```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

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

```
library(tidyr)
library(scales)
```

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

```
library(ggcorrplot)
```

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

```
#Import dataset into R
thedata <- read.csv("C:/Users/gabri/Documents/online_shoppers_intention.csv")
```

```
head(thedata)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                       0             0                      0
## 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                0.000000  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
```

```
#Stucture of dataset
str(thedata)
```

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 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                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ 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            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
#Sumamry of dataset
summary(thedata)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   7.50         Median : 0.0000
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##
```

```
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##
##      Month      OperatingSystems    Browser           Region
##  May    :3364   Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Nov    :2998   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mar    :1907   Median :2.000    Median : 2.000   Median :3.000
##  Dec    :1727   Mean   :2.124    Mean   : 2.357   Mean   :3.147
##  Oct    : 549   3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##  Sep    : 448   Max.   :8.000    Max.   :13.000   Max.   :9.000
##  (Other):1337
##   TrafficType                VisitorType     Weekend          Revenue
##  Min.   : 1.00   New_Visitor      : 1694   Mode :logical   Mode :logical
##  1st Qu.: 2.00   Other            :   85   FALSE:9462      FALSE:10422
##  Median : 2.00   Returning_Visitor:10551   TRUE :2868      TRUE :1908
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
##
```

```
#Check to see if dataset has NAs
sum(is.na(thedata))
```

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

```
sapply(thedata, function(x) sum(is.na(x)))
```

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

```
#Dimenstion of the dataset
dim(thedata)
```

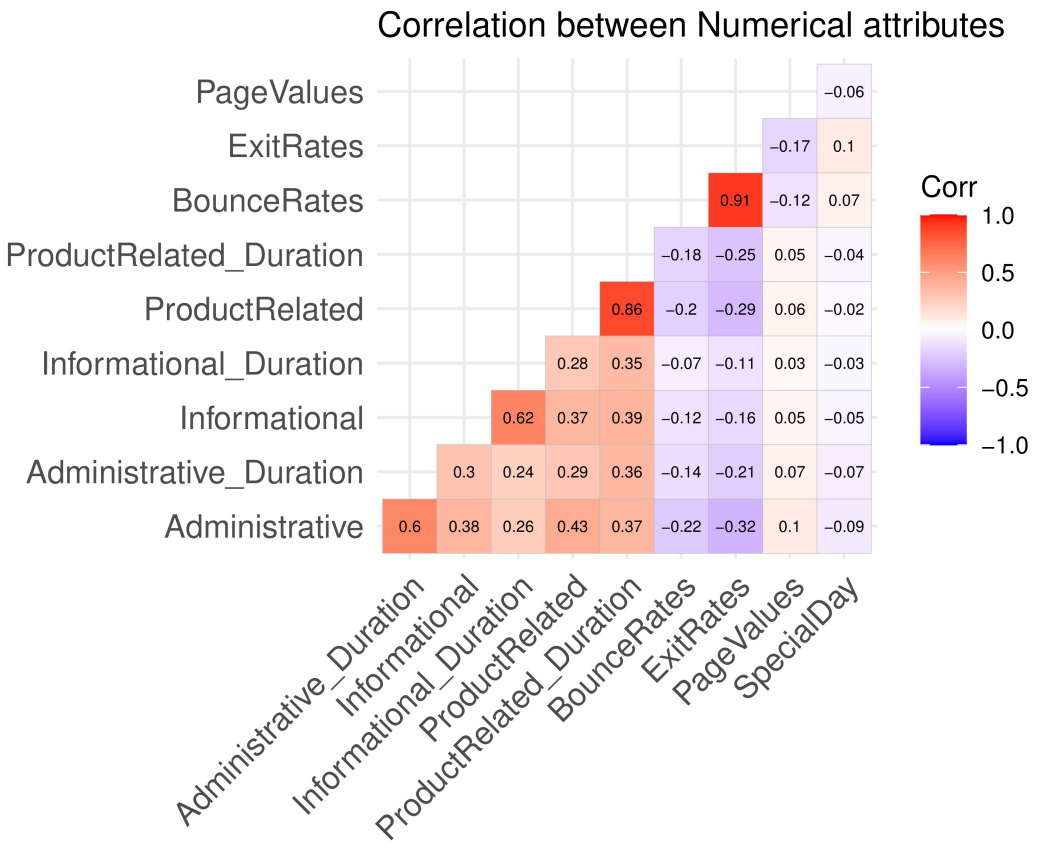
```
## [1] 12330    18
```

```r
#Correlation  betweem numerical attributes
variablesnum <- c("Administrative", "Administrative_Duration", "Informational", "Informational_Duration
data2 <- (thedata[variablesnum])
cor(data2)
```

