.04739015
## SpecialDay                -0.097072098            -0.074736885   -0.04937677
## OperatingSystems          -0.006697922            -0.007610715   -0.00962587
## Browser                   -0.025763658            -0.015833675   -0.03876681
## Region                    -0.007262053            -0.006723711   -0.03047732
## TrafficType               -0.034784126            -0.015075015   -0.03518669
##                          Informational_Duration ProductRelated
## Administrative                      0.254786021    0.428191515
```

```
## Administrative_Duration            0.237189860    0.286783914
## Informational                      0.618677947    0.372604721
## Informational_Duration             1.000000000    0.279061948
## ProductRelated                     0.279061948    1.000000000
## ProductRelated_Duration            0.346580691    0.860308186
## BounceRates                       -0.070159472   -0.193515772
## ExitRates                         -0.102932678   -0.286163211
## PageValues                         0.030064160    0.054115494
## SpecialDay                        -0.031293040   -0.025930622
## OperatingSystems                  -0.009749983    0.004090351
## Browser                           -0.019609349   -0.013706213
## Region                            -0.027920098   -0.040106501
## TrafficType                       -0.025163571   -0.044344333
##                         ProductRelated_Duration  BounceRates    ExitRates
## Administrative                       0.371027224 -0.213666635 -0.311274132
## Administrative_Duration              0.353513793 -0.137333397 -0.202024452
## Informational                        0.386083717 -0.109505298 -0.159566815
## Informational_Duration               0.346580691 -0.070159472 -0.102932678
## ProductRelated                       0.860308186 -0.193515772 -0.286163211
## ProductRelated_Duration              1.000000000 -0.174375499 -0.245334012
## BounceRates                         -0.174375499  1.000000000  0.903358192
## ExitRates                           -0.245334012  0.903358192  1.000000000
## PageValues                           0.050840624 -0.115991977 -0.173571542
## SpecialDay                          -0.038210652  0.087839995  0.116783762
## OperatingSystems                     0.002775788  0.026839839  0.016482012
## Browser                             -0.007838332 -0.016018380 -0.003565541
## Region                              -0.034862498  0.001432015 -0.001837556
## TrafficType                         -0.037506944  0.089199039  0.087386232
##                          PageValues   SpecialDay OperatingSystems      Browser
## Administrative           0.09692097 -0.097072098     -0.006697922 -0.025763658
## Administrative_Duration  0.06616837 -0.074736885     -0.007610715 -0.015833675
## Informational            0.04739015 -0.049376774     -0.009625870 -0.038766808
## Informational_Duration   0.03006416 -0.031293040     -0.009749983 -0.019609349
## ProductRelated           0.05411549 -0.025930622      0.004090351 -0.013706213
## ProductRelated_Duration  0.05084062 -0.038210652      0.002775788 -0.007838332
## BounceRates             -0.11599198  0.087839995      0.026839839 -0.016018380
## ExitRates               -0.17357154  0.116783762      0.016482012 -0.003565541
## PageValues               1.00000000 -0.064532709      0.018583782  0.045845065
## SpecialDay              -0.06453271  1.000000000      0.012757766  0.003465984
## OperatingSystems         0.01858378  0.012757766      1.000000000  0.212244823
## Browser                  0.04584506  0.003465984      0.212244823  1.000000000
## Region                   0.01059087 -0.016452464      0.071953240  0.091889464
## TrafficType              0.01223694  0.052827944      0.182874100  0.102886237
##                                Region TrafficType
## Administrative           -0.007262053 -0.03478413
## Administrative_Duration  -0.006723711 -0.01507502
## Informational            -0.030477323 -0.03518669
## Informational_Duration   -0.027920098 -0.02516357
## ProductRelated           -0.040106501 -0.04434433
## ProductRelated_Duration  -0.034862498 -0.03750694
## BounceRates               0.001432015  0.08919904
## ExitRates                -0.001837556  0.08738623
## PageValues                0.010590868  0.01223694
## SpecialDay               -0.016452464  0.05282794
```

