ECommerceCustomersAnlaysis

Joy Machuka

9/3/2021

PROBLEM DEFINITION

Defining the Question Kira Plastinina is a Russian brand whose sales and Marketing team would like to understand their customer's behavior from data that they have collected over the past year.

Metrics of Success Our analysis will be considered successful when we are able to draw insights from the cluster analysis performed on the data.

Context Kira Plastinina is a Russian fashion designer whose brand is sold through a defunct chain of retail stores in Russia, Ukraine, Kazakhstan, Belarus, China, Philippines, and Armenia. The marketing team wants an analysis carried out on their customers and insights drawn from various attributes and features of their customers.

Experimental Design

Defining the Question Data preparation Data Cleaning Univariate Analysis Bivariate Analysis Clustering Conclusion

Data Sourcing(Loading dataset)

```
packages<-function(x){
  x<-as.character(match.call()[[2]])
  if (!require(x,character.only=TRUE)){
    install.packages(pkgs=x,repos="http://cran.r-project.org")
    require(x,character.only=TRUE)
  }
}

#importing libraries
library(tidyverse) # data manipulation
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library("factoextra")
```

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

```
library(cluster) # clustering algorithms
```

```
customers <- read.csv("http://bit.ly/EcommerceCustomersDataset")
```

```
#Checking head of our dataset
head(customers)
```

```
##   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
```

```
class(customers)
```

```
## [1] "data.frame"
```

We have a data frame

```
dim(customers)
```

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

Our dataset has 12330 rows and 18 columns

```
#checking column names
names(customers)
```

```
##  [1] "Administrative"          "Administrative_Duration"
##  [3] "Informational"           "Informational_Duration"
##  [5] "ProductRelated"          "ProductRelated_Duration"
##  [7] "BounceRates"             "ExitRates"
##  [9] "PageValues"              "SpecialDay"
## [11] "Month"                   "OperatingSystems"
## [13] "Browser"                 "Region"
## [15] "TrafficType"             "VisitorType"
## [17] "Weekend"                 "Revenue"
```

Above are our column names.

```
str(customers)
```

```
## '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 ...
```

Our columns are categorical num, int and characters.

```
str(customers)
```

```
## '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 ...
```

Data Cleaning

```
anyNA(customers)
```

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

We have missing values. We go ahead and check number

```
#checking for null values per column
colSums(is.na(customers))
```

```
##          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 have a number of nulls that we decided to drop since they are minimal

```
#dropping nulls
customers = na.omit(customers)
#Confirming nulls after dropping
anyNA(customers)
```

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

There aren't any more nulls.

```
#checking for duplicates
duplicated_rows <- customers[duplicated(customers),]
duplicated_rows
```

```
##       Administrative Administrative_Duration Informational
## 159                0                       0             0
## 179                0                       0             0
## 419                0                       0             0
## 457                0                       0             0
## 484                0                       0             0
## 513                0                       0             0
## 555                0                       0             0
## 590                0                       0             0
## 660                0                       0             0
## 775                0                       0             0
## 873                0                       0             0
## 890                0                       0             0
## 923                0                       0             0
## 948                0                       0             0
## 975                0                       0             0
## 1035               0                       0             0
## 1120               0                       0             0
## 1171               0                       0             0
## 1177               0                       0             0
## 1214               0                       0             0
## 1215               0                       0             0
## 1292               0                       0             0
## 1326               0                       0             0
## 1357               0                       0             0
## 1367               0                       0             0
## 1382               0                       0             0
## 1391               0                       0             0
## 1395               0                       0             0
## 1437               0                       0             0
## 1454               0                       0             0
## 1516               0                       0             0
## 1574               0                       0             0
```

```
## 1609              0                 0              0
## 1698              0                 0              0
## 1776              0                 0              0
## 1805              0                 0              0
## 1840              0                 0              0
## 1867              0                 0              0
## 1926              0                 0              0
## 1934              0                 0              0
## 1950              0                 0              0
## 2057              0                 0              0
## 2058              0                 0              0
## 2236              0                 0              0
## 2622              0                 0              0
## 2740              0                 0              0
## 3232              0                 0              0
## 3273              0                 0              0
## 3282              0                 0              0
## 3578              0                 0              0
## 3651              0                 0              0
## 3664              0                 0              0
## 3722              0                 0              0
## 3892              0                 0              0
## 4164              0                 0              0
## 4183              0                 0              0
## 4232              0                 0              0
## 4344              0                 0              0
## 4375              0                 0              0
## 4404              0                 0              0
## 4427              0                 0              0
## 4464              0                 0              0
## 4490              0                 0              0
## 4553              0                 0              0
## 4818              0                 0              0
## 4884              0                 0              0
## 4914              0                 0              0
## 5039              0                 0              0
## 5044              0                 0              0
## 5057              0                 0              0
## 5119              0                 0              0
## 5199              0                 0              0
## 5200              0                 0              0
## 5255              0                 0              0
## 5277              0                 0              0
## 5287              0                 0              0
## 5356              0                 0              0
## 5408              0                 0              0
## 6930              0                 0              0
## 7152              0                 0              0
## 7636              0                 0              0
## 8545              0                 0              0
## 9307              0                 0              0
## 9495              0                 0              0
## 9552              0                 0              0
## 9569              0                 0              0
```

```
## 9582                0                       0              0
## 9719                0                       0              0
## 9770                0                       0              0
## 9879                0                       0              0
## 9908                0                       0              0
## 10147               0                       0              0
## 10223               0                       0              0
## 10270               0                       0              0
## 10573               0                       0              0
## 10632               0                       0              0
## 10752               0                       0              0
## 10796               0                       0              0
## 10842               0                       0              0
## 10989               0                       0              0
## 11044               0                       0              0
## 11206               0                       0              0
## 11405               0                       0              0
## 11524               0                       0              0
## 11582               0                       0              0
## 11625               0                       0              0
## 11659               0                       0              0
## 11734               0                       0              0
## 11748               0                       0              0
## 11802               0                       0              0
## 11814               0                       0              0
## 11828               0                       0              0
## 11935               0                       0              0
## 11939               0                       0              0
## 12160               0                       0              0
## 12181               0                       0              0
## 12186               0                       0              0
##       Informational_Duration ProductRelated ProductRelated_Duration BounceRates
## 159                        0              1                       0         0.2
## 179                        0              1                       0         0.2
## 419                        0              1                       0         0.2
## 457                        0              1                       0         0.2
## 484                        0              1                       0         0.2
## 513                        0              1                       0         0.2
## 555                        0              1                       0         0.2
## 590                        0              1                       0         0.2
## 660                        0              2                       0         0.2
## 775                        0              1                       0         0.2
## 873                        0              1                       0         0.2
## 890                        0              1                       0         0.2
## 923                        0              1                       0         0.2
## 948                        0              1                       0         0.2
## 975                        0              1                       0         0.2
## 1035                       0              1                       0         0.2
## 1120                       0              1                       0         0.2
## 1171                       0              1                       0         0.2
## 1177                       0              1                       0         0.2
## 1214                       0              1                       0         0.2
## 1215                       0              1                       0         0.2
## 1292                       0              2                       0         0.2
```

```
## 1326          0           1           0         0.2
## 1357          0           2           0         0.2
## 1367          0           1           0         0.2
## 1382          0           1           0         0.2
## 1391          0           1           0         0.2
## 1395          0           1           0         0.2
## 1437          0           1           0         0.2
## 1454          0           1           0         0.2
## 1516          0           1           0         0.2
## 1574          0           1           0         0.2
## 1609          0           1           0         0.2
## 1698          0           1           0         0.2
## 1776          0           1           0         0.2
## 1805          0           1           0         0.2
## 1840          0           1           0         0.2
## 1867          0           1           0         0.2
## 1926          0           1           0         0.2
## 1934          0           1           0         0.2
## 1950          0           1           0         0.2
## 2057          0           1           0         0.2
## 2058          0           1           0         0.2
## 2236          0           1           0         0.2
## 2622          0           1           0         0.2
## 2740          0           1           0         0.2
## 3232          0           1           0         0.2
## 3273          0           1           0         0.2
## 3282          0           1           0         0.2
## 3578          0           1           0         0.2
## 3651          0           1           0         0.2
## 3664          0           1           0         0.2
## 3722          0           1           0         0.2
## 3892          0           1           0         0.2
## 4164          0           1           0         0.2
## 4183          0           1           0         0.2
## 4232          0           1           0         0.2
## 4344          0           1           0         0.2
## 4375          0           1           0         0.2
## 4404          0           1           0         0.2
## 4427          0           1           0         0.2
## 4464          0           1           0         0.2
## 4490          0           1           0         0.2
## 4553          0           2           0         0.2
## 4818          0           1           0         0.2
## 4884          0           1           0         0.2
## 4914          0           1           0         0.2
## 5039          0           1           0         0.2
## 5044          0           1           0         0.2
## 5057          0           1           0         0.2
## 5119          0           1           0         0.2
## 5199          0           1           0         0.2
## 5200          0           2           0         0.2
## 5255          0           1           0         0.2
## 5277          0           1           0         0.2
## 5287          0           1           0         0.2
```

```
## 5356                    0             1                    0         0.2
## 5408                    0             1                    0         0.2
## 6930                    0             1                    0         0.2
## 7152                    0             1                    0         0.2
## 7636                    0             1                    0         0.2
## 8545                    0             1                    0         0.2
## 9307                    0             1                    0         0.2
## 9495                    0             1                    0         0.2
## 9552                    0             1                    0         0.2
## 9569                    0             1                    0         0.2
## 9582                    0             1                    0         0.2
## 9719                    0             1                    0         0.2
## 9770                    0             1                    0         0.2
## 9879                    0             1                    0         0.2
## 9908                    0             1                    0         0.2
## 10147                   0             1                    0         0.2
## 10223                   0             2                    0         0.2
## 10270                   0             1                    0         0.2
## 10573                   0             1                    0         0.2
## 10632                   0             1                    0         0.2
## 10752                   0             1                    0         0.2
## 10796                   0             1                    0         0.2
## 10842                   0             1                    0         0.2
## 10989                   0             1                    0         0.2
## 11044                   0             1                    0         0.2
## 11206                   0             1                    0         0.2
## 11405                   0             1                    0         0.2
## 11524                   0             1                    0         0.2
## 11582                   0             1                    0         0.2
## 11625                   0             1                    0         0.2
## 11659                   0             1                    0         0.2
## 11734                   0             1                    0         0.2
## 11748                   0             1                    0         0.2
## 11802                   0             1                    0         0.2
## 11814                   0             1                    0         0.2
## 11828                   0             1                    0         0.2
## 11935                   0             1                    0         0.2
## 11939                   0             1                    0         0.2
## 12160                   0             1                    0         0.2
## 12181                   0             1                    0         0.2
## 12186                   0             1                    0         0.2
##       ExitRates PageValues SpecialDay Month OperatingSystems Browser Region
## 159         0.2          0       0.0   Feb                1       1      1
## 179         0.2          0       0.0   Feb                3       2      3
## 419         0.2          0       0.0   Mar                1       1      1
## 457         0.2          0       0.0   Mar                2       2      4
## 484         0.2          0       0.0   Mar                3       2      3
## 513         0.2          0       0.0   Mar                2       2      1
## 555         0.2          0       0.0   Mar                2       2      1
## 590         0.2          0       0.0   Mar                2       2      1
## 660         0.2          0       0.0   Mar                2       5      1
## 775         0.2          0       0.0   Mar                2       2      4
## 873         0.2          0       0.0   Mar                3       2      3
## 890         0.2          0       0.0   Mar                1       1      2
```

```
## 923        0.2          0        0.0   Mar                3       2       2
## 948        0.2          0        0.0   Mar                2       2       1
## 975        0.2          0        0.0   Mar                2       2       1
## 1035       0.2          0        0.0   Mar                2       2       1
## 1120       0.2          0        0.0   Mar                2       2       1
## 1171       0.2          0        0.0   Mar                3       2       1
## 1177       0.2          0        0.0   Mar                2       4       1
## 1214       0.2          0        0.0   Mar                3       2       3
## 1215       0.2          0        0.0   Mar                1       1       1
## 1292       0.2          0        0.0   Mar                2       2       1
## 1326       0.2          0        0.0   Mar                1       1       3
## 1357       0.2          0        0.0   Mar                1       1       1
## 1367       0.2          0        0.0   Mar                1       1       8
## 1382       0.2          0        0.0   Mar                1       1       4
## 1391       0.2          0        0.0   Mar                2       2       1
## 1395       0.2          0        0.0   Mar                2       2       1
## 1437       0.2          0        0.0   Mar                3       2       3
## 1454       0.2          0        0.0   Mar                2       2       1
## 1516       0.2          0        0.0   Mar                1       1       1
## 1574       0.2          0        0.0   Mar                2       2       1
## 1609       0.2          0        0.0   Mar                2       2       7
## 1698       0.2          0        0.0   Mar                2       2       2
## 1776       0.2          0        0.0   Mar                3       2       1
## 1805       0.2          0        0.0   Mar                1       1       8
## 1840       0.2          0        0.0   Mar                2       2       1
## 1867       0.2          0        0.0   Mar                1       1       1
## 1926       0.2          0        0.0   Mar                3       2       1
## 1934       0.2          0        0.0   Mar                2       2       1
## 1950       0.2          0        0.0   Mar                2       2       1
## 2057       0.2          0        0.0   Mar                3       2       3
## 2058       0.2          0        0.0   Mar                2       4       1
## 2236       0.2          0        0.0   May                1       1       4
## 2622       0.2          0        0.0   May                1       1       1
## 2740       0.2          0        0.0   May                2       2       1
## 3232       0.2          0        0.0   May                2       4       1
## 3273       0.2          0        0.0   May                1       1       3
## 3282       0.2          0        0.0   May                1       1       1
## 3578       0.2          0        0.0   May                2       2       1
## 3651       0.2          0        0.0   May                2       2       4
## 3664       0.2          0        0.0   May                1       1       1
## 3722       0.2          0        0.0   May                1       1       4
## 3892       0.2          0        0.0   May                2       2       7
## 4164       0.2          0        0.0   May                1       1       4
## 4183       0.2          0        0.0   May                1       1       1
## 4232       0.2          0        0.0   May                2       2       2
## 4344       0.2          0        0.0   May                3       2       1
## 4375       0.2          0        0.0   May                2       2       1
## 4404       0.2          0        0.0   May                2       2       1
## 4427       0.2          0        0.0   May                2       2       1
## 4464       0.2          0        0.0   May                1       1       1
## 4490       0.2          0        0.0   May                3       2       9
## 4553       0.2          0        0.0   May                2       2       2
## 4818       0.2          0        0.0   May                2       2       1
## 4884       0.2          0        0.0   May                2       2       1
```

```
## 4914       0.2          0        0.8   May                2       2       1
## 5039       0.2          0        0.0   May                3       2       3
## 5044       0.2          0        0.0   May                2       2       1
## 5057       0.2          0        0.0   May                2       2       6
## 5119       0.2          0        0.0   May                1       1       6
## 5199       0.2          0        0.0   May                2       2       1
## 5200       0.2          0        0.0   May                2       2       2
## 5255       0.2          0        0.6   May                2       2       1
## 5277       0.2          0        0.0   May                3       2       3
## 5287       0.2          0        0.0   May                1       1       3
## 5356       0.2          0        0.0   May                1       1       3
## 5408       0.2          0        0.0   May                2       4       1
## 6930       0.2          0        0.0  June                2       2       1
## 7152       0.2          0        0.0  June                2       2       1
## 7636       0.2          0        0.0  June                3       2       3
## 8545       0.2          0        0.0   Nov                3       2       3
## 9307       0.2          0        0.0   Dec                3       2       3
## 9495       0.2          0        0.0   Dec                2       2       1
## 9552       0.2          0        0.0   Nov                3       2       4
## 9569       0.2          0        0.0   Dec                2       2       8
## 9582       0.2          0        0.0   Nov                2       2       1
## 9719       0.2          0        0.0   Nov                3       2       7
## 9770       0.2          0        0.0   Dec                2       2       2
## 9879       0.2          0        0.0   Dec                2       2       6
## 9908       0.2          0        0.0   Dec                2       2       1
## 10147      0.2          0        0.0   Dec                8      13       9
## 10223      0.2          0        0.0   Nov                1       1       1
## 10270      0.2          0        0.0   Nov                1       1       3
## 10573      0.2          0        0.0   Nov                2       2       3
## 10632      0.2          0        0.0   Nov                2       2       1
## 10752      0.2          0        0.0   Dec                1       1       1
## 10796      0.2          0        0.0   Nov                1       1       4
## 10842      0.2          0        0.0   Nov                2       2       3
## 10989      0.2          0        0.0   Nov                2       4       3
## 11044      0.2          0        0.0   Dec                3       2       6
## 11206      0.2          0        0.0   Dec                8      13       9
## 11405      0.2          0        0.0   Nov                3       2       1
## 11524      0.2          0        0.0   Dec                2       2       1
## 11582      0.2          0        0.0   Dec                8      13       9
## 11625      0.2          0        0.0   Nov                3       2       1
## 11659      0.2          0        0.0   Dec                1       1       1
## 11734      0.2          0        0.0   Nov                2       2       1
## 11748      0.2          0        0.0   Nov                1       1       3
## 11802      0.2          0        0.0   Dec                1       1       4
## 11814      0.2          0        0.0   Dec                2       2       1
## 11828      0.2          0        0.0   Dec                2       2       1
## 11935      0.2          0        0.0   Dec                1       1       1
## 11939      0.2          0        0.0   Dec                1       1       4
## 12160      0.2          0        0.0   Dec                1       1       1
## 12181      0.2          0        0.0   Dec                1      13       9
## 12186      0.2          0        0.0   Dec                8      13       9
##       TrafficType       VisitorType Weekend Revenue
## 159             3 Returning_Visitor   FALSE   FALSE
## 179             3 Returning_Visitor   FALSE   FALSE
```

```
## 419             1 Returning_Visitor    TRUE   FALSE
## 457             1 Returning_Visitor   FALSE   FALSE
## 484             1 Returning_Visitor   FALSE   FALSE
## 513             1 Returning_Visitor   FALSE   FALSE
## 555             1 Returning_Visitor   FALSE   FALSE
## 590             1 Returning_Visitor   FALSE   FALSE
## 660             1 Returning_Visitor   FALSE   FALSE
## 775             1 Returning_Visitor   FALSE   FALSE
## 873             1 Returning_Visitor   FALSE   FALSE
## 890             1 Returning_Visitor   FALSE   FALSE
## 923             1 Returning_Visitor   FALSE   FALSE
## 948             1 Returning_Visitor   FALSE   FALSE
## 975             1 Returning_Visitor   FALSE   FALSE
## 1035            1 Returning_Visitor   FALSE   FALSE
## 1120            1 Returning_Visitor   FALSE   FALSE
## 1171            1 Returning_Visitor   FALSE   FALSE
## 1177            1 Returning_Visitor   FALSE   FALSE
## 1214            1 Returning_Visitor   FALSE   FALSE
## 1215            3 Returning_Visitor   FALSE   FALSE
## 1292            1 Returning_Visitor   FALSE   FALSE
## 1326            3 Returning_Visitor   FALSE   FALSE
## 1357            1 Returning_Visitor   FALSE   FALSE
## 1367            1 Returning_Visitor   FALSE   FALSE
## 1382            1 Returning_Visitor   FALSE   FALSE
## 1391            1 Returning_Visitor   FALSE   FALSE
## 1395            1 Returning_Visitor   FALSE   FALSE
## 1437            1 Returning_Visitor   FALSE   FALSE
## 1454            1 Returning_Visitor   FALSE   FALSE
## 1516            3 Returning_Visitor    TRUE   FALSE
## 1574            1 Returning_Visitor   FALSE   FALSE
## 1609            1 Returning_Visitor   FALSE   FALSE
## 1698            1 Returning_Visitor   FALSE   FALSE
## 1776            1 Returning_Visitor   FALSE   FALSE
## 1805            1 Returning_Visitor   FALSE   FALSE
## 1840            3 Returning_Visitor   FALSE   FALSE
## 1867            9 Returning_Visitor    TRUE   FALSE
## 1926            1 Returning_Visitor   FALSE   FALSE
## 1934            1 Returning_Visitor   FALSE   FALSE
## 1950            1 Returning_Visitor   FALSE   FALSE
## 2057            1 Returning_Visitor   FALSE   FALSE
## 2058            1 Returning_Visitor   FALSE   FALSE
## 2236            3 Returning_Visitor   FALSE   FALSE
## 2622            3 Returning_Visitor   FALSE   FALSE
## 2740            1 Returning_Visitor   FALSE   FALSE
## 3232            3 Returning_Visitor   FALSE   FALSE
## 3273            3 Returning_Visitor   FALSE   FALSE
## 3282            3 Returning_Visitor   FALSE   FALSE
## 3578            4 Returning_Visitor   FALSE   FALSE
## 3651            1 Returning_Visitor   FALSE   FALSE
## 3664            3 Returning_Visitor   FALSE   FALSE
## 3722            3 Returning_Visitor   FALSE   FALSE
## 3892            4 Returning_Visitor   FALSE   FALSE
## 4164            3 Returning_Visitor   FALSE   FALSE
## 4183            3 Returning_Visitor   FALSE   FALSE
```

```
## 4232           1 Returning_Visitor   FALSE   FALSE
## 4344          13 Returning_Visitor   FALSE   FALSE
## 4375           3 Returning_Visitor   FALSE   FALSE
## 4404           3 Returning_Visitor   FALSE   FALSE
## 4427           3 Returning_Visitor   FALSE   FALSE
## 4464           3 Returning_Visitor   FALSE   FALSE
## 4490           3 Returning_Visitor   FALSE   FALSE
## 4553           3 Returning_Visitor   FALSE   FALSE
## 4818           3 Returning_Visitor   FALSE   FALSE
## 4884           3 Returning_Visitor   FALSE   FALSE
## 4914           1 Returning_Visitor   FALSE   FALSE
## 5039           3 Returning_Visitor   FALSE   FALSE
## 5044           3 Returning_Visitor   FALSE   FALSE
## 5057           3 Returning_Visitor   FALSE   FALSE
## 5119           4 Returning_Visitor    TRUE   FALSE
## 5199          13 Returning_Visitor   FALSE   FALSE
## 5200           3 Returning_Visitor   FALSE   FALSE
## 5255           1 Returning_Visitor   FALSE   FALSE
## 5277          13 Returning_Visitor   FALSE   FALSE
## 5287          15 Returning_Visitor   FALSE   FALSE
## 5356           3 Returning_Visitor   FALSE   FALSE
## 5408           6 Returning_Visitor   FALSE   FALSE
## 6930           1 Returning_Visitor   FALSE   FALSE
## 7152           1 Returning_Visitor   FALSE   FALSE
## 7636          13 Returning_Visitor   FALSE   FALSE
## 8545           3 Returning_Visitor   FALSE   FALSE
## 9307           1 Returning_Visitor    TRUE   FALSE
## 9495           3 Returning_Visitor   FALSE   FALSE
## 9552           3 Returning_Visitor   FALSE   FALSE
## 9569           1 Returning_Visitor   FALSE   FALSE
## 9582           1 Returning_Visitor   FALSE   FALSE
## 9719          13 Returning_Visitor   FALSE   FALSE
## 9770           1 Returning_Visitor   FALSE   FALSE
## 9879          13 Returning_Visitor   FALSE   FALSE
## 9908          13 Returning_Visitor   FALSE   FALSE
## 10147         20             Other   FALSE   FALSE
## 10223          1 Returning_Visitor   FALSE   FALSE
## 10270          2 Returning_Visitor   FALSE   FALSE
## 10573          1 Returning_Visitor   FALSE   FALSE
## 10632          1 Returning_Visitor   FALSE   FALSE
## 10752          1 Returning_Visitor    TRUE   FALSE
## 10796          1 Returning_Visitor   FALSE   FALSE
## 10842          1 Returning_Visitor   FALSE   FALSE
## 10989          3 Returning_Visitor   FALSE   FALSE
## 11044          1 Returning_Visitor   FALSE   FALSE
## 11206         20             Other   FALSE   FALSE
## 11405         13 Returning_Visitor   FALSE   FALSE
## 11524         13 Returning_Visitor   FALSE   FALSE
## 11582         20             Other   FALSE   FALSE
## 11625          1 Returning_Visitor   FALSE   FALSE
## 11659          1 Returning_Visitor    TRUE   FALSE
## 11734          1 Returning_Visitor   FALSE   FALSE
## 11748          3 Returning_Visitor   FALSE   FALSE
## 11802          1 Returning_Visitor    TRUE   FALSE
```

```
## 11814           1 Returning_Visitor   FALSE   FALSE
## 11828           1 Returning_Visitor   FALSE   FALSE
## 11935           2       New_Visitor   FALSE   FALSE
## 11939           1 Returning_Visitor    TRUE   FALSE
## 12160           3 Returning_Visitor   FALSE   FALSE
## 12181          20 Returning_Visitor   FALSE   FALSE
## 12186          20             Other   FALSE   FALSE
```

We have 117 duplicated rows that we are going to delete and print out only the unique items.

```
customers <- customers[!duplicated(customers), ]
dim(customers)
```

```
## [1] 12199    18
```

After deleting we are left with 12199 rows

```
duplicated_rows <- customers[duplicated(customers),]
# duplicated_rows
```

```
names(customers)
```

```
##  [1] "Administrative"          "Administrative_Duration"
##  [3] "Informational"           "Informational_Duration"
##  [5] "ProductRelated"          "ProductRelated_Duration"
##  [7] "BounceRates"             "ExitRates"
##  [9] "PageValues"              "SpecialDay"
## [11] "Month"                   "OperatingSystems"
## [13] "Browser"                 "Region"
## [15] "TrafficType"             "VisitorType"
## [17] "Weekend"                 "Revenue"
```

We remove the created column duplicated_rows.

```
# Dplyr remove a column by name:
# library("dplyr")
# select(customers, -duplicated_rows)
```

Anomalies

Next we convert the negative values we noticed in the duration columns while viewing the head of our dataset to nulls.

```
#replacing negatives with nulls
customers[customers<0] <- NA
```

```
#checking created nulls
anyNA(customers)
```

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

We will replace the created nulls with mode.

```
#mode function
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
#apply it on the duration columns
getmode(customers$Administrative_Duration)
```

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

```
getmode(customers$Informational_Duration)
```

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

```
getmode(customers$ProductRelated_Duration)
```

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

```
#Replacing nulls created with mode gotten above
customers$Administrative_Duration[is.na(customers$Administrative_Duration)] <- 0
customers$Informational_Duration[is.na(customers$Informational_Duration)] <- 0
customers$ProductRelated_Duration[is.na(customers$ProductRelated_Duration)] <- 0
```

```
#Confirming we have no more nulls
anyNA(customers)
```

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

We convert all char datatypes to factors so we can check for outliers and for better modelling.

