Study_On_Consumer_Behaviour

Shoyeb_Khan

01/11/2021

```
#"Administrative", "Administrative Duration", "Informational", "Informational
#Duration", "Product Related" and "Product Related Duration" represent the number
#of different types of pages visited by the visitor in that session and total time
#spent in each of these page categories. The values of these features are derived
#from the URL information of the pages visited by the user and updated in real
#time when a user takes an action, e.g. moving from one page to another.
#The "Bounce Rate", "Exit Rate" and "Page Value" features represent the
#metrics measured by "Google Analytics" for each page in the e-commerce site.
#The value of "Bounce Rate" feature for a web page refers to the percentage of
#visitors who enter the site from that page and then leave ("bounce") without
#triggering any other requests to the analytics server during that session.
#The value of "Exit Rate" feature for a specific web page is calculated as for
#all pageviews to the page, the percentage that were the last in the session.
#The "Page Value" feature represents the average value for a web page that a
#user visited before completing an e-commerce transaction. The "Special Day"
#feature indicates the closeness of the site visiting time to a specific special day
#(e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be
#finalized with transaction. The value of this attribute is determined by
#considering the dynamics of e-commerce such as the duration between the order
#date and delivery date. For example, for Valentin's day, this value takes a
#nonzero value between February 2 and February 12, zero before and after this
#date unless it is close to another special day, and its maximum value of 1 on
#February 8. The dataset also includes operating system, browser, region,
#traffic type, visitor type as returning or new visitor, a Boolean value
#indicating whether the date of the visit is weekend, and month of the year.

###

## Loading the Libraries.

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(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.0

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

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.1.1

## corrplot 0.90 loaded
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.1

## Loading required package: lattice

##
## Attaching package: 'caret'

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

```
library(janitor)
```

```
## Warning: package 'janitor' was built under R version 4.1.1

##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.1
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(reshape)
```

```
## Warning: package 'reshape' was built under R version 4.1.1
```

```
##
## Attaching package: 'reshape'

## The following objects are masked from 'package:tidyr':
##
##     expand, smiths

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

```
library(moments)
```

```
## Warning: package 'moments' was built under R version 4.1.1
```

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.1.1
```

```
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.1.1
```

```
library(aod)
```

```
## Warning: package 'aod' was built under R version 4.1.1
```

```
library(gplots)
```

```
## Warning: package 'gplots' was built under R version 4.1.1

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
## Reading the CSV file.

data_set_cst_ol <- read.csv('online_shoppers_intention.csv')

#View(data_set_cst_ol)

data_set_cst_ol %>%  glimpse()
```

```
## Rows: 12,330
## Columns: 18
## $ Administrative          <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 2~
## $ Administrative_Duration <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5~
## $ Informational           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Informational_Duration  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ProductRelated          <int> 1, 2, 1, 2, 10, 19, 1, 0, 2, 3, 3, 16, 7, 6, 2~
## $ ProductRelated_Duration <dbl> 0.000000, 64.000000, 0.000000, 2.666667, 627.5~
## $ BounceRates             <dbl> 0.200000000, 0.000000000, 0.200000000, 0.05000~
## $ ExitRates               <dbl> 0.200000000, 0.100000000, 0.200000000, 0.14000~
## $ PageValues              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ SpecialDay              <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.4, 0.0, 0.8, 0~
## $ Month                   <chr> "Feb", "Feb", "Feb", "Feb", "Feb", "Feb", "Feb~
## $ OperatingSystems        <int> 1, 2, 4, 3, 3, 2, 2, 1, 2, 2, 1, 1, 1, 2, 3, 1~
## $ Browser                 <int> 1, 2, 1, 2, 3, 2, 4, 2, 2, 4, 1, 1, 1, 5, 2, 1~
## $ Region                  <int> 1, 1, 9, 2, 1, 1, 3, 1, 2, 1, 3, 4, 1, 1, 3, 9~
## $ TrafficType             <int> 1, 2, 3, 4, 4, 3, 3, 5, 3, 2, 3, 3, 3, 3, 3, 3~
## $ VisitorType             <chr> "Returning_Visitor", "Returning_Visitor", "Ret~
## $ Weekend                 <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE~
## $ Revenue                 <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALS~
```

```r
data_set_cst_ol_table <- data_set_cst_ol

## Creating the Dummy Variable on the features "Visitor Type" and "Weekend".

data_set_categorical <- caret::dummyVars("~ VisitorType+Weekend",data = data_set_cst_ol)

data_set_categorical <- data.frame(predict(data_set_categorical,newdata = data_set_cst_ol))

data_set_categorical %>%  glimpse()
```

```
## Rows: 12,330
## Columns: 5
## $ VisitorTypeNew_Visitor       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ VisitorTypeOther             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ VisitorTypeReturning_Visitor <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ WeekendFALSE                 <dbl> 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1,~
## $ WeekendTRUE                  <dbl> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,~
```

```r
data_set_cst_ol <- data_set_cst_ol %>% select(-c(Month,VisitorType,Weekend))

data_set_cst_ol['New_Visitor'] <- data_set_categorical$VisitorTypeNew_Visitor

data_set_cst_ol['Other'] <- data_set_categorical$VisitorTypeOther
```

```
data_set_cst_ol['Returning_Visitor'] <- data_set_categorical$VisitorTypeReturning_Visitor

data_set_cst_ol['Weekend_False'] <- data_set_categorical$WeekendFALSE

data_set_cst_ol['Weekend_True'] <- data_set_categorical$WeekendTRUE

data_set_cst_ol$Revenue <- ifelse(data_set_cst_ol$Revenue==TRUE,1,0)

data_set_cst_ol %>%  glimpse()
```

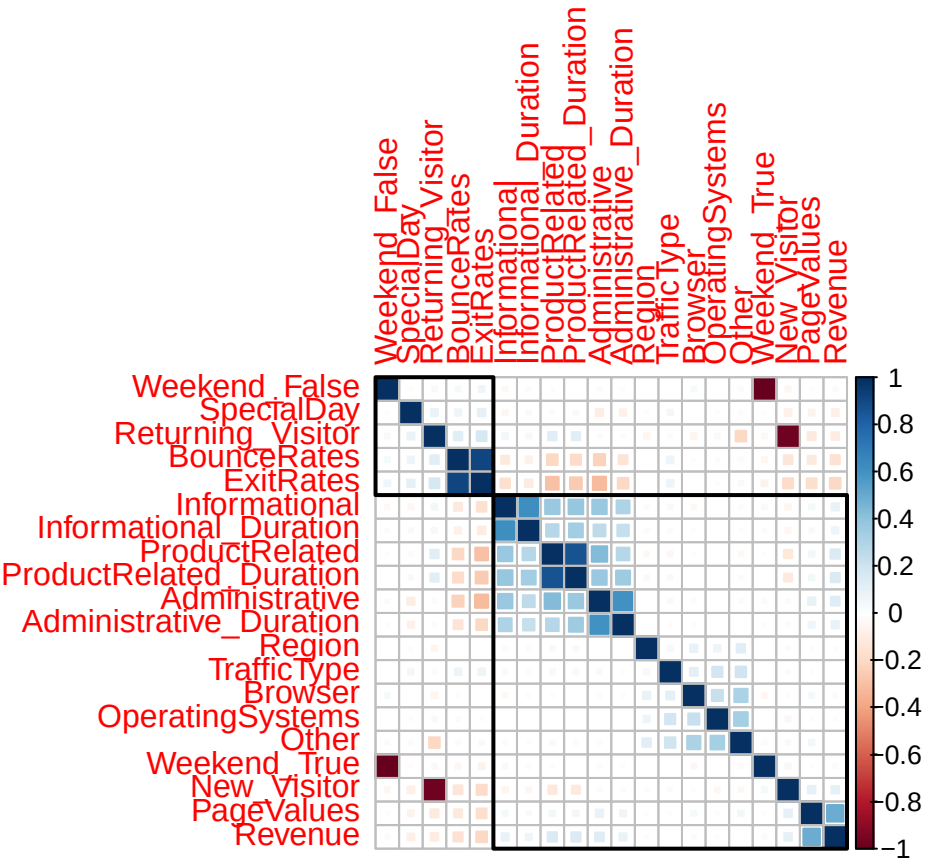
```
## Rows: 12,330
## Columns: 20
## $ Administrative          <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 2~
## $ Administrative_Duration <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5~
## $ Informational           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Informational_Duration  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ProductRelated          <int> 1, 2, 1, 2, 10, 19, 1, 0, 2, 3, 3, 16, 7, 6, 2~
## $ ProductRelated_Duration <dbl> 0.000000, 64.000000, 0.000000, 2.666667, 627.5~
## $ BounceRates             <dbl> 0.200000000, 0.000000000, 0.200000000, 0.05000~
## $ ExitRates               <dbl> 0.200000000, 0.100000000, 0.200000000, 0.14000~
## $ PageValues              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ SpecialDay              <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.4, 0.0, 0.8, 0~
## $ OperatingSystems        <int> 1, 2, 4, 3, 3, 2, 2, 1, 2, 2, 1, 1, 1, 2, 3, 1~
## $ Browser                 <int> 1, 2, 1, 2, 3, 2, 4, 2, 2, 4, 1, 1, 1, 5, 2, 1~
## $ Region                  <int> 1, 1, 9, 2, 1, 1, 3, 1, 2, 1, 3, 4, 1, 1, 3, 9~
## $ TrafficType             <int> 1, 2, 3, 4, 4, 3, 3, 5, 3, 2, 3, 3, 3, 3, 3, 3~
## $ Revenue                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ New_Visitor             <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Other                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Returning_Visitor       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Weekend_False           <dbl> 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1~
## $ Weekend_True            <dbl> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0~
```

```
describe(data_set_cst_ol)
```

```
##                          vars     n    mean      sd median trimmed    mad min
## Administrative              1 12330    2.32    3.32   1.00    1.63   1.48   0
## Administrative_Duration     2 12330   80.82  176.78   7.50   42.10  11.12   0
## Informational               3 12330    0.50    1.27   0.00    0.18   0.00   0
## Informational_Duration      4 12330   34.47  140.75   0.00    3.59   0.00   0
## ProductRelated              5 12330   31.73   44.48  18.00   22.75  19.27   0
## ProductRelated_Duration     6 12330 1194.75 1913.67 598.94  820.08 742.69   0
## BounceRates                 7 12330    0.02    0.05   0.00    0.01   0.00   0
## ExitRates                   8 12330    0.04    0.05   0.03    0.03   0.02   0
## PageValues                  9 12330    5.89   18.57   0.00    1.29   0.00   0
## SpecialDay                 10 12330    0.06    0.20   0.00    0.00   0.00   0
## OperatingSystems           11 12330    2.12    0.91   2.00    2.06   0.00   1
## Browser                    12 12330    2.36    1.72   2.00    2.00   0.00   1
## Region                     13 12330    3.15    2.40   3.00    2.79   2.97   1
## TrafficType                14 12330    4.07    4.03   2.00    3.21   1.48   1
## Revenue                    15 12330    0.15    0.36   0.00    0.07   0.00   0
## New_Visitor                16 12330    0.14    0.34   0.00    0.05   0.00   0
```

```
## Other                     17 12330    0.01    0.08   0.00    0.00   0.00   0
## Returning_Visitor         18 12330    0.86    0.35   1.00    0.94   0.00   0
## Weekend_False             19 12330    0.77    0.42   1.00    0.83   0.00   0
## Weekend_True              20 12330    0.23    0.42   0.00    0.17   0.00   0
##                             max    range  skew kurtosis    se
## Administrative            27.00    27.00  1.96     4.70  0.03
## Administrative_Duration 3398.75  3398.75  5.61    50.53  1.59
## Informational             24.00    24.00  4.04    26.92  0.01
## Informational_Duration  2549.38  2549.38  7.58    76.27  1.27
## ProductRelated           705.00   705.00  4.34    31.19  0.40
## ProductRelated_Duration 63973.52 63973.52  7.26   137.10 17.23
## BounceRates                0.20     0.20  2.95     7.72  0.00
## ExitRates                  0.20     0.20  2.15     4.01  0.00
## PageValues               361.76   361.76  6.38    65.60  0.17
## SpecialDay                 1.00     1.00  3.30     9.91  0.00
## OperatingSystems           8.00     7.00  2.07    10.45  0.01
## Browser                   13.00    12.00  3.24    12.74  0.02
## Region                     9.00     8.00  0.98    -0.15  0.02
## TrafficType               20.00    19.00  1.96     3.48  0.04
## Revenue                    1.00     1.00  1.91     1.64  0.00
## New_Visitor                1.00     1.00  2.11     2.44  0.00
## Other                      1.00     1.00 11.92   140.04  0.00
## Returning_Visitor          1.00     1.00 -2.02     2.10  0.00
## Weekend_False              1.00     1.00 -1.27    -0.40  0.00
## Weekend_True               1.00     1.00  1.27    -0.40  0.00
```

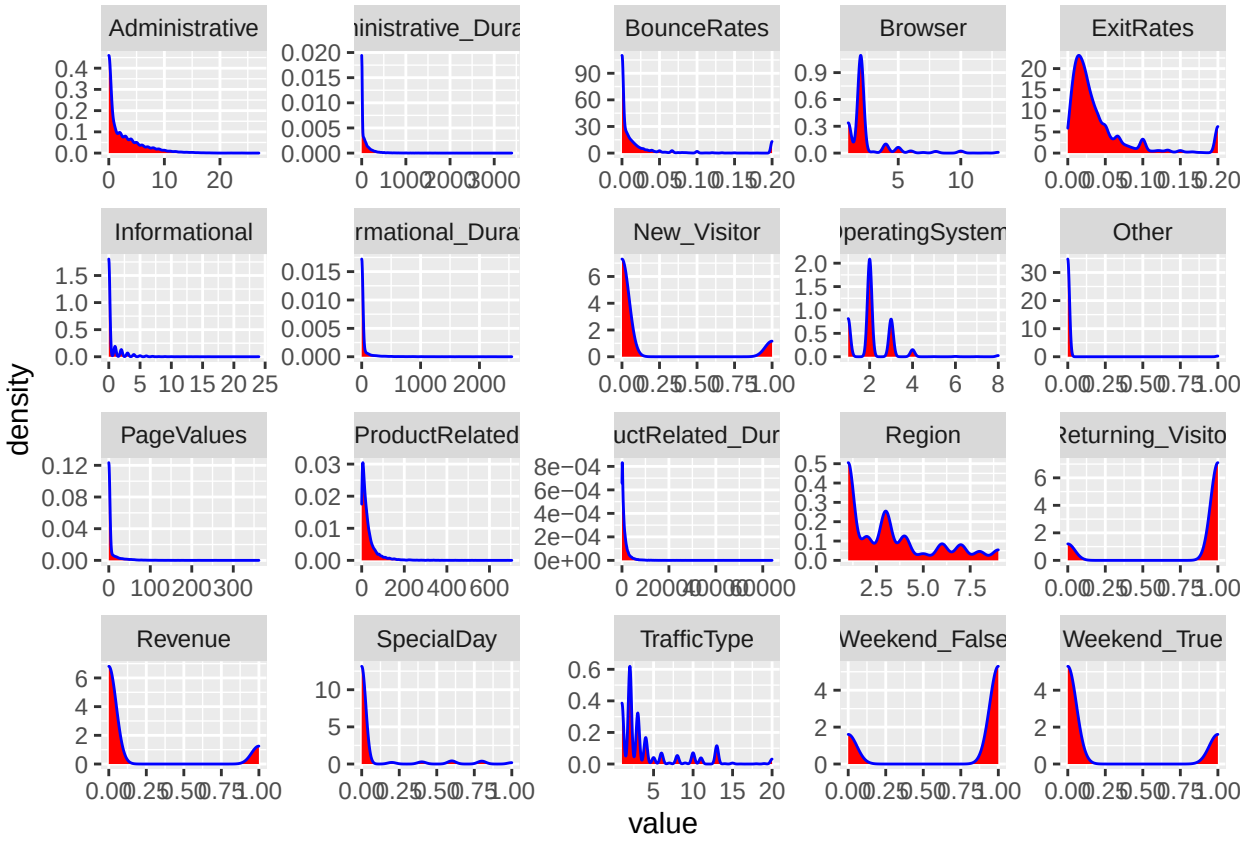
```r
corr_matrix <- cor(data_set_cst_ol)

corrplot(corr_matrix,method = 'square',order = 'hclust',addrect =  2,addCoefasPercent = TRUE)
```



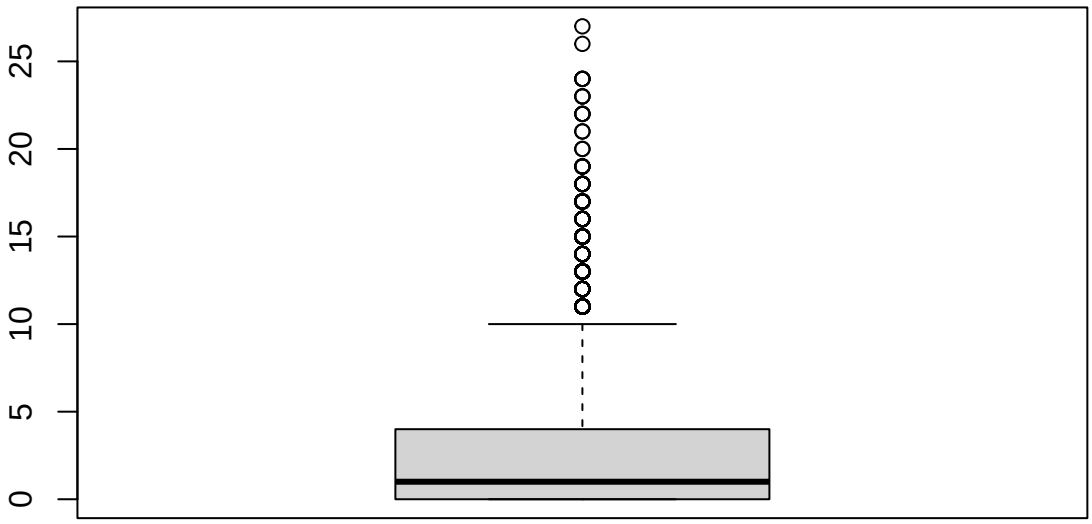