```
## OperatingSystems         0.071953240  0.18287410
## Browser                  0.091889464  0.10288624
## Region                   1.000000000  0.04252523
## TrafficType              0.042525234  1.00000000
```

6.Modelling

##Implementing the Solution

K Means

```
head(df_new)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                      -1             0                     -1
## 4              0                       0             0                      0
## 5              0                       0             0                      0
## 6              0                       0             0                      0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000  0.20000000 0.2000000          0
## 2              2               64.000000  0.00000000 0.1000000          0
## 3              1               -1.000000  0.20000000 0.2000000          0
## 4              2                2.666667  0.05000000 0.1400000          0
## 5             10              627.500000  0.02000000 0.0500000          0
## 6             19              154.216667  0.01578947 0.0245614          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType
## 1          0   Feb                1       1      1           1
## 2          0   Feb                2       2      1           2
## 3          0   Feb                4       1      9           3
## 4          0   Feb                3       2      2           4
## 5          0   Feb                3       3      1           4
## 6          0   Feb                2       2      1           3
##         VisitorType Weekend Revenue
## 1 Returning_Visitor   FALSE   FALSE
## 2 Returning_Visitor   FALSE   FALSE
## 3 Returning_Visitor   FALSE   FALSE
## 4 Returning_Visitor   FALSE   FALSE
## 5 Returning_Visitor    TRUE   FALSE
## 6 Returning_Visitor   FALSE   FALSE
```

```
#transforming the logical variables
df_new$Revenue <- as.numeric(df_new$Revenue)
df_new$Weekend <- as.numeric(df_new$Weekend)
head(df_new)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
```

```
## 3              0                       -1            0                      -1
## 4              0                        0            0                       0
## 5              0                        0            0                       0
## 6              0                        0            0                       0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000  0.20000000 0.2000000          0
## 2              2               64.000000  0.00000000 0.1000000          0
## 3              1               -1.000000  0.20000000 0.2000000          0
## 4              2                2.666667  0.05000000 0.1400000          0
## 5             10              627.500000  0.02000000 0.0500000          0
## 6             19              154.216667  0.01578947 0.0245614          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType
## 1          0   Feb                1       1      1           1
## 2          0   Feb                2       2      1           2
## 3          0   Feb                4       1      9           3
## 4          0   Feb                3       2      2           4
## 5          0   Feb                3       3      1           4
## 6          0   Feb                2       2      1           3
##         VisitorType Weekend Revenue
## 1 Returning_Visitor       0       0
## 2 Returning_Visitor       0       0
## 3 Returning_Visitor       0       0
## 4 Returning_Visitor       0       0
## 5 Returning_Visitor       1       0
## 6 Returning_Visitor       0       0
```

```
#Label encoding the categorical variables
library(CatEncoders)
```

```
##
## Attaching package: 'CatEncoders'
```

```
## The following object is masked from 'package:base':
##
##     transform
```

```
encode <- LabelEncoder.fit(df_new$VisitorType)
df_new$VisitorType <- transform(encode,df_new$VisitorType)
encode <- LabelEncoder.fit(df_new$Month)
df_new$Month <- transform(encode,df_new$Month)
print(unique(df_new$Month))
```

```
##  [1]  3  6  7  9  5  4  1  8 10  2
```

```
print(unique(df_new$VisitorType))
```

```
## [1] 3 1 2
```

```
#checking if the variables have been encoded
head(df_new)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                      -1             0                     -1
## 4              0                       0             0                      0
## 5              0                       0             0                      0
## 6              0                       0             0                      0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000  0.20000000 0.2000000          0
## 2              2               64.000000  0.00000000 0.1000000          0
## 3              1               -1.000000  0.20000000 0.2000000          0
## 4              2                2.666667  0.05000000 0.1400000          0
## 5             10              627.500000  0.02000000 0.0500000          0
## 6             19              154.216667  0.01578947 0.0245614          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType VisitorType
## 1          0     3                1       1      1           1           3
## 2          0     3                2       2      1           2           3
## 3          0     3                4       1      9           3           3
## 4          0     3                3       2      2           4           3
## 5          0     3                3       3      1           4           3
## 6          0     3                2       2      1           3           3
##   Weekend Revenue
## 1       0       0
## 2       0       0
## 3       0       0
## 4       0       0
## 5       1       0
## 6       0       0
```