```
# convert into a factor
customers$VisitorType <- factor(customers$VisitorType)
head(customers$VisitorType)
```

```
## [1] Returning_Visitor Returning_Visitor Returning_Visitor Returning_Visitor
## [5] Returning_Visitor Returning_Visitor
## Levels: New_Visitor Other Returning_Visitor
```

```
customers$Weekend <- factor(customers$Weekend)
head(customers$Weekend)
```

```
## [1] FALSE FALSE FALSE FALSE TRUE  FALSE
## Levels: FALSE TRUE
```

```
customers$Revenue <- factor(customers$Revenue)
head(customers$Revenue)
```

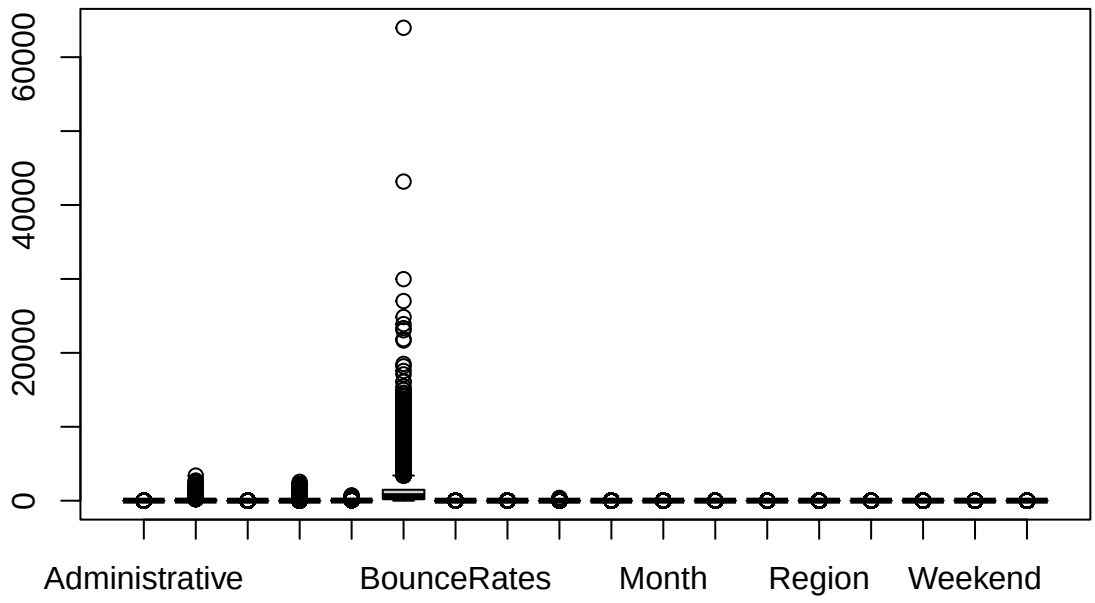
```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE
## Levels: FALSE TRUE
```