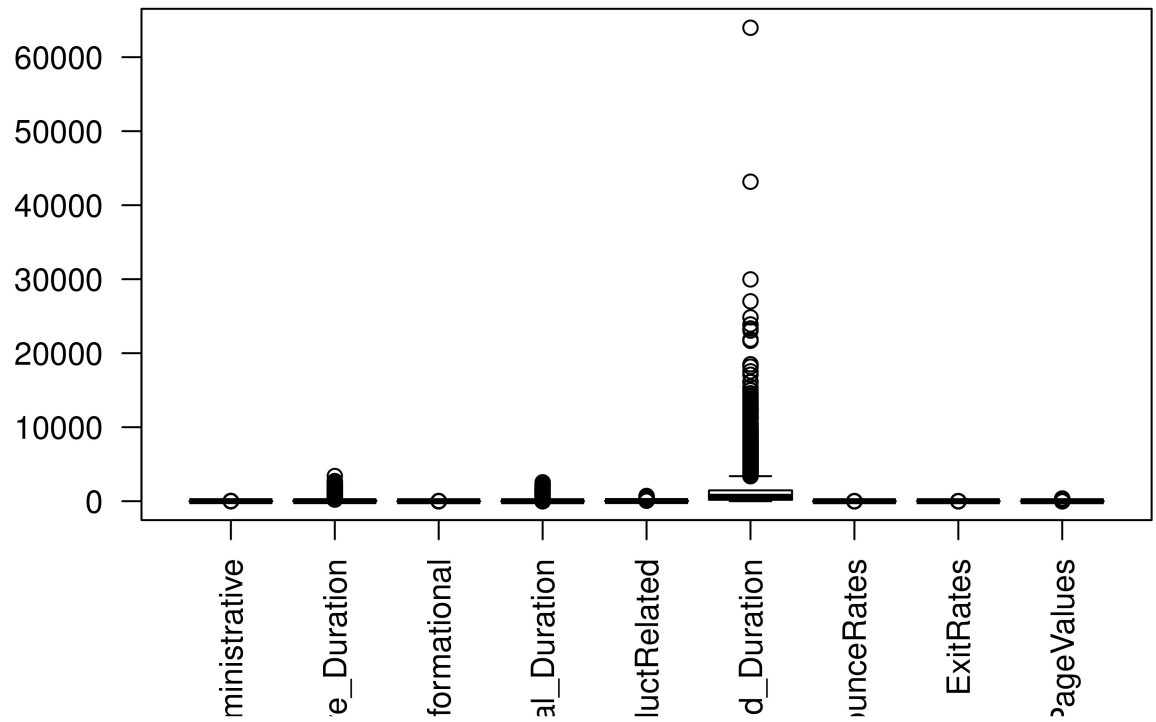
```
##                          Administrative Administrative_Duration Informational
## Administrative                1.00000000              0.60158334    0.37685043
## Administrative_Duration       0.60158334              1.00000000    0.30270971
## Informational                 0.37685043              0.30270971    1.00000000
## Informational_Duration        0.25584814              0.23803079    0.61895486
## ProductRelated                0.43111934              0.28908662    0.37416429
## ProductRelated_Duration       0.37393901              0.35542195    0.38750531
## BounceRates                  -0.22356263             -0.14417041   -0.11611362
## ExitRates                    -0.31648300             -0.20579776   -0.16366606
## PageValues                    0.09898959              0.06760848    0.04863169
## SpecialDay                   -0.09477760             -0.07330372   -0.04821925
##                          Informational_Duration ProductRelated
## Administrative                      0.25584814     0.43111934
## Administrative_Duration             0.23803079     0.28908662
## Informational                       0.61895486     0.37416429
## Informational_Duration              1.00000000     0.28004627
## ProductRelated                      0.28004627     1.00000000
## ProductRelated_Duration             0.34736358     0.86092684
## BounceRates                        -0.07406661    -0.20457763
## ExitRates                          -0.10527568    -0.29252628
## PageValues                          0.03086087     0.05628179
## SpecialDay                         -0.03057655    -0.02395817
##                          ProductRelated_Duration BounceRates  ExitRates
## Administrative                        0.37393901 -0.22356263 -0.3164830
## Administrative_Duration               0.35542195 -0.14417041 -0.2057978
## Informational                         0.38750531 -0.11611362 -0.1636661
## Informational_Duration                0.34736358 -0.07406661 -0.1052757
## ProductRelated                        0.86092684 -0.20457763 -0.2925263
## ProductRelated_Duration               1.00000000 -0.18454112 -0.2519841
## BounceRates                          -0.18454112  1.00000000  0.9130044
## ExitRates                            -0.25198410  0.91300440  1.0000000
## PageValues                            0.05282306 -0.11938603 -0.1744983
## SpecialDay                           -0.03637985  0.07270225  0.1022418
##                           PageValues  SpecialDay
## Administrative            0.09898959 -0.09477760
## Administrative_Duration   0.06760848 -0.07330372
## Informational             0.04863169 -0.04821925
## Informational_Duration    0.03086087 -0.03057655
## ProductRelated            0.05628179 -0.02395817
## ProductRelated_Duration   0.05282306 -0.03637985
## BounceRates              -0.11938603  0.07270225
## ExitRates                -0.17449831  0.10224180
## PageValues                1.00000000 -0.06354127
## SpecialDay               -0.06354127  1.00000000
```

```r
#Correlation matrix betweem numerical attributes
variables_correlations <- cor(data2)
ggcorrplot(variables_correlations,type="lower", lab = TRUE, lab_size = 2) +
```