```
det(corr_matrix)
```

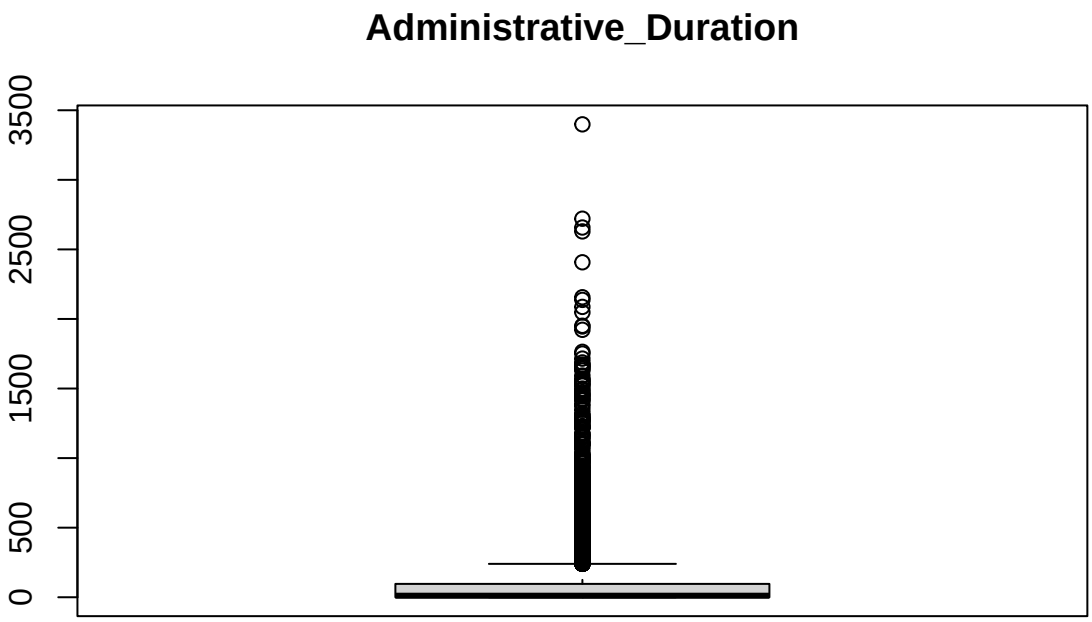
```
## [1] -1.812407e-34
```

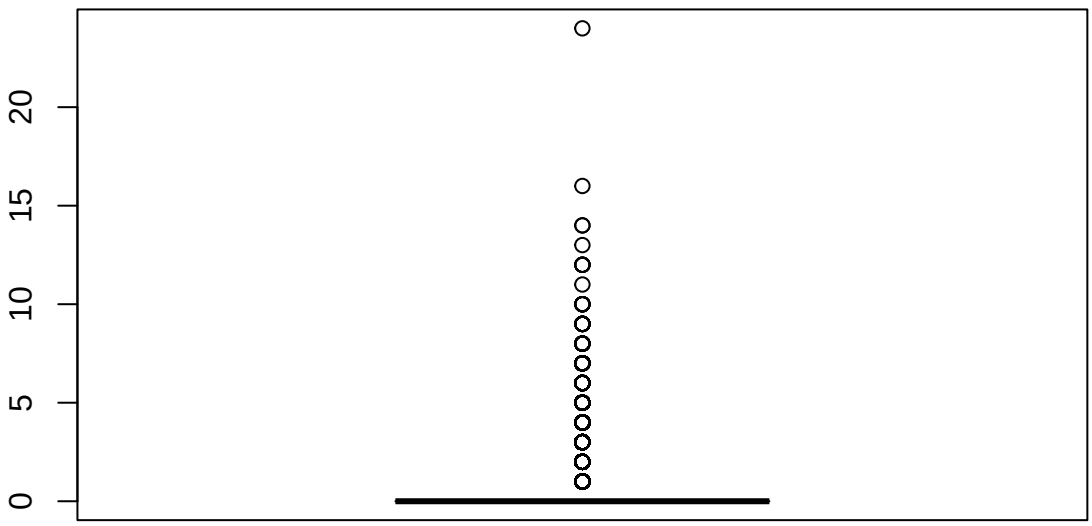
```
##########################################Density Plots#############################
data_set_cst_ol %>% keep(is.numeric) %>% gather() %>% ggplot(aes(value)) +
facet_wrap(~ key, scales = "free")+ geom_density(color = "blue",fill = "red")
```

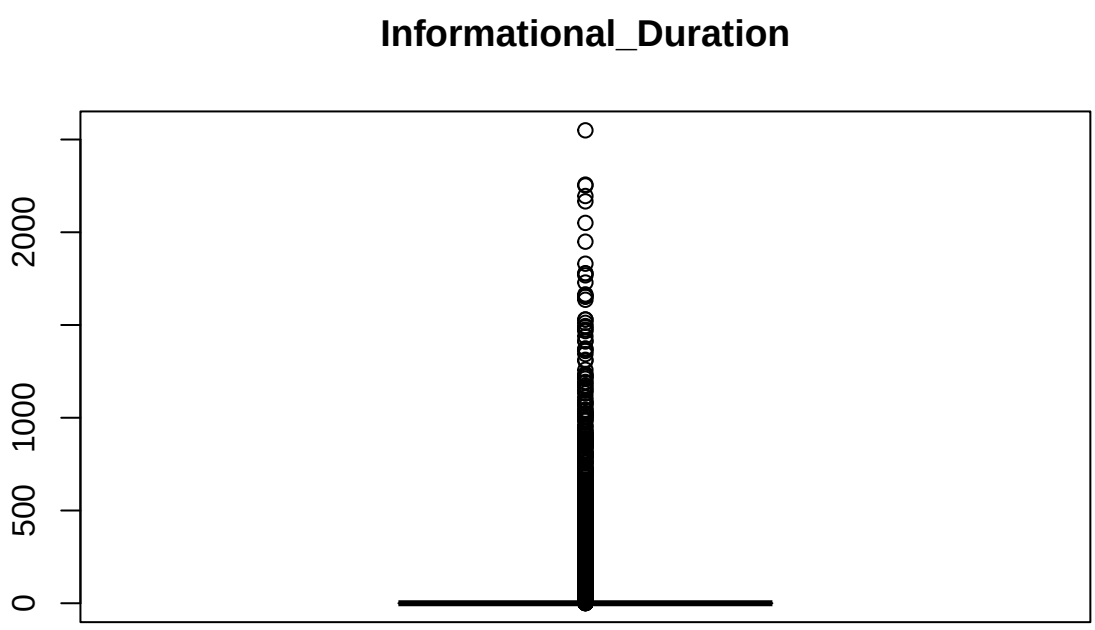


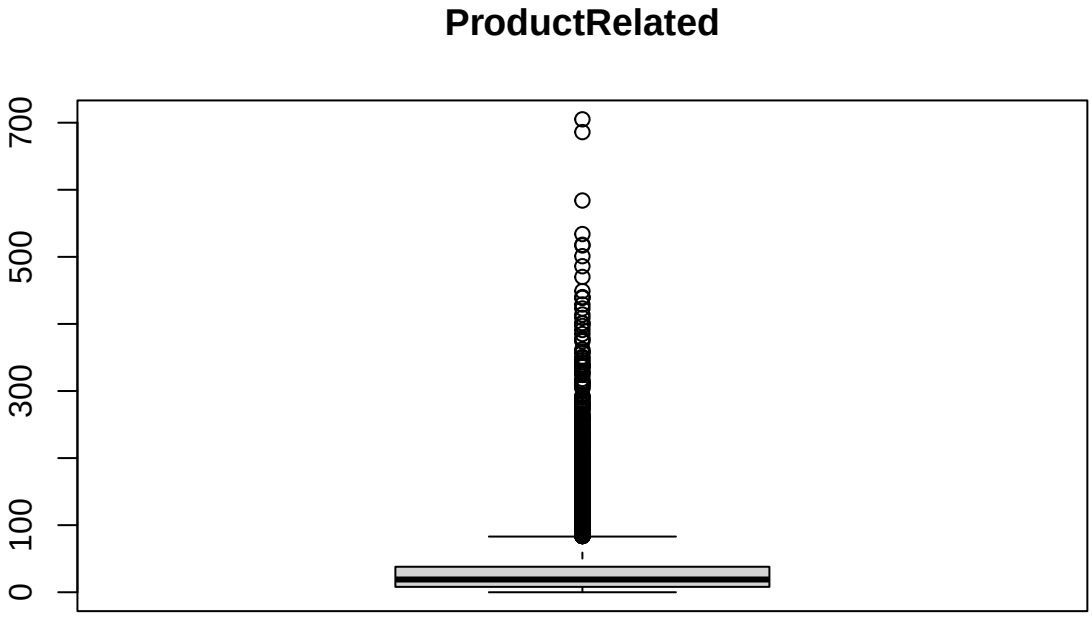

```
##############################Remove the Duplicates Values#########################
duplicates <- duplicated(data_set_cst_ol)
data_set_cst_ol_2 <- data_set_cst_ol[!duplicates,]

#View(data_set_cst_ol_2)

#######Identify and Remove the outliers keeping 99.5% of the data ##############
x_prev <- data_set_cst_ol_2[,1:20]
x_prev <- x_prev[,-15]

y_prev <- data_set_cst_ol_2[,15]