```
customers$Month <- factor(customers$Month)
head(customers$Month)
```

```
## [1] Feb Feb Feb Feb Feb Feb
## Levels: Aug Dec Feb Jul June Mar May Nov Oct Sep
```

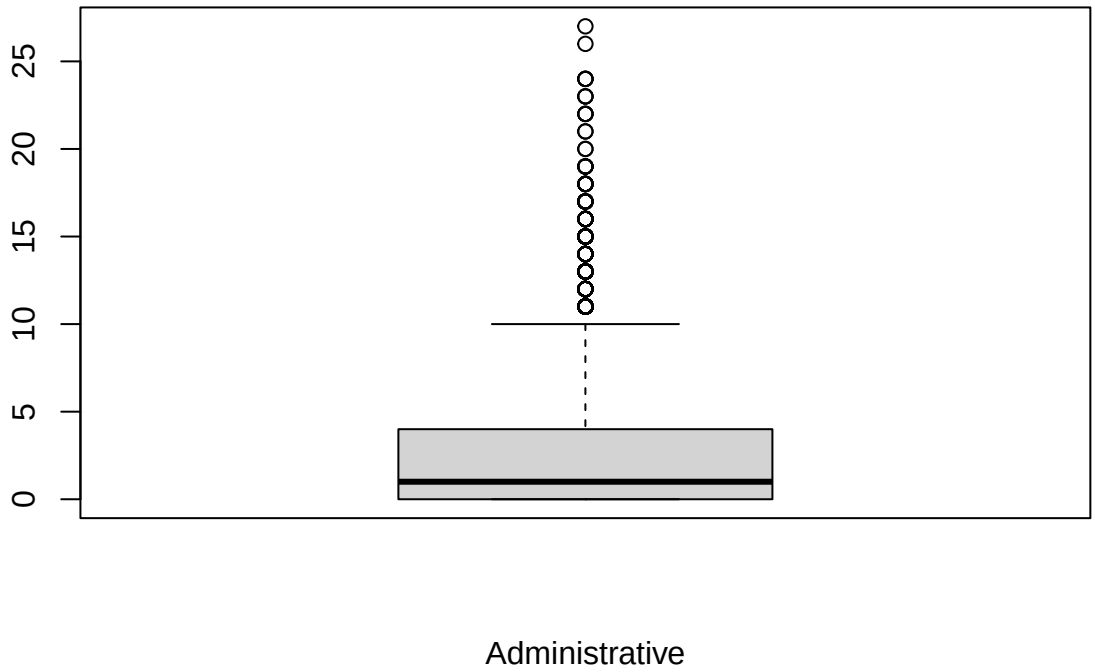
Outliers

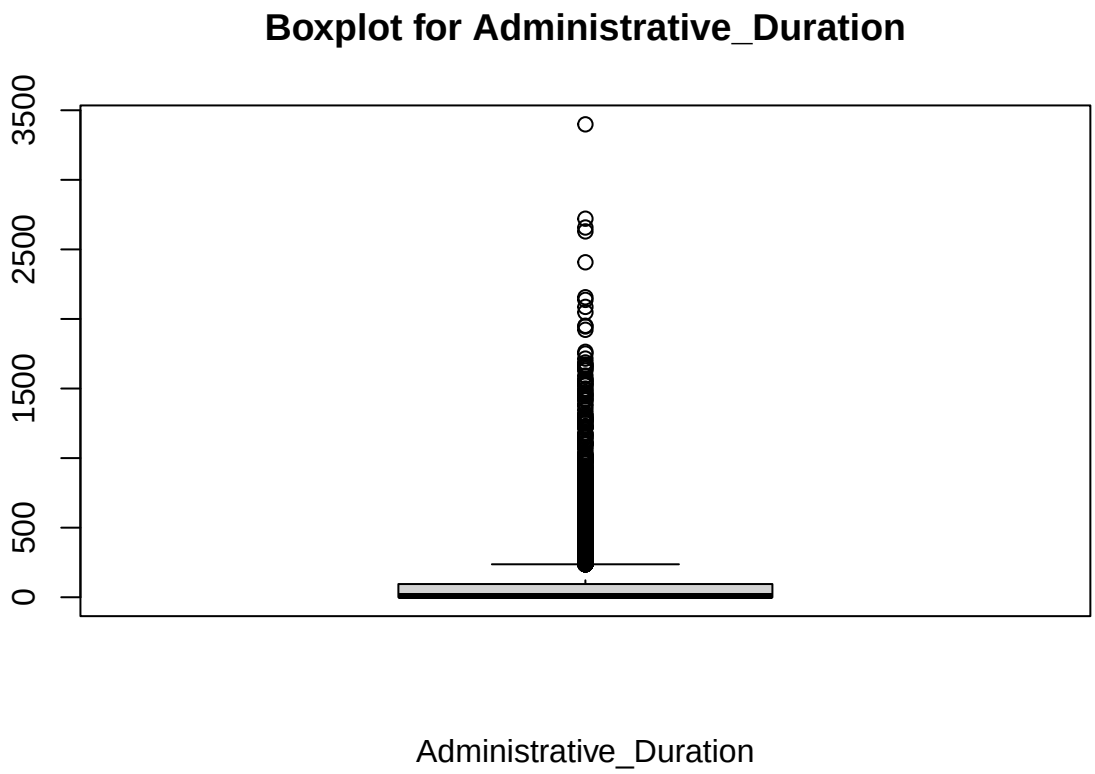
```
#boxplot for whole dataset
boxplot(customers)
```

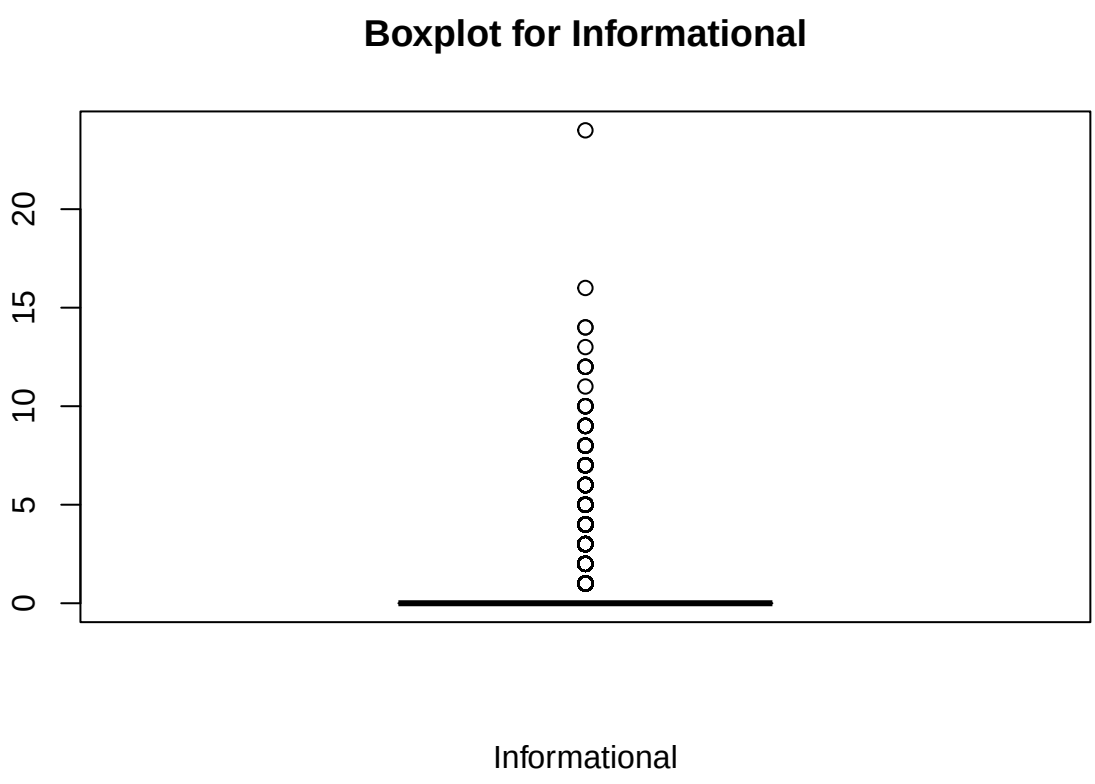


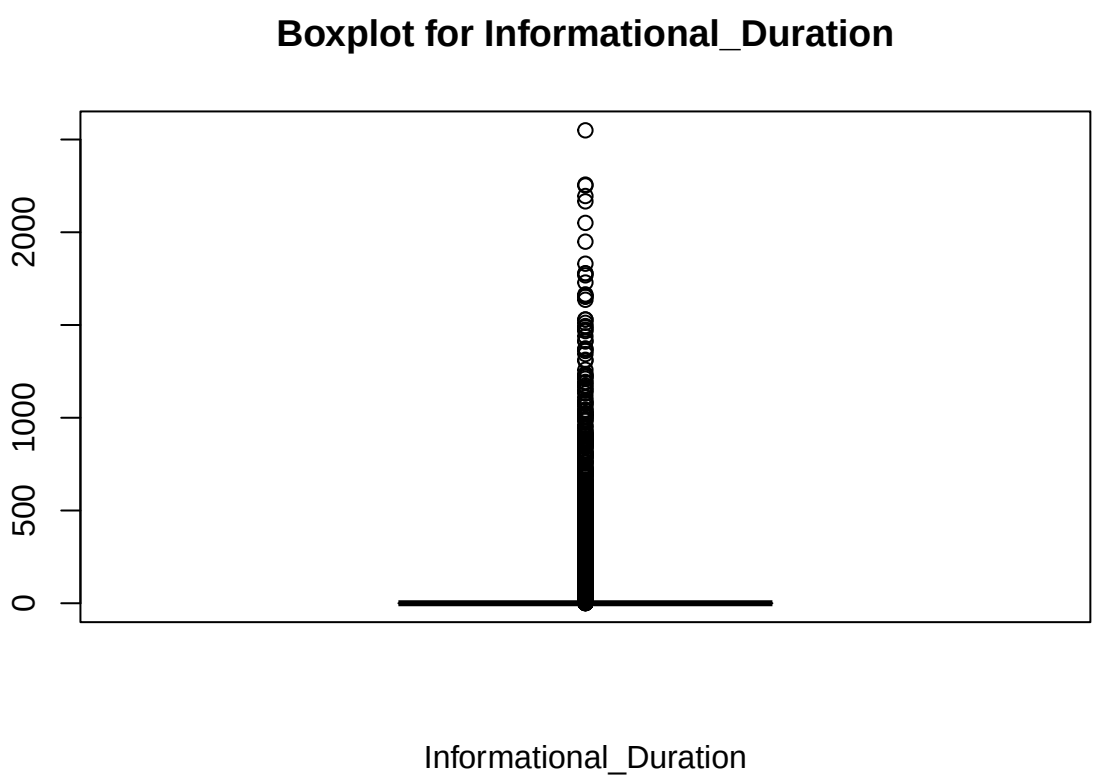
We have outliers in several columns. We plot them individually to check for specific columns clearly.

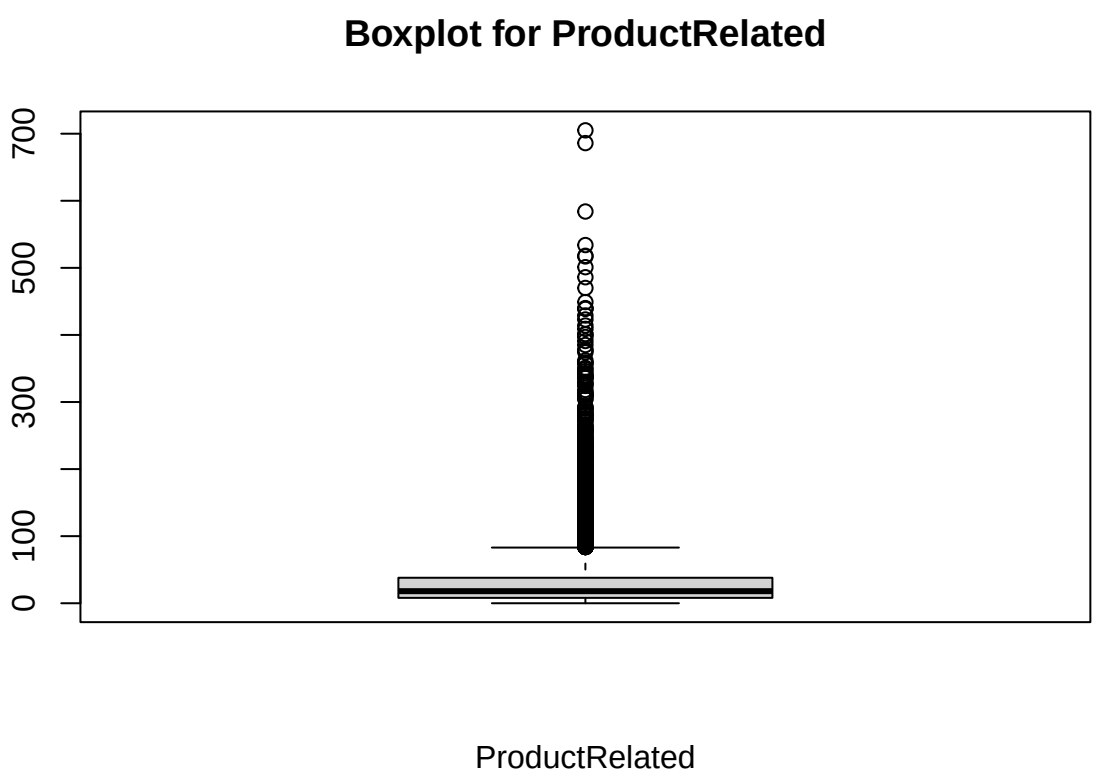
```
num_col <- customers[ ,c(1,2,3,4,5,6,7,8,9,10,12,13,14,15)]
outliers = function(x){
  for(i in colnames(x)){
    boxplot(customers[[i]], xlab=i, main=paste0("Boxplot for ",i))
  }
}
outliers(num_col)
```

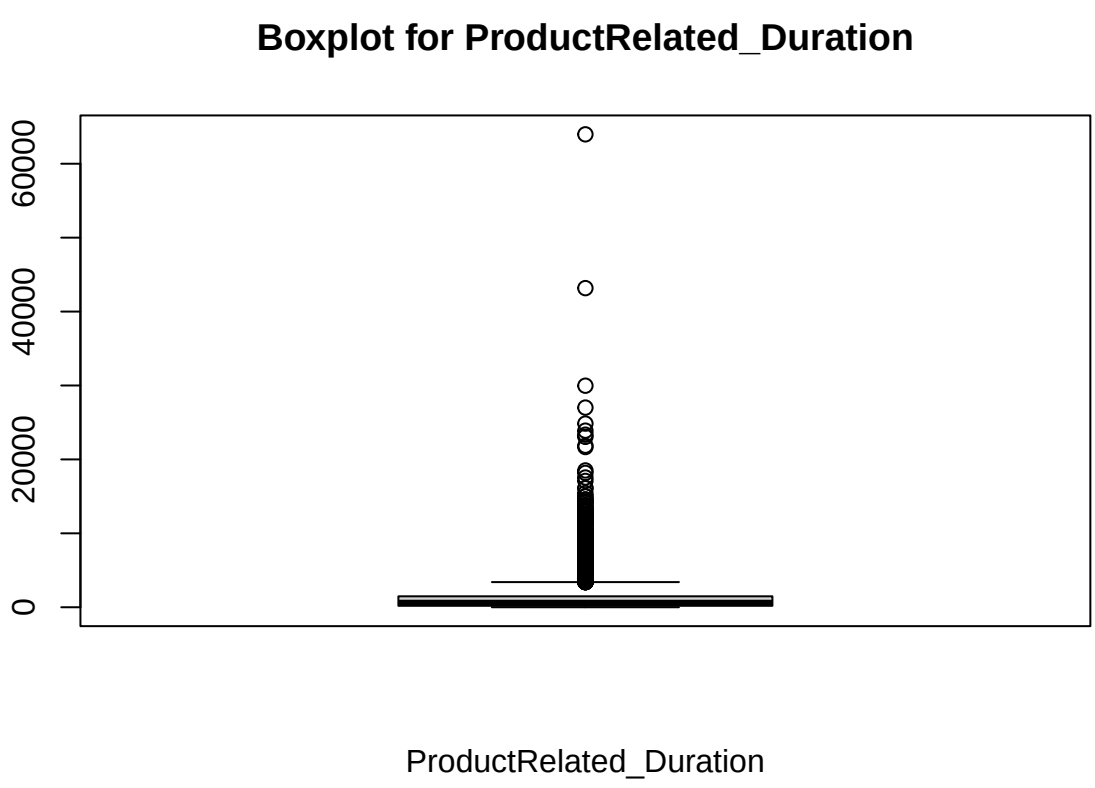

Boxplot for Administrative
0
5
10
15
20
25
Administrative

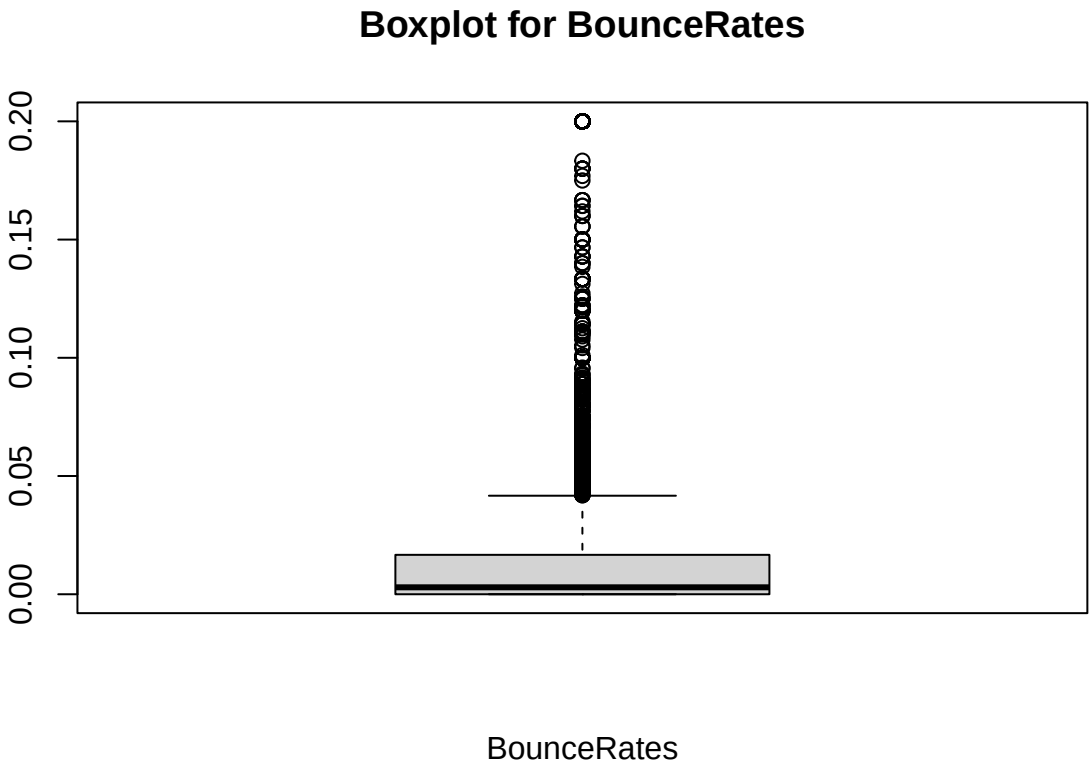

Boxplot for Administrative_Duration
0
500
1500
2500
3500
Administrative_Duration

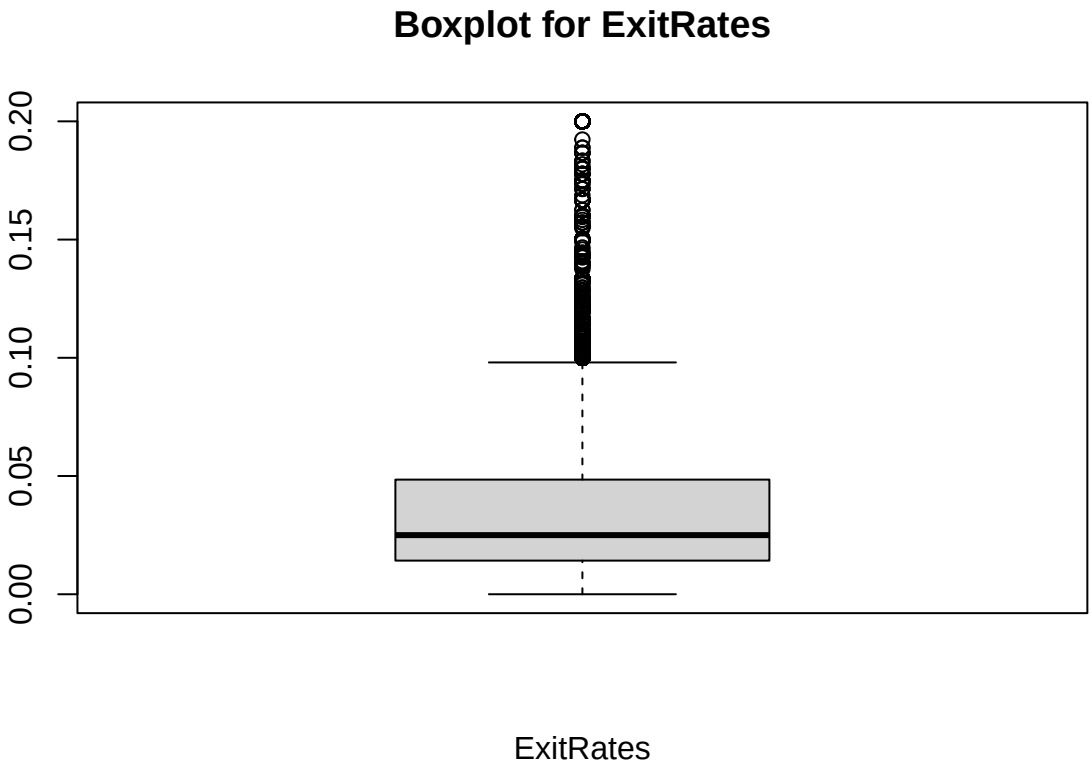

Boxplot for Informational
0
5
10
15
20
Informational

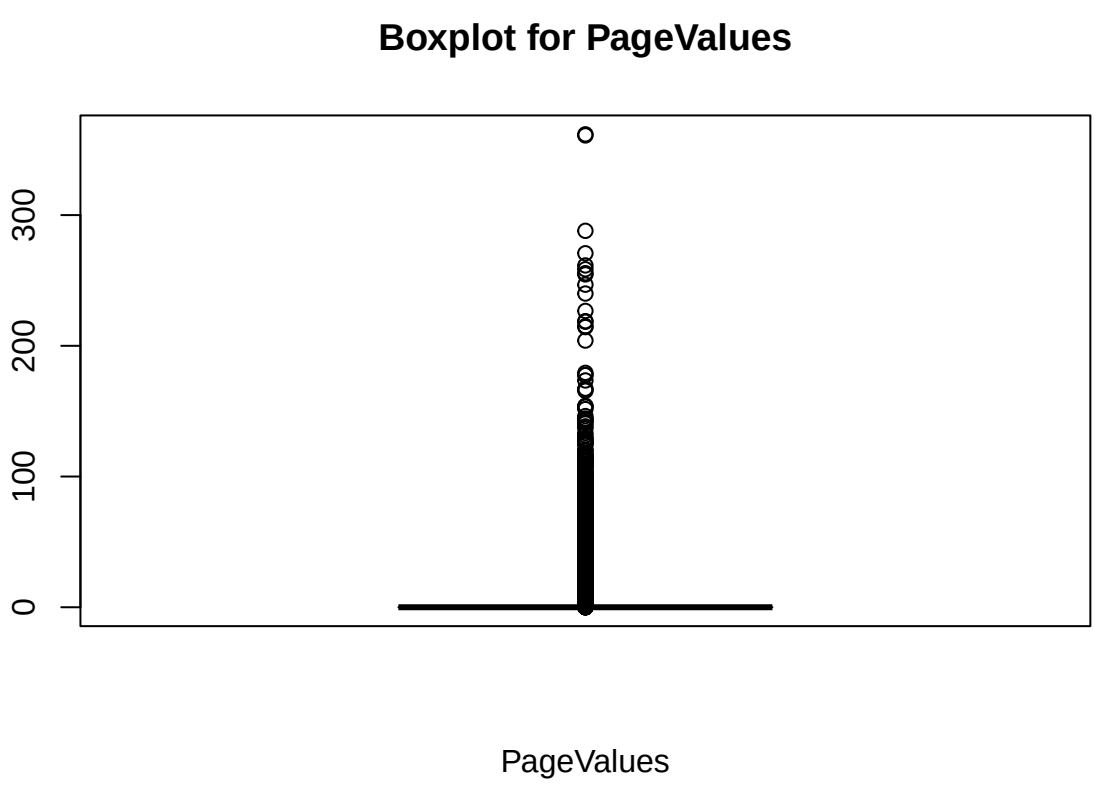

Boxplot for Informational_Duration
0
500
1000
2000
Informational_Duration

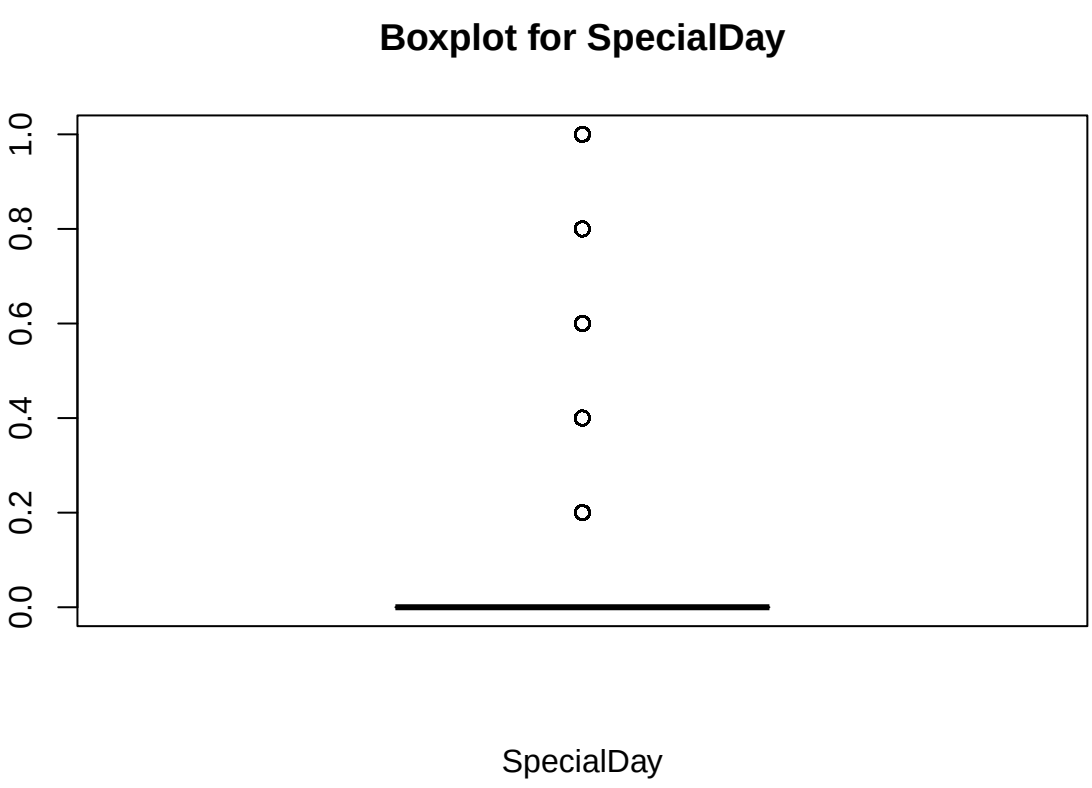

Boxplot for ProductRelated
0
100
300
500
700
ProductRelated

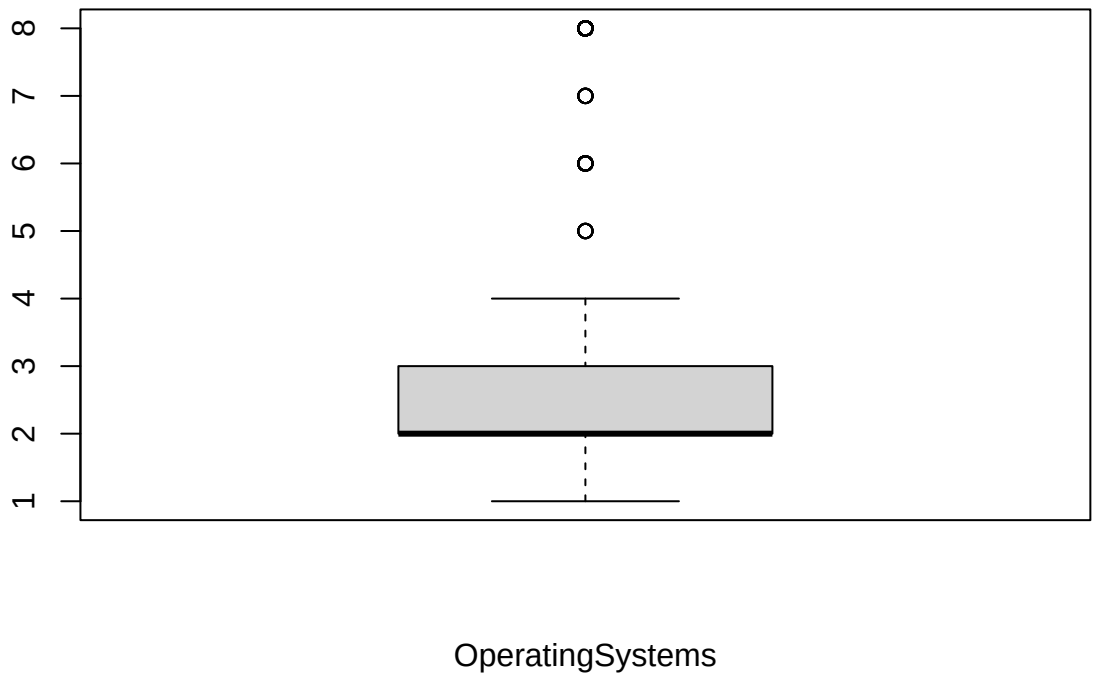

Boxplot for ProductRelated_Duration
0
20000
40000
60000
ProductRelated_Duration

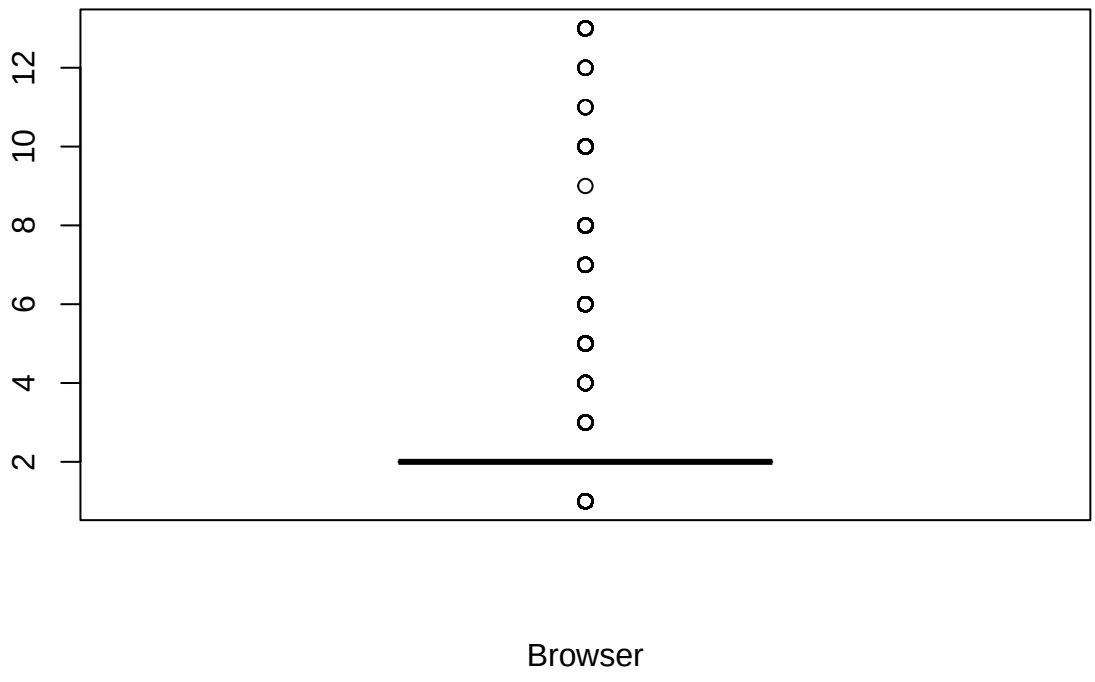

Boxplot for BounceRates
0.00
0.05
0.10
0.15
0.20
BounceRates

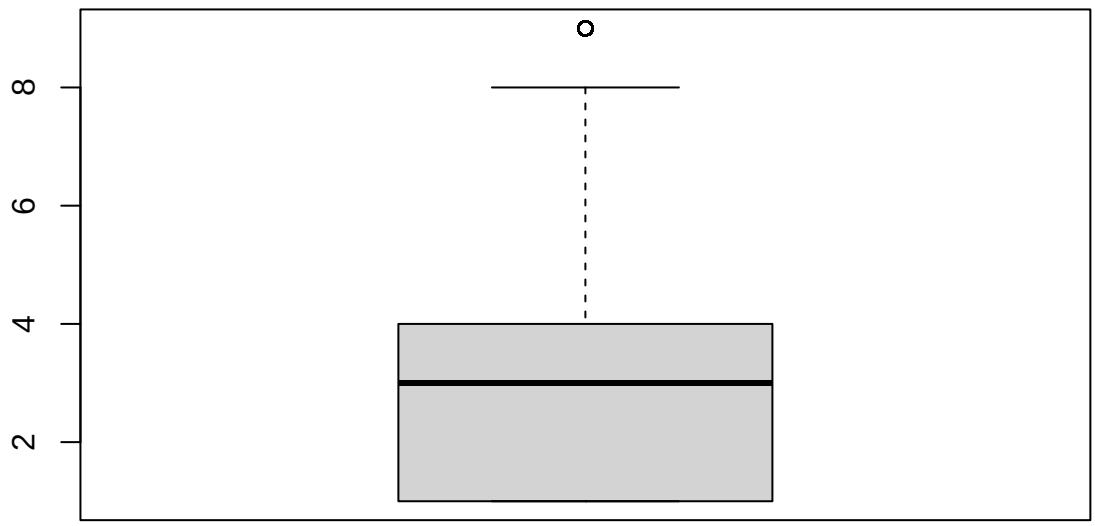

Boxplot for ExitRates
0.00
0.05
0.10
0.15
0.20
ExitRates

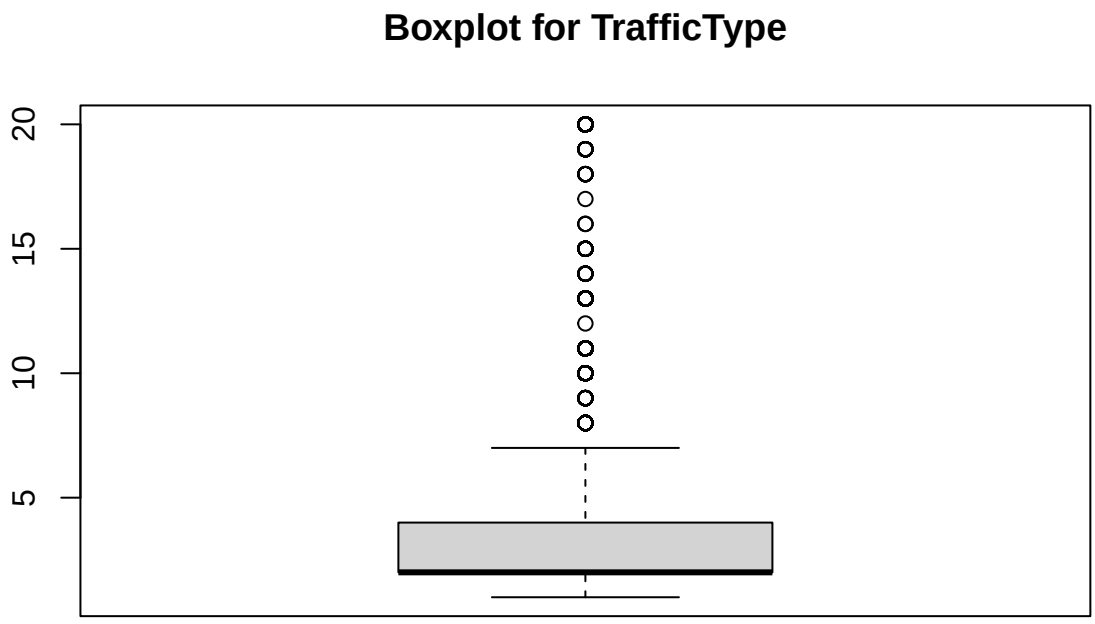

Boxplot for PageValues
0
100
200
300
PageValues


Boxplot for SpecialDay
0.0
0.2
0.4
0.6
0.8
1.0
SpecialDay


Boxplot for OperatingSystems
1
2
3
4
5
6
7
8
OperatingSystems


Boxplot for Browser
2
4
6
8
10
12
Browser

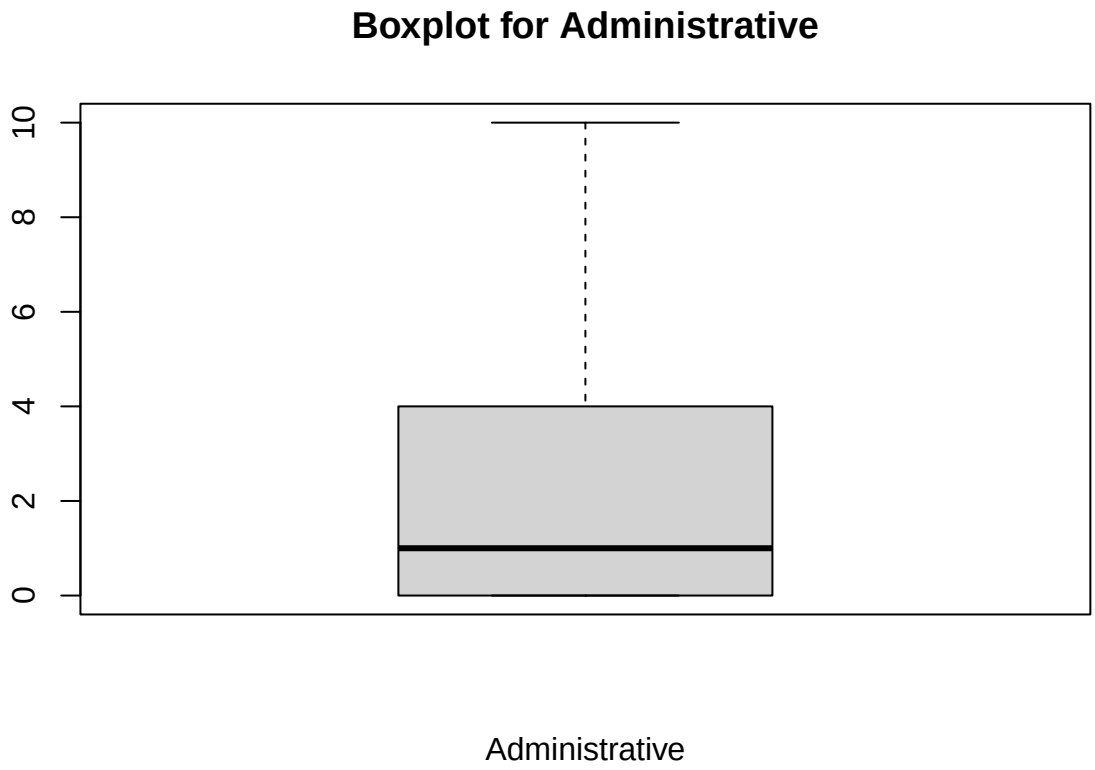

Boxplot for Region
2
4
6
8
Region

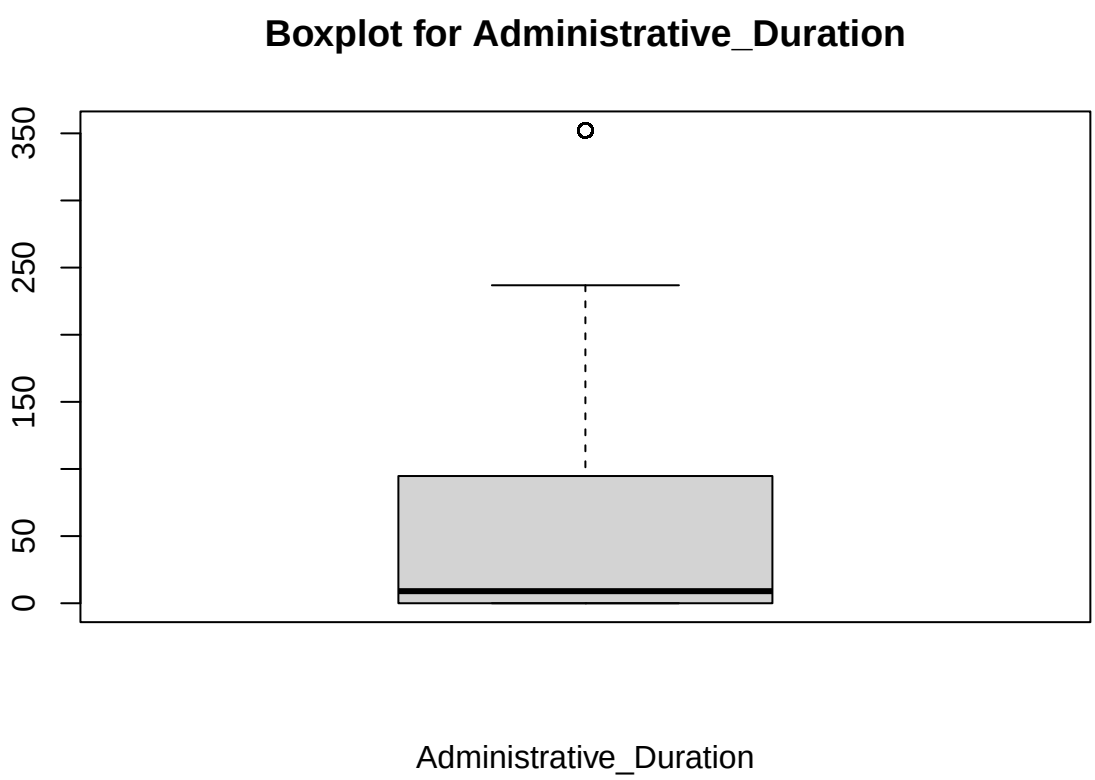


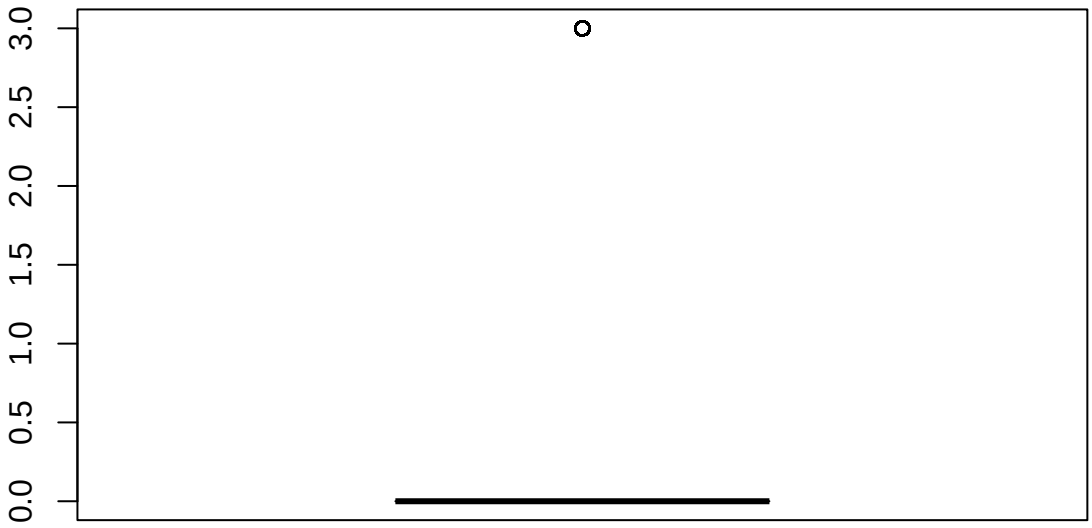
We can see the ouliers more evidently. We will replace outliers with 5th and 95th percentile

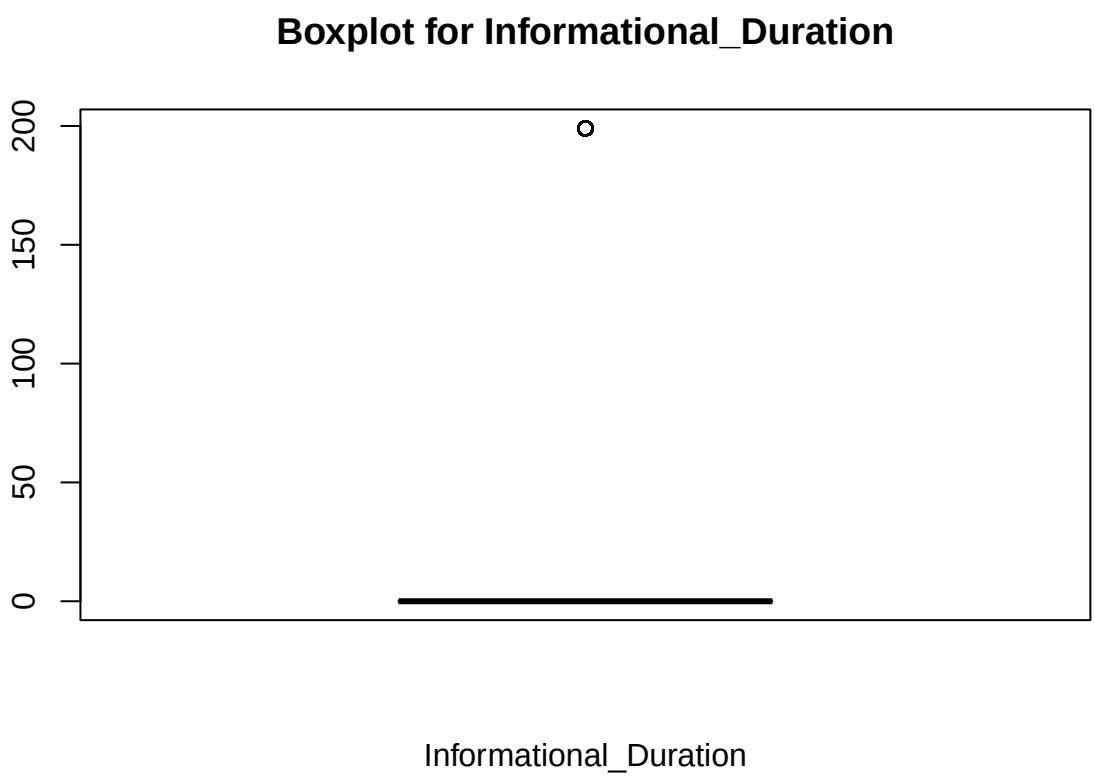
```
outreplace <- function(x){
   qnt <- quantile(x, probs=c(.25, .75), na.rm = T)
   caps <- quantile(x, probs=c(.05, .95), na.rm = T)
   H <- 1.5 * IQR(x, na.rm = T)
   x[x < (qnt[1] - H)] <- caps[1]
   x[x > (qnt[2] + H)] <- caps[2]
   return(x)
}
customers$Administrative <- outreplace(customers$Administrative)
customers$Administrative_Duration <-outreplace(customers$Administrative_Duration)
customers$Informational <- outreplace(customers$Informational)
customers$Informational_Duration <- outreplace(customers$Informational_Duration )
customers$ProductRelated <- outreplace(customers$ProductRelated)
customers$ProductRelated_Duration <- outreplace(customers$ProductRelated_Duration)
customers$BounceRates <- outreplace(customers$BounceRates)
customers$ExitRates <- outreplace(customers$ExitRates)
customers$PageValues <- outreplace(customers$PageValues)
customers$SpecialDay <- outreplace(customers$SpecialDay)
customers$OperatingSystems <- outreplace(customers$OperatingSystems)
customers$Browser <- outreplace(customers$Browser)
customers$Region <- outreplace(customers$Region)
customers$TrafficType <- outreplace(customers$TrafficType)
```

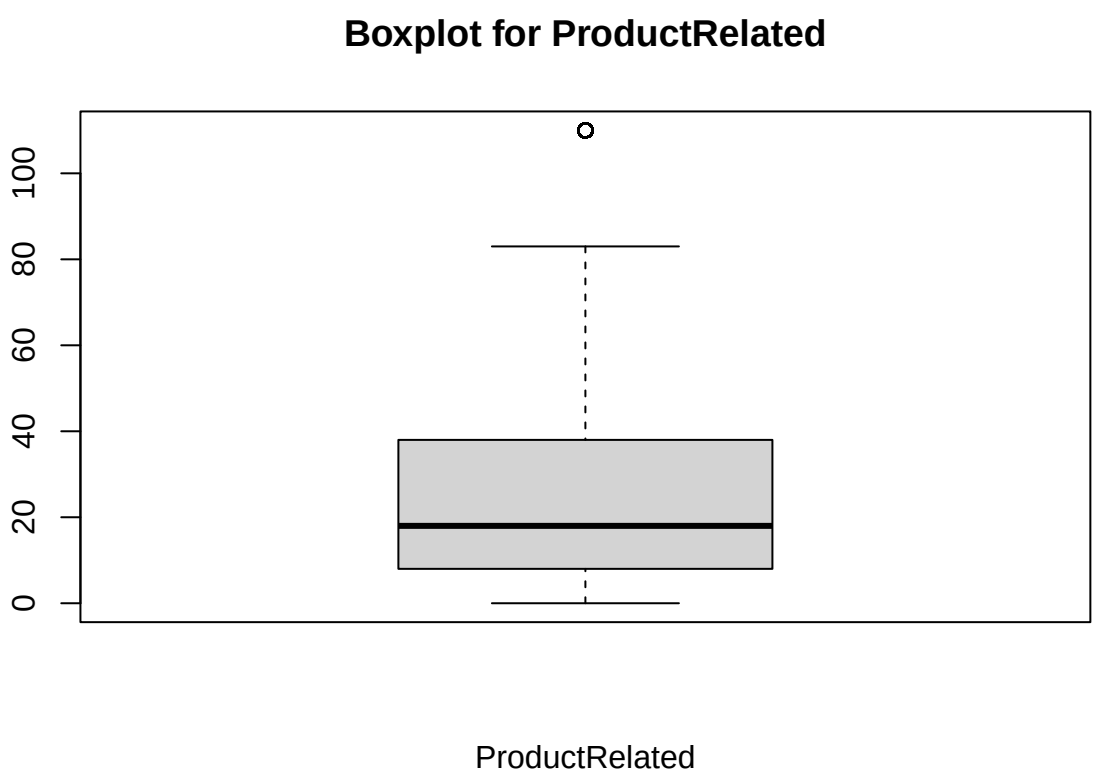
```
outliers(num_col)
```

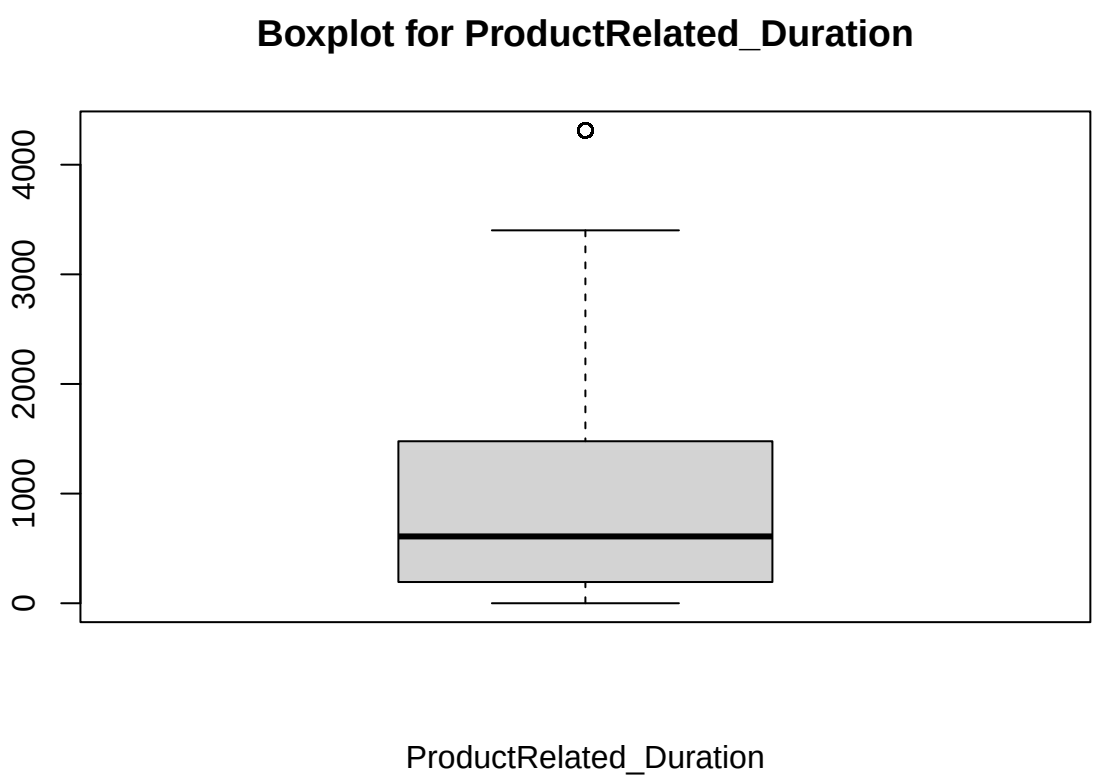


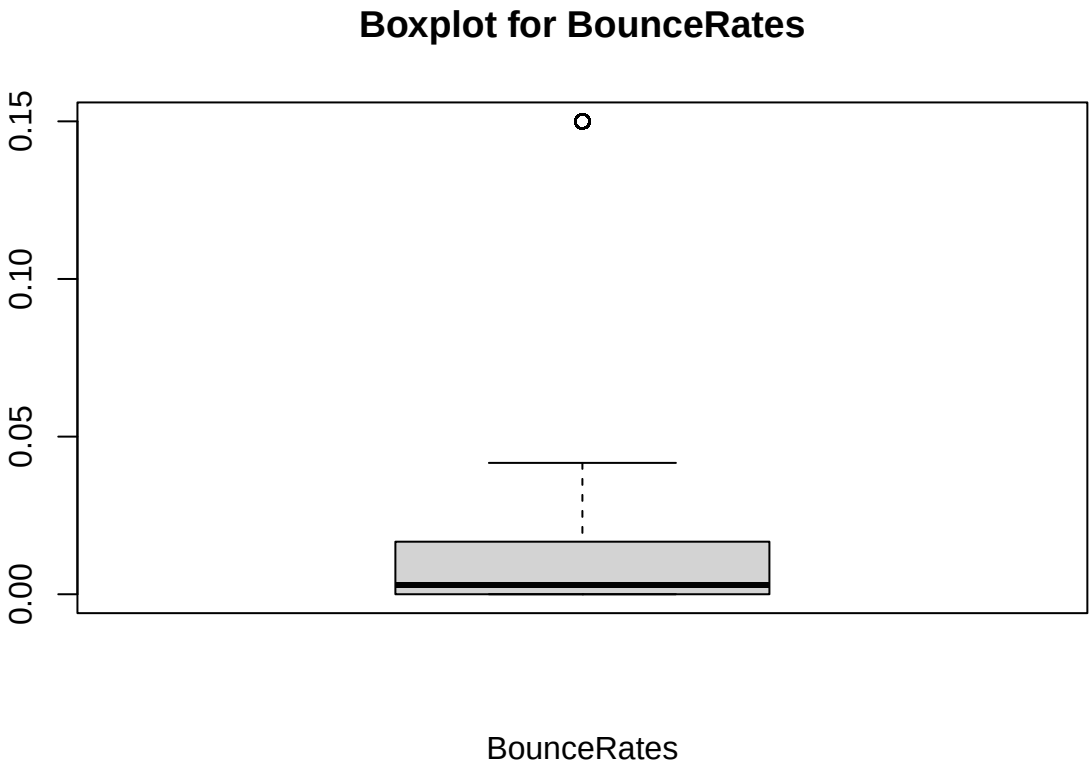

Boxplot for Administrative_Duration
0
50
150
250
350
Administrative_Duration

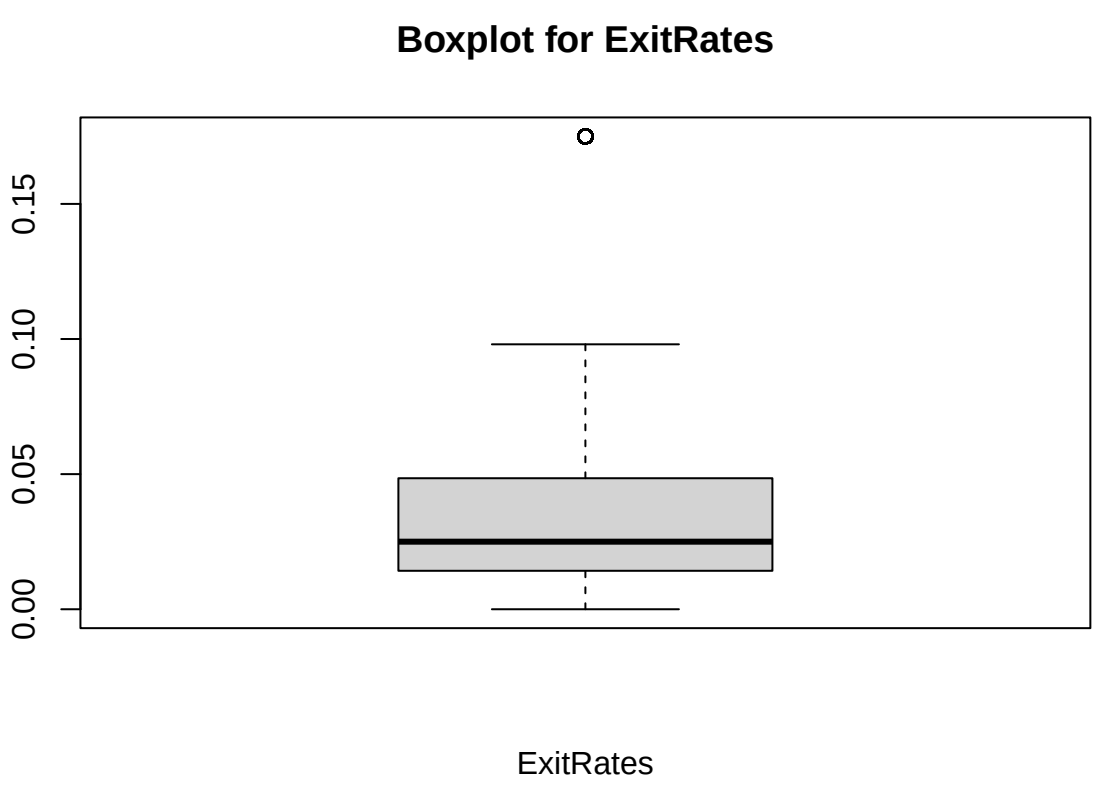

Boxplot for Informational
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Informational

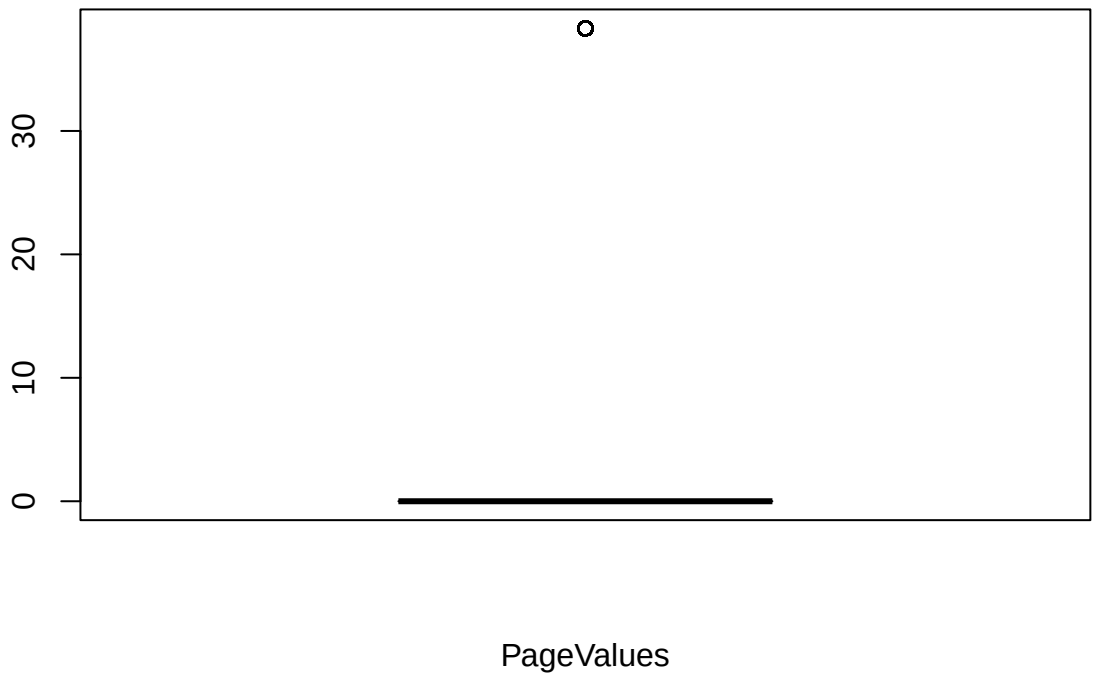

Boxplot for Informational_Duration
0
50
100
150
200
Informational_Duration

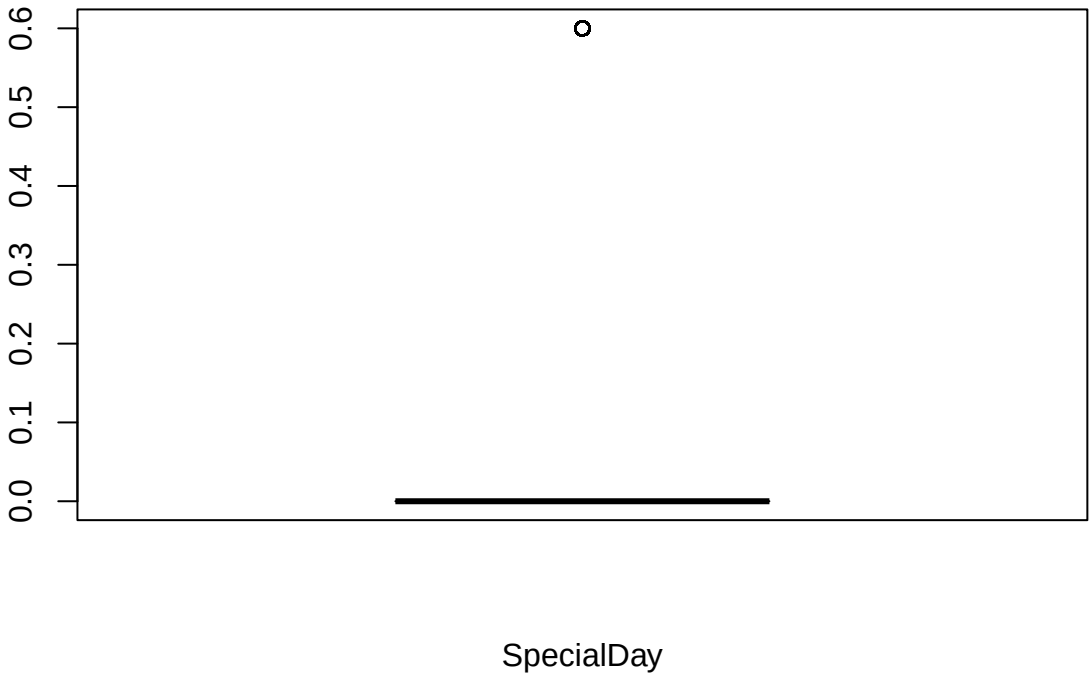

Boxplot for ProductRelated
0
20
40
60
80
100
ProductRelated

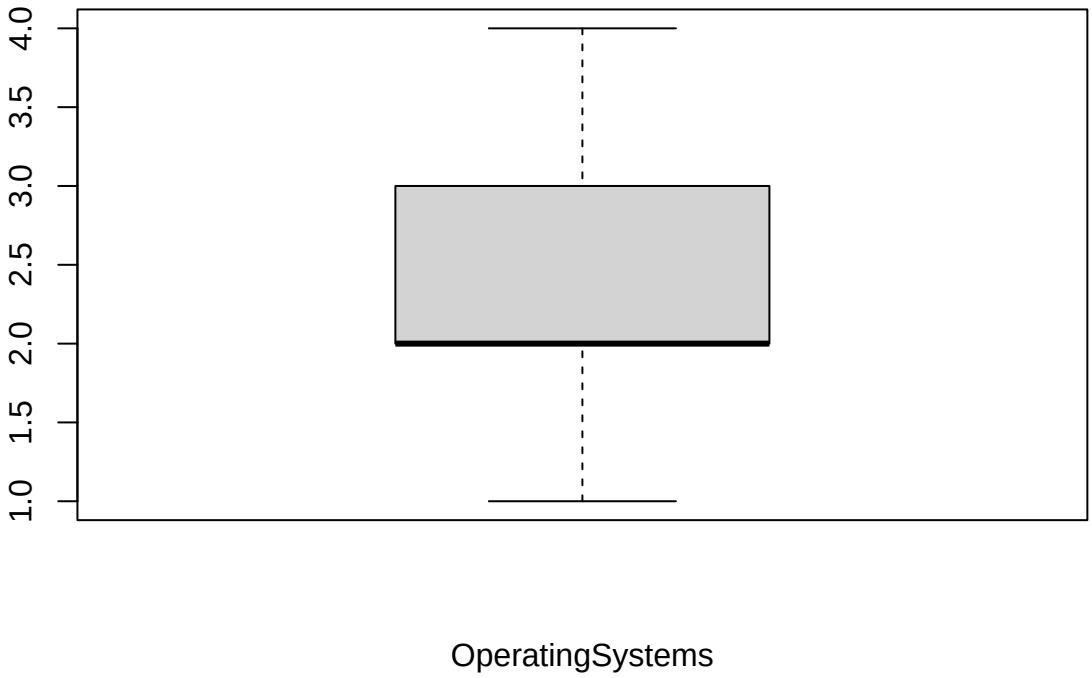

Boxplot for ProductRelated_Duration
0
1000
2000
3000
4000
ProductRelated_Duration

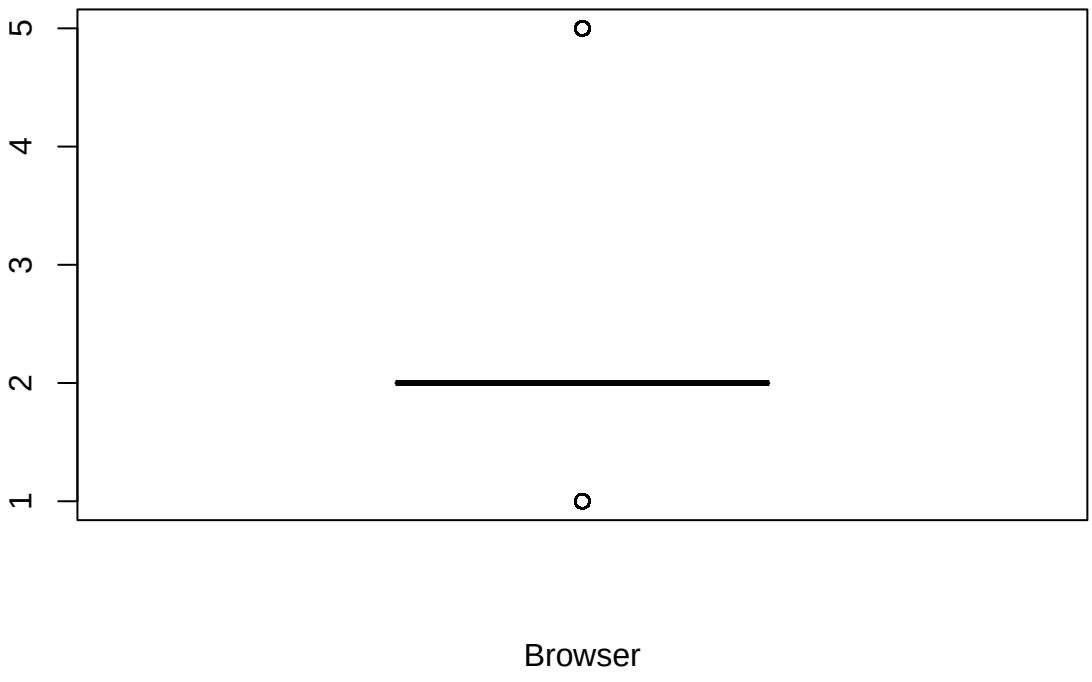

Boxplot for BounceRates
0.15
0.10
0.05
0.00
BounceRates

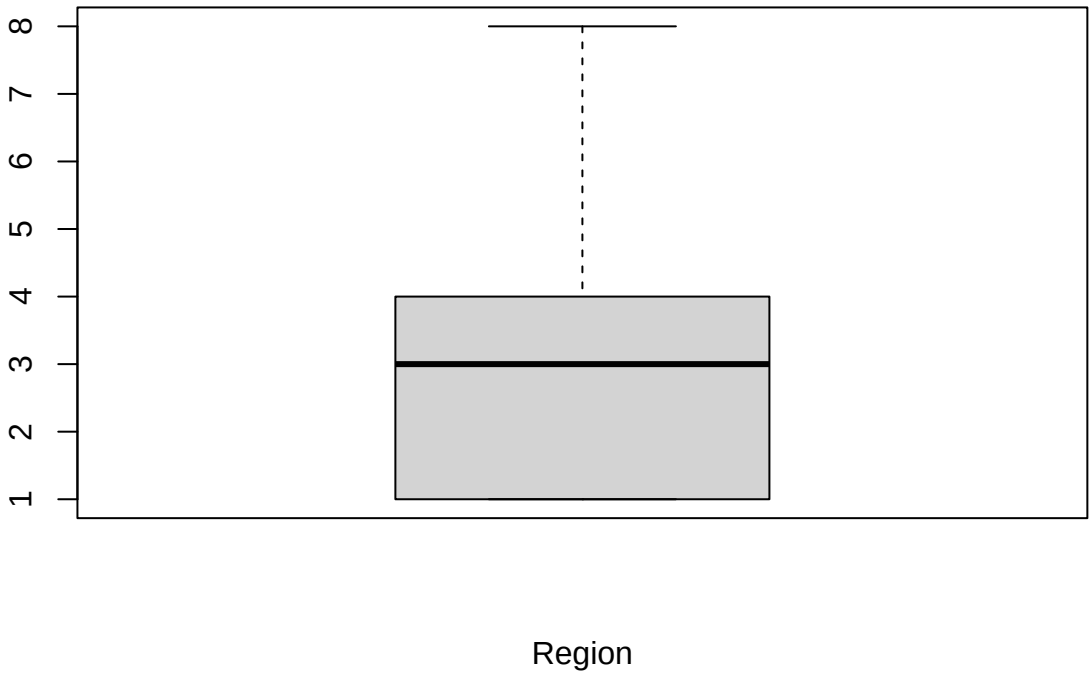

Boxplot for ExitRates
0.00
0.05
0.10
0.15
ExitRates

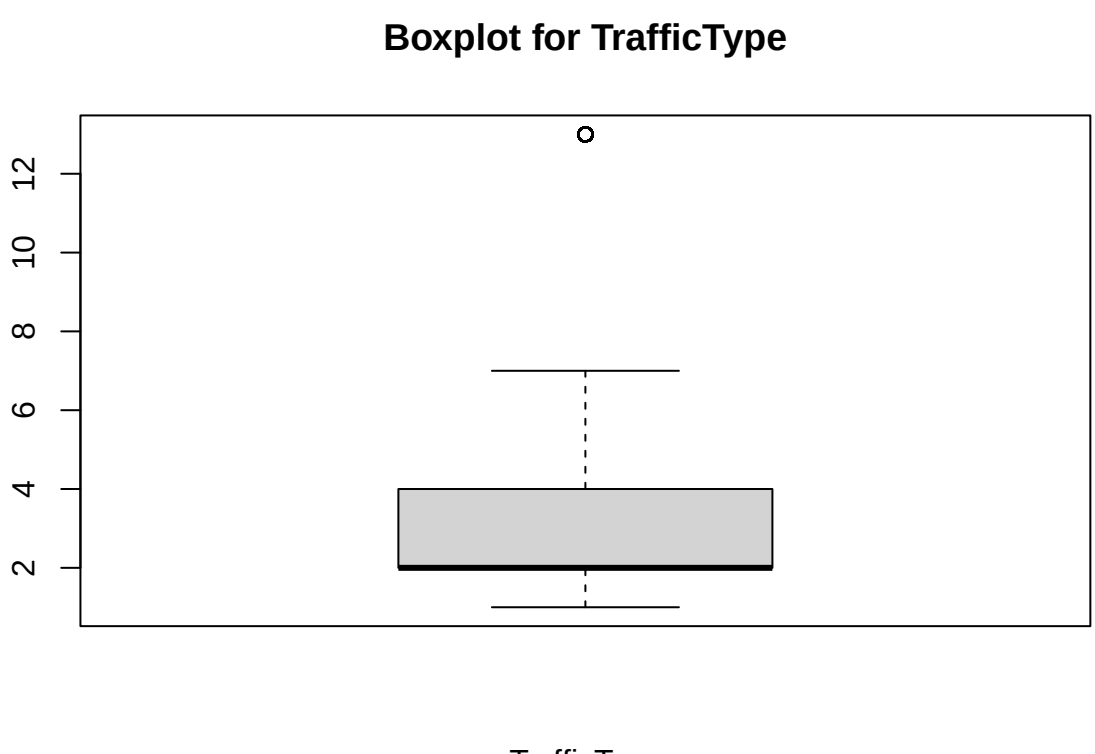

Boxplot for PageValues
0
10
20
30
PageValues


Boxplot for SpecialDay
0.0
0.1
0.2
0.3
0.4
0.5
0.6
SpecialDay


Boxplot for OperatingSystems
4.0
3.5
3.0
2.5
2.0
1.5
1.0
OperatingSystems


Boxplot for Browser
5
4
3
2
1
Browser


Boxplot for Region
1
2
3
4
5
6
7
8
Region



Most of the outliers are replaced. We decided toleave the remaining ones.