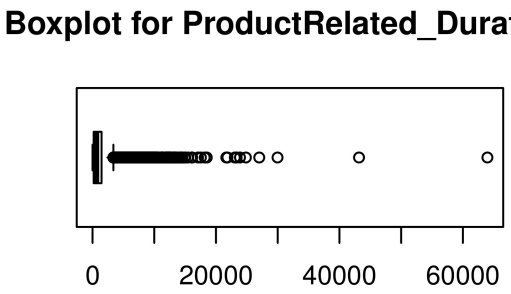
```
  ggtitle("Correlation between Numerical attributes")
```

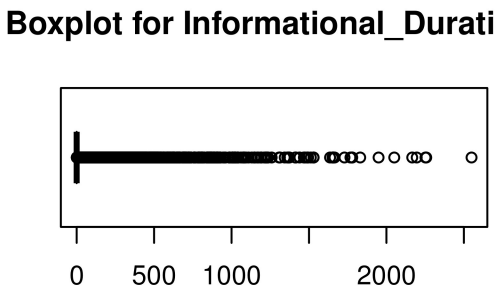


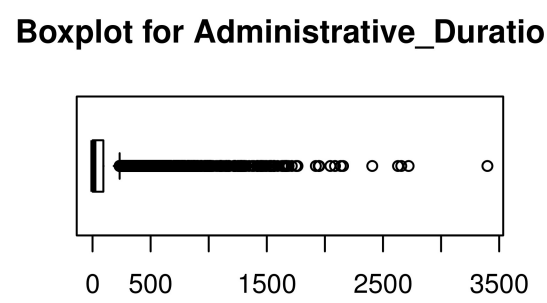

```
#Boxplots

#Boxplot for all numerical attributes

outliersnum <-boxplot(thedata[1:9], las = 2, xlab ="")
```

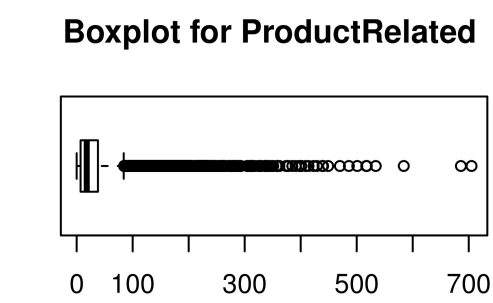



```
#Boxplot for numerical attributes
par(mfcol=c(2,2))

outliersProductRelated_Duration <- boxplot(thedata$ProductRelated_Duration, horizontal = TRUE, outline =
outliersAdministrative_Duration <- boxplot(thedata$Administrative_Duration, horizontal = TRUE, outline =

outliersInformational_Duration <- boxplot(thedata$Informational_Duration, horizontal = TRUE, outline = '

outliersProductRelated <- boxplot(thedata$ProductRelated, horizontal = TRUE, outline = TRUE, main = "Bo
```