for( i in 1 : 19){
  boxplot(x_prev[,i],main = names(x_prev)[i])
}
```

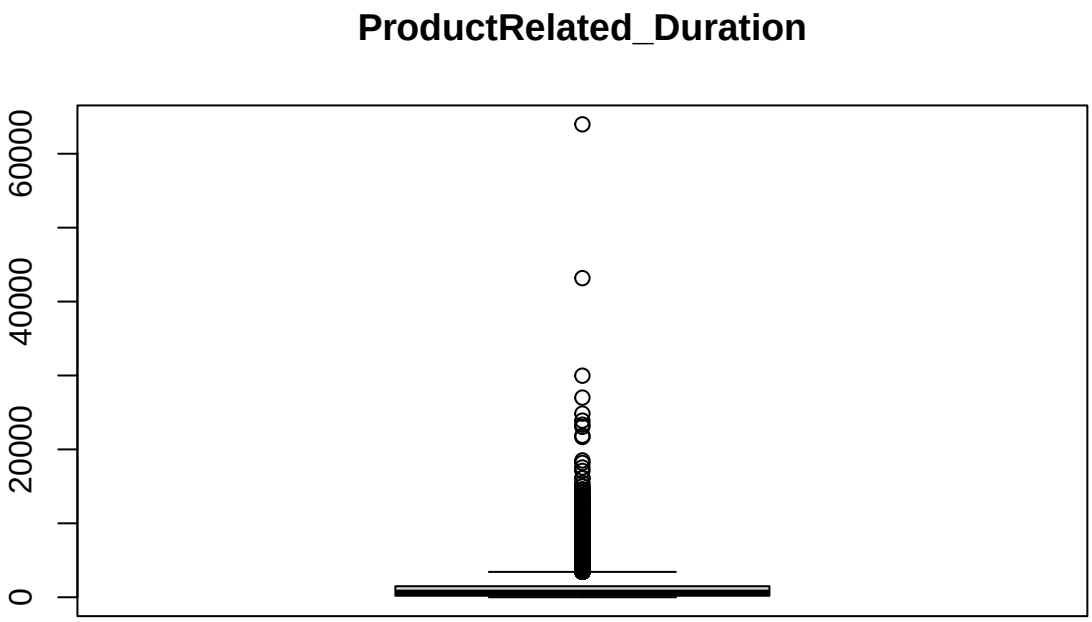

Administrative
0
5
10
15
20
25

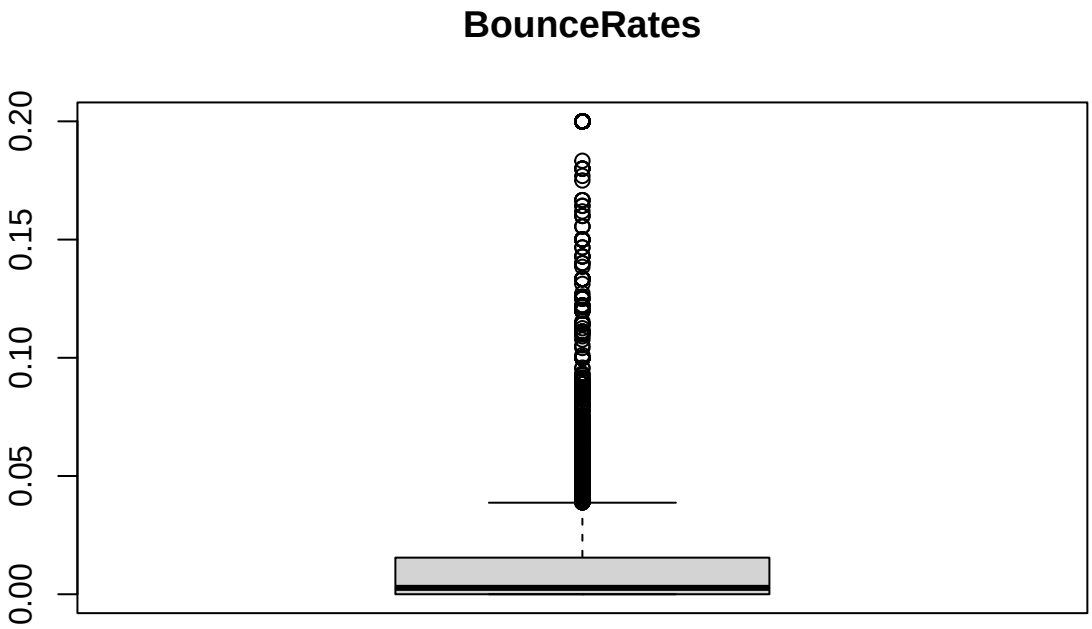

Administrative_Duration
0
500
1500
2500
3500

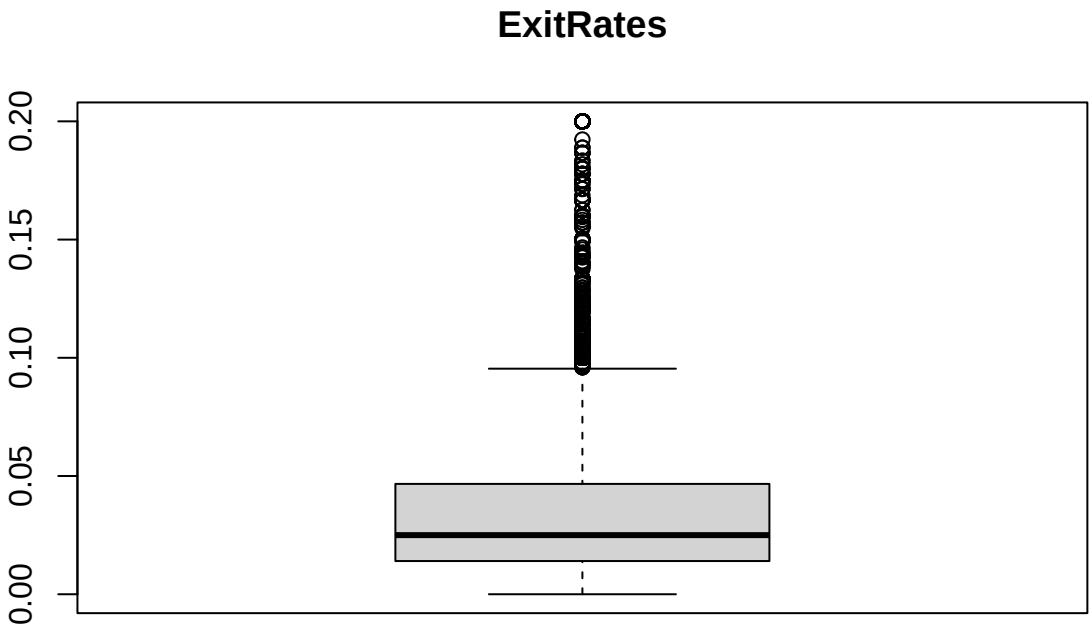

Informational
0
5
10
15
20

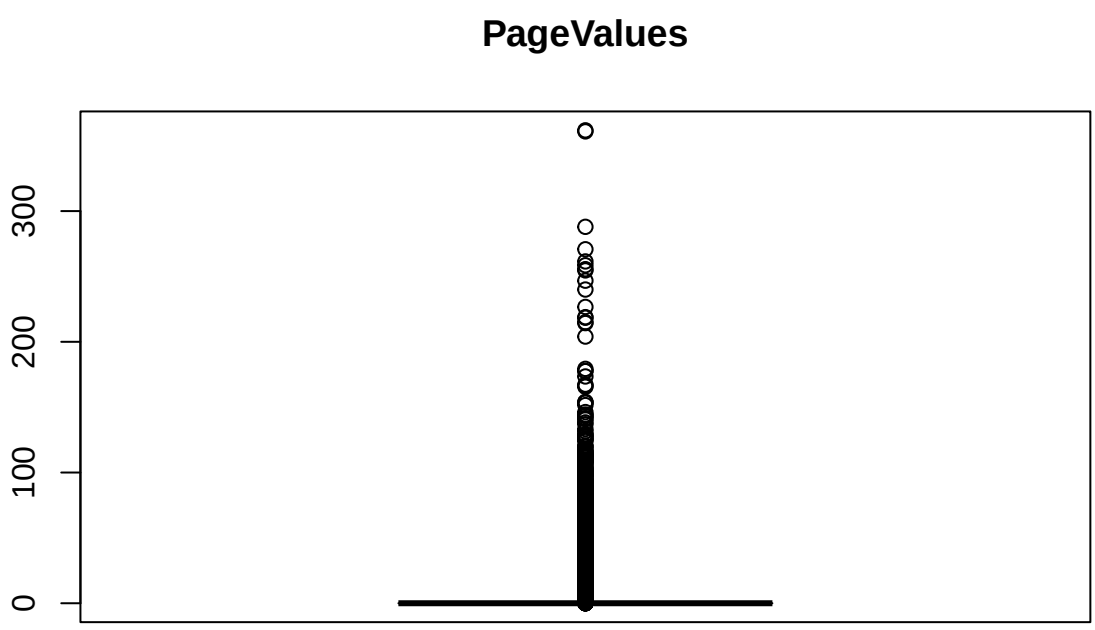

Informational_Duration
0
500
1000
2000

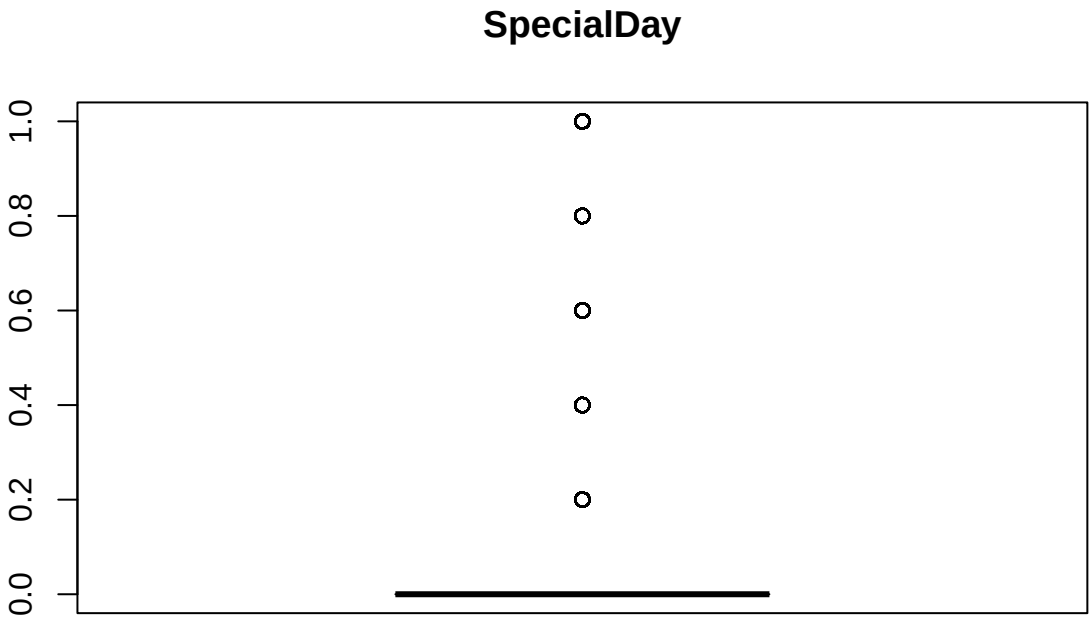

ProductRelated
0
100
300
500
700

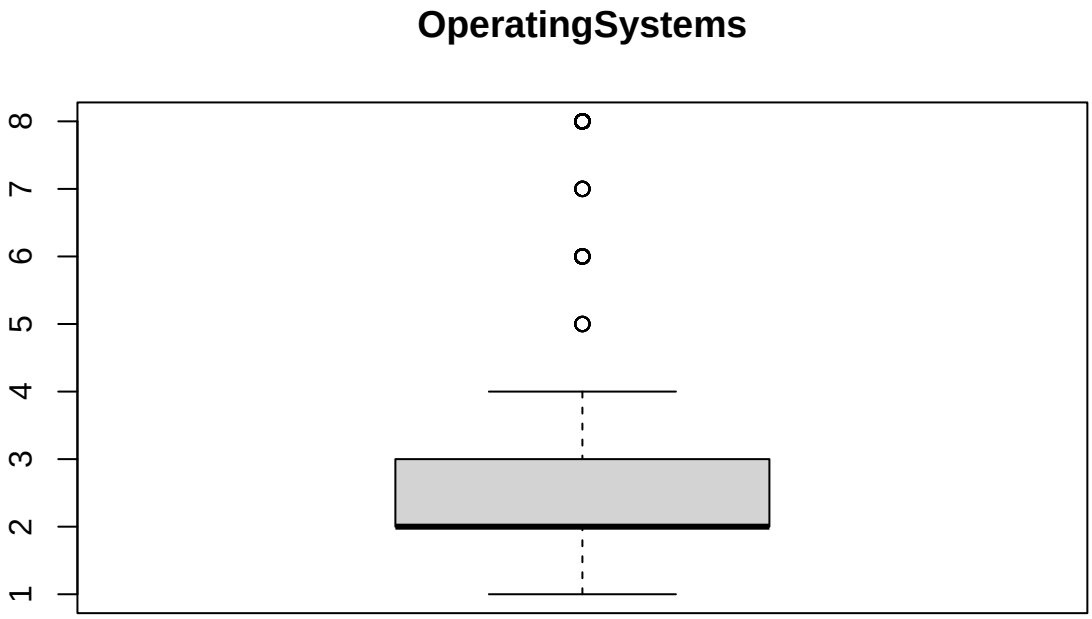

ProductRelated_Duration
0
20000
40000
60000

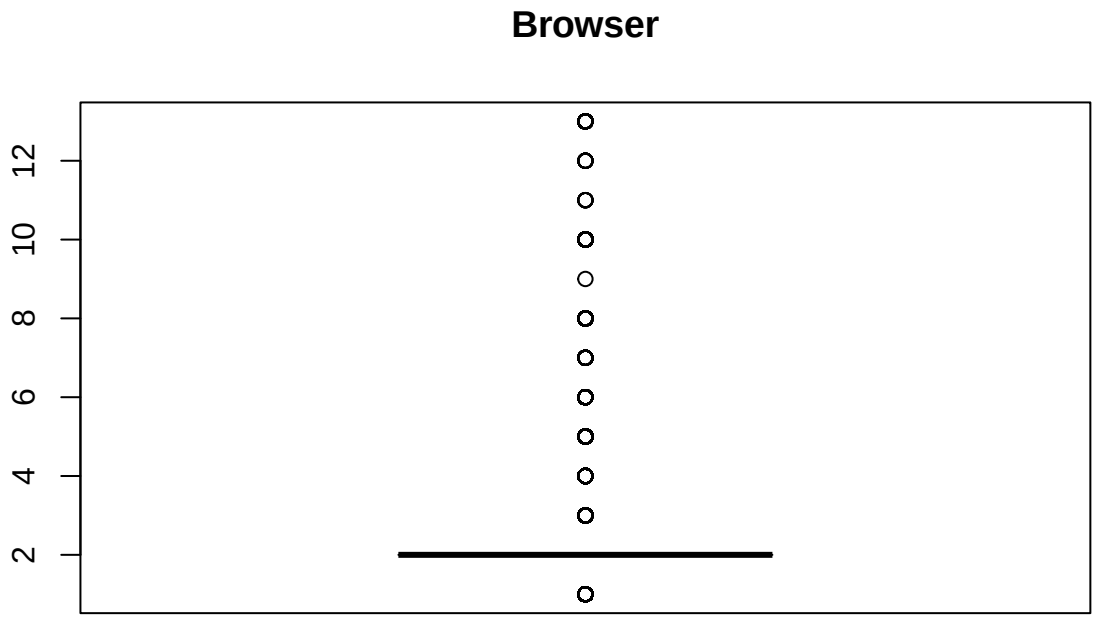

BounceRates
0.00
0.05
0.10
0.15
0.20

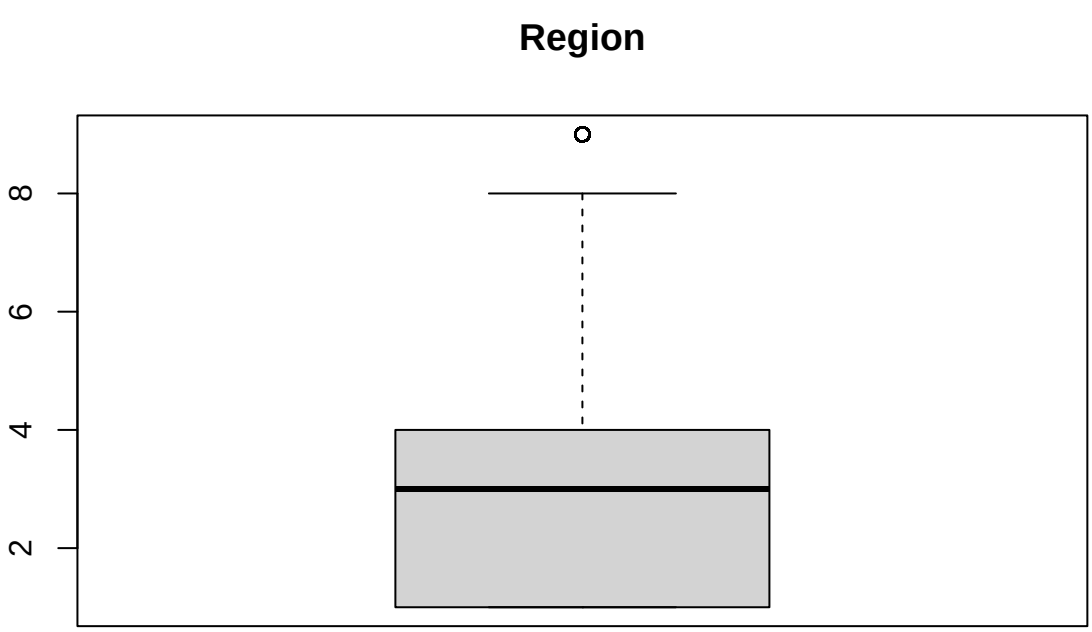

ExitRates
0.00
0.05
0.10
0.15
0.20

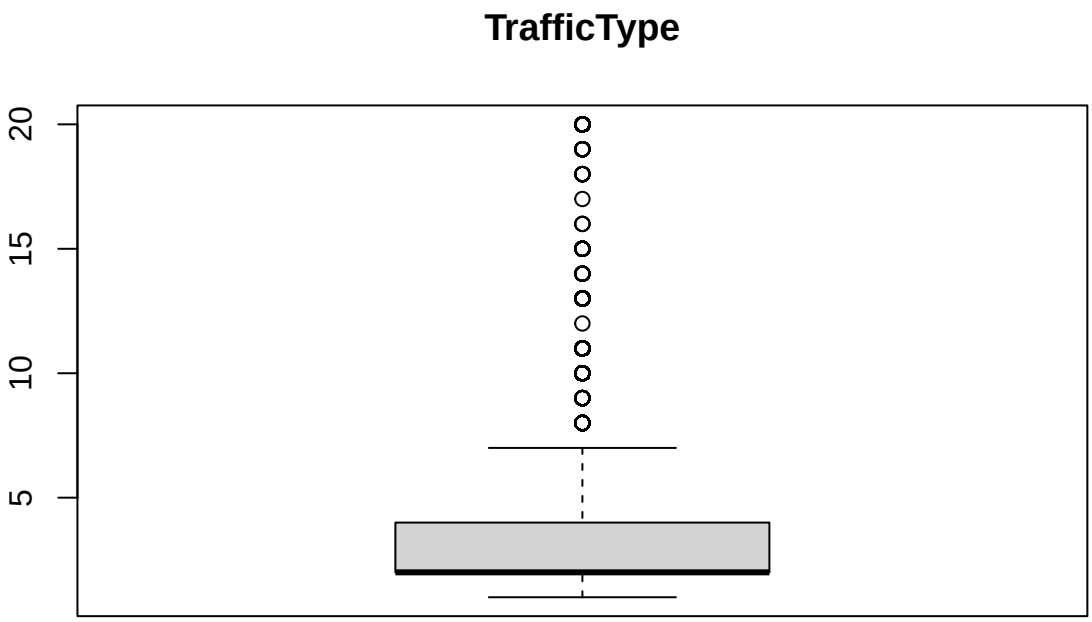

PageValues
0
100
200
300

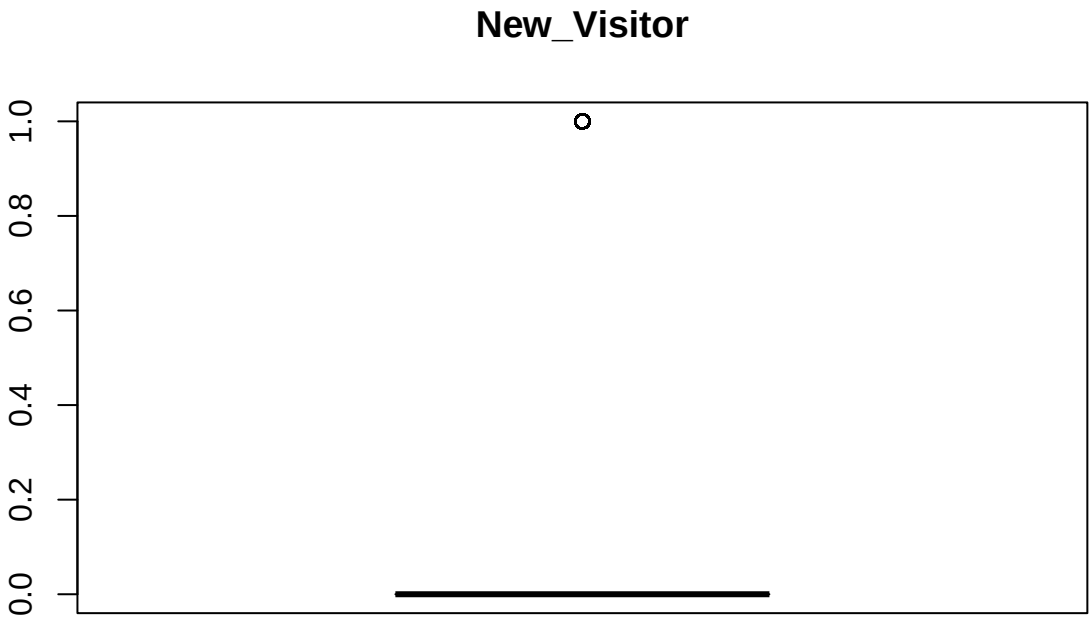

SpecialDay
0.0
0.2
0.4
0.6
0.8
1.0

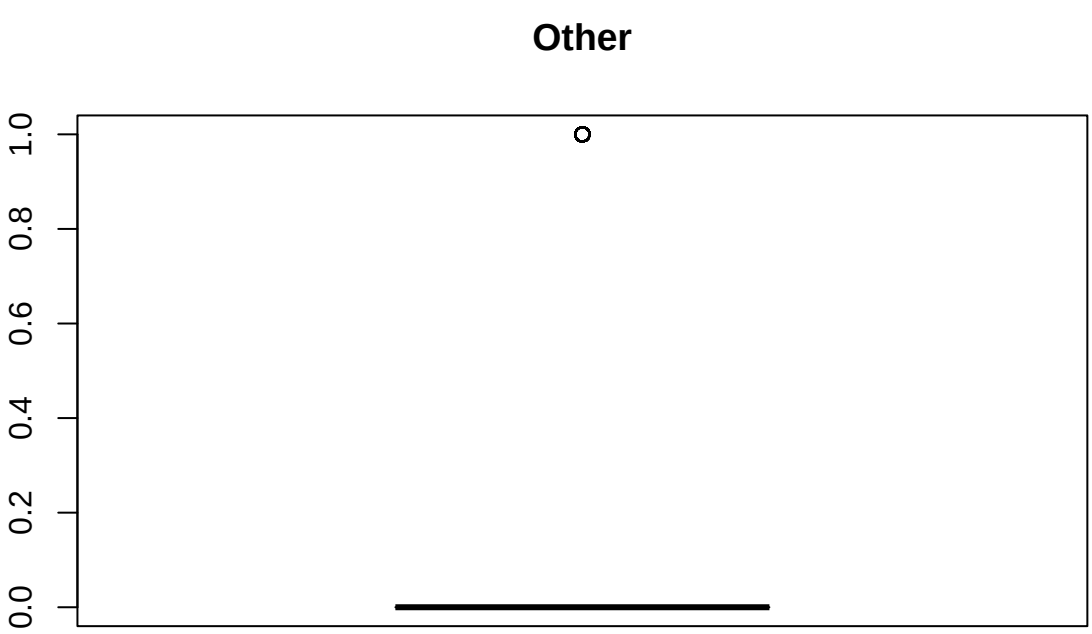

OperatingSystems
1
2
3
4
5
6
7
8

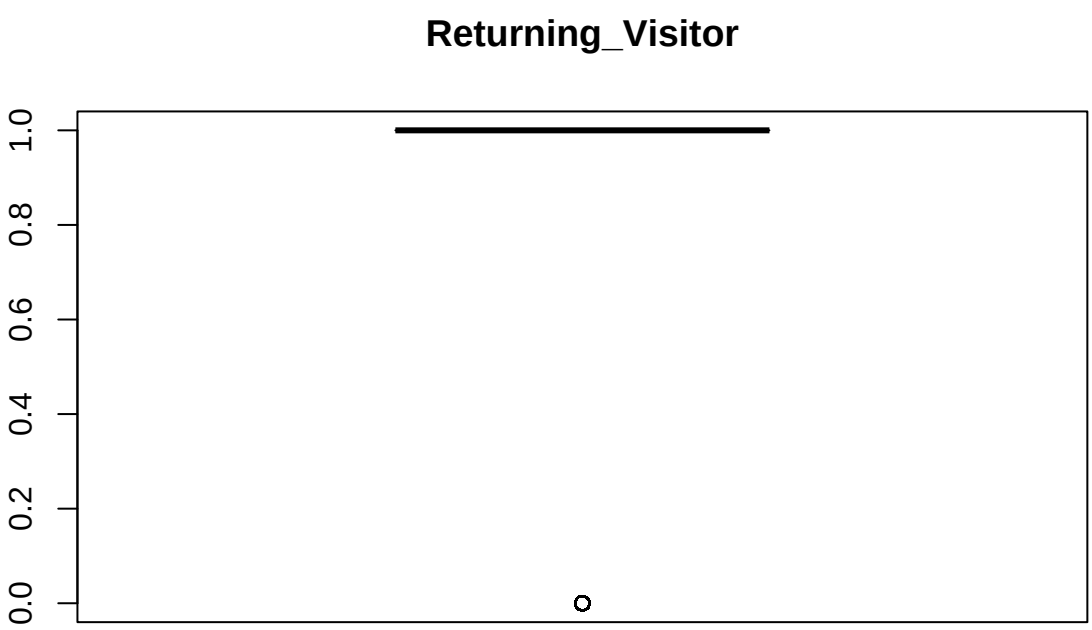

Browser
2
4
6
8
10
12


Region
2
4
6
8

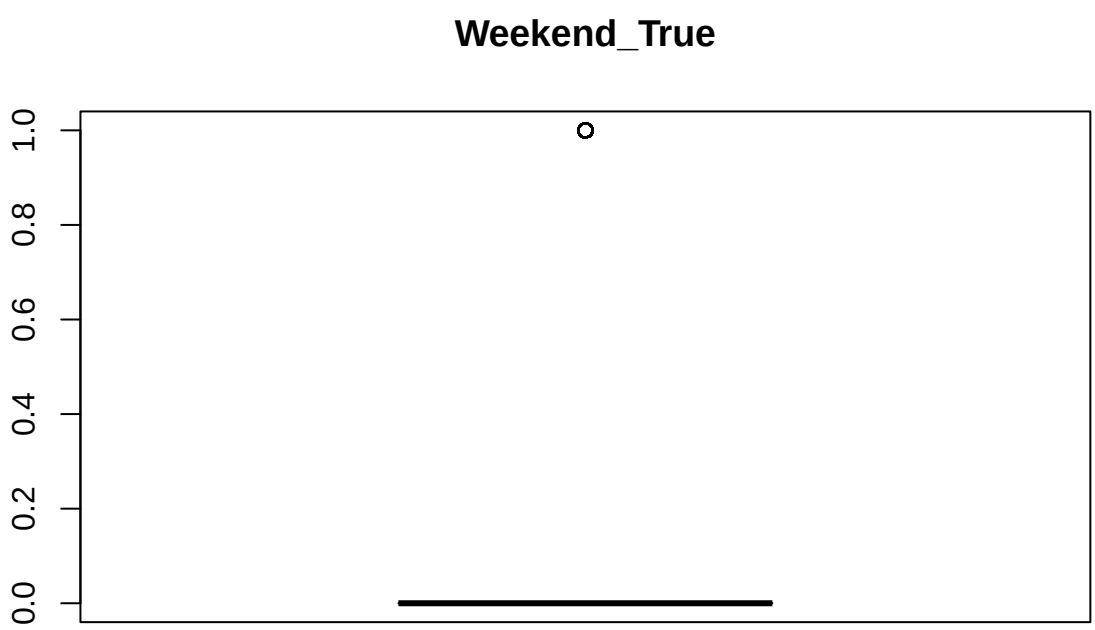

TrafficType
20
15
10
5


New_Visitor
0.0
0.2
0.4
0.6
0.8
1.0


Other
0.0
0.2
0.4
0.6
0.8
1.0


Returning_Visitor
0.0
0.2
0.4
0.6
0.8
1.0


Weekend_False
0.0
0.2
0.4
0.6
0.8
1.0



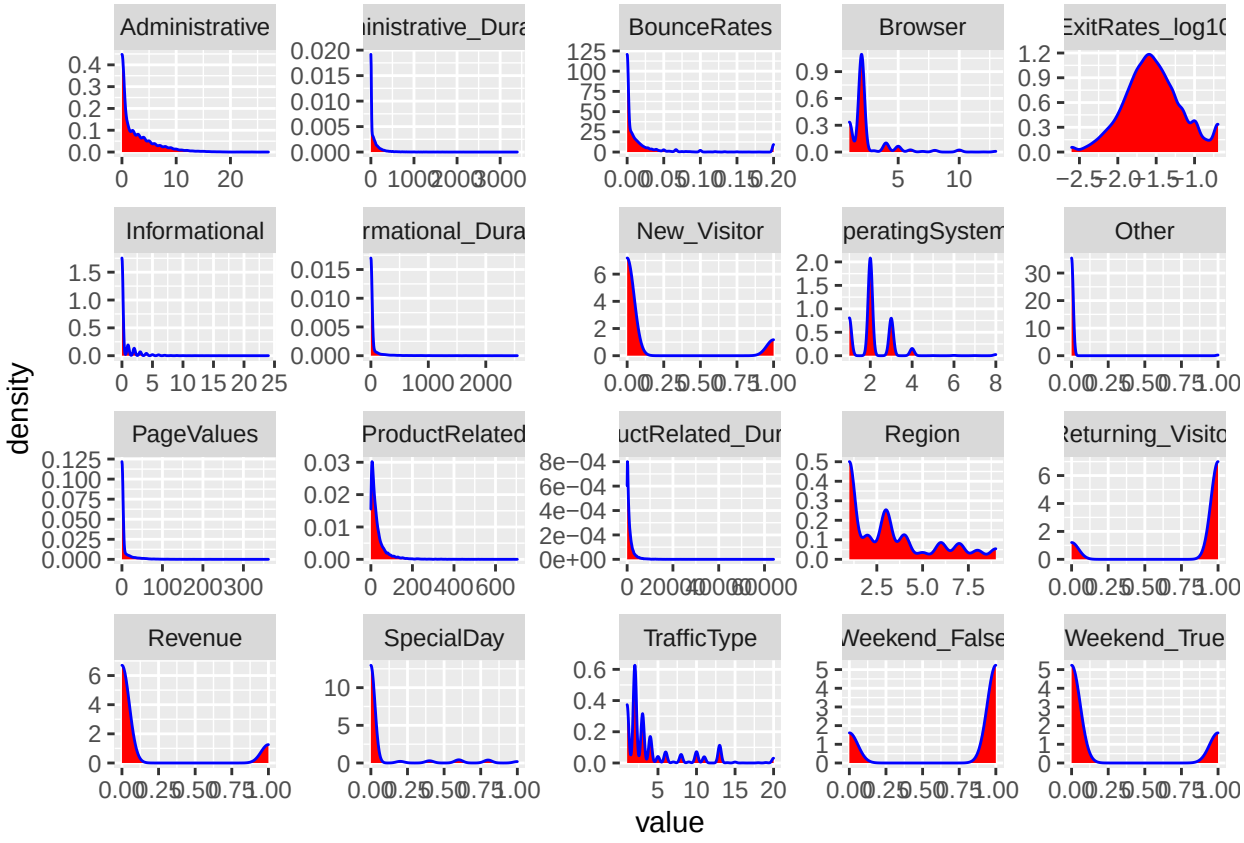