```
dim(customers)
```

```
## [1] 12199    18
```

Our final cleaned data.frame is left with 12199 rows and 18 columns.

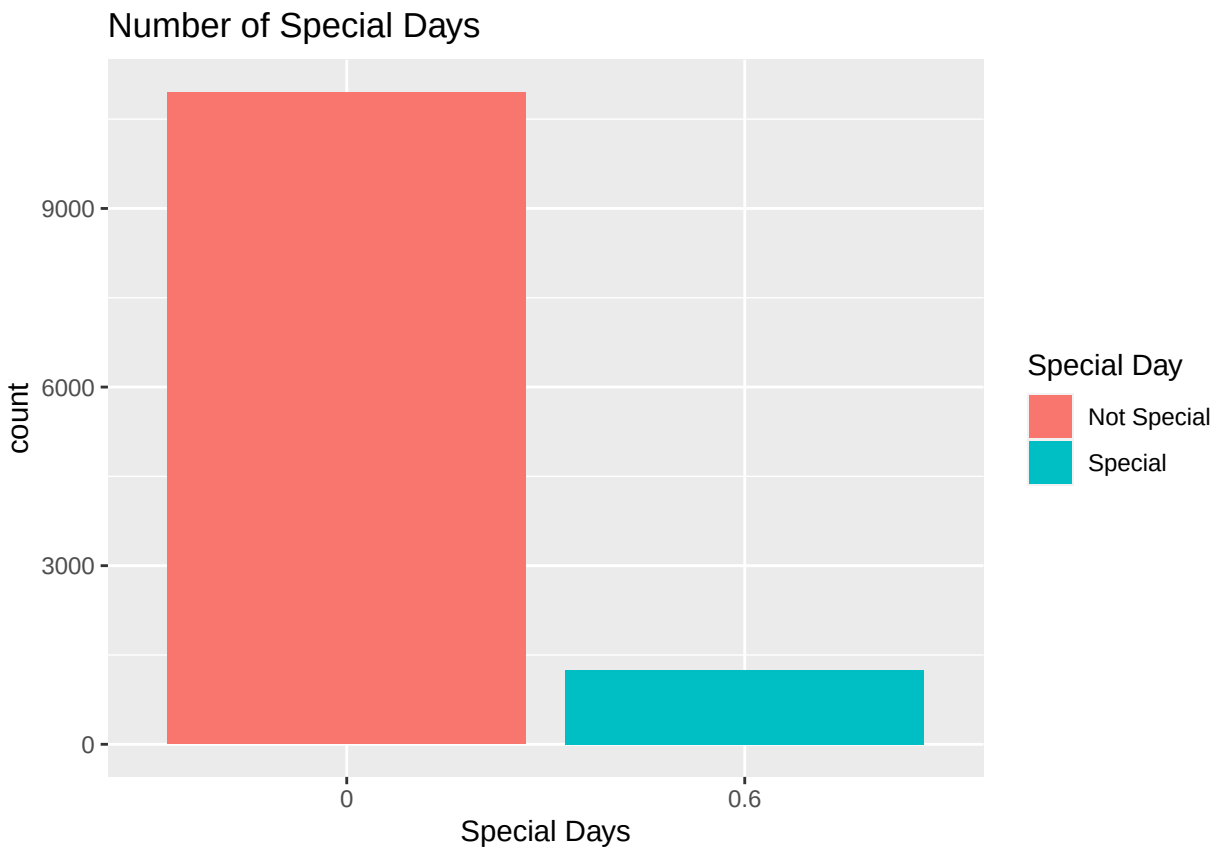
Univariate Analysis