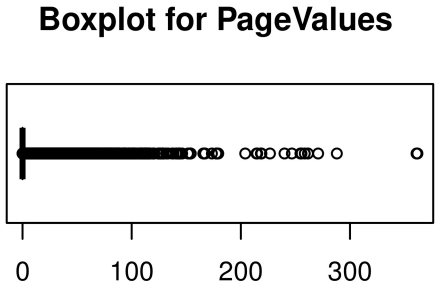


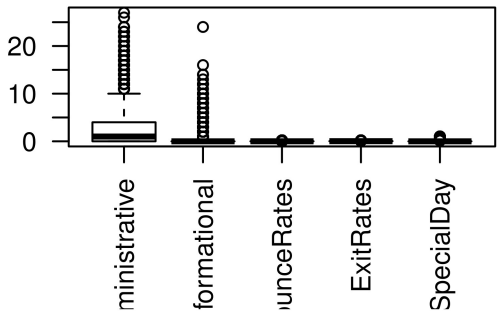



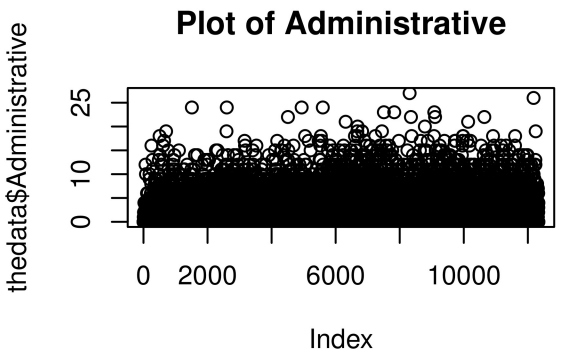

```
outliersPageValues <- boxplot(thedata$PageValues, horizontal = TRUE, outline = TRUE, main = "Boxplot fo

temp <- c("Administrative","Informational","BounceRates","ExitRates","SpecialDay")
tempdata<-thedata[temp]
outlierstemp <-boxplot(tempdata, las = 2, xlab ="")

#out <- boxplot.stats(thedata$ProductRelated_Duration)$out
#out_ind <- which(thedata$ProductRelated_Duration %in% c(out))
#thedata[out_ind, ]
```

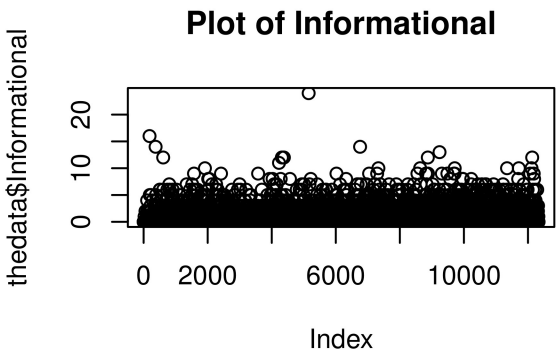


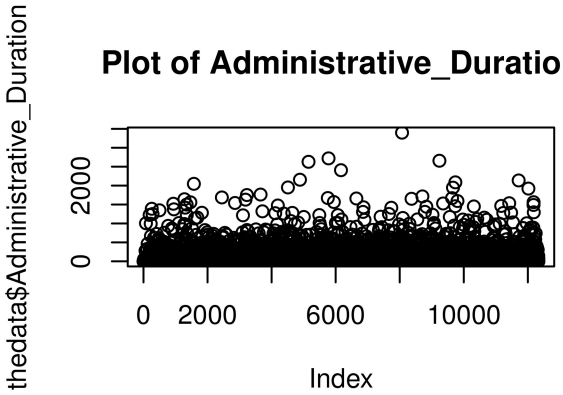



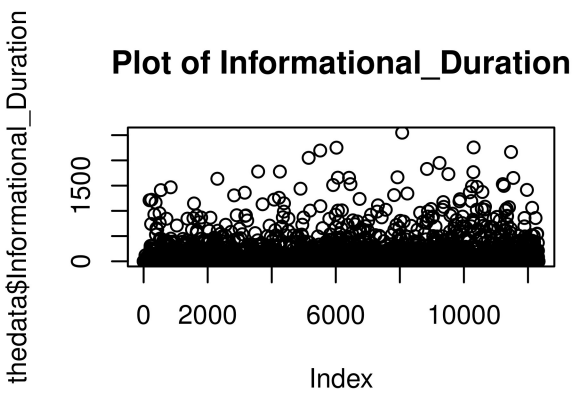

```
# As we can se above, there are numerous outliers presented in our data, if we remove them there would
```

```
#plots for numerical attributes
library(ggplot2)

par(mfcol=c(2,2))


plot(thedata$Administrative,main="Plot of Administrative" )
plot(thedata$Administrative_Duration,main="Plot of Administrative_Duration")
plot(thedata$Informational,main="Plot of Informational")
plot(thedata$Informational_Duration,main="Plot of Informational_Duration")
```