```r
# normalize the dataset
normal <- function(x){
  return ((x-min(x)) / (max(x)-min(x)))
}
df_new$Administrative<- normal(df_new$Administrative)
df_new$Administrative_Duration<- normal(df_new$Administrative_Duration)
df_new$Informational<- normal(df_new$Informational)
df_new$Informational_Duration<- normal(df_new$Informational_Duration)
df_new$ProductRelated<- normal(df_new$ProductRelated)
df_new$ProductRelated_Duration<- normal(df_new$ProductRelated_Duration)
df_new$BounceRates<- normal(df_new$BounceRates)
df_new$ExitRates<- normal(df_new$ExitRates)
df_new$PageValues<- normal(df_new$PageValues)
df_new$SpecialDay<- normal(df_new$SpecialDay)
df_new$OperatingSystems<- normal(df_new$OperatingSystems)
df_new$Browser<- normal(df_new$Browser)
df_new$Region<- normal(df_new$Region)
df_new$TrafficType<- normal(df_new$TrafficType)
head(df_new)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0           0.0002941393             0           0.0003920992
## 2              0           0.0002941393             0           0.0003920992
## 3              0           0.0000000000             0           0.0000000000
## 4              0           0.0002941393             0           0.0003920992
```

```
## 5              0            0.0002941393             0           0.0003920992
## 6              0            0.0002941393             0           0.0003920992
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1    0.001418440            1.563122e-05  1.00000000  1.000000          0
## 2    0.002836879            1.016029e-03  0.00000000  0.500000          0
## 3    0.001418440            0.000000e+00  1.00000000  1.000000          0
## 4    0.002836879            5.731448e-05  0.25000000  0.700000          0
## 5    0.014184397            9.824223e-03  0.10000000  0.250000          0
## 6    0.026950355            2.426226e-03  0.07894737  0.122807          0
##   SpecialDay Month OperatingSystems    Browser Region TrafficType VisitorType
## 1          0     3        0.0000000 0.00000000  0.000  0.00000000           3
## 2          0     3        0.1428571 0.08333333  0.000  0.05263158           3
## 3          0     3        0.4285714 0.00000000  1.000  0.10526316           3
## 4          0     3        0.2857143 0.08333333  0.125  0.15789474           3
## 5          0     3        0.2857143 0.16666667  0.000  0.15789474           3
## 6          0     3        0.1428571 0.08333333  0.000  0.10526316           3
##   Weekend Revenue
## 1       0       0
## 2       0       0
## 3       0       0
## 4       0       0
## 5       1       0
## 6       0       0
```

```
#setting revenue as the target variable
class <- df_new[,18]
df1<-df_new[,-18]
df_kmeans <- kmeans(df1,centers=2)
```

```
# Previewing the no. of records in each cluster

df_kmeans$size
```

Computing KNN

```
## [1] 9446 2753
```

The first cluster had 9446 records and the second cluster had 2753 clusters.