```
#summary of the descriptive staistics of the columns
summary(customers)
```

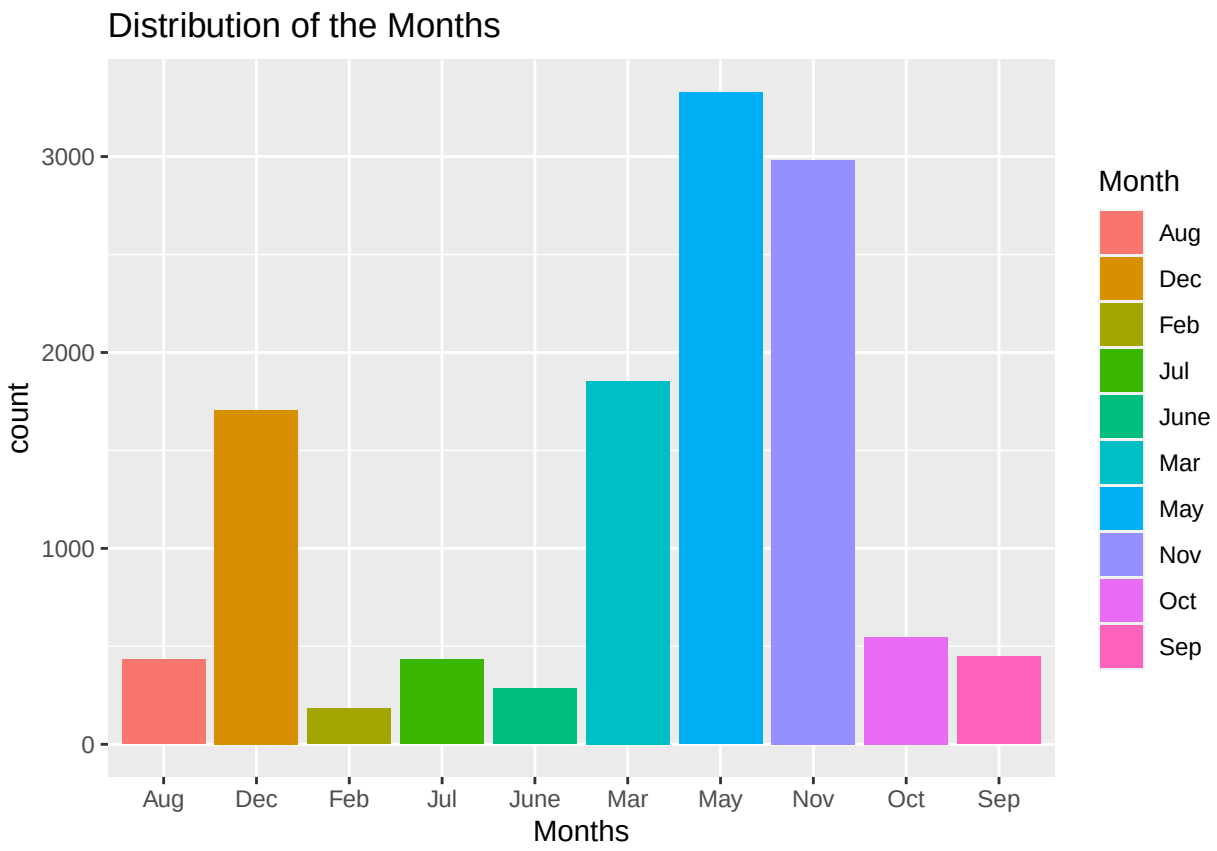
```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :  0.00          Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:  0.00          1st Qu.:0.0000
##  Median : 1.000   Median :  9.00          Median :0.0000
##  Mean   : 2.189   Mean   : 68.78          Mean   :0.6468
##  3rd Qu.: 4.000   3rd Qu.: 94.75          3rd Qu.:0.0000
##  Max.   :10.000   Max.   :352.23          Max.   :3.0000
##
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :  0.00         Min.   :  0.00   Min.   :   0.0
##  1st Qu.:  0.00         1st Qu.:  8.00   1st Qu.: 193.6
##  Median :  0.00         Median : 18.00   Median : 609.5
##  Mean   : 39.22         Mean   : 29.07   Mean   :1072.7
```

```
##  3rd Qu.:  0.00         3rd Qu.: 38.00   3rd Qu.:1477.6
##  Max.   :199.00         Max.   :110.00   Max.   :4313.5
##
##   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   :0.02329   Mean   :0.04363   Mean   : 8.574   Mean   :0.06143
##  3rd Qu.:0.01667   3rd Qu.:0.04848   3rd Qu.: 0.000   3rd Qu.:0.00000
##  Max.   :0.15000   Max.   :0.17500   Max.   :38.312   Max.   :0.60000
##
##      Month      OperatingSystems    Browser          Region
##  May    :3328   Min.   :1.000    Min.   :1.000   Min.   :1.000
##  Nov    :2983   1st Qu.:2.000    1st Qu.:2.000   1st Qu.:1.000
##  Mar    :1853   Median :2.000    Median :2.000   Median :3.000
##  Dec    :1706   Mean   :2.086    Mean   :2.267   Mean   :3.112
##  Oct    : 549   3rd Qu.:3.000    3rd Qu.:2.000   3rd Qu.:4.000
##  Sep    : 448   Max.   :4.000    Max.   :5.000   Max.   :8.000
##  (Other):1332
##   TrafficType                VisitorType     Weekend      Revenue
##  Min.   : 1.000   New_Visitor      : 1693   FALSE:9343   FALSE:10291
##  1st Qu.: 2.000   Other            :   81   TRUE :2856   TRUE : 1908
##  Median : 2.000   Returning_Visitor:10425
##  Mean   : 4.249
##  3rd Qu.: 4.000
##  Max.   :13.000
##
```

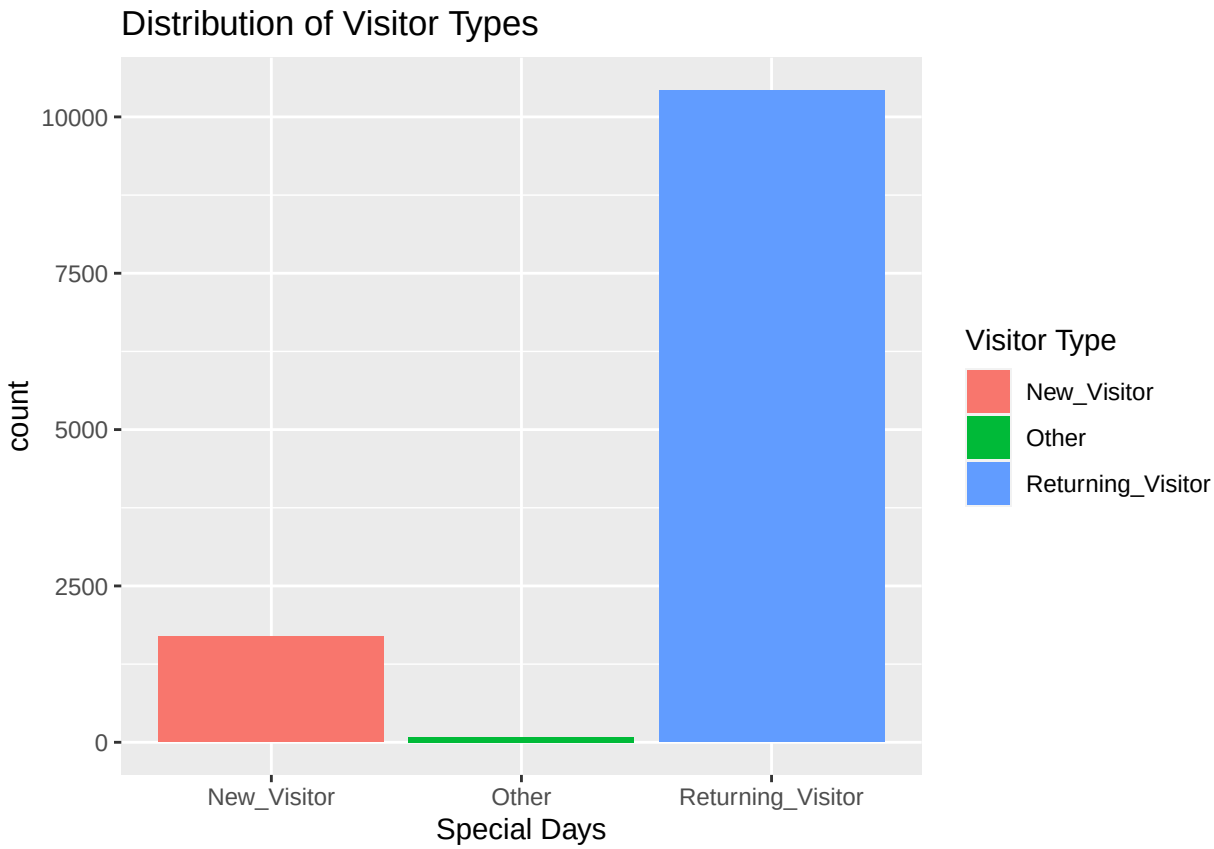
```
#plotting barplots of the categorical columns
library("ggplot2" )

# Group Special Days
specialday <- ggplot(customers, aes(x=factor(SpecialDay), fill = factor(SpecialDay))) + geom_bar()
specialday +  scale_fill_discrete(name = "Special Day", labels = c("Not Special","Special"))+ labs(titl
```



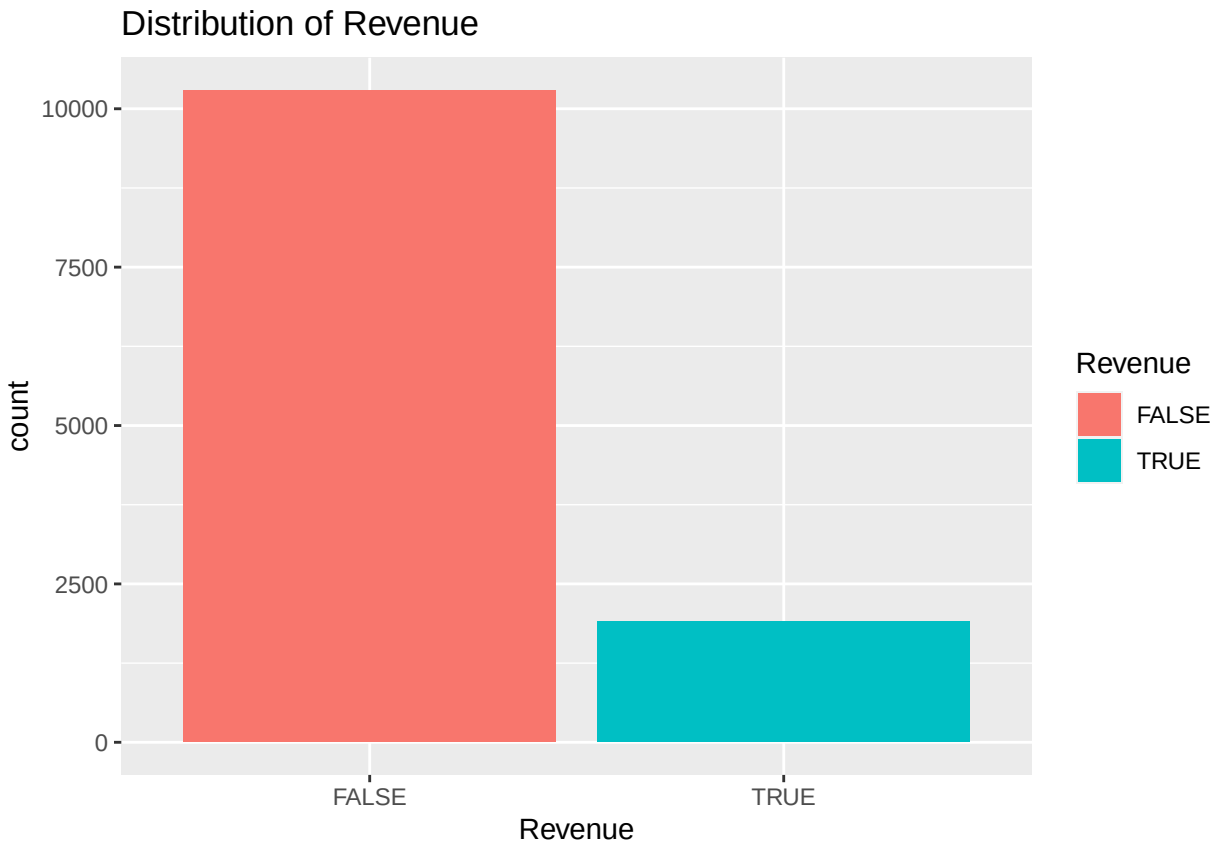

```
# Count the Months
months <- ggplot(customers ,aes(x=Month , fill=factor(Month))) + geom_bar() + labs(title = "Distributio
months +scale_fill_discrete(name = "Month")
```



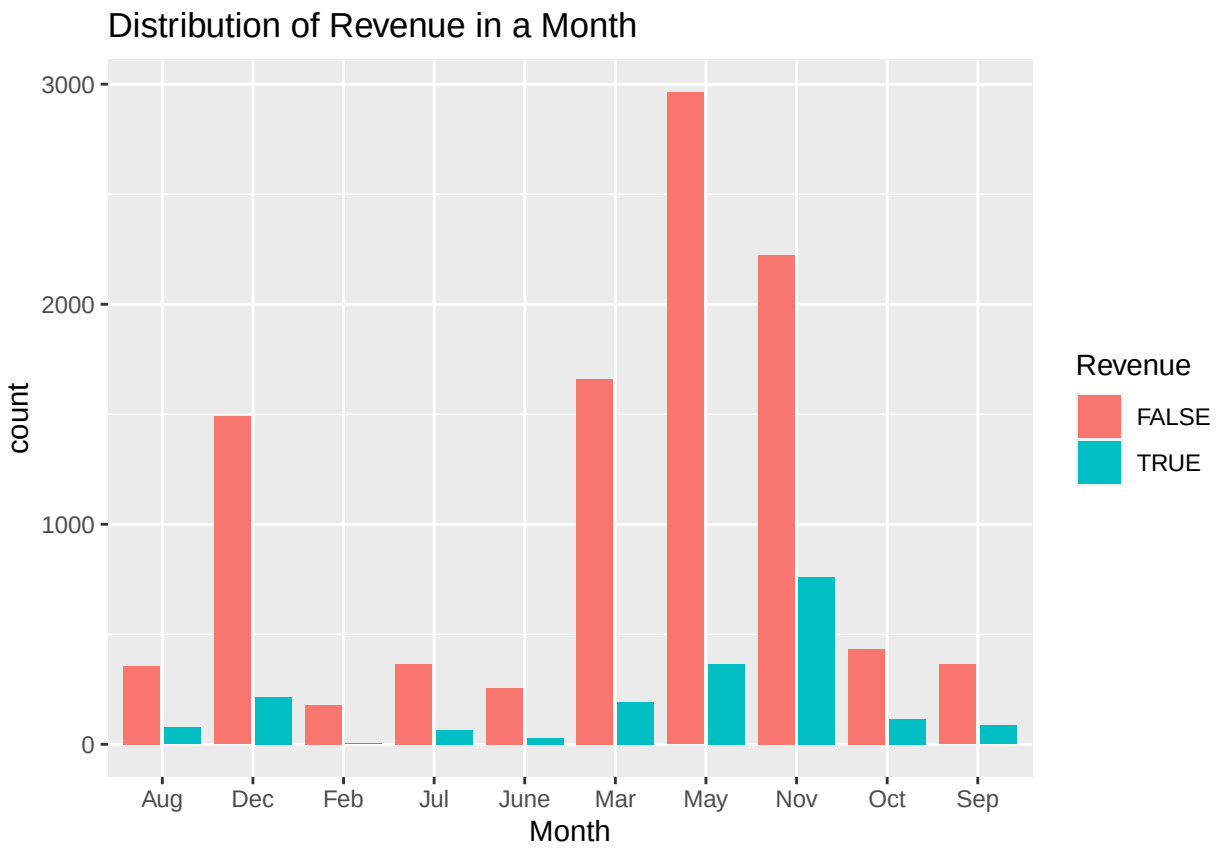

```
# Count on Visitor Type
visitor <- ggplot(customers, aes(VisitorType, fill=factor(VisitorType)))+ geom_bar() + labs(title = "Dis
visitor + scale_fill_discrete(name = "Visitor Type")
```



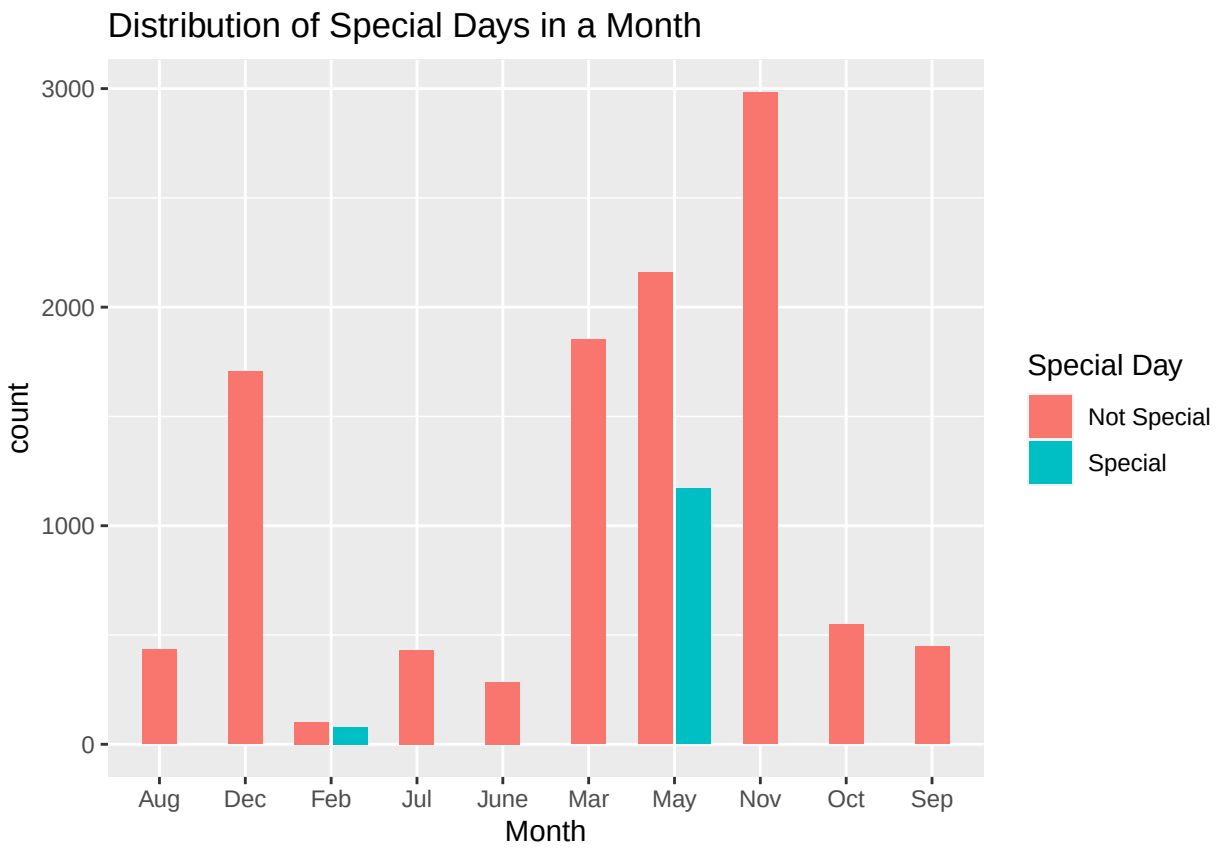

```
# Count on Revenue
revenue <- ggplot(customers, aes(Revenue, fill=factor(Revenue))) +geom_bar() + labs(title = "Distributi
revenue +scale_fill_discrete(name = "Revenue")
```



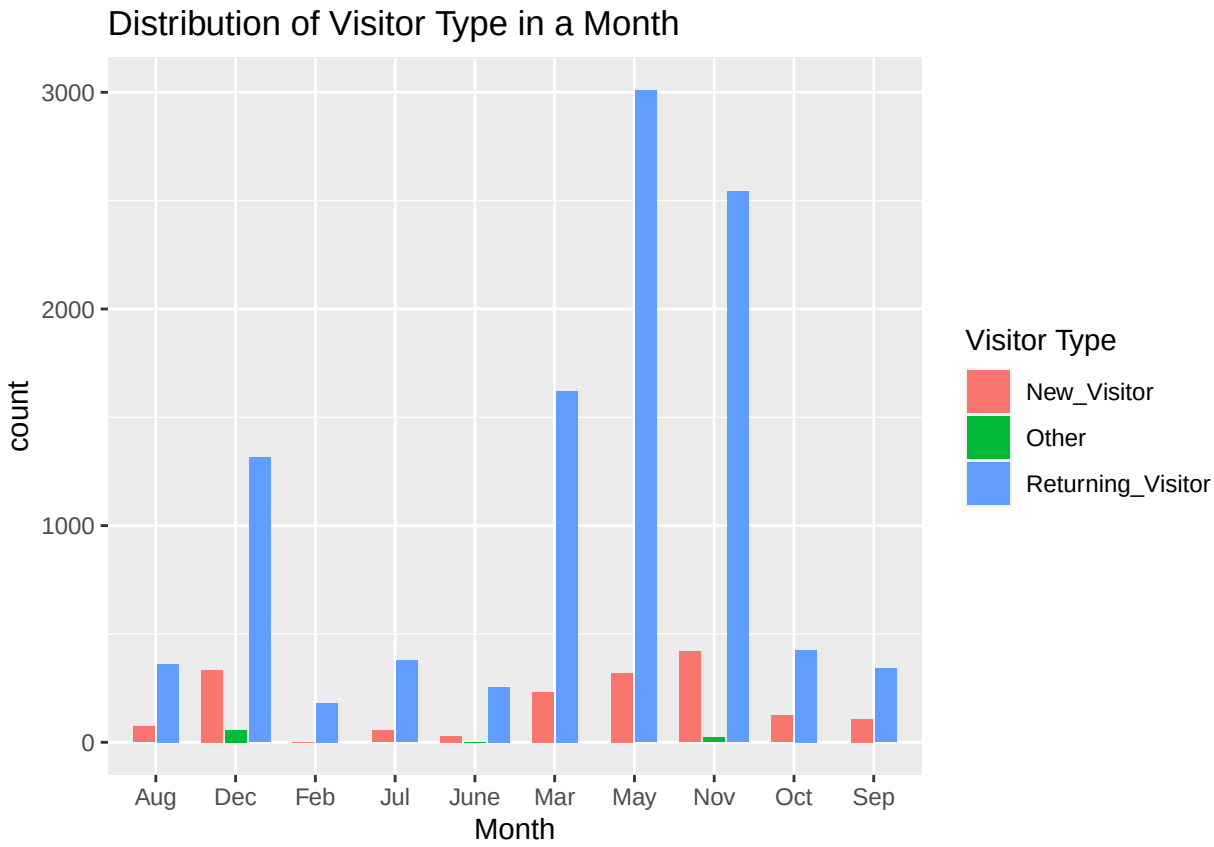

```
# Group Revenue by Months
revenue1 <- ggplot(customers, aes(x=Month, fill= factor(Revenue)))+ geom_bar(position=position_dodge2(w
revenue1 + labs(title = "Distribution of Revenue in a Month") +scale_fill_discrete(name = "Revenue")
```



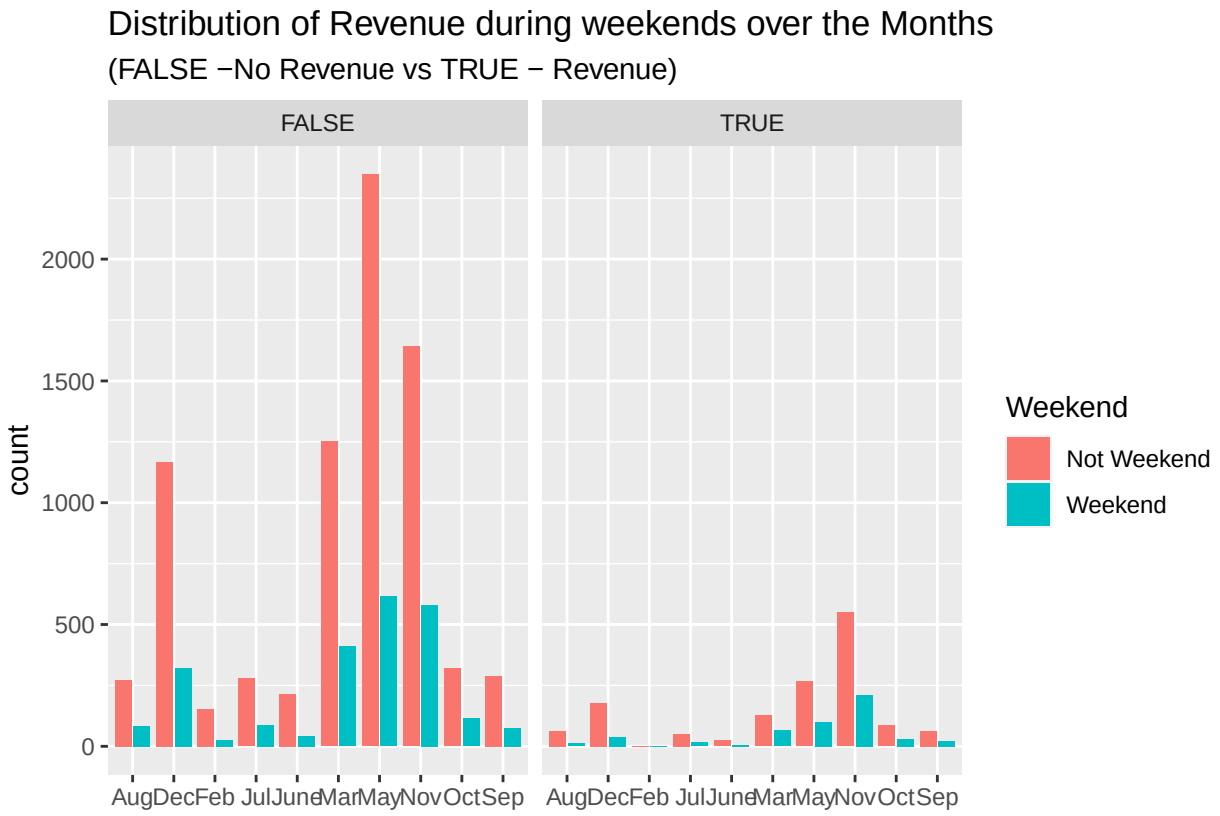