```
#####################################Identifying the Outliers##################
data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Administrative
                  > quantile(data_set_cst_ol_2$Administrative,probs = c(0.01,0.99))[2]
|data_set_cst_ol_2$Administrative < quantile(data_set_cst_ol_2$Administrative,probs = c(0.01,0.99))[1]

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Administrative_Duration
                                                > quantile(data_set_cst_ol_2$Administrative_Duration,pr

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Informational > quantile(data_set_cst_

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Informational_Duration > quantile(data

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$ProductRelated > quantile(data_set_cs

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$ProductRelated_Duration > quantile(da

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$BounceRates > quantile(data_set_cst_o

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$ExitRates > quantile(data_set_cst_ol_

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$PageValues > quantile(data_set_cst_ol_

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$SpecialDay > quantile(data_set_cst_ol_

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$OperatingSystems > quantile(data_set_
```

```r
data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Browser >
quantile(data_set_cst_ol_2$Browser,probs = c(0.01,0.99))[2] | data_set_cst_ol_2$Browser <
quantile(data_set_cst_ol_2$Browser,probs = c(0.01,0.99))[1] ))

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Region >
quantile(data_set_cst_ol_2$Region,probs = c(0.01,0.99))[2]| data_set_cst_ol_2$Region <
quantile(data_set_cst_ol_2$Region,probs = c(0.01,0.99))[1]))

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$TrafficType > quantile(data_set_cst_ol

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$New_Visitor > quantile(data_set_cst_ol

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Other >
quantile(data_set_cst_ol_2$Other,probs = c(0.01,0.99))[2] | data_set_cst_ol_2$Other < quantile(data_set_

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Returning_Visitor > quantile(data_set_
data_set_cst_ol_2$Returning_Visitor < quantile(data_set_cst_ol_2$Returning_Visitor,probs = c(0.01,0.99))

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Weekend_False > quantile(data_set_cst_

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Weekend_True >
quantile(data_set_cst_ol_2$Weekend_True,probs = c(0.01,0.99))[2] | data_set_cst_ol_2$Weekend_True < quar

data_set_cst_ol_3 <- subset(data_set_cst_ol_2,!(data_set_cst_ol_2$Revenue >
quantile(data_set_cst_ol_2$Revenue,probs = c(0.01,0.99))[2] | data_set_cst_ol_2$Revenue <
quantile(data_set_cst_ol_2$Revenue,probs = c(0.01,0.99))[1]))

#View(data_set_cst_ol_3)

##################Removing the Skewness in the Data#################################

data_set_cst_ol_3$ExitRates_log10 <- log10(0.0025+ data_set_cst_ol_3$ExitRates)
data_set_cst_ol_3$ExitRates_log10 <- data_set_cst_ol_3$ExitRates_log10
data_set_cst_ol_3 <- select(data_set_cst_ol_3,-ExitRates)

names(data_set_cst_ol_3)
```

```
##  [1] "Administrative"          "Administrative_Duration"
##  [3] "Informational"           "Informational_Duration"
##  [5] "ProductRelated"          "ProductRelated_Duration"
##  [7] "BounceRates"             "PageValues"
##  [9] "SpecialDay"              "OperatingSystems"
## [11] "Browser"                 "Region"
## [13] "TrafficType"             "Revenue"
## [15] "New_Visitor"             "Other"
## [17] "Returning_Visitor"       "Weekend_False"
## [19] "Weekend_True"            "ExitRates_log10"
```