```
# Getting the value of cluster center datapoint
df_kmeans$centers
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1     0.08723269              0.02449753    0.02138471             0.01383154
## 2     0.08472912              0.02371116    0.02056847             0.01480752
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1     0.04631117              0.01928934   0.1016487 0.2063983 0.01647697
## 2     0.04259699              0.01752184   0.1042409 0.2112088 0.01637569
```

```
##   SpecialDay    Month OperatingSystems   Browser    Region TrafficType
## 1 0.07548169 7.317806        0.1586008 0.1097731 0.2630611   0.1659015
## 2 0.01561932 2.222666        0.1675575 0.1248638 0.2900926   0.1478196
##   VisitorType   Weekend
## 1    2.736714 0.2399958
## 2    2.644025 0.2139484
```

```
# Getting the cluster vector that shows the cluster where each record falls
#df_kmeans$cluster
```

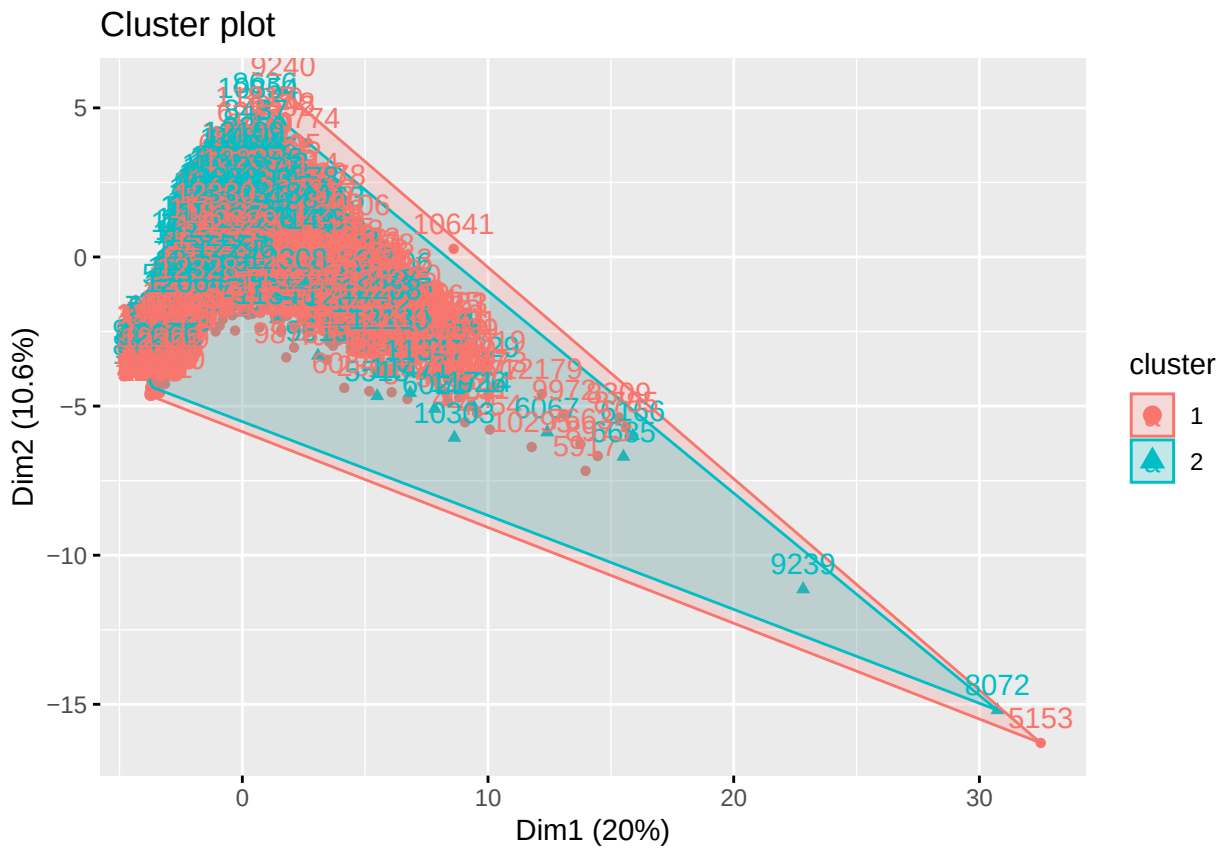
```
library(factoextra)
```

Visualizing the kmeans clusters

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_cluster(df_kmeans,data= df1)
```



The Kmeans did not bring a clear results.