```
# Group Special Days by Month
specialday1 <- ggplot(customers, aes(x=Month, fill= factor(SpecialDay)))+ geom_bar(position = position_
specialday1 + scale_fill_discrete(name = "Special Day", labels = c("Not Special","Special")) + labs(tit
```



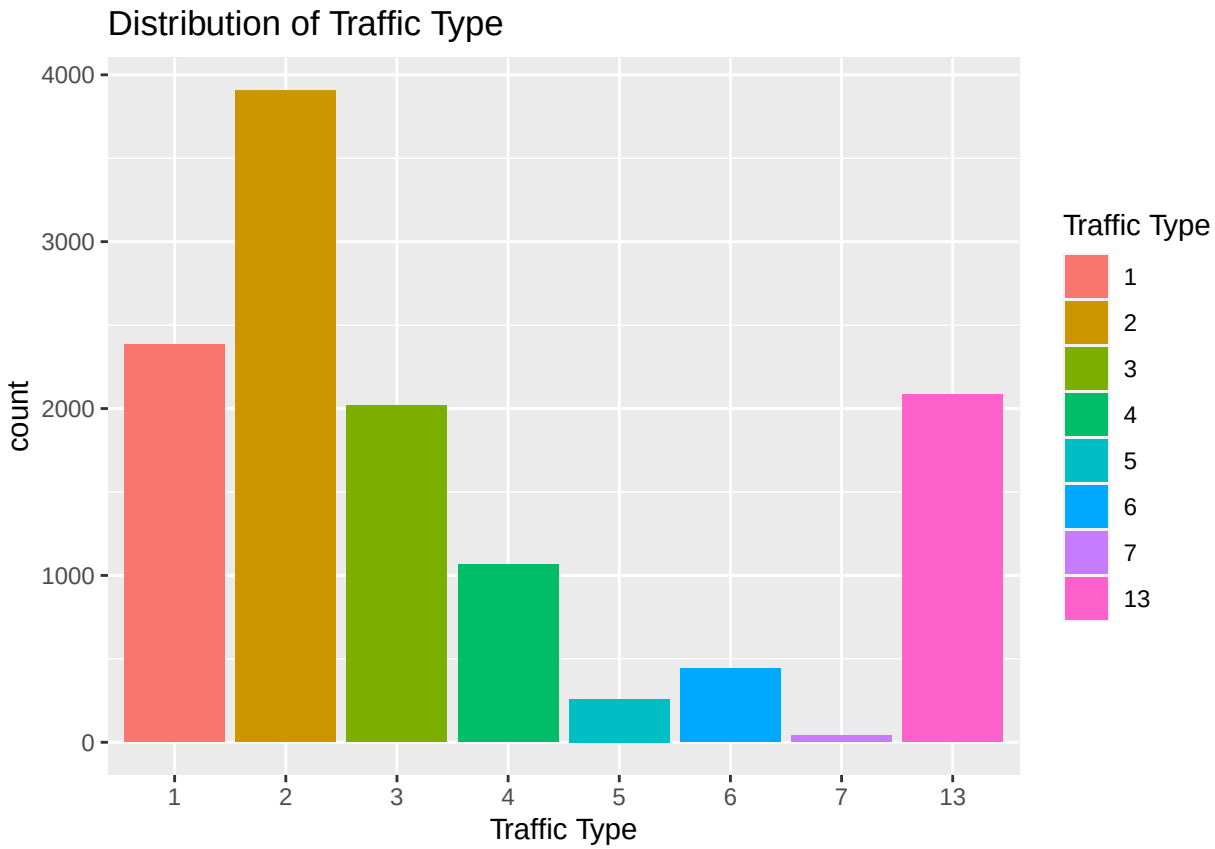

```
# Group Visitor Type by Month
visitor1 <- ggplot(customers, aes(x=Month, fill=factor(VisitorType)))+geom_bar(position=position_dodge2
visitor1 + scale_fill_discrete(name = "Visitor Type")
```



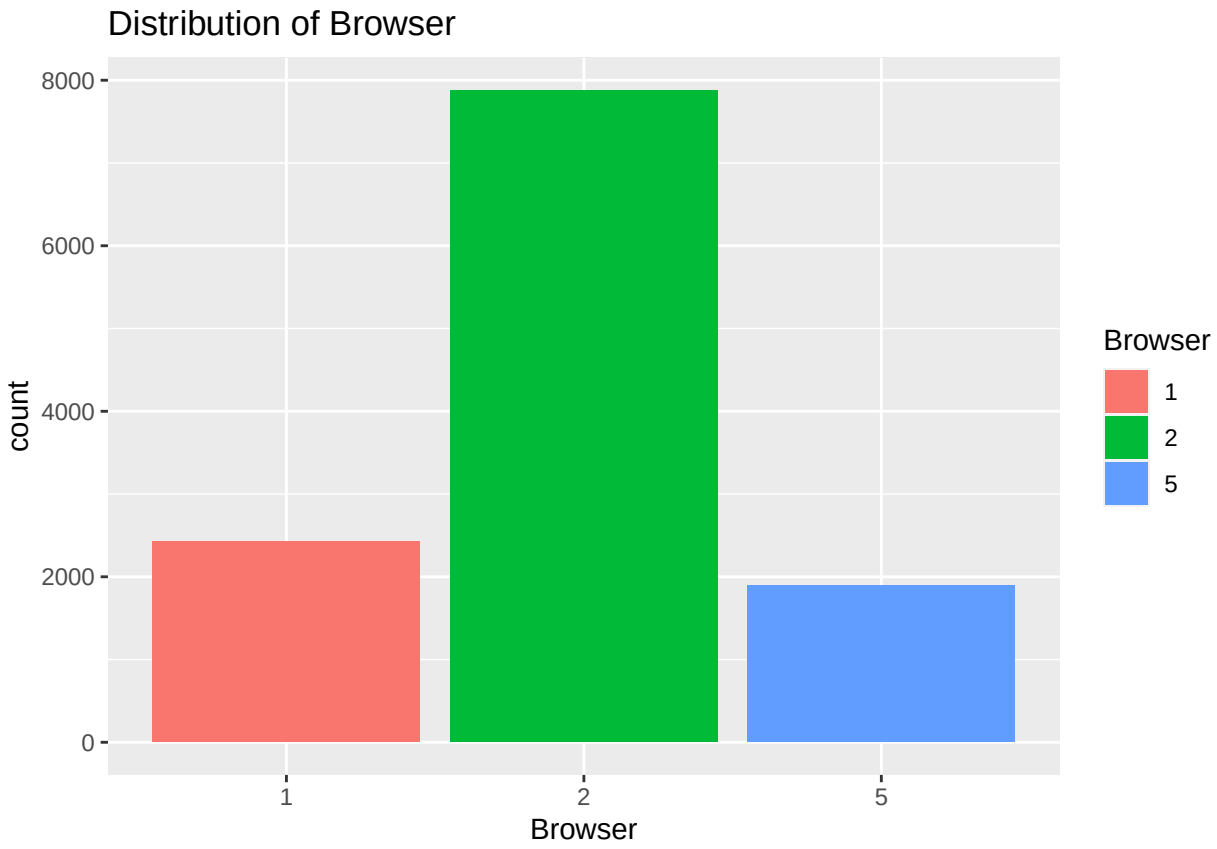

```
# Group Weekend by Month
weekend <- ggplot(customers, aes(x=Month, fill=factor(Weekend)))+geom_bar(position=position_dodge2(widt
weekend + scale_fill_discrete(name = "Weekend", labels = c("Not Weekend","Weekend"))
```



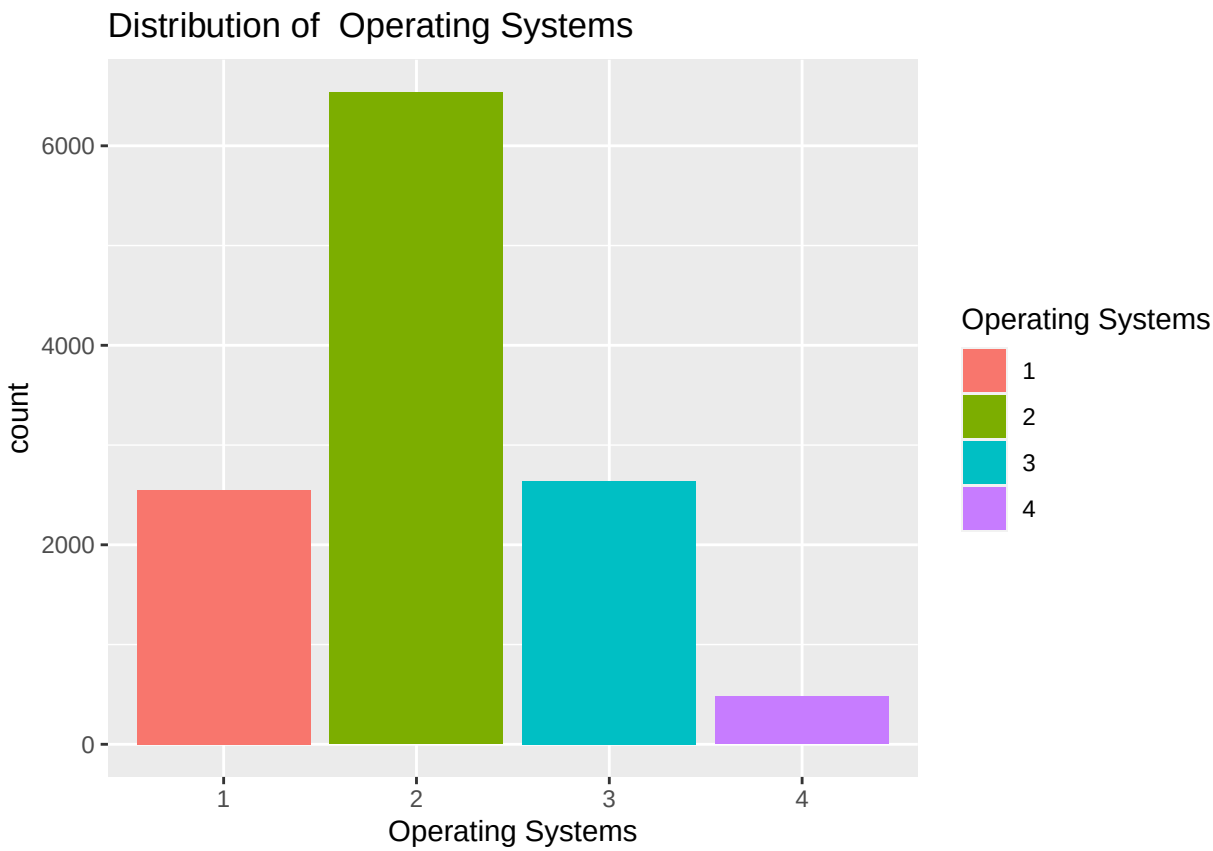

```
# Distribution of Traffic Type
traffictype <- ggplot(customers, aes(x=factor(TrafficType), fill=factor(TrafficType)))+ geom_bar()+labs
traffictype +scale_fill_discrete(name = "Traffic Type")
```



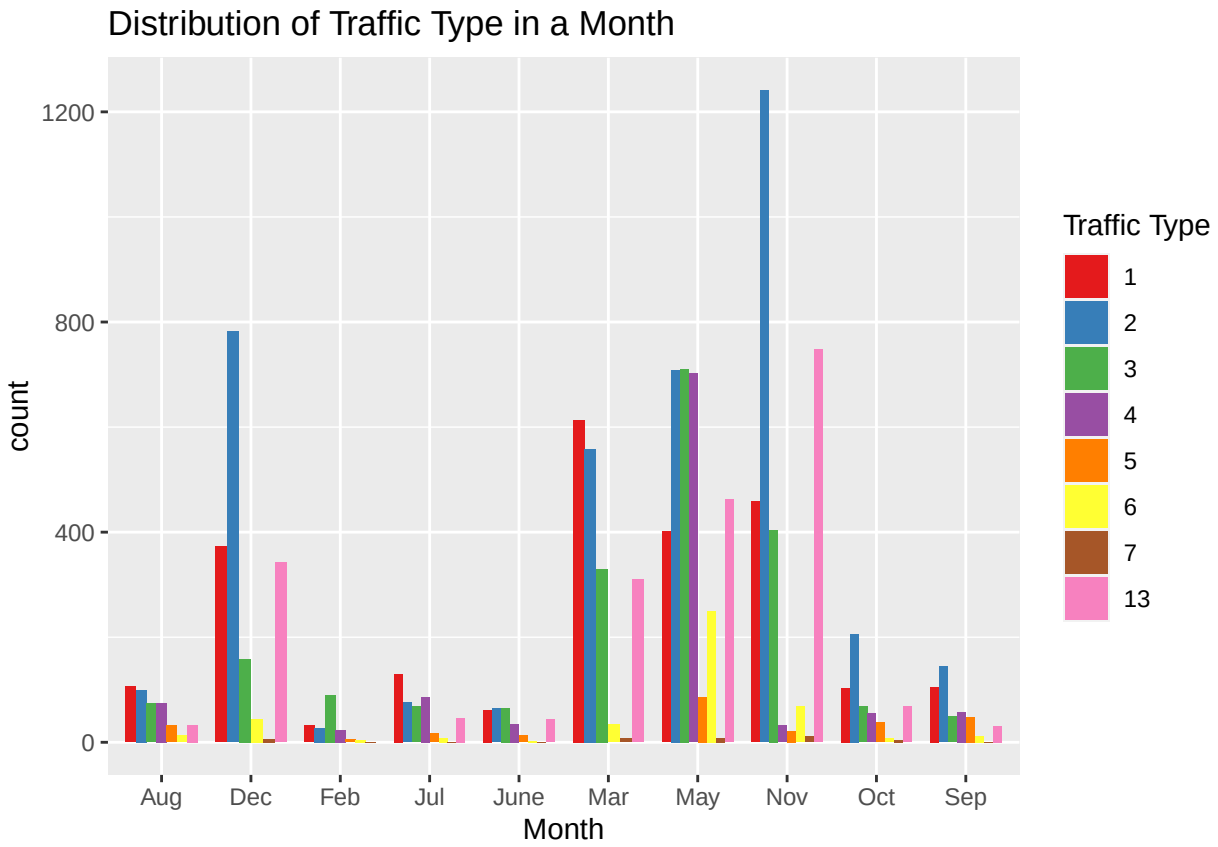

```
# Distribution of Browser
browser <-ggplot(customers, aes(x=factor(Browser), fill=factor(Browser)))+ geom_bar()+labs(title="Distr
browser +scale_fill_discrete(name = "Browser")
```




```
# Distribution of Operating System
os <- ggplot(customers, aes(x=factor(OperatingSystems), fill= factor(OperatingSystems)))+ geom_bar()+lab
os +scale_fill_discrete(name = "Operating Systems")
```

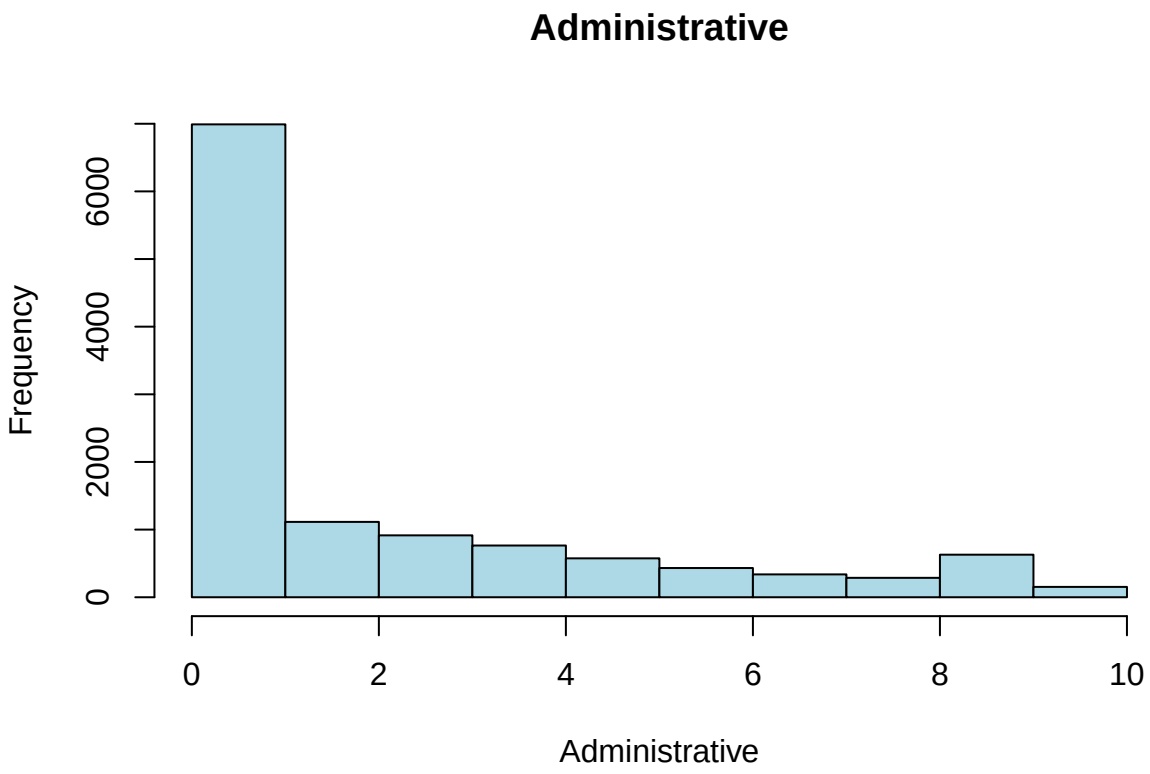


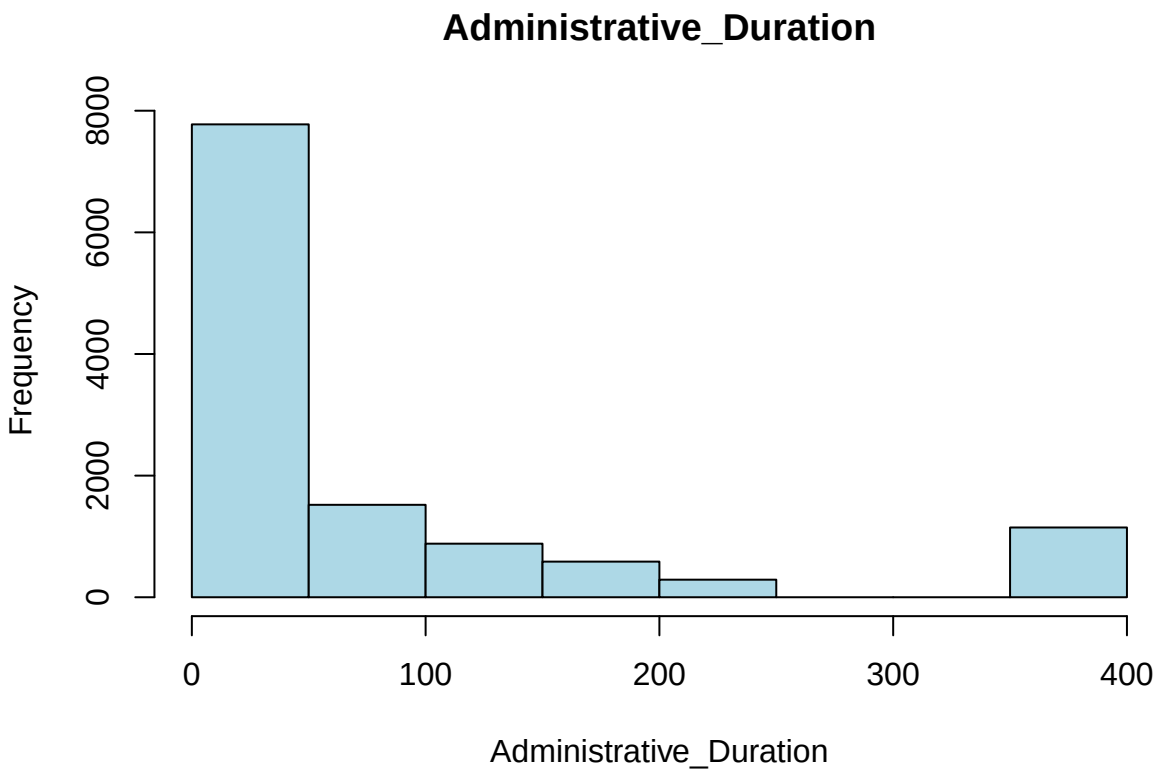