```r
data_set_cst_ol_3 <- select(data_set_cst_ol_3,c(Administrative,Administrative_Duration,Informational,
                                                Informational_Duration,ProductRelated,ProductRelated_Du
                                                PageValues,BounceRates,SpecialDay,OperatingSystems,Brows
                                                TrafficType,New_Visitor,Returning_Visitor,Other,Weekend_
                                                Weekend_True,ExitRates_log10,BounceRates,Revenue))
```

```
data_set_cst_ol_3 %>% keep(is.numeric) %>% gather() %>% ggplot(aes(value)) + facet_wrap(~ key, scales =
  geom_density(color = "blue",fill = "red")
```

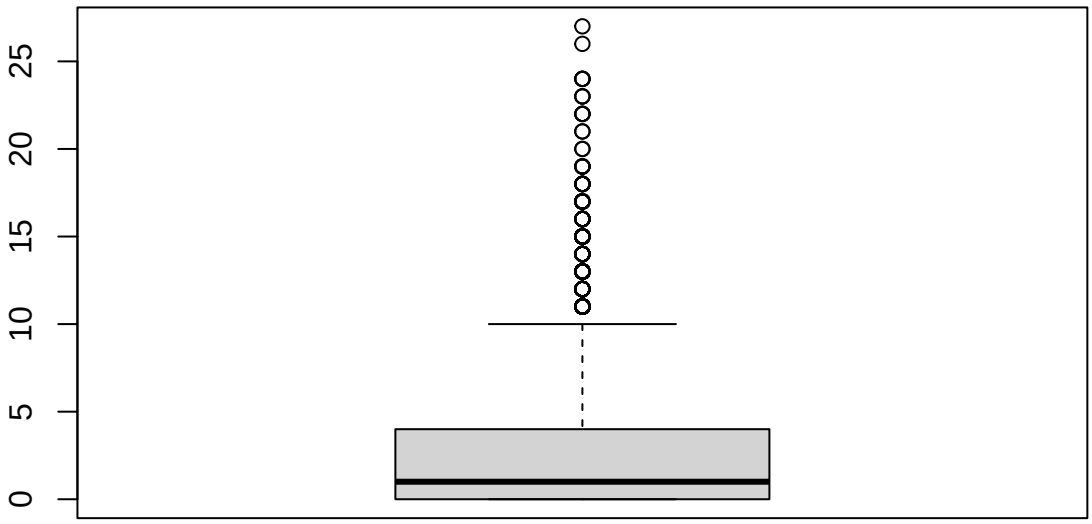


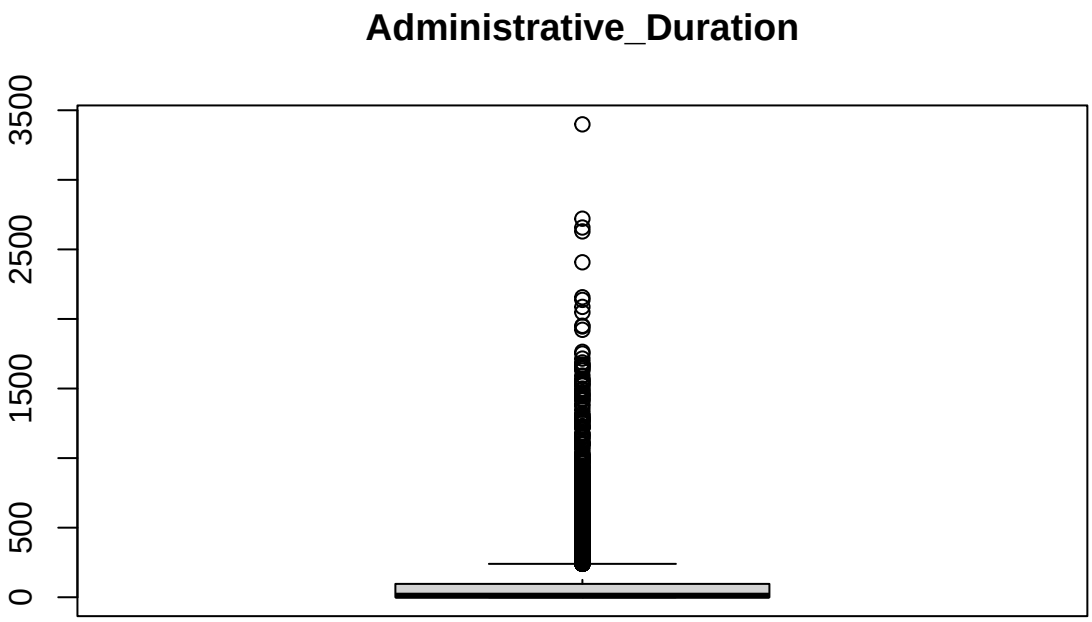