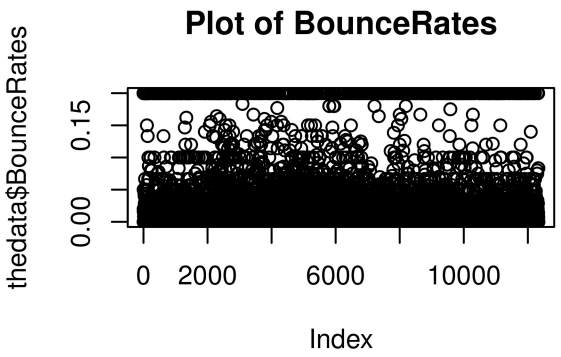


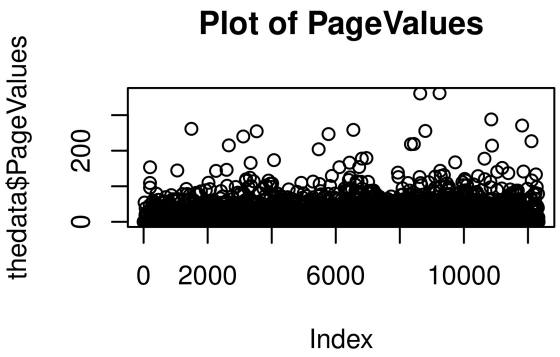



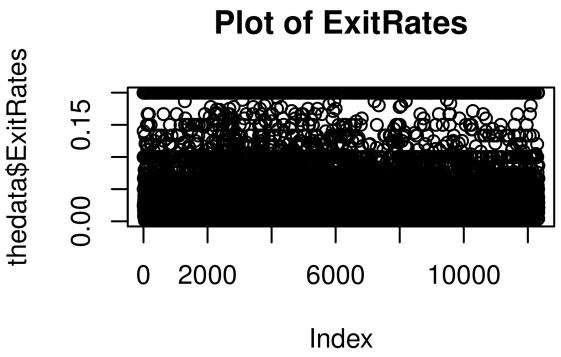



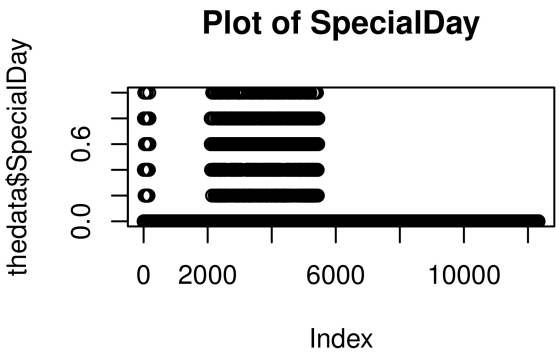




```r
plot(thedata$BounceRates,main="Plot of BounceRates")
plot(thedata$ExitRates,main="Plot of ExitRates")
plot(thedata$PageValues,main="Plot of PageValues")
plot(thedata$SpecialDay,main="Plot of SpecialDay")
```