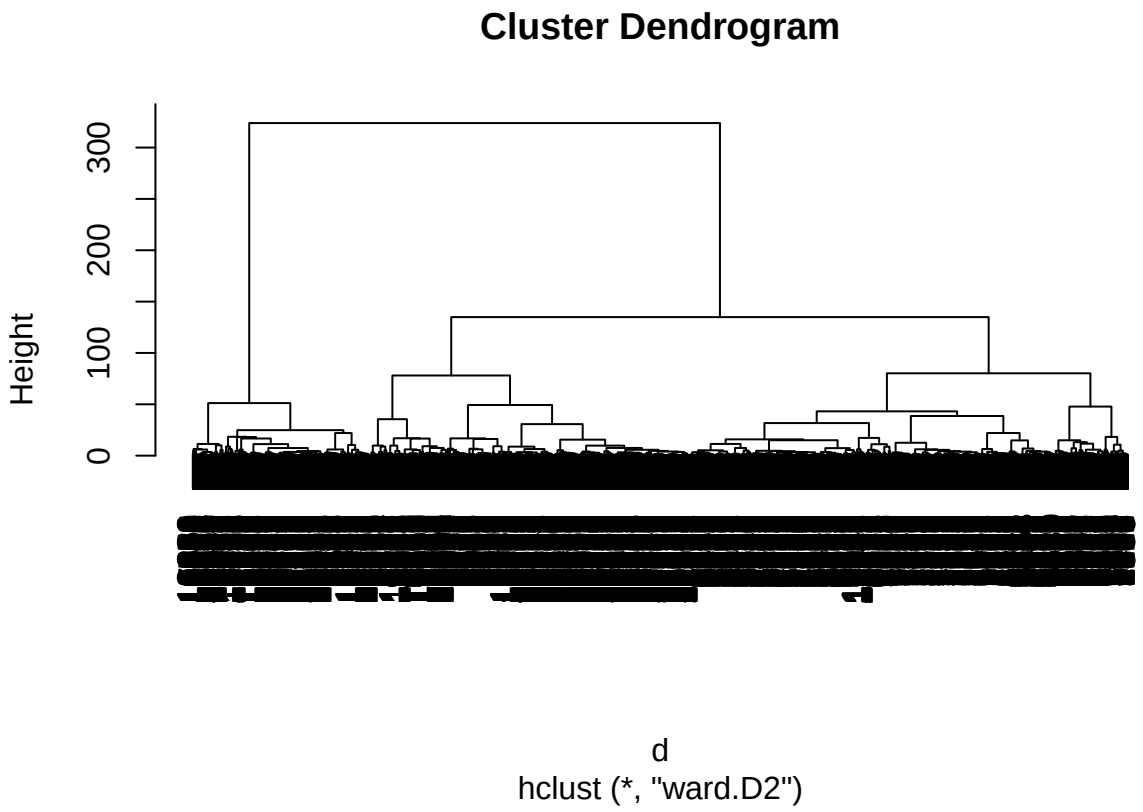
##CHallenging the solution

challenging the solution by using other clustering algorithms

Hierachical Clustering

```
# Compute distances using euclidean
d <- dist(df1, method = "euclidean")
#hierarchical clustering using ward.d2
hc <- hclust(d, method = "ward.D2")
```

```
#plotting a dendogram
plot(hc, xlim = c(1, 20), ylim = c(1,8))
```



The Hierarchical clustering did not perform well since we used a huge dataset.