```
describe(data_set_cst_ol_3)
```

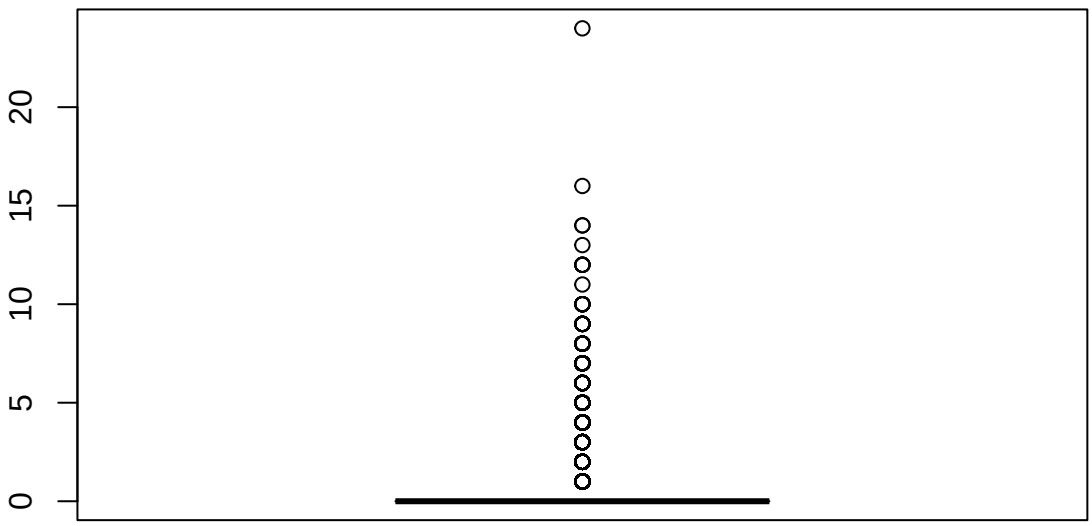
```
##                           vars     n    mean       sd median trimmed    mad  min
## Administrative               1 12079    2.36     3.34   1.00    1.68   1.48  0.0
## Administrative_Duration      2 12079   82.50   178.22  11.00   43.58  16.31  0.0
## Informational                3 12079    0.51     1.28   0.00    0.19   0.00  0.0
## Informational_Duration       4 12079   35.19   142.12   0.00    3.86   0.00  0.0
## ProductRelated               5 12079   32.37    44.71  19.00   23.35  19.27  0.0
## ProductRelated_Duration      6 12079 1219.57  1925.61 623.72  843.76 752.26  0.0
## PageValues                   7 12079    6.01    18.74   0.00    1.37   0.00  0.0
## BounceRates                  8 12079    0.02     0.04   0.00    0.01   0.00  0.0
## SpecialDay                   9 12079    0.06     0.20   0.00    0.00   0.00  0.0
## OperatingSystems            10 12079    2.13     0.91   2.00    2.06   0.00  1.0
## Browser                     11 12079    2.36     1.71   2.00    2.01   0.00  1.0
## Region                      12 12079    3.16     2.40   3.00    2.80   2.97  1.0
## TrafficType                 13 12079    4.08     4.01   2.00    3.23   1.48  1.0
## New_Visitor                 14 12079    0.14     0.35   0.00    0.05   0.00  0.0
## Returning_Visitor           15 12079    0.85     0.35   1.00    0.94   0.00  0.0
## Other                       16 12079    0.01     0.08   0.00    0.00   0.00  0.0
## Weekend_False               17 12079    0.76     0.42   1.00    0.83   0.00  0.0
```

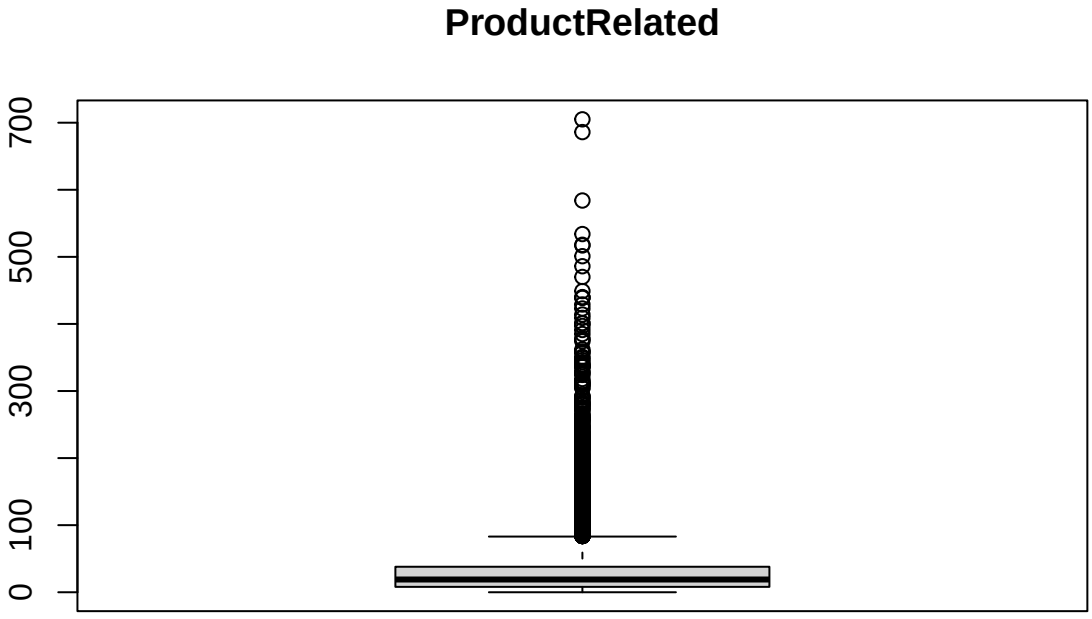
```
## Weekend_True              18 12079    0.24    0.42   0.00    0.17   0.00  0.0
## ExitRates_log10           19 12079   -1.54    0.38  -1.56   -1.55   0.35 -2.6
## Revenue                   20 12079    0.16    0.36   0.00    0.07   0.00  0.0
##                              max    range  skew kurtosis    se
## Administrative             27.00    27.00  1.93     4.58  0.03
## Administrative_Duration  3398.75  3398.75  5.57    49.68  1.62
## Informational              24.00    24.00  3.99    26.37  0.01
## Informational_Duration   2549.38  2549.38  7.50    74.70  1.29
## ProductRelated            705.00   705.00  4.32    30.92  0.41
## ProductRelated_Duration 63973.52 63973.52  7.24   136.12 17.52
## PageValues                361.76   361.76  6.32    64.32  0.17
## BounceRates                 0.20     0.20  3.41    11.40  0.00
## SpecialDay                  1.00     1.00  3.26     9.65  0.00
## OperatingSystems            8.00     7.00  2.03    10.26  0.01
## Browser                    13.00    12.00  3.21    12.49  0.02
## Region                      9.00     8.00  0.97    -0.17  0.02
## TrafficType                20.00    19.00  1.96     3.47  0.04
## New_Visitor                 1.00     1.00  2.07     2.30  0.00
## Returning_Visitor           1.00     1.00 -2.00     1.98  0.00
## Other                       1.00     1.00 12.16   145.97  0.00
## Weekend_False               1.00     1.00 -1.25    -0.45  0.00
## Weekend_True                1.00     1.00  1.25    -0.45  0.00
## ExitRates_log10            -0.69     1.91  0.11     0.03  0.00
## Revenue                     1.00     1.00  1.88     1.52  0.00
```

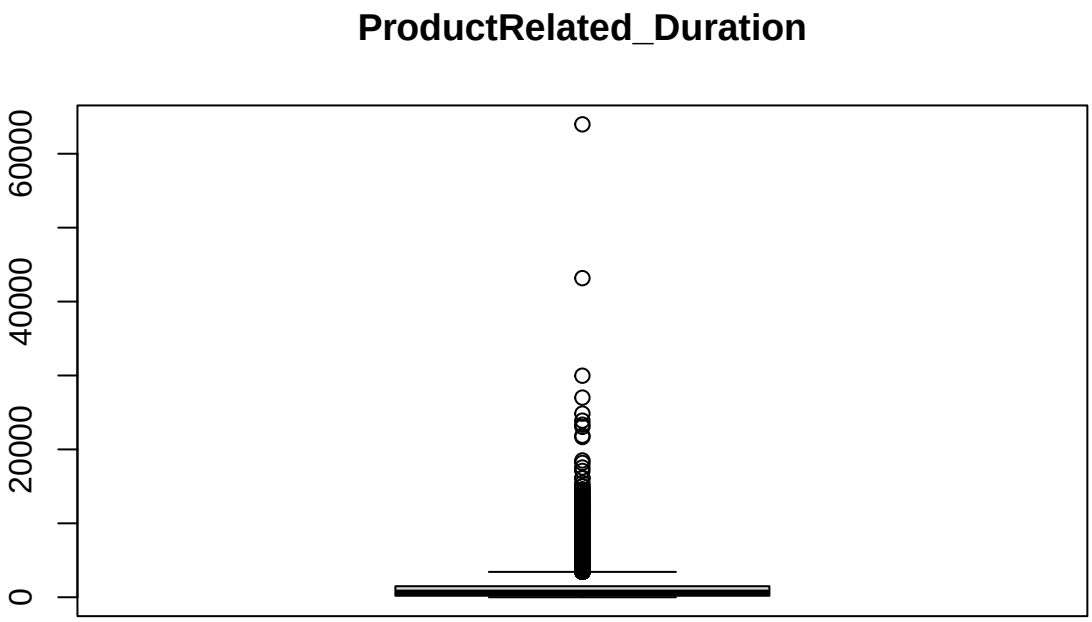
```
#########################Box Plot ################################################
x <- data_set_cst_ol_3[,1:20]
x <- x[,-4]
y <- data_set_cst_ol_3[,4]
for( i in 1 : 19){
  boxplot(x[,i],main = names(x)[i])
}
```

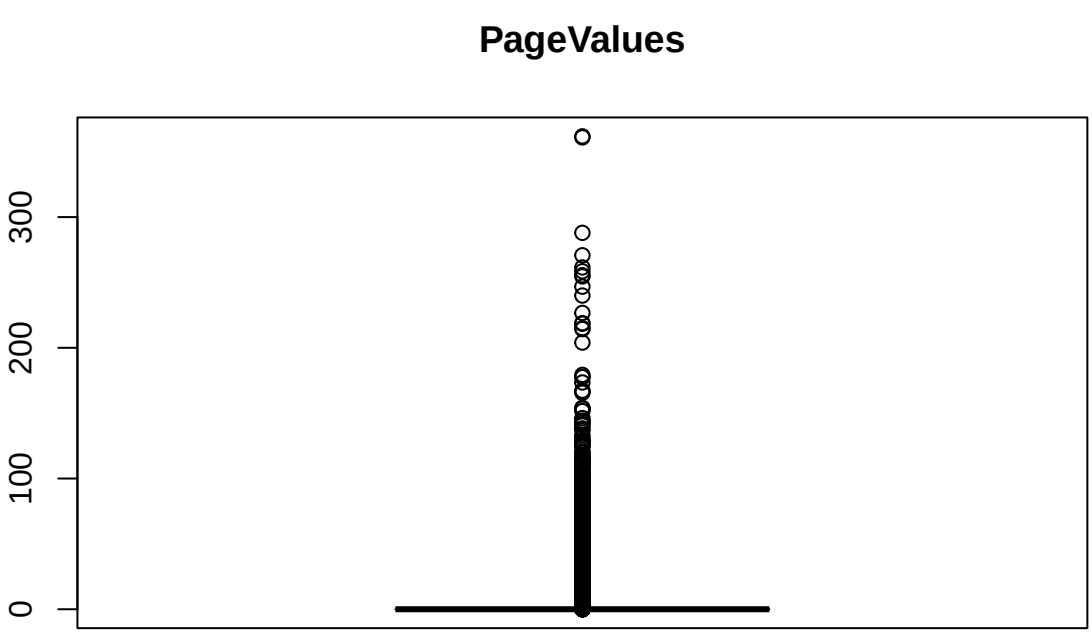

Administrative
0
5
10
15
20
25

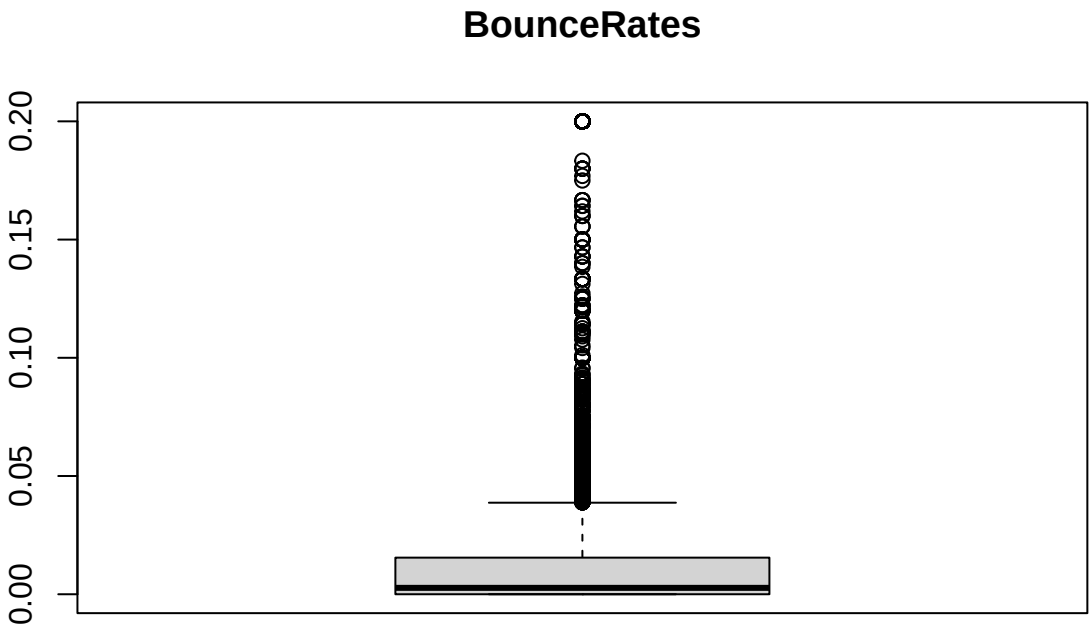

Administrative_Duration
0
500
1500
2500
3500

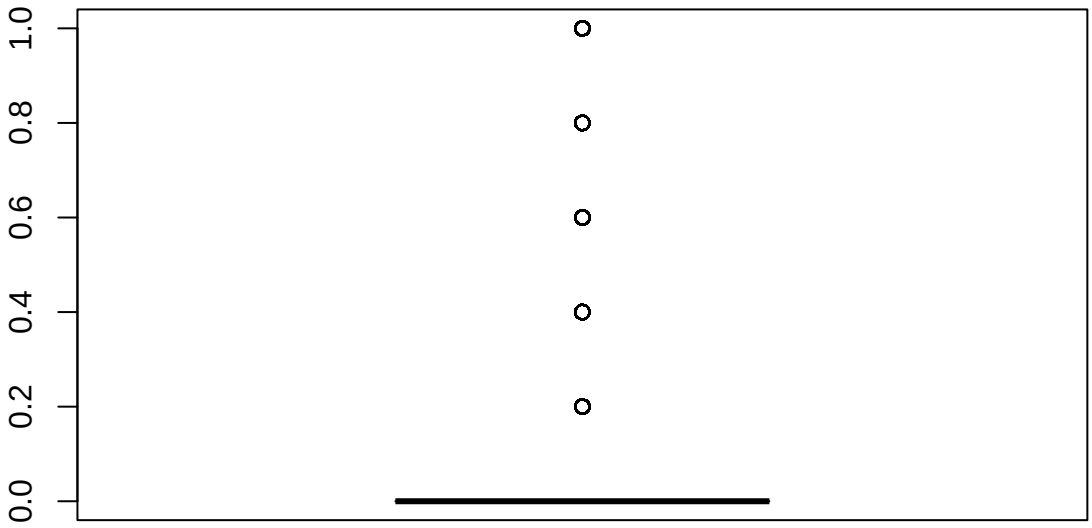

Informational
0
5
10
15
20

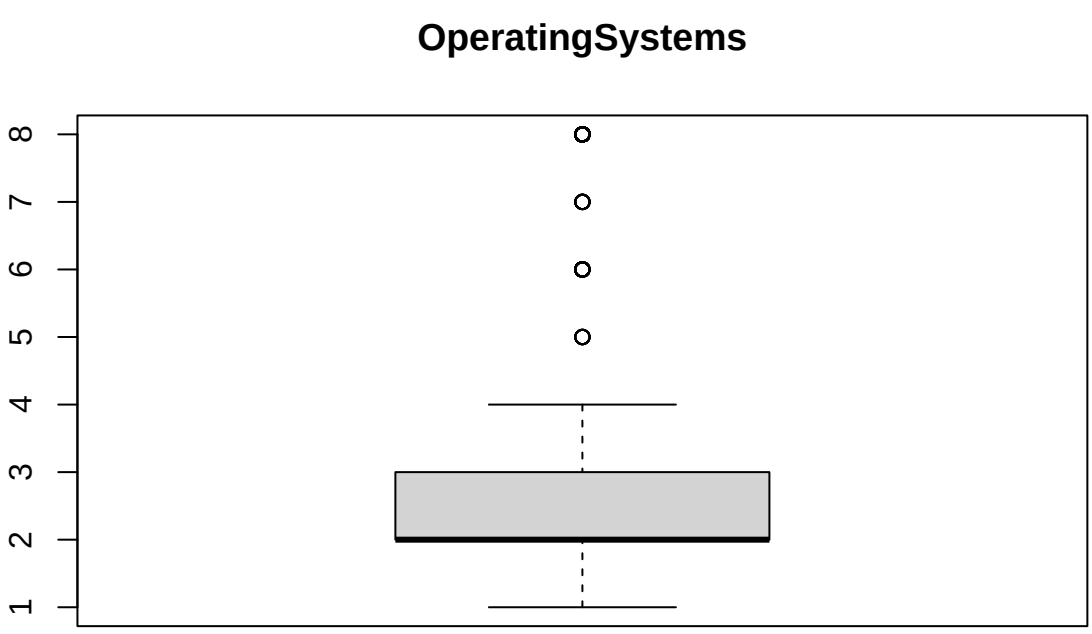

ProductRelated
0
100
300
500
700

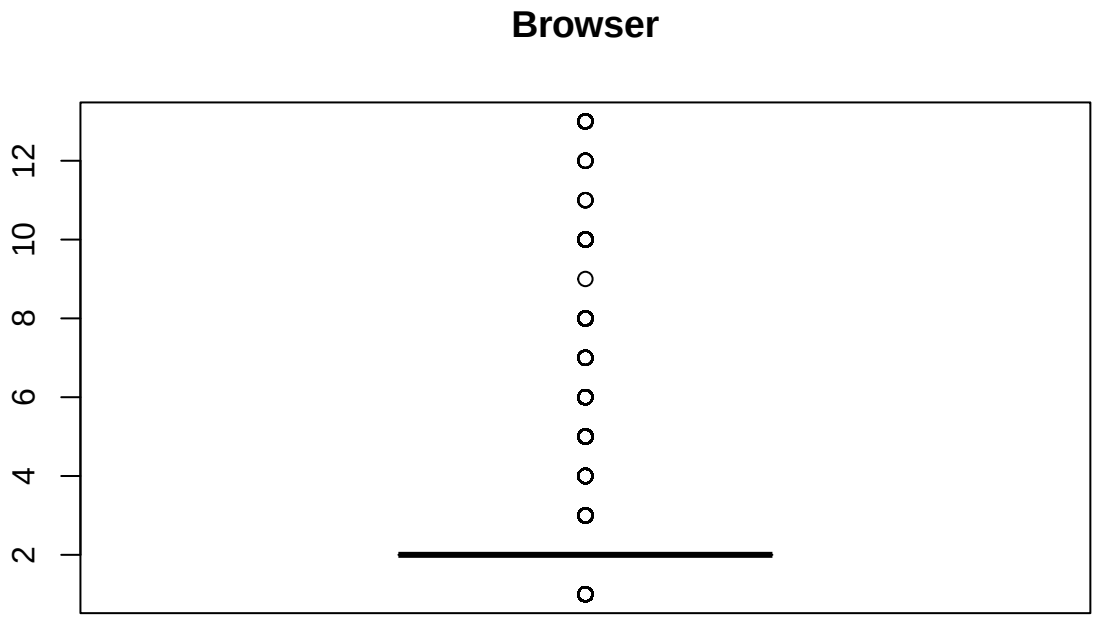

ProductRelated_Duration
0
20000
40000
60000

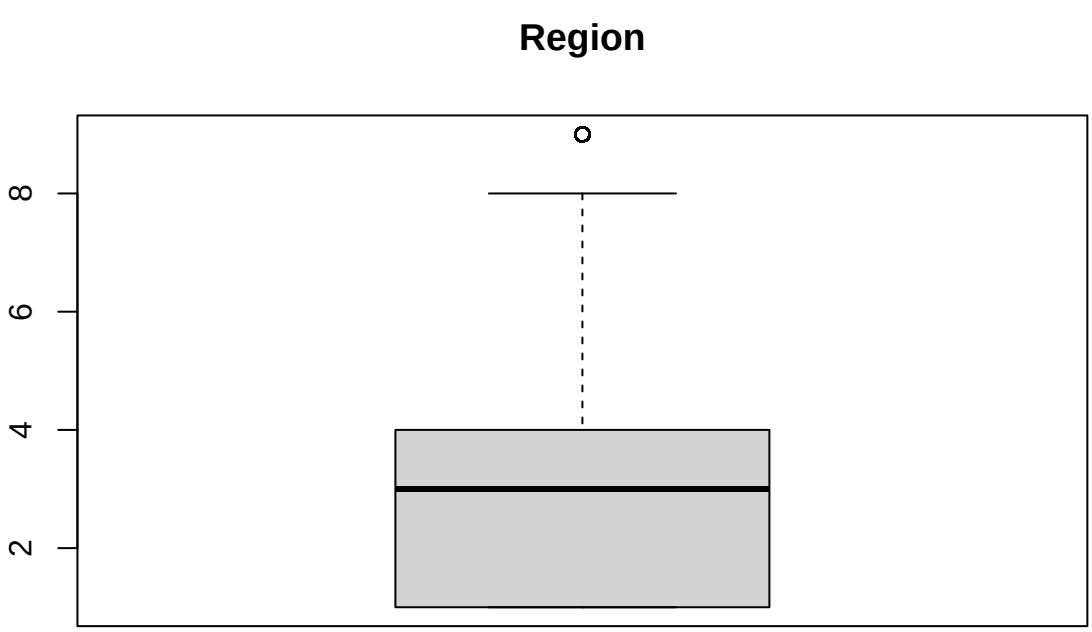

PageValues
0
100
200
300

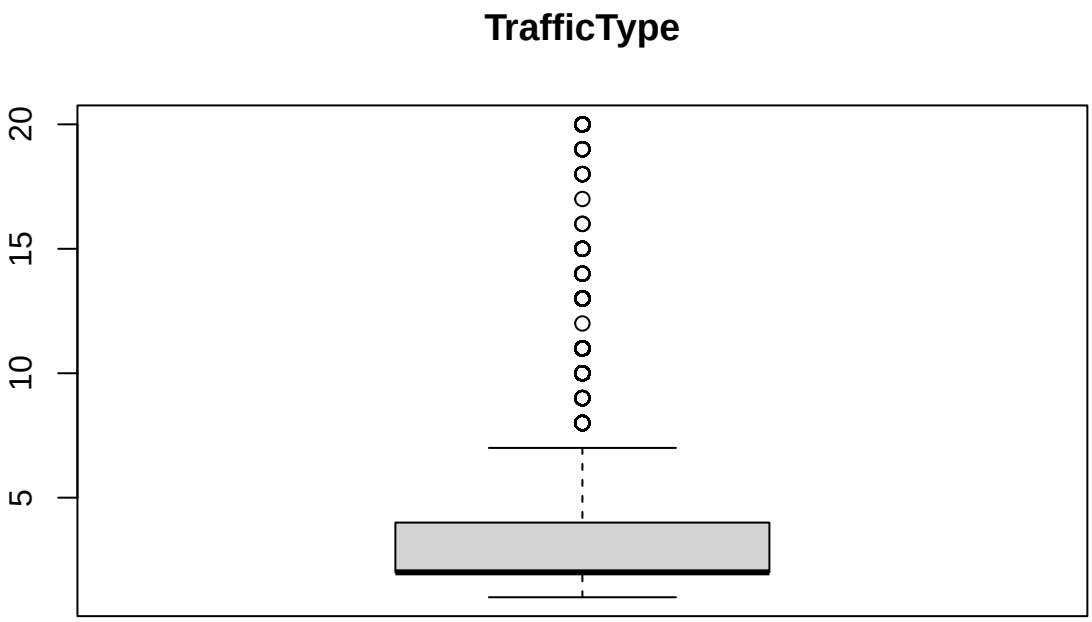

BounceRates
0.00
0.05
0.10
0.15
0.20

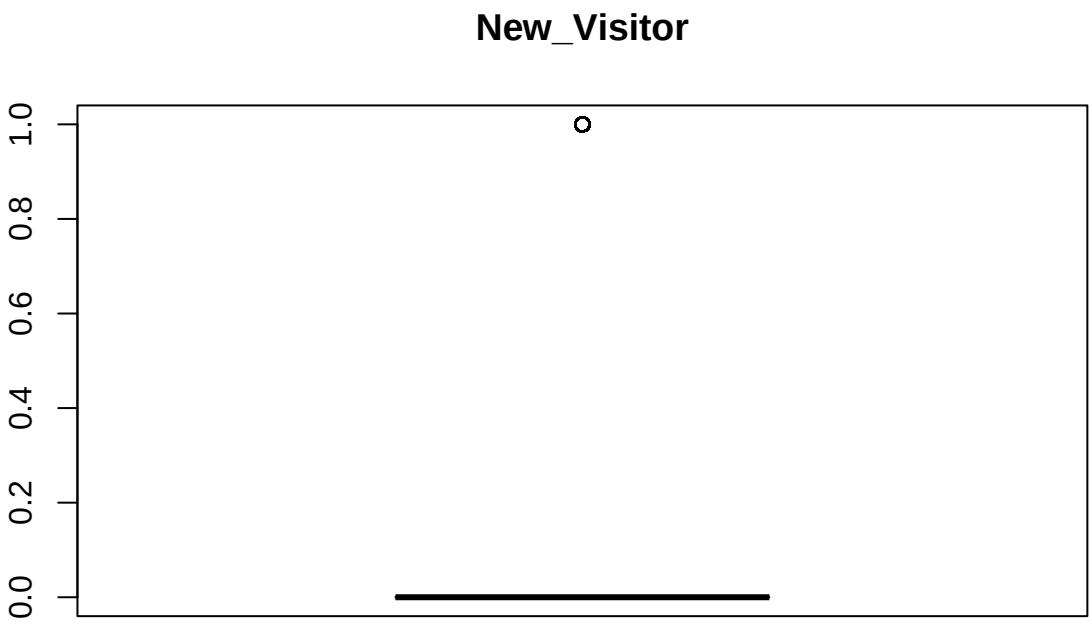

SpecialDay
0.0
0.2
0.4
0.6
0.8
1.0

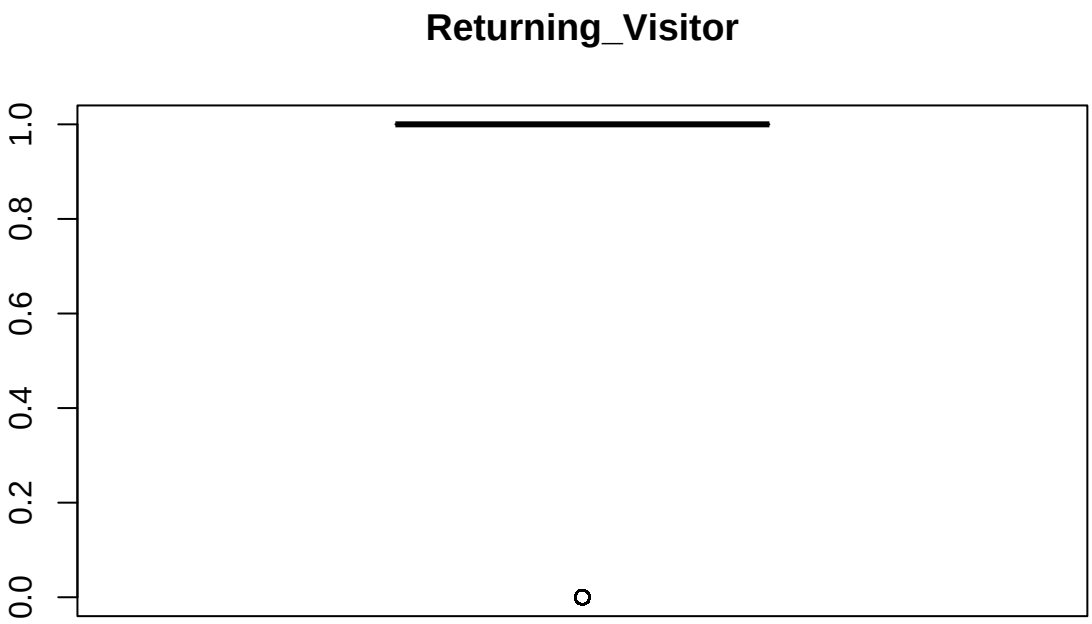

OperatingSystems
1
2
3
4
5
6
7
8


Browser
2
4
6
8
10
12


Region
2
4
6
8

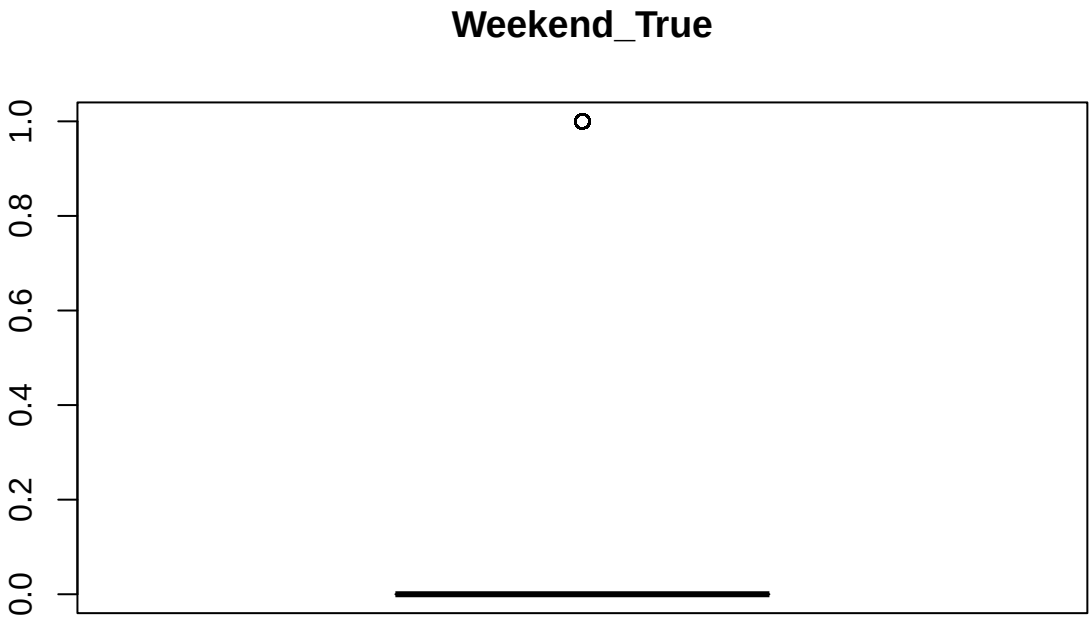

TrafficType
20
15
10
5


New_Visitor
0.0
0.2
0.4
0.6
0.8
1.0

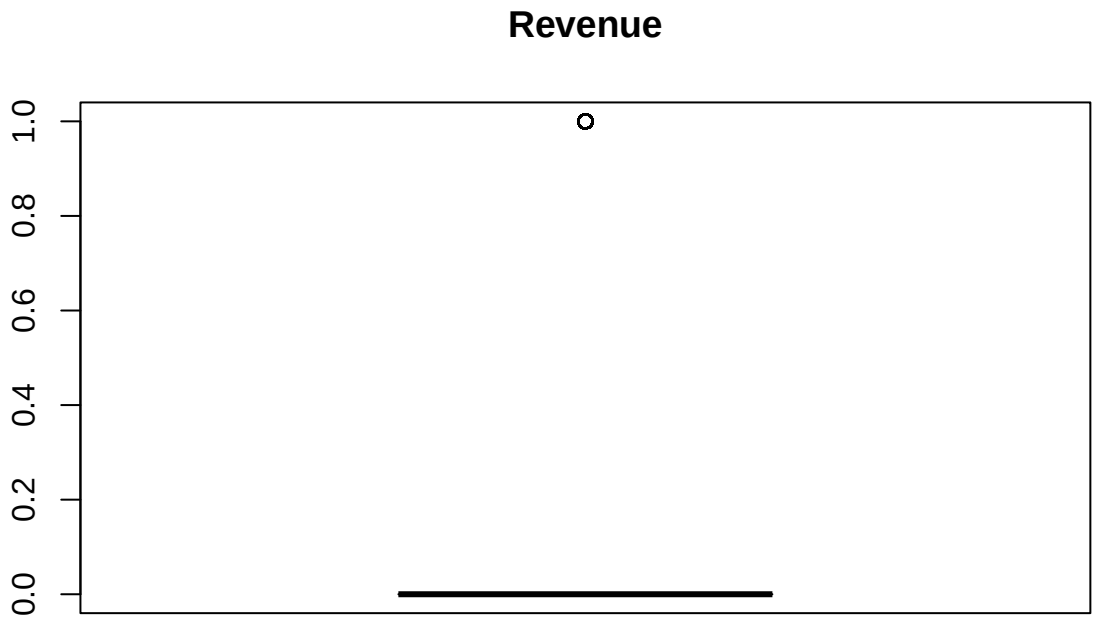

Returning_Visitor
0.0
0.2
0.4
0.6
0.8
1.0


Other
0.0
0.2
0.4
0.6
0.8
1.0

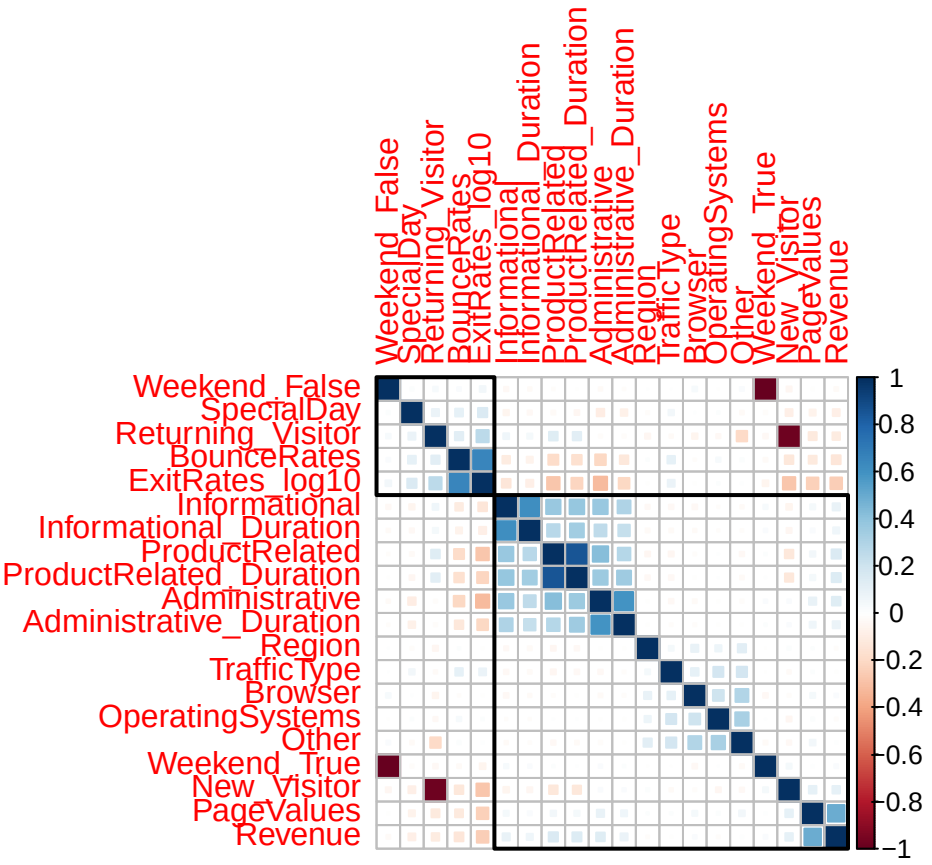

Weekend_False
0.0
0.2
0.4
0.6
0.8
1.0


Weekend_True
0.0
0.2
0.4
0.6
0.8
1.0


ExitRates_log10
−1.0
−1.5
−2.0
−2.5




