Analysis of Behaviour

Machocho Mengo

3/9/2020

PROBLEM DEFINITION

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

DATA SOURCING

[http://bit.ly/EcommerceCustomersDataset -This is where our data is collected from.

The dataset consists of 10 numerical and 8 categorical attributes. The 'Revenue' attribute can be used as the class label.

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represents 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 the "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 the "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that was 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 the 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 Valentina'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 the 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.

CHECKING THE DATA

library importation.

```
install.packages("readr")
install.packages("tidyverse")

library("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse
1.3.0 --

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

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

library(grid)
library(devtools)

## Loading required package: usethis
```

Loading the data from the csv document

```
#loading the dataset
#install.packages("readr")
library("readr")
shoppers=read.csv('http://bit.ly/EcommerceCustomersDataset')
```

Previewing the dataset: first observations

```
head(shoppers)

##   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
```

Previewing the last observations

```
tail(shoppers)
```

```
##       Administrative Administrative_Duration Informational
## 12325              0                       0             1
## 12326              3                     145             0
## 12327              0                       0             0
## 12328              0                       0             0
## 12329              4                      75             0
## 12330              0                       0             0
##       Informational_Duration ProductRelated ProductRelated_Duration
BounceRates
## 12325                      0             16                 503.000
0.000000000
## 12326                      0             53                1783.792
0.007142857
## 12327                      0              5                 465.750
0.000000000
## 12328                      0              6                 184.250
```

```
0.083333333
## 12329                     0             15                  346.000
0.000000000
## 12330                     0              3                   21.250
0.000000000
##        ExitRates PageValues SpecialDay Month OperatingSystems Browser
Region
## 12325 0.03764706    0.00000          0   Nov                2       2
1
## 12326 0.02903061   12.24172          0   Dec                4       6
1
## 12327 0.02133333    0.00000          0   Nov                3       2
1
## 12328 0.08666667    0.00000          0   Nov                3       2
1
## 12329 0.02105263    0.00000          0   Nov                2       2
3
## 12330 0.06666667    0.00000          0   Nov                3       2
1
##       TrafficType       VisitorType Weekend Revenue
## 12325           1 Returning_Visitor   FALSE   FALSE
## 12326           1 Returning_Visitor    TRUE   FALSE
## 12327           8 Returning_Visitor    TRUE   FALSE
## 12328          13 Returning_Visitor    TRUE   FALSE
## 12329          11 Returning_