```
# Distribution of TRaffic Type in a Month
traffictype1 <- ggplot(customers, aes(x=Month, fill=factor(TrafficType))) +geom_bar(width = 0.8,positio
traffictype1 + scale_fill_brewer(name ="Traffic Type",palette="Set1")
```

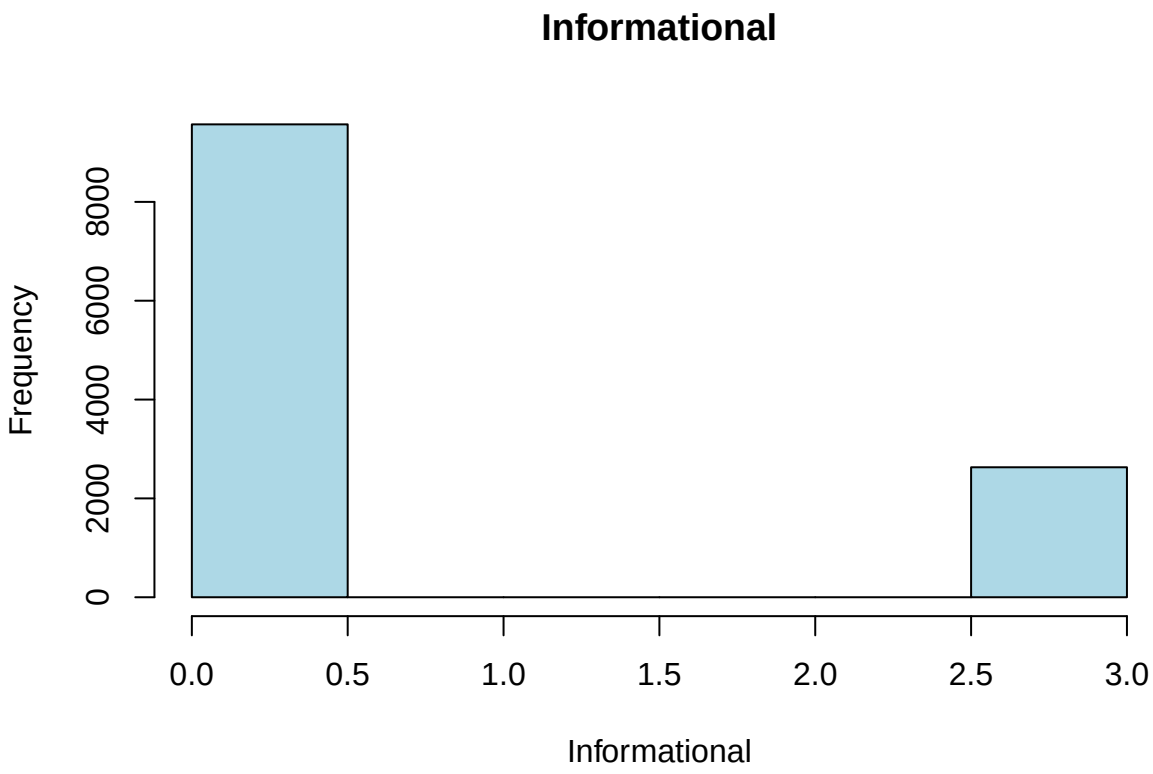


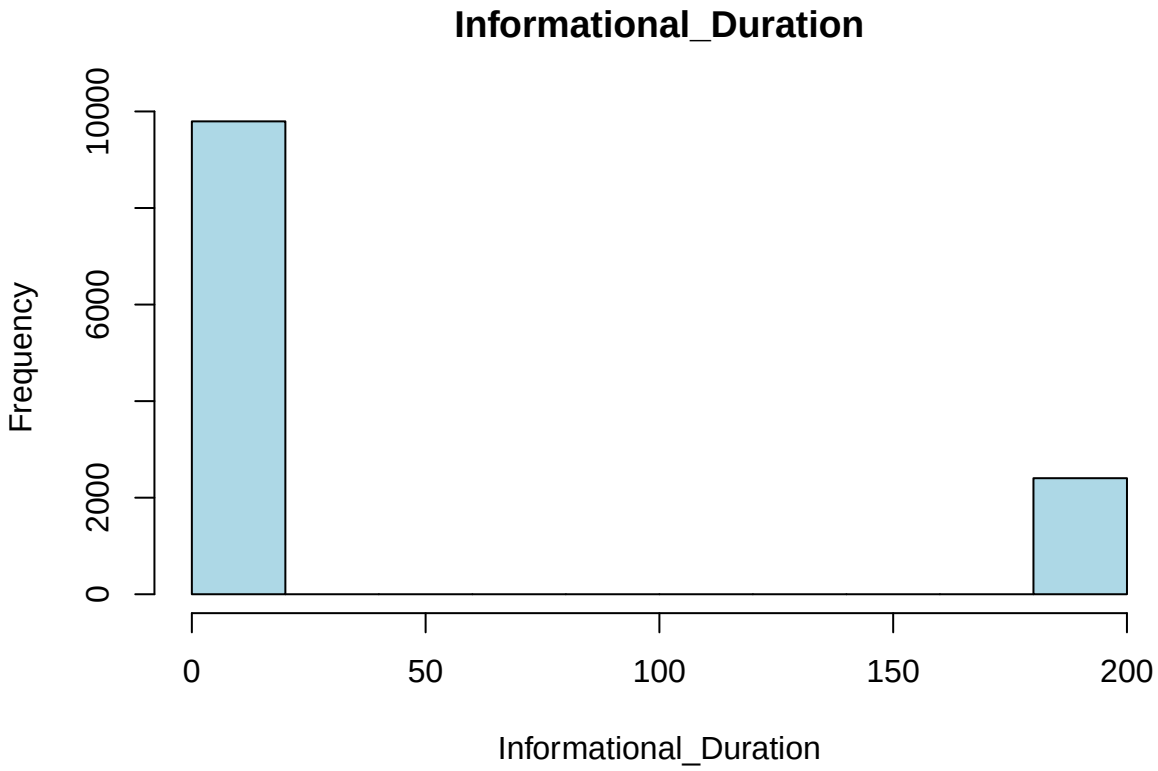
There was most engagement on not special days than special days as from the first plot. There was most customer engagement in the months of March, May and November both for true and false revenue and returning visitors. Most customers were returning visitor types. our class attribute revenue had most not revenue engagements. May and February were the only months with special days engagement. Traffic type 2 was most popular. Browser type 2 was most used. Operating system 2 was most used. Generally for all attributed May and November led in the distribution.

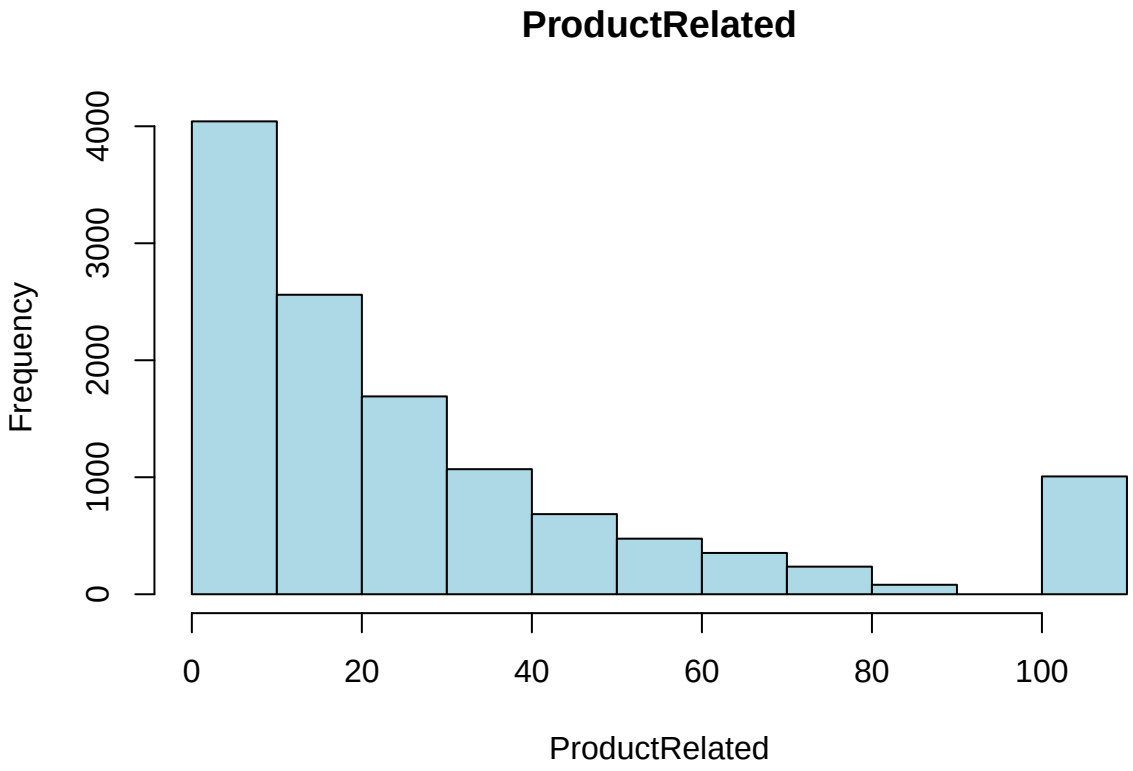
```
#plotting histograms of the numerical columns
histogram = function(x){
  for(i in colnames(x)){
    hist(customers[[i]], breaks = 10,main =i,xlab = i,col = "lightblue")
  }
}
histogram(num_col)
```

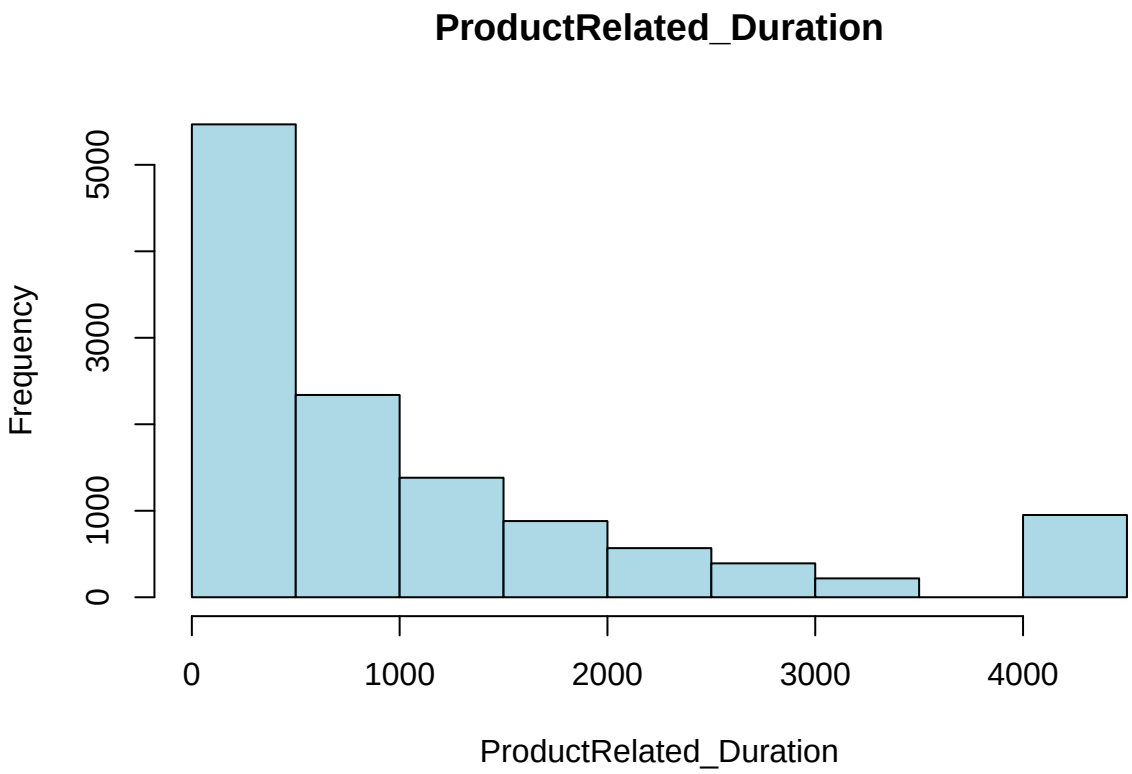

Administrative
Frequency
0
2000
4000
6000
0
2
4
6
8
10
Administrative

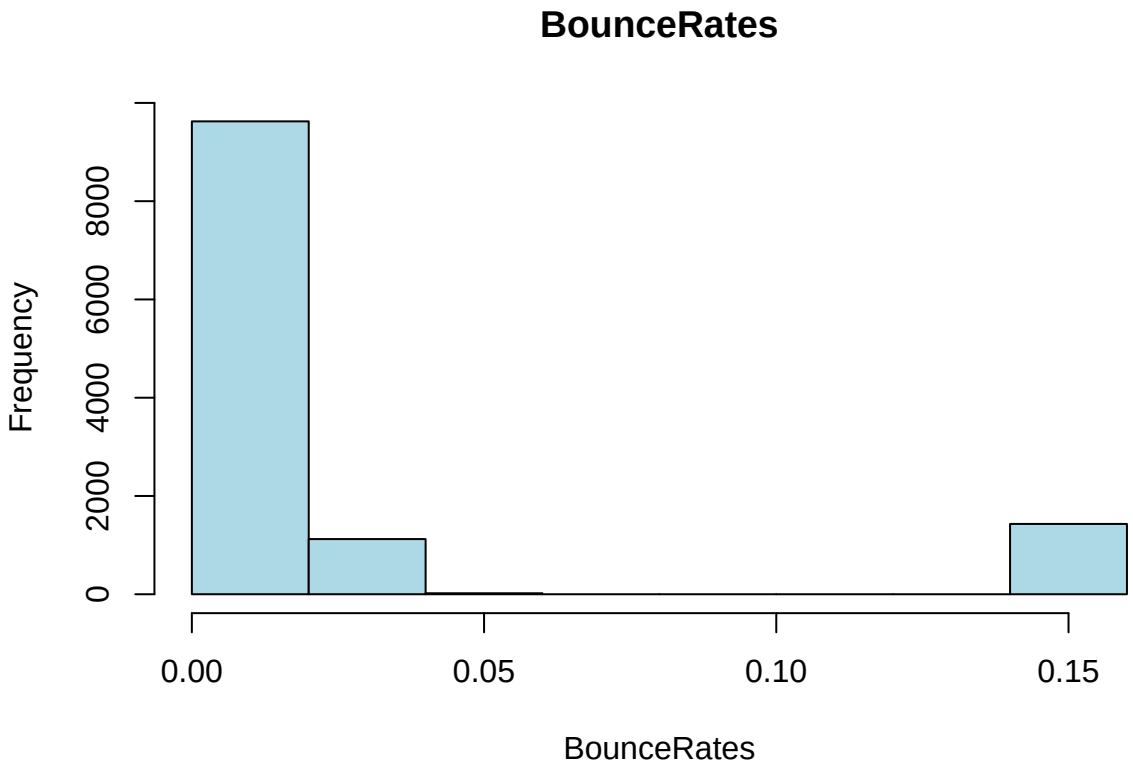

Administrative_Duration
Frequency
0
2000
4000
6000
8000
0
100
200
300
400
Administrative_Duration

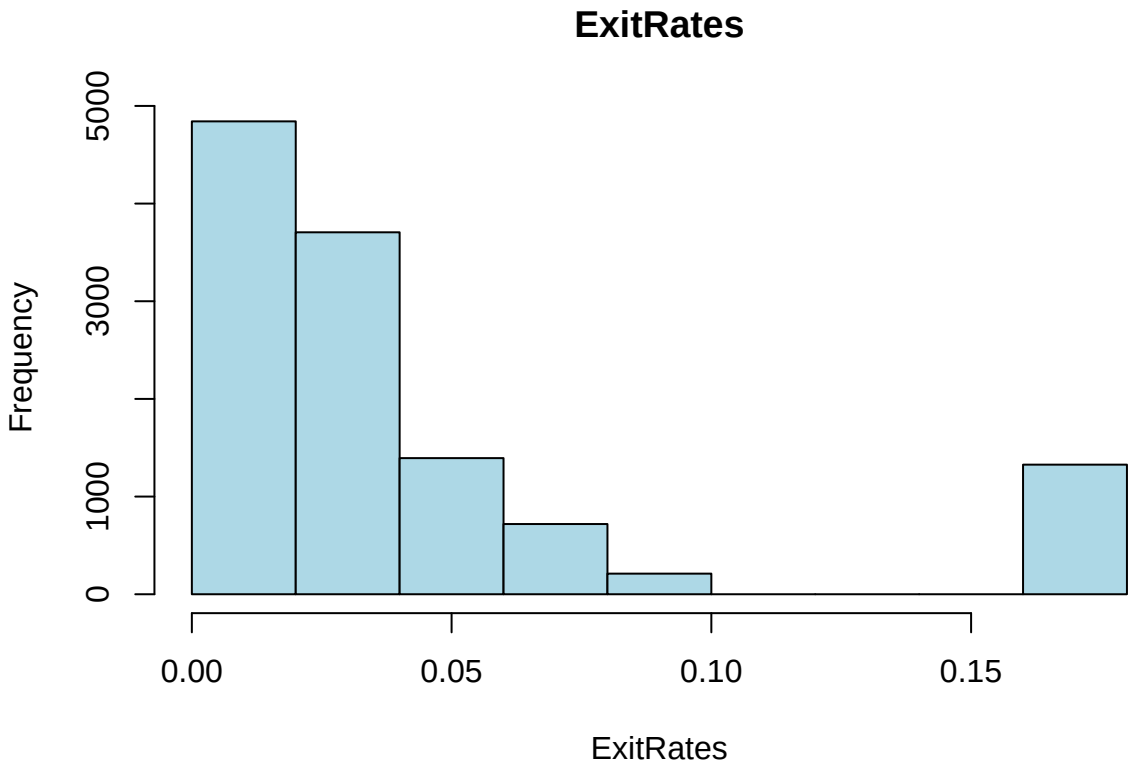

Informational
Frequency
0
2000
4000
6000
8000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Informational

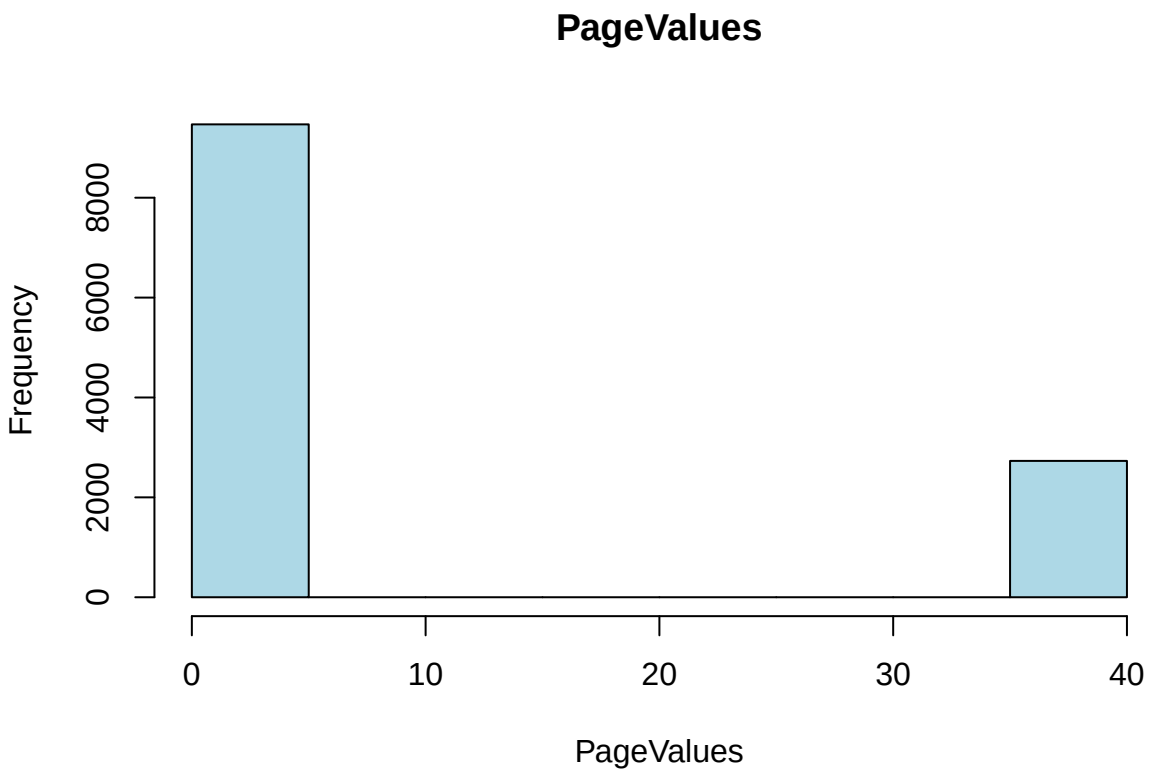

Informational_Duration
Frequency
0
2000
6000
10000
0
50
100
150
200
Informational_Duration

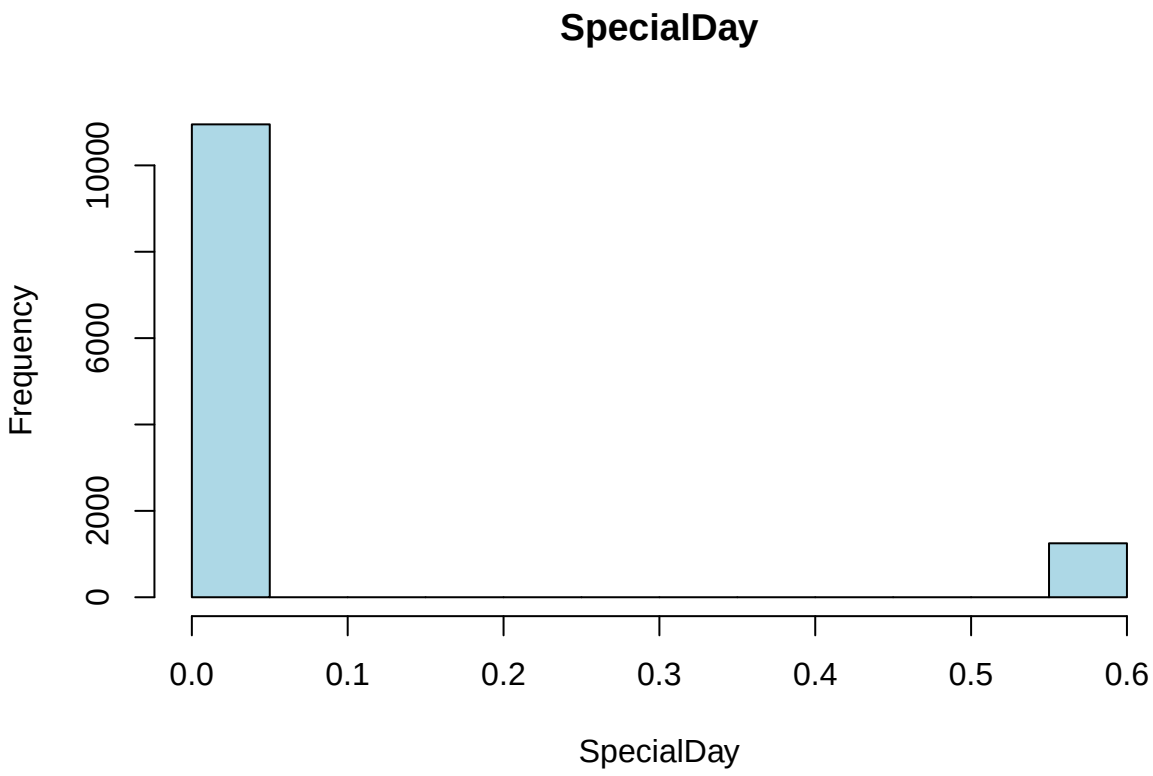

ProductRelated
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
ProductRelated

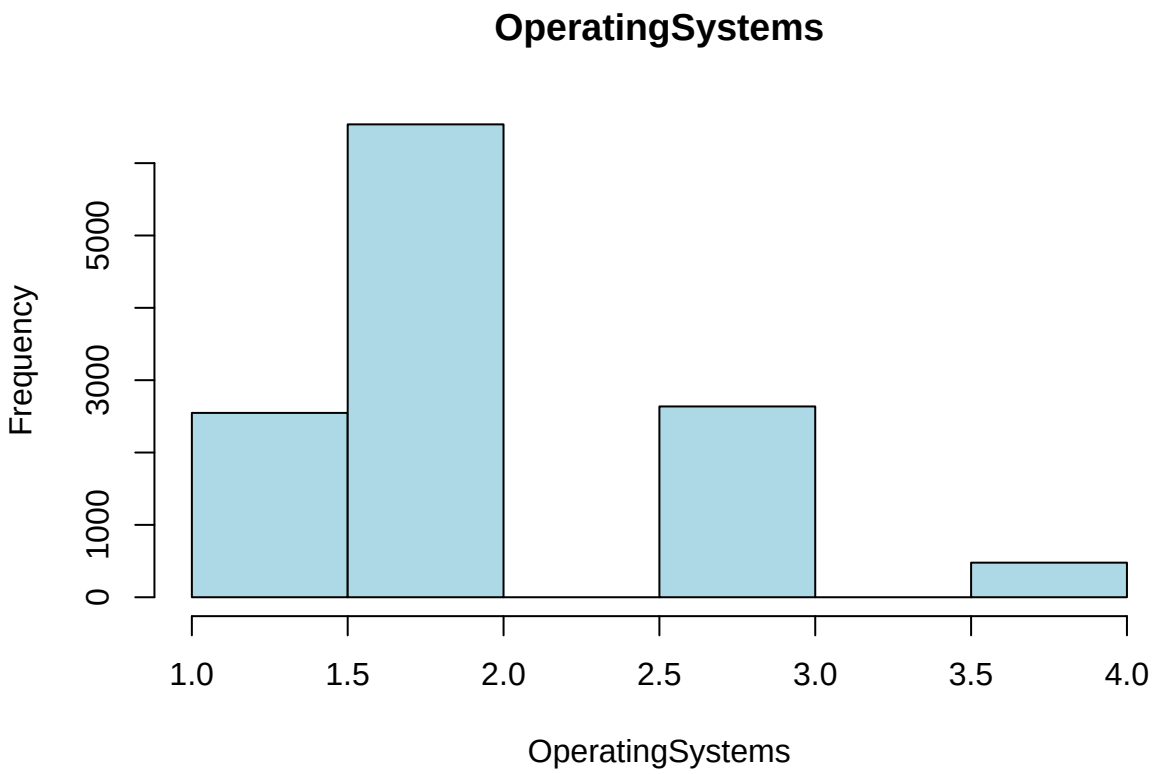

ProductRelated_Duration
Frequency
0
1000
3000
5000
0
1000
2000
3000
4000
ProductRelated_Duration

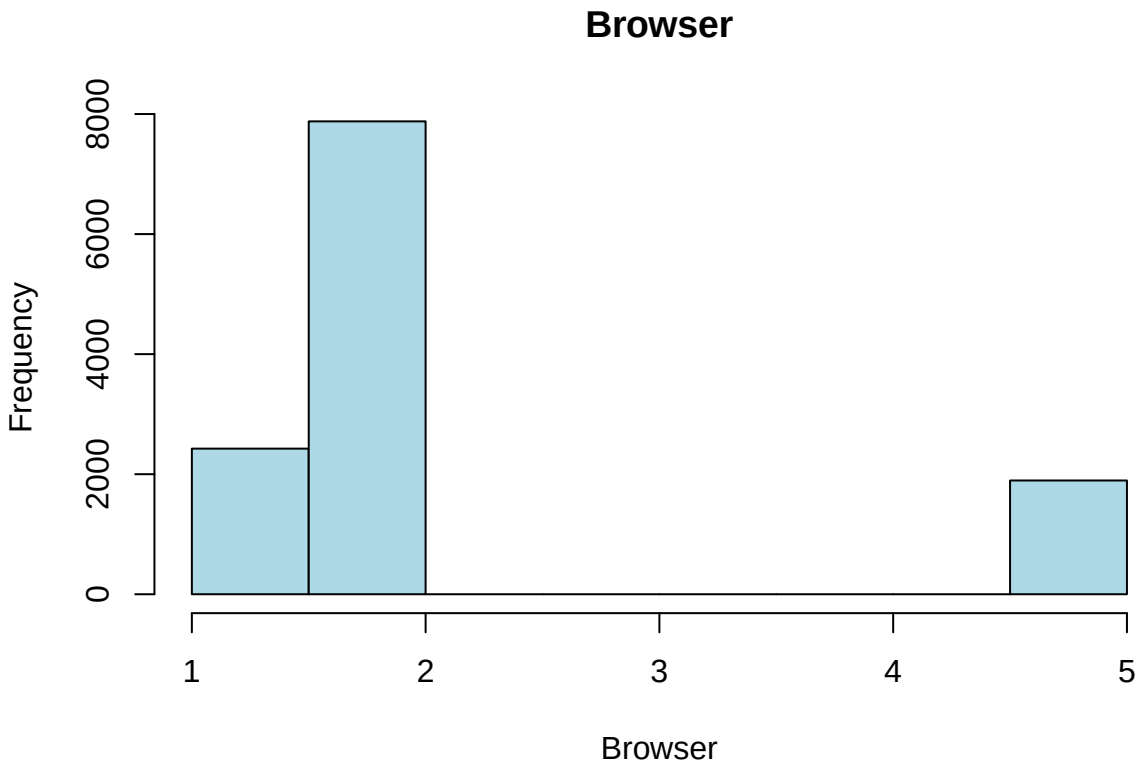

BounceRates
Frequency
0
2000
4000
6000
8000
0.00
0.05
0.10
0.15
BounceRates

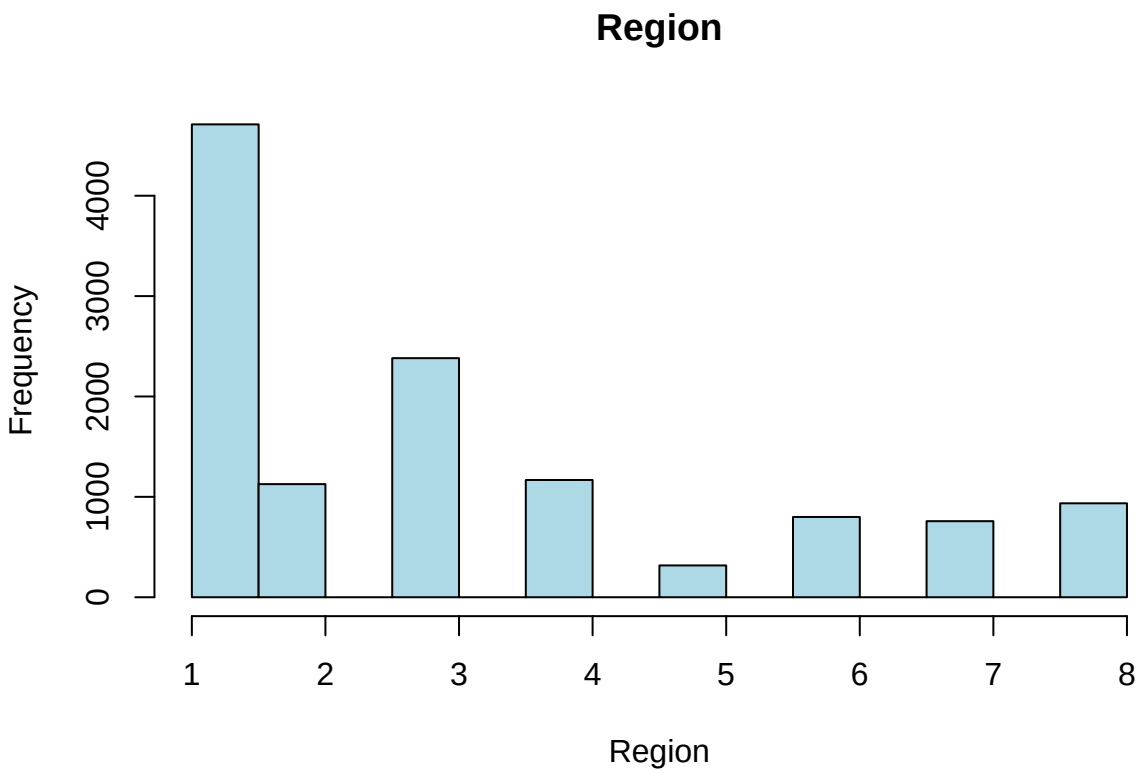

ExitRates
Frequency
0
1000
3000
5000
0.00
0.05
0.10
0.15
ExitRates

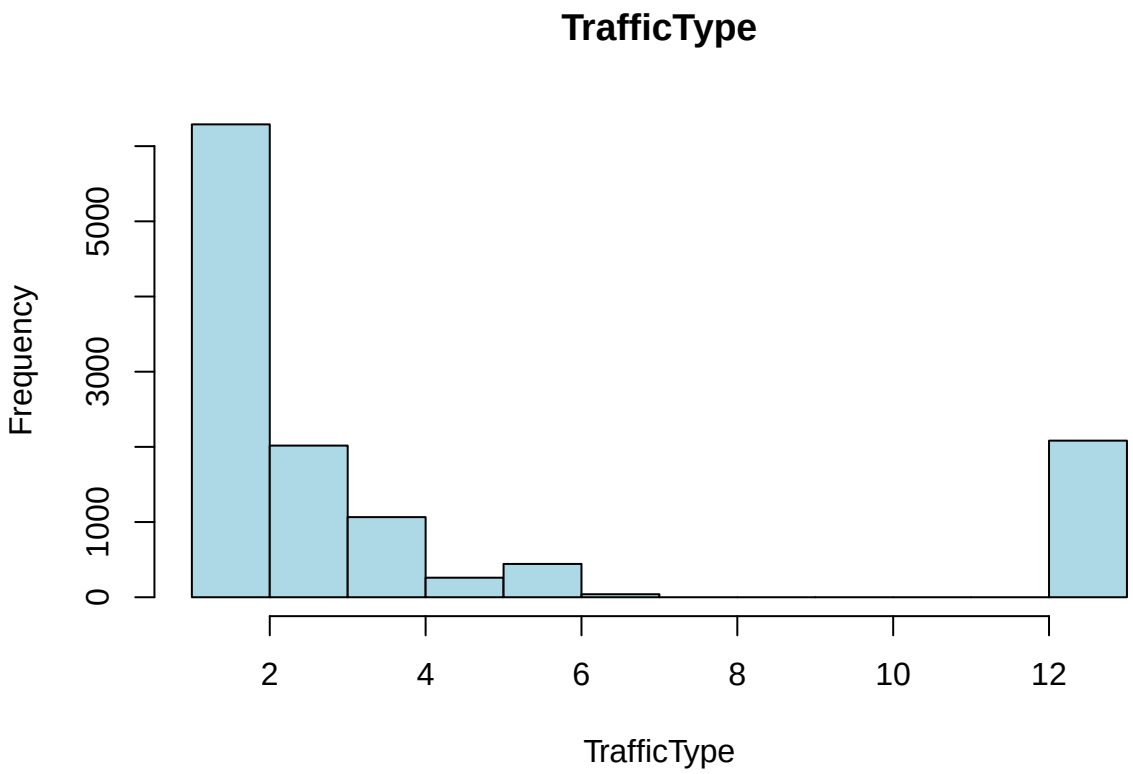

PageValues
Frequency
0
2000
4000
6000
8000
0
10
20
30
40
PageValues


SpecialDay
Frequency
0
2000
6000
10000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
SpecialDay


OperatingSystems
Frequency
0
1000
3000
5000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
OperatingSystems


Browser
Frequency
0
2000
4000
6000
8000
1
2
3
4
5
Browser


Region
Frequency
0
1000
2000
3000
4000
1
2
3
4
5
6
7
8
Region



Bivariate Analysis