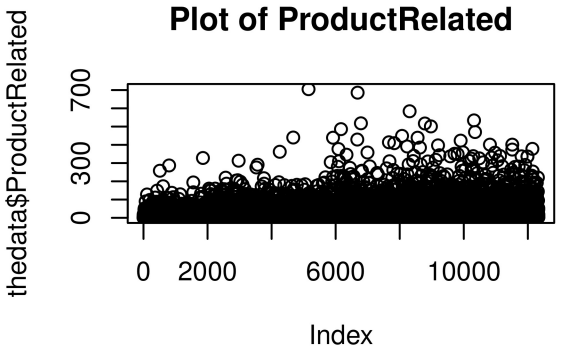


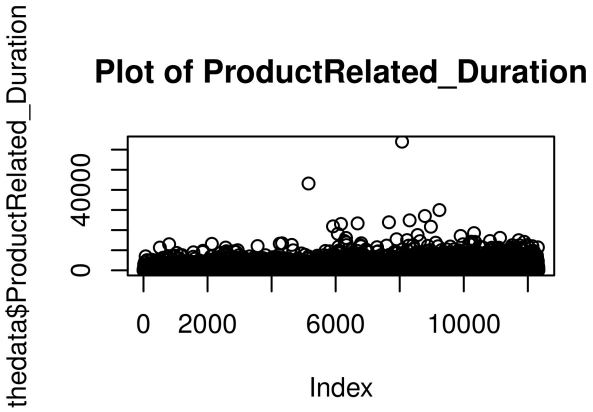



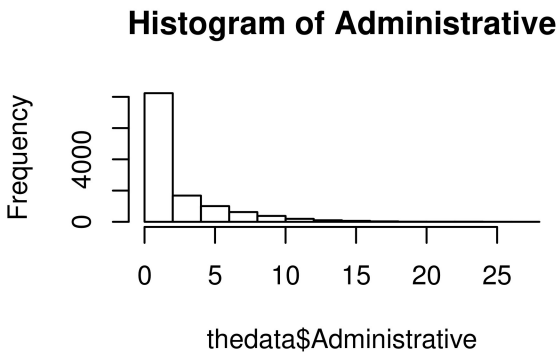



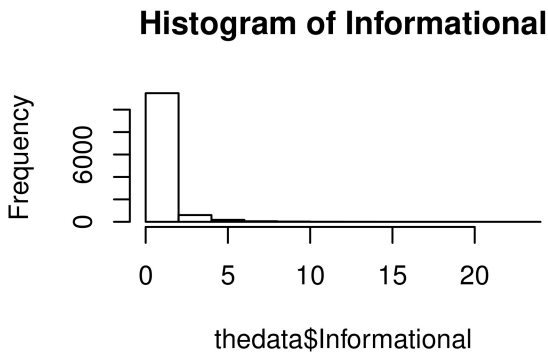

```
plot(thedata$ProductRelated,main="Plot of ProductRelated")
plot(thedata$ProductRelated_Duration, main="Plot of ProductRelated_Duration")
```

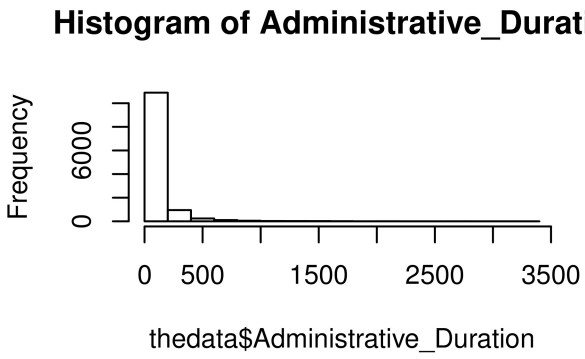


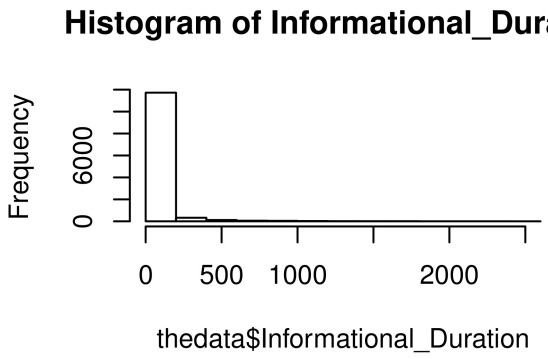




```r
#Histograms for Numerical attributes
par(mfcol=c(2,2))


hist(thedata$Administrative,main="Histogram of Administrative" )
hist(thedata$Administrative_Duration,main="Histogram of Administrative_Duration")
hist(thedata$Informational,main="Histogram of Informational")
hist(thedata$Informational_Duration,main="Histogram of Informational_Duration")
```