DBSCAN Clustering

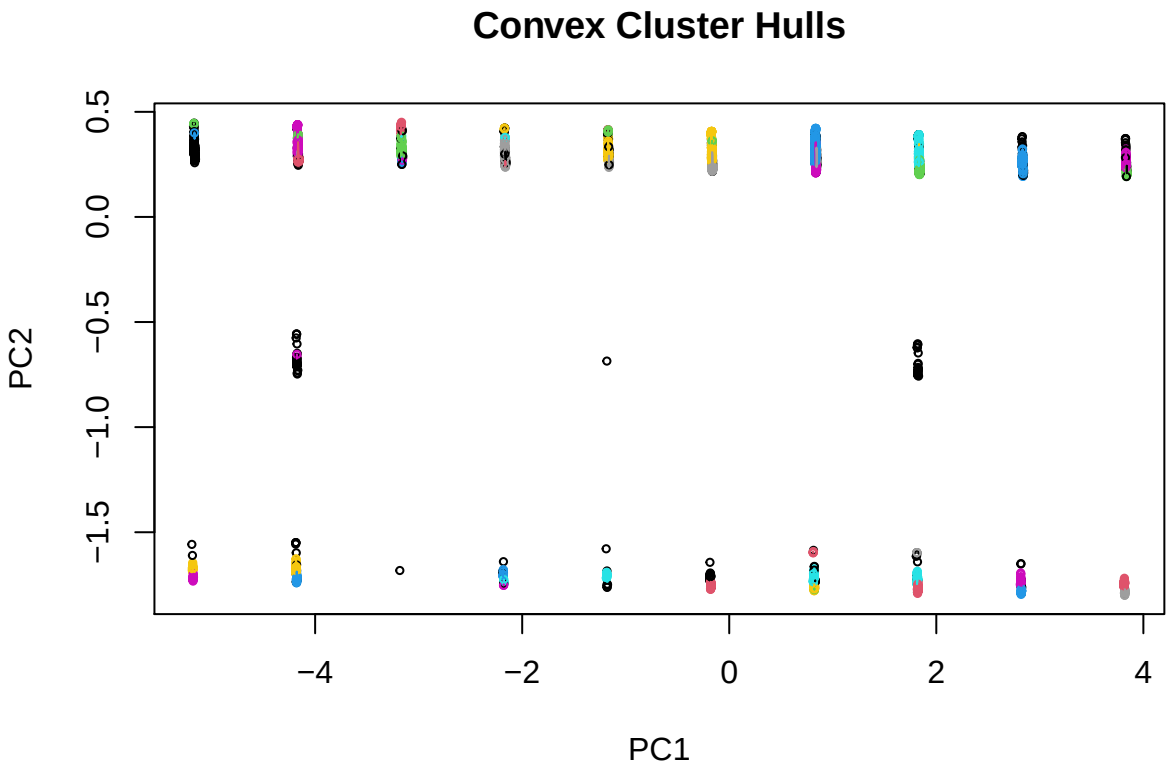
```
# minimum 4 points with in a distance of eps(0.4)
library("dbscan")
db<-dbscan(df1,eps=0.4,minPts = 4)
# print the clustering results
print(db)
```

```
## DBSCAN clustering for 12199 objects.
## Parameters: eps = 0.4, minPts = 4
## The clustering contains 63 cluster(s) and 422 noise points.
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##  422   26  122    8    5    4 1225  363  138   87   16 2354  217  479   70  126
##   16   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31
##    4    5    4  303   23   79  165  261   60  125  624   87 1856  250   46   70
##   32   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47
##  272   84   59   36  269   24   20   26    5   10    8   38    6    5    8   21
##   48   49   50   51   52   53   54   55   56   57   58   59   60   61   62   63
##    4    4    6    4    6 1007  249    4   40  255   16   63   13    4    4    5
##
## Available fields: cluster, eps, minPts
```

```
# plot our clusters
hullplot(df1,db$cluster)
```

```
## Warning in hullplot(df1, db$cluster): Not enough colors. Some colors will be
## reused.
```



Conclusion $ Recommendation K means clustering performed well and so we recommended the use of Kmeans in learning the characteristics of customer groups.