```
#########################Correlation Plot#########################################
corr_matrix <- cor(data_set_cst_ol_3)
corrplot(corr_matrix,method = 'square',order = 'hclust',addrect =  2)
```

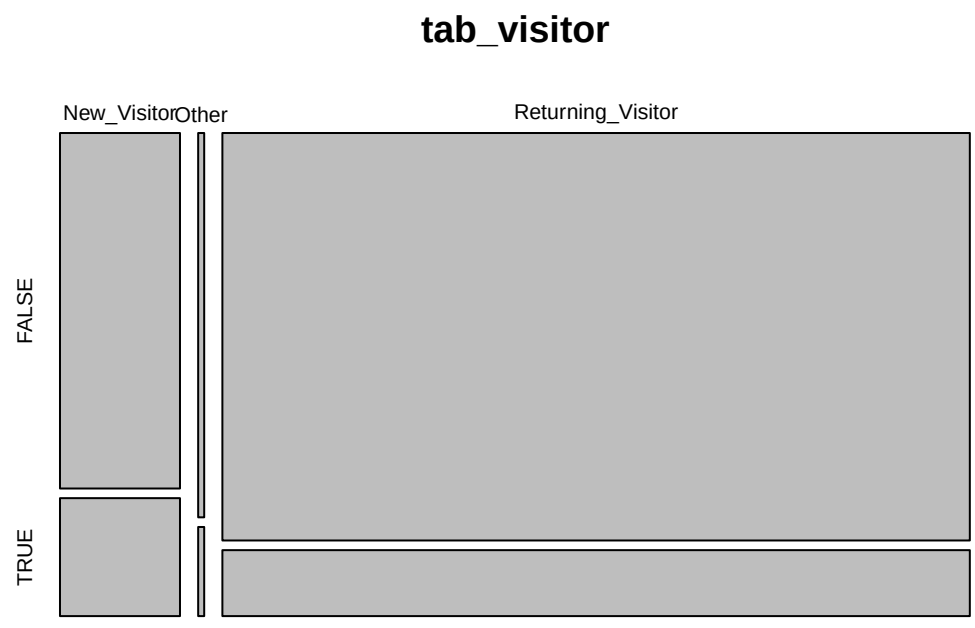



```
describe(data_set_cst_ol_3)
```

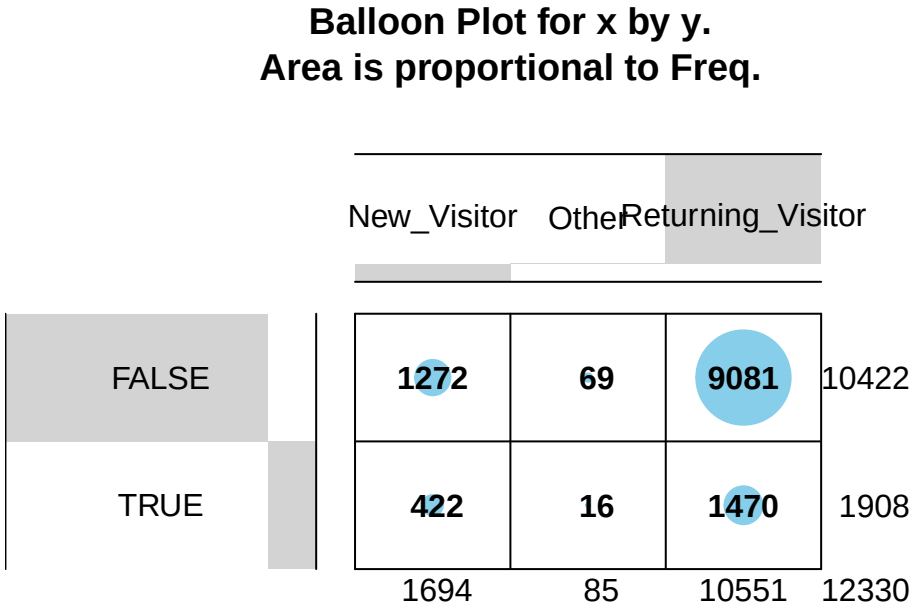
```
##                          vars     n    mean       sd median trimmed    mad  min
## Administrative              1 12079    2.36     3.34   1.00    1.68   1.48  0.0
## Administrative_Duration     2 12079   82.50   178.22  11.00   43.58  16.31  0.0
## Informational               3 12079    0.51     1.28   0.00    0.19   0.00  0.0
## Informational_Duration      4 12079   35.19   142.12   0.00    3.86   0.00  0.0
## ProductRelated              5 12079   32.37    44.71  19.00   23.35  19.27  0.0
## ProductRelated_Duration     6 12079 1219.57  1925.61 623.72  843.76 752.26  0.0
## PageValues                  7 12079    6.01    18.74   0.00    1.37   0.00  0.0
## BounceRates                 8 12079    0.02     0.04   0.00    0.01   0.00  0.0
## SpecialDay                  9 12079    0.06     0.20   0.00    0.00   0.00  0.0
## OperatingSystems           10 12079    2.13     0.91   2.00    2.06   0.00  1.0
## Browser                    11 12079    2.36     1.71   2.00    2.01   0.00  1.0
## Region                     12 12079    3.16     2.40   3.00    2.80   2.97  1.0
## TrafficType                13 12079    4.08     4.01   2.00    3.23   1.48  1.0
## New_Visitor                14 12079    0.14     0.35   0.00    0.05   0.00  0.0
## Returning_Visitor          15 12079    0.85     0.35   1.00    0.94   0.00  0.0
## Other                      16 12079    0.01     0.08   0.00    0.00   0.00  0.0
## Weekend_False              17 12079    0.76     0.42   1.00    0.83   0.00  0.0
## Weekend_True               18 12079    0.24     0.42   0.00    0.17   0.00  0.0
## ExitRates_log10            19 12079   -1.54     0.38  -1.56   -1.55   0.35 -2.6
## Revenue                    20 12079    0.16     0.36   0.00    0.07   0.00  0.0
##                             max    range  skew kurtosis    se
## Administrative            27.00    27.00  1.93     4.58  0.03
```