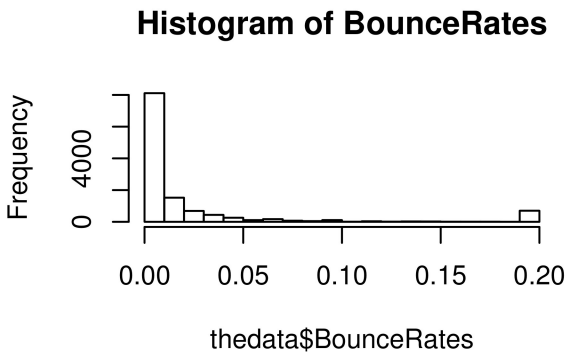


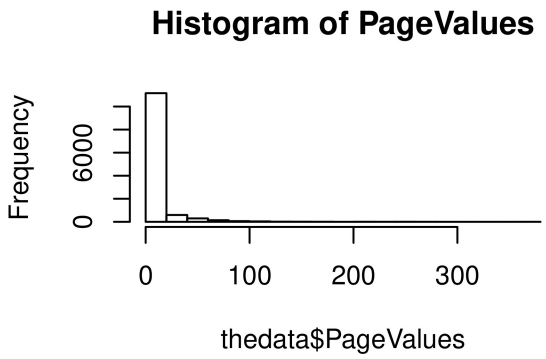



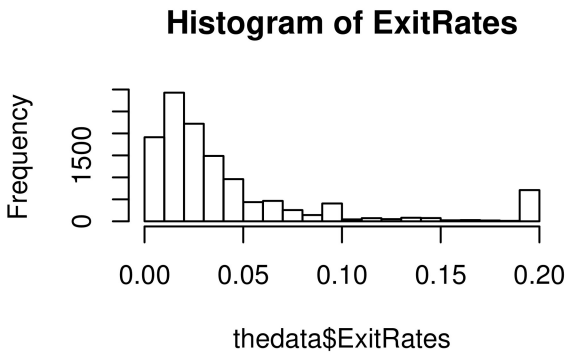



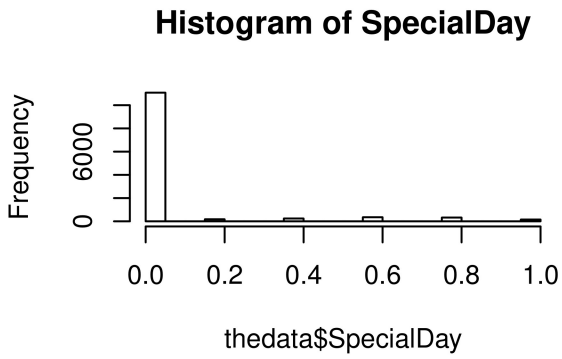




```r
hist(thedata$BounceRates,main="Histogram of BounceRates")
hist(thedata$ExitRates,main="Histogram of ExitRates")
hist(thedata$PageValues,main="Histogram of PageValues")
hist(thedata$SpecialDay,main="Histogram of SpecialDay")
```

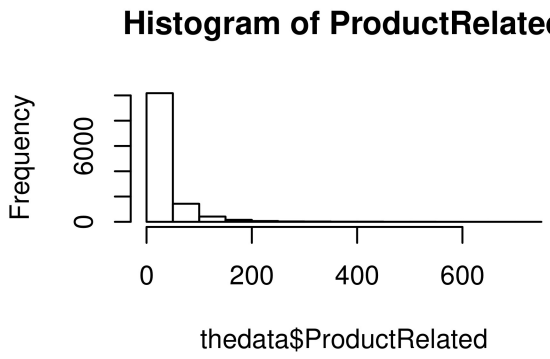


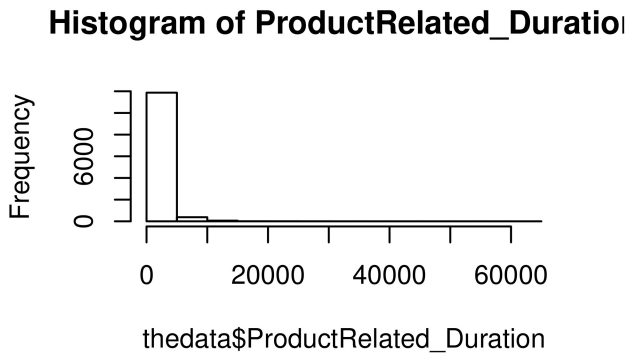



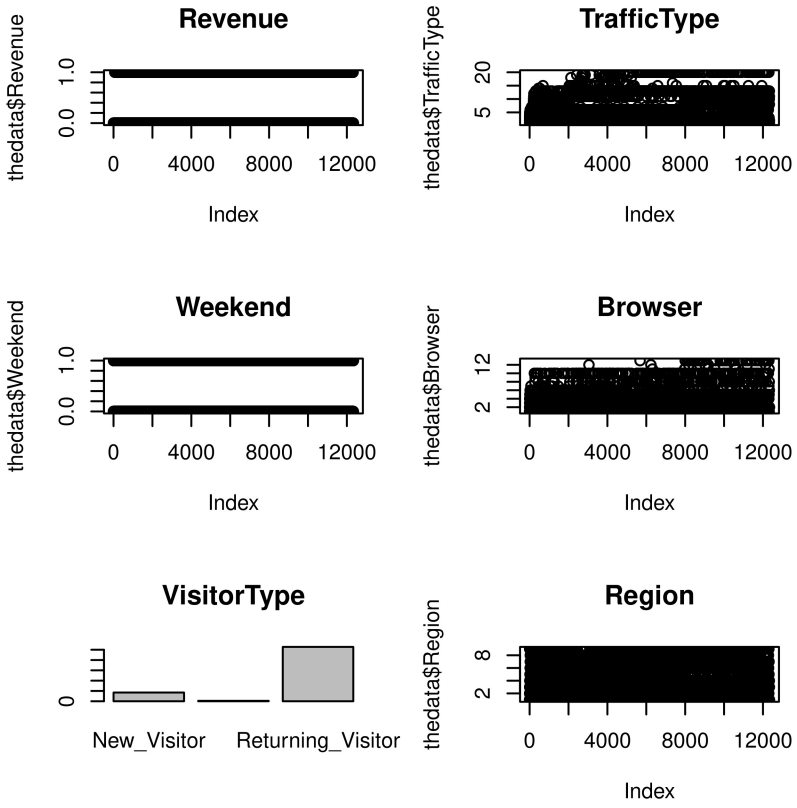




```r
hist(thedata$ProductRelated,main="Histogram of ProductRelated")
hist(thedata$ProductRelated_Duration, main="Histogram of ProductRelated_Duration")
```