```
# Covariance
covariance = cov(num_col)
View(round(covariance,2))
```

```
# Convert Revenue Column to Numeric for correlation checking
customers$Revenue <- as.numeric(customers$Revenue)
numcorr <- customers[ ,c(1,2,3,4,5,6,7,8,9,10,12,13,14,15,18)]
# Correlation matrix
corr_matrix = cor(numcorr)
corr <- as.data.frame(round(corr_matrix,2))
corr
```

```
##                          Administrative Administrative_Duration Informational
## Administrative                     1.00                    0.77          0.37
## Administrative_Duration            0.77                    1.00          0.32
## Informational                      0.37                    0.32          1.00
## Informational_Duration             0.37                    0.32          0.94
## ProductRelated                     0.44                    0.33          0.38
## ProductRelated_Duration            0.39                    0.34          0.37
## BounceRates                       -0.25                   -0.20         -0.14
## ExitRates                         -0.35                   -0.29         -0.20
## PageValues                         0.35                    0.29          0.23
## SpecialDay                        -0.11                   -0.10         -0.05
## OperatingSystems                   0.00                   -0.01          0.00
```

```
## Browser                          -0.03                   -0.04         -0.03
## Region                            0.00                    0.01         -0.02
## TrafficType                      -0.03                   -0.02         -0.03
## Revenue                           0.14                    0.13          0.11
##                          Informational_Duration ProductRelated
## Administrative                             0.37           0.44
## Administrative_Duration                    0.32           0.33
## Informational                              0.94           0.38
## Informational_Duration                     1.00           0.37
## ProductRelated                             0.37           1.00
## ProductRelated_Duration                    0.37           0.85
## BounceRates                               -0.14          -0.26
## ExitRates                                 -0.21          -0.37
## PageValues                                 0.24           0.34
## SpecialDay                                -0.05          -0.03
## OperatingSystems                           0.00           0.03
## Browser                                   -0.03           0.00
## Region                                    -0.01          -0.04
## TrafficType                               -0.03          -0.05
## Revenue                                    0.11           0.17
##                          ProductRelated_Duration BounceRates ExitRates
## Administrative                              0.39       -0.25     -0.35
## Administrative_Duration                     0.34       -0.20     -0.29
## Informational                               0.37       -0.14     -0.20
## Informational_Duration                      0.37       -0.14     -0.21
## ProductRelated                              0.85       -0.26     -0.37
## ProductRelated_Duration                     1.00       -0.24     -0.34
## BounceRates                                -0.24        1.00      0.79
## ExitRates                                  -0.34        0.79      1.00
## PageValues                                  0.34       -0.19     -0.26
## SpecialDay                                 -0.05        0.14      0.13
## OperatingSystems                            0.03        0.04      0.01
## Browser                                     0.01       -0.03     -0.01
## Region                                     -0.02       -0.01     -0.01
## TrafficType                                -0.05        0.09      0.07
## Revenue                                     0.18       -0.16     -0.21
##                          PageValues SpecialDay OperatingSystems Browser Region
## Administrative                 0.35      -0.11             0.00   -0.03   0.00
## Administrative_Duration        0.29      -0.10            -0.01   -0.04   0.01
## Informational                  0.23      -0.05             0.00   -0.03  -0.02
## Informational_Duration         0.24      -0.05             0.00   -0.03  -0.01
## ProductRelated                 0.34      -0.03             0.03    0.00  -0.04
## ProductRelated_Duration        0.34      -0.05             0.03    0.01  -0.02
## BounceRates                   -0.19       0.14             0.04   -0.03  -0.01
## ExitRates                     -0.26       0.13             0.01   -0.01  -0.01
## PageValues                     1.00      -0.07            -0.01    0.02  -0.01
## SpecialDay                    -0.07       1.00             0.02    0.01  -0.02
## OperatingSystems              -0.01       0.02             1.00    0.14   0.01
## Browser                        0.02       0.01             0.14    1.00   0.05
## Region                        -0.01      -0.02             0.01    0.05   1.00
## TrafficType                   -0.03       0.04             0.10    0.00   0.01
## Revenue                        0.60      -0.09            -0.02    0.02  -0.01
##                          TrafficType Revenue
## Administrative                 -0.03    0.14
```

```
## Administrative_Duration       -0.02    0.13
## Informational                 -0.03    0.11
## Informational_Duration        -0.03    0.11
## ProductRelated                -0.05    0.17
## ProductRelated_Duration       -0.05    0.18
## BounceRates                    0.09   -0.16
## ExitRates                      0.07   -0.21
## PageValues                    -0.03    0.60
## SpecialDay                     0.04   -0.09
## OperatingSystems               0.10   -0.02
## Browser                        0.00    0.02
## Region                         0.01   -0.01
## TrafficType                    1.00    0.00
## Revenue                        0.00    1.00
```

```
names(customers)
```

```
##  [1] "Administrative"          "Administrative_Duration"
##  [3] "Informational"           "Informational_Duration"
##  [5] "ProductRelated"          "ProductRelated_Duration"
##  [7] "BounceRates"             "ExitRates"
##  [9] "PageValues"              "SpecialDay"
## [11] "Month"                   "OperatingSystems"
## [13] "Browser"                 "Region"
## [15] "TrafficType"             "VisitorType"
## [17] "Weekend"                 "Revenue"
```

Clustering

```
# Transform Factors to Numeric
customers$Month <- as.numeric(customers$Month)
customers$VisitorType <- as.numeric(customers$VisitorType)
customers$Weekend <- as.numeric(customers$Weekend)
str(customers)
```

```
## 'data.frame':    12199 obs. of  18 variables:
##  $ Administrative         : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : num  1 2 1 2 10 19 1 1 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
##  $ BounceRates            : num  0.15 0 0.15 0.15 0.02 ...
##  $ ExitRates              : num  0.175 0.175 0.175 0.175 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.6 0 0.6 0.6 ...
##  $ Month                  : num  3 3 3 3 3 3 3 3 3 3 ...
##  $ OperatingSystems       : num  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : num  1 2 1 2 5 2 5 2 2 5 ...
##  $ Region                 : num  1 1 8 2 1 1 3 1 2 1 ...
##  $ TrafficType            : num  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : num  3 3 3 3 3 3 3 3 3 3 ...
##  $ Weekend                : num  1 1 1 1 2 1 1 2 1 1 ...
```

```
##  $ Revenue                : num  1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:14] 1066 1133 1134 1135 1136 1137 1474 1475 1476 1477 .
##   ..- attr(*, "names")= chr [1:14] "1066" "1133" "1134" "1135" ...
```

```
#normalize data
customersNorm <- as.data.frame(scale(customers))
head(customersNorm)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1     -0.7687743              -0.6426881    -0.5242359               -0.49539
## 2     -0.7687743              -0.6426881    -0.5242359               -0.49539
## 3     -0.7687743              -0.6426881    -0.5242359               -0.49539
## 4     -0.7687743              -0.6426881    -0.5242359               -0.49539
## 5     -0.7687743              -0.6426881    -0.5242359               -0.49539
## 6     -0.7687743              -0.6426881    -0.5242359               -0.49539
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1     -0.9235168              -0.8830025  2.69141950  2.651251 -0.5369223
## 2     -0.8906220              -0.8303206 -0.49464207  2.651251 -0.5369223
## 3     -0.9235168              -0.8830025  2.69141950  2.651251 -0.5369223
## 4     -0.8906220              -0.8808074  2.69141950  2.651251 -0.5369223
## 5     -0.6274643              -0.3664732 -0.06983386  0.128461 -0.5369223
## 6     -0.3314118              -0.7560584 -0.15926716 -0.384949 -0.5369223
##   SpecialDay     Month OperatingSystems    Browser     Region TrafficType
## 1 -0.3377197 -1.333953         -1.43238 -1.0258790 -0.9149958 -0.78397272
## 2 -0.3377197 -1.333953         -0.11302 -0.2162854 -0.9149958 -0.54264839
## 3 -0.3377197 -1.333953          2.52570 -1.0258790  2.1178123 -0.30132406
## 4 -0.3377197 -1.333953          1.20634 -0.2162854 -0.4817375 -0.05999973
## 5 -0.3377197 -1.333953          1.20634  2.2124954 -0.9149958 -0.05999973
## 6 -0.3377197 -1.333953         -0.11302 -0.2162854 -0.9149958 -0.30132406
##   VisitorType    Weekend    Revenue
## 1    0.409771 -0.5528638 -0.4305688
## 2    0.409771 -0.5528638 -0.4305688
## 3    0.409771 -0.5528638 -0.4305688
## 4    0.409771 -0.5528638 -0.4305688
## 5    0.409771  1.8086156 -0.4305688
## 6    0.409771 -0.5528638 -0.4305688
```

```
customers.new<- customersNorm[, c(1, 2, 3, 4,5,6,7,8,9,10,11,12,13,14,15,16,17)]
customers.class<- customersNorm[, "Revenue"]
```

```
str(customers.new)
```

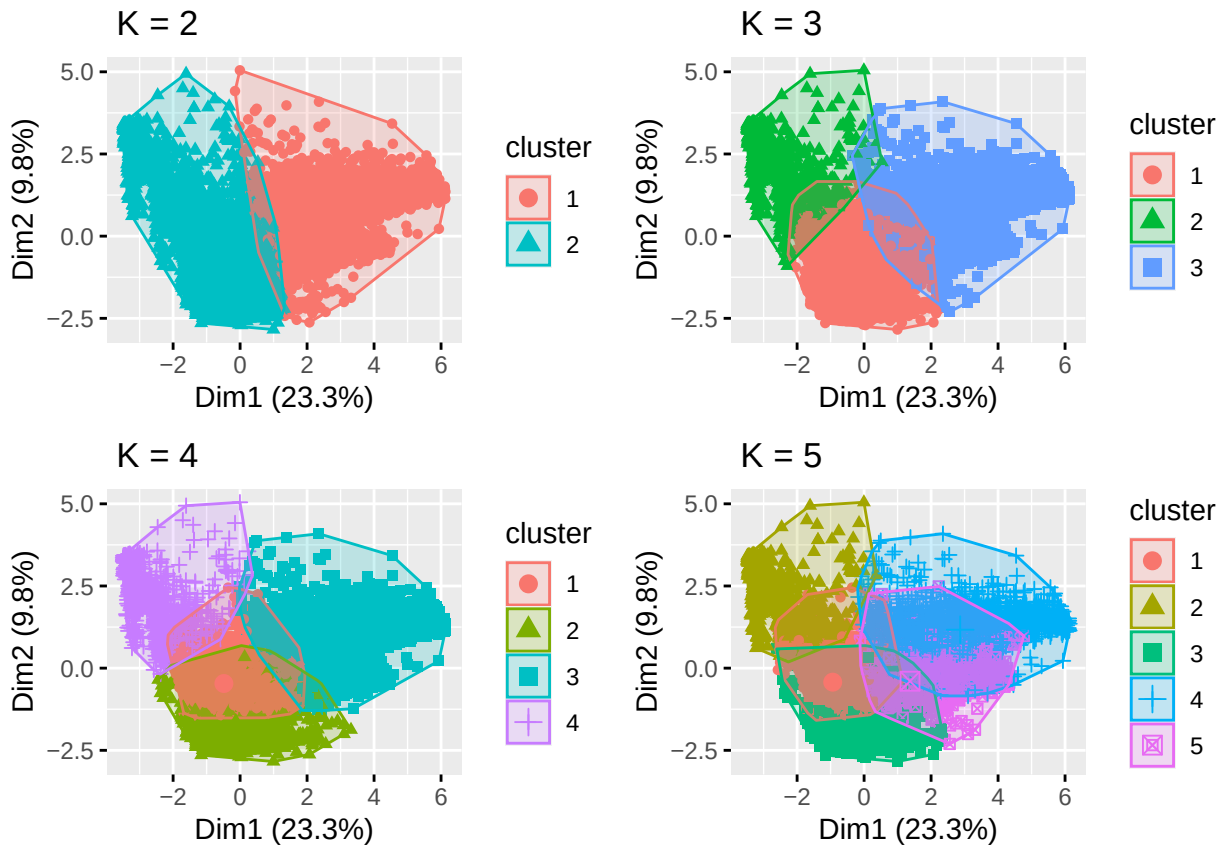
```
## 'data.frame':    12199 obs. of  17 variables:
##  $ Administrative         : num  -0.769 -0.769 -0.769 -0.769 -0.769 ...
##  $ Administrative_Duration: num  -0.643 -0.643 -0.643 -0.643 -0.643 ...
##  $ Informational          : num  -0.524 -0.524 -0.524 -0.524 -0.524 ...
##  $ Informational_Duration : num  -0.495 -0.495 -0.495 -0.495 -0.495 ...
##  $ ProductRelated         : num  -0.924 -0.891 -0.924 -0.891 -0.627 ...
##  $ ProductRelated_Duration: num  -0.883 -0.83 -0.883 -0.881 -0.366 ...
##  $ BounceRates            : num  2.6914 -0.4946 2.6914 2.6914 -0.0698 ...
##  $ ExitRates              : num  2.651 2.651 2.651 2.651 0.128 ...
##  $ PageValues             : num  -0.537 -0.537 -0.537 -0.537 -0.537 ...
##  $ SpecialDay             : num  -0.338 -0.338 -0.338 -0.338 -0.338 ...
```

```
##  $ Month                  : num  -1.33 -1.33 -1.33 -1.33 -1.33 ...
##  $ OperatingSystems       : num  -1.432 -0.113 2.526 1.206 1.206 ...
##  $ Browser                : num  -1.026 -0.216 -1.026 -0.216 2.212 ...
##  $ Region                 : num  -0.915 -0.915 2.118 -0.482 -0.915 ...
##  $ TrafficType            : num  -0.784 -0.543 -0.301 -0.06 -0.06 ...
##  $ VisitorType            : num  0.41 0.41 0.41 0.41 0.41 ...
##  $ Weekend                : num  -0.553 -0.553 -0.553 -0.553 1.809 ...
```

```
set.seed(123)
customers_K2 <- kmeans(customers.new, centers = 2, nstart = 25)
```

```
customers_K3 <- kmeans(customers.new, centers = 3, nstart = 25)
customers_K4 <- kmeans(customers.new, centers = 4, nstart = 25)
customers_K5 <- kmeans(customers.new, centers = 5, nstart = 25)
```

```
p1 <- fviz_cluster(customers_K2, geom = "point", data = customers.new) + ggtitle(" K = 2")
p2 <- fviz_cluster(customers_K3, geom = "point", data = customers.new) + ggtitle(" K = 3")
p3 <- fviz_cluster(customers_K4, geom = "point", data = customers.new) + ggtitle(" K = 4")
p4 <- fviz_cluster(customers_K5, geom = "point", data = customers.new) + ggtitle(" K = 5")

grid.arrange(p1, p2, p3, p4, nrow = 2)
```




```
#getting the performance of various values of K using the BSS to TSS ratio
customers_K2$betweenss/customers_K2$totss
```

```
## [1] 0.1622397
```

```
customers_K3$betweenss/customers_K3$totss
```

```
## [1] 0.2560234
```

```
customers_K4$betweenss/customers_K4$totss
```

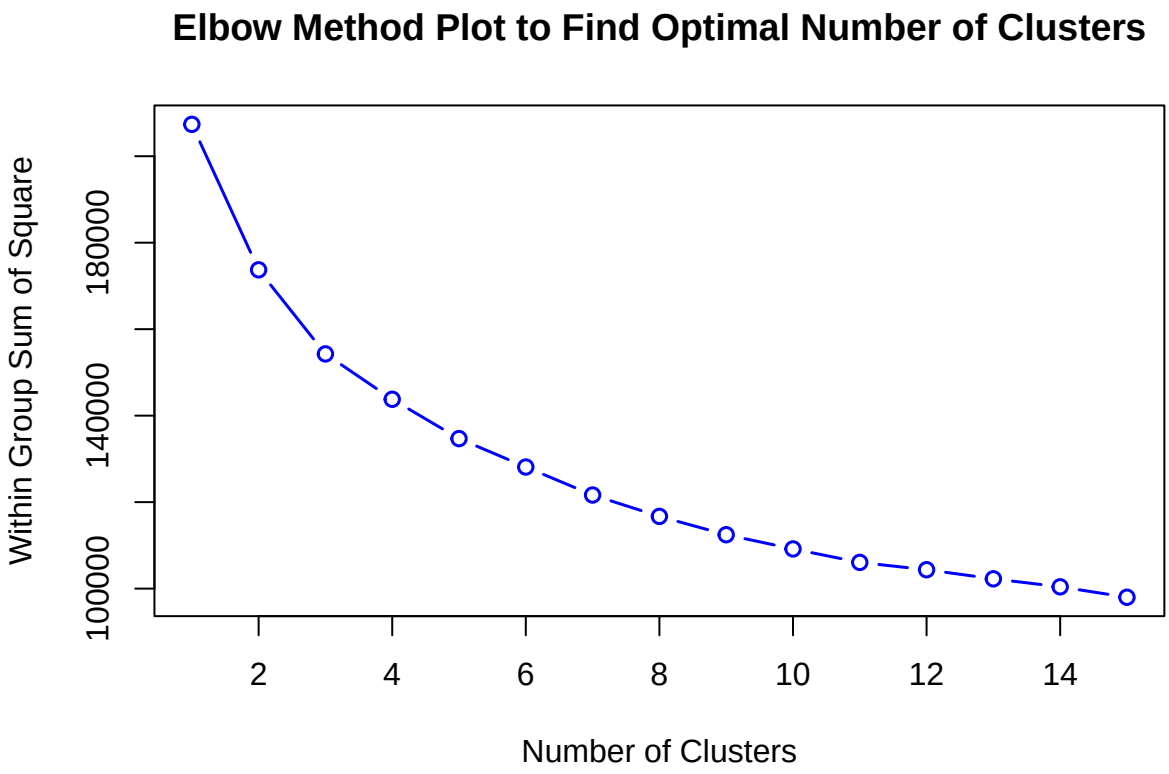
```
## [1] 0.3066127
```

```
customers_K5$betweenss/customers_K5$totss
```

```
## [1] 0.3504694
```

We find the K=5 having the highest ratio of BSS to TSS hence being the best performed model for Kmeans. We try and find optimal number of clusters using elbow method.

```
wssplot <- function(data, nc = 15, set.seed = 1234){
  wss <- (nrow(data) - 1)*sum(apply(data, 2, var))
  for(i in 2:nc) {
    set.seed(1234)
    wss[i] <- sum(kmeans(x = data, centers = i, nstart = 25)$withinss)
  }
  plot(1:nc, wss, type = 'b', xlab = 'Number of Clusters', ylab = 'Within Group Sum of Square',
       main = 'Elbow Method Plot to Find Optimal Number of Clusters', frame.plot = T,
       col = 'blue', lwd = 1.5)
}

wssplot(customers.new)
```

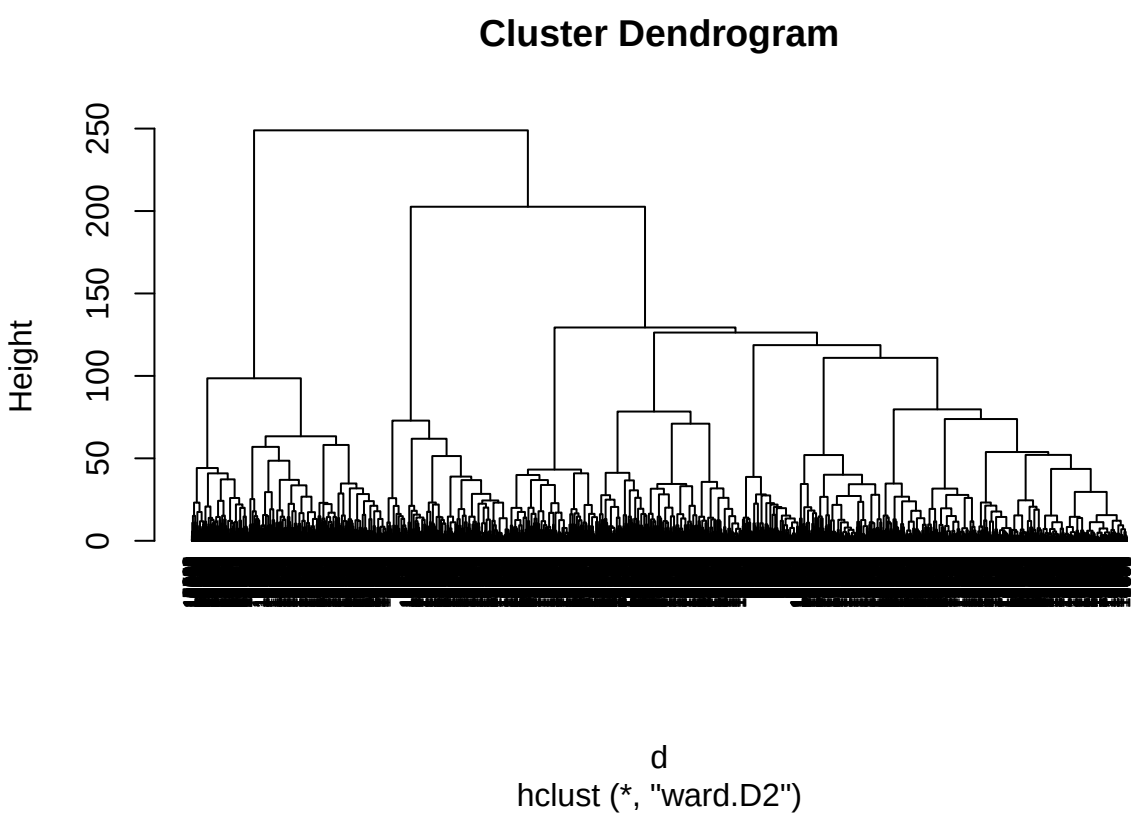


Hierarchical Clustering

```
d <- dist(customers.new, method = "euclidean")
```

```
res.hc <- hclust(d, method = "ward.D2" )
```

```
plot(res.hc, cex = 0.6, hang = -1)
```



CONCLUSION and Recommendation

We used the ward.D2 method for our hierarchical clustering. It appears to perform better than the KMeans clustering. Our KMeans of k=5 had the highest BSS to TSS ratio which is what we are seeking to achieve. Despite this, it wasn't the best performed as an accuracy of 35% is still low. We recommend trying other unsupervised models or optimizing the KMeans.