```
## Administrative_Duration  3398.75  3398.75  5.57    49.68  1.62
## Informational              24.00    24.00  3.99    26.37  0.01
## Informational_Duration   2549.38  2549.38  7.50    74.70  1.29
## ProductRelated            705.00   705.00  4.32    30.92  0.41
## ProductRelated_Duration 63973.52 63973.52  7.24   136.12 17.52
## PageValues                361.76   361.76  6.32    64.32  0.17
## BounceRates                 0.20     0.20  3.41    11.40  0.00
## SpecialDay                  1.00     1.00  3.26     9.65  0.00
## OperatingSystems            8.00     7.00  2.03    10.26  0.01
## Browser                    13.00    12.00  3.21    12.49  0.02
## Region                      9.00     8.00  0.97    -0.17  0.02
## TrafficType                20.00    19.00  1.96     3.47  0.04
## New_Visitor                 1.00     1.00  2.07     2.30  0.00
## Returning_Visitor           1.00     1.00 -2.00     1.98  0.00
## Other                       1.00     1.00 12.16   145.97  0.00
## Weekend_False               1.00     1.00 -1.25    -0.45  0.00
## Weekend_True                1.00     1.00  1.25    -0.45  0.00
## ExitRates_log10            -0.69     1.91  0.11     0.03  0.00
## Revenue                     1.00     1.00  1.88     1.52  0.00
```

```
##########################Building Contingency Table)##############################
#View(data_set_cst_ol_table)

#Contingency table for Visitor Type.
tab_visitor <- table(data_set_cst_ol_table$VisitorType,data_set_cst_ol_table$Revenue)
mosaicplot(tab_visitor)
```



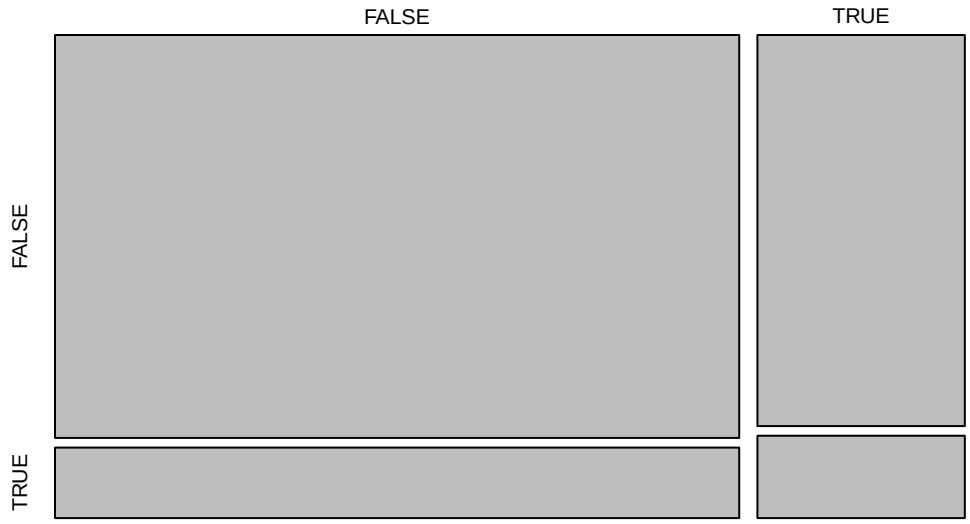
```
balloonplot(tab_visitor,show.margins = TRUE)
```



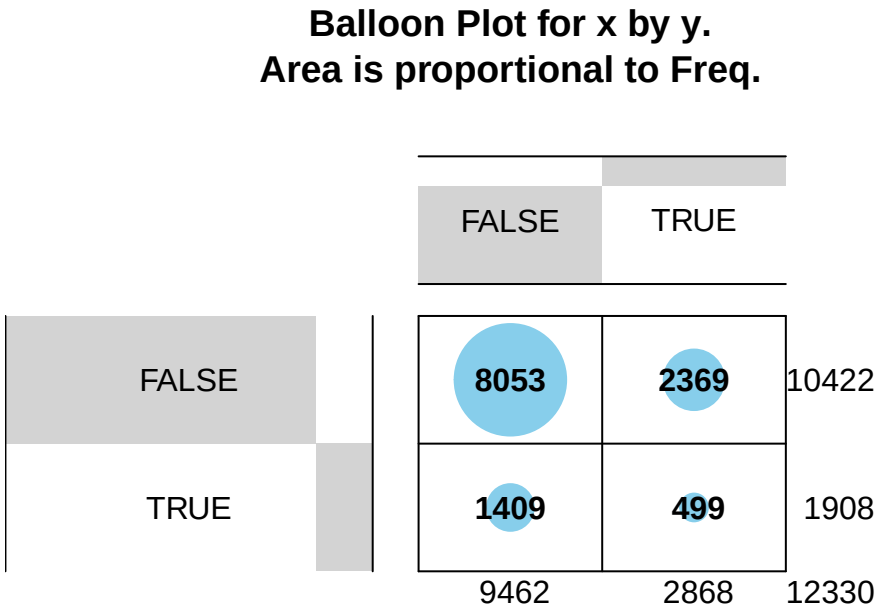

```
chisq.test(tab_visitor)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab_visitor
## X-squared = 135.25, df = 2, p-value < 2.2e-16
```

```
#Contingency table for Weekend.
tab_weekend <- table(data_set_cst_ol_table$Weekend,data_set_cst_ol_table$Revenue)
mosaicplot(tab_weekend)
```




```
balloonplot(tab_weekend)
```

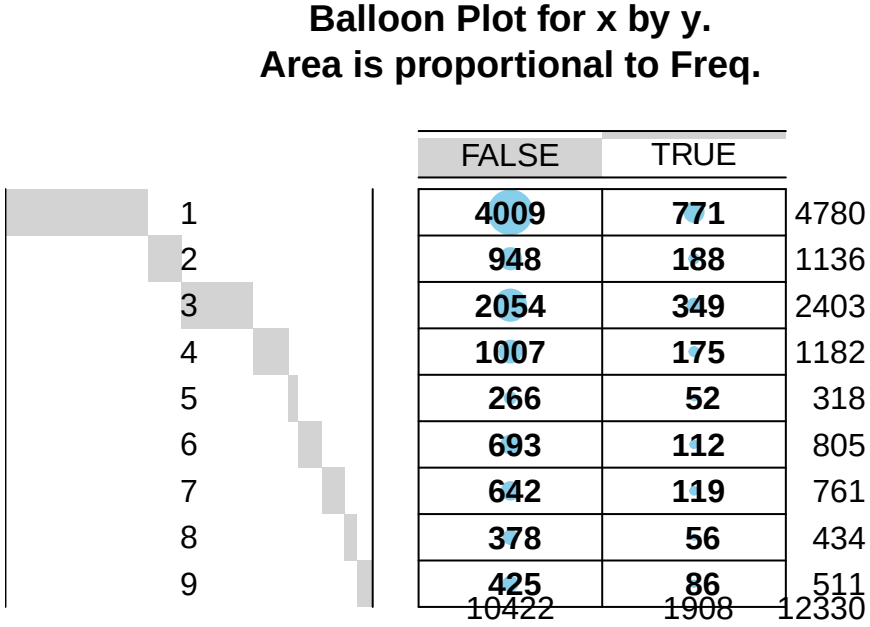



```
chisq.test(tab_weekend)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab_weekend
## X-squared = 10.391, df = 1, p-value = 0.001266
```

```
#Contingency table for Region.
tab_region <- table(data_set_cst_ol_table$Revenue,data_set_cst_ol_table$Region)
mosaicplot(tab_region)
```

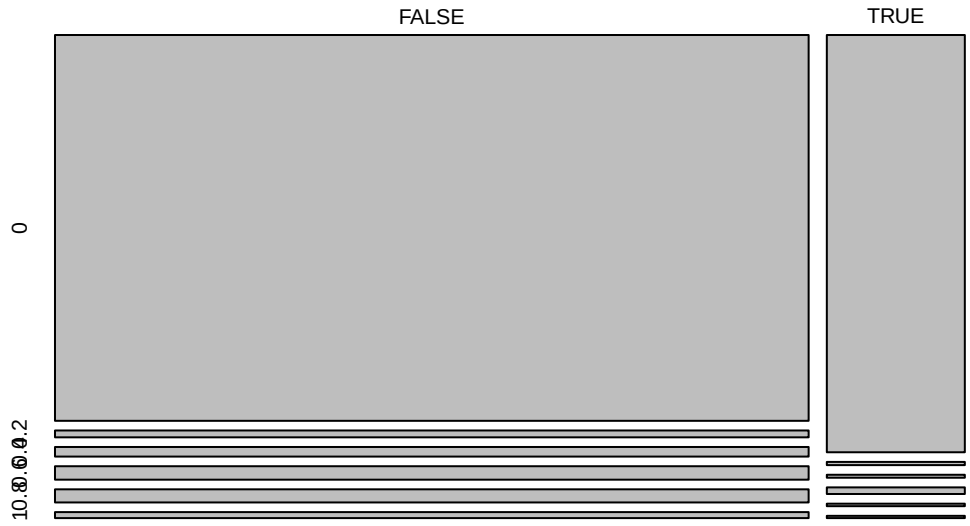

```
balloonplot(tab_region)
```



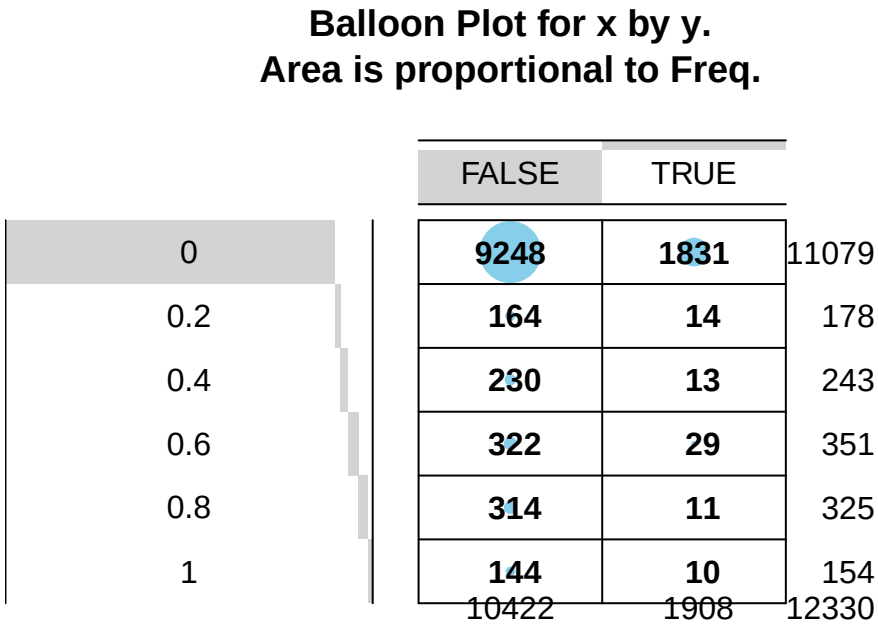

```
chisq.test(tab_region)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab_region
## X-squared = 9.2528, df = 8, p-value = 0.3214
```

```
#Contingency table for Special Day.
tab_special_day <- table(data_set_cst_ol_table$Revenue,data_set_cst_ol_table$SpecialDay)
mosaicplot(tab_special_day)
```




```
balloonplot(tab_special_day)
```




```
chisq.test(tab_special_day)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab_special_day
## X-squared = 96.077, df = 5, p-value < 2.2e-16
```

```
############################Creating Training and Testing Data set################
names(data_set_cst_ol_3)
```

```
##  [1] "Administrative"          "Administrative_Duration"
##  [3] "Informational"           "Informational_Duration"
##  [5] "ProductRelated"          "ProductRelated_Duration"
##  [7] "PageValues"              "BounceRates"
##  [9] "SpecialDay"              "OperatingSystems"
## [11] "Browser"                 "Region"
## [13] "TrafficType"             "New_Visitor"
## [15] "Returning_Visitor"       "Other"
## [17] "Weekend_False"           "Weekend_True"
## [19] "ExitRates_log10"         "Revenue"
```

```
split_values <- sample.split(data_set_cst_ol_3$Revenue,SplitRatio = 0.65)
```

```r
train_set <- subset(data_set_cst_ol_3,split_values == 1)
test_set <- subset(data_set_cst_ol_3,split_values == 0)
test_set_01 <- select(test_set,-c(Revenue))

write.csv(test_set_01,"test_set_01.csv")

train_set_01 <- select(train_set,-c(Other,Weekend_False))

model_one <- glm(Revenue ~., data = train_set_01, family = "binomial")

summary(model_one)
```

```
##
## Call:
## glm(formula = Revenue ~ ., family = "binomial", data = train_set_01)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.5668  -0.4678  -0.3808  -0.2331   3.2736
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.817e+00  8.261e-01  -4.620 3.84e-06 ***
## Administrative            7.657e-03  1.388e-02   0.552 0.581041
## Administrative_Duration -2.400e-04  2.518e-04  -0.953 0.340438
## Informational             4.462e-02  3.329e-02   1.340 0.180097
## Informational_Duration  -1.144e-04  2.786e-04  -0.411 0.681372
## ProductRelated            4.817e-03  1.393e-03   3.458 0.000544 ***
## ProductRelated_Duration   4.494e-05  3.289e-05   1.366 0.171810
## PageValues                8.219e-02  2.998e-03  27.416  < 2e-16 ***
## BounceRates              -1.331e+01  3.152e+00  -4.221 2.43e-05 ***
## SpecialDay               -9.857e-01  2.781e-01  -3.545 0.000393 ***
## OperatingSystems         -8.775e-02  4.881e-02  -1.798 0.072230 .
## Browser                   2.554e-02  2.356e-02   1.084 0.278326
## Region                   -9.809e-03  1.609e-02  -0.610 0.542127
## TrafficType               1.200e-02  1.028e-02   1.167 0.243280
## New_Visitor               9.741e-01  7.517e-01   1.296 0.195032
## Returning_Visitor         6.021e-01  7.470e-01   0.806 0.420184
## Weekend_True              9.037e-02  8.733e-02   1.035 0.300721
## ExitRates_log10          -4.835e-01  1.519e-01  -3.182 0.001463 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6849.9  on 7850  degrees of freedom
## Residual deviance: 4787.3  on 7833  degrees of freedom
## AIC: 4823.3
##
## Number of Fisher Scoring iterations: 7
```

```
confint(model_one)
```

```
## Waiting for profiling to be done...

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##                                 2.5 %        97.5 %
## (Intercept)             -5.601431e+00 -2.3456454658
## Administrative          -1.973548e-02  0.0346702301
## Administrative_Duration -7.484630e-04  0.0002391382
## Informational           -2.120657e-02  0.1093704750
## Informational_Duration  -6.784425e-04  0.0004161737
## ProductRelated           2.033304e-03  0.0075078655
## ProductRelated_Duration -1.720618e-05  0.0001123237
## PageValues               7.641698e-02  0.0881699268
## BounceRates             -1.992181e+01 -7.5688708102
## SpecialDay              -1.557087e+00 -0.4643448712
## OperatingSystems        -1.842533e-01  0.0070716323
## Browser                 -2.154559e-02  0.0708711857
## Region                  -4.157162e-02  0.0215217384
## TrafficType             -8.426363e-03  0.0318881298
## New_Visitor             -3.361892e-01  2.6237395228
## Returning_Visitor       -6.963703e-01  2.2439996905
## Weekend_True            -8.212478e-02  0.2603132781
## ExitRates_log10         -7.808247e-01 -0.1851527682
```

```
wald.test(b= coef(model_one),Sigma = vcov(model_one), Terms = 1:18)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 2807.9, df = 18, P(> X2) = 0.0
```

```
test_set_01 <- select(test_set,-c(Revenue,Weekend_False))

prob_one <- predict(model_one,newdata = test_set_01,type = "response")

glm_pred <- ifelse(prob_one > 0.5,1,0)

cmp <- data.frame(glm_pred,test_set$Revenue)

#Building Classification Model
model_class <- rpart(Revenue ~.,data = train_set)

result_class <- predict(model_class,test_set)

cmp <- data.frame(round(result_class),test_set$Revenue)

################################################################################
################################################################################
```