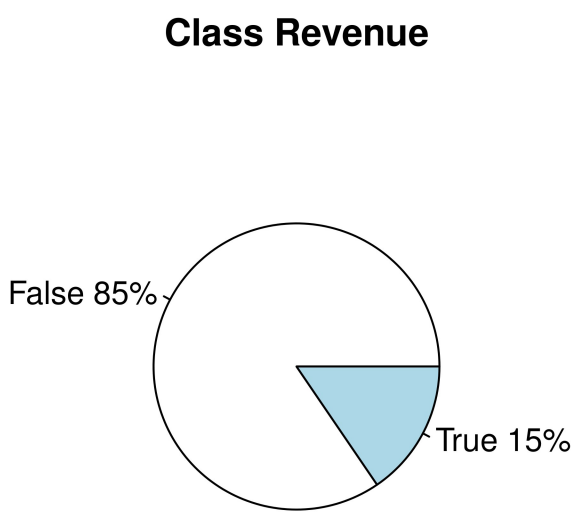

```r
#Plots For categorical attributes


par(mfcol=c(3,3))

plot(thedata$Revenue,main="Revenue")
plot(thedata$Weekend,main="Weekend")
plot(thedata$VisitorType,main="VisitorType")
plot(thedata$TrafficType,main="TrafficType")
plot(thedata$Browser,main="Browser")
plot(thedata$Region,main="Region")
```




```
#Count values for the Revenue attribute
count(thedata, Revenue, name = 'Class.count')
```

```
## # A tibble: 2 x 2
##   Revenue Class.count
##   <lgl>         <int>
## 1 FALSE         10422
## 2 TRUE           1908
```

```
#Pie chart to show Class Revenue

par(mfrow=c(1,2))
slices <- c(10422, 1908)
lbls <-c("False", "True")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct)
lbls <- paste(lbls,"%",sep="")
pie(slices, labels = lbls, main = "Class Revenue")

#This chart shows how the data is unbalanced
```




```
# Standard deviation for numerical attributes:

sd_administrative<- sd(thedata$Administrative)
sd_Administrative_Duration<- sd(thedata$Administrative_Duration)
sd_Informational<- sd(thedata$Informational)
sd_Informational_Duration<- sd(thedata$Informational_Duration)
sd_ProductRelated<- sd(thedata$ProductRelated)
sd_ProductRelated_Duration<- sd(thedata$ProductRelated_Duration)
sd_BounceRates<- sd(thedata$BounceRates)
sd_ExitRates<- sd(thedata$ExitRates)
sd_PageValues<- sd(thedata$PageValues)
sd_SpecialDay<- sd(thedata$SpecialDay)
sd_administrative
```

```
## [1] 3.321784
```

```
sd_Administrative_Duration
```

```
## [1] 176.7791
```

```
sd_Informational
```

```
## [1] 1.270156
```

```
sd_Informational_Duration
```

```
## [1] 140.7493
```

```
sd_ProductRelated
```

```
## [1] 44.4755
```

```
sd_ProductRelated_Duration
```

```
## [1] 1913.669
```

```
sd_BounceRates
```

```
## [1] 0.04848832
```

```
sd_ExitRates
```

```
## [1] 0.04859654
```

```
sd_PageValues
```

```
## [1] 18.56844
```

```
sd_SpecialDay
```

```
## [1] 0.1989173
```

```
#Correlation between Revenue and Region:
#we cannot reject the Nypotesis the results are not statiscally significant

chisq.test(thedata$Revenue, thedata$Region)
```

```
##
##  Pearson's Chi-squared test
##
## data:  thedata$Revenue and thedata$Region
## X-squared = 9.2528, df = 8, p-value = 0.3214
```

```
#library(readr)
#write_csv(thedata, path="C:/Users/gabri/Desktop/dataset_1.csv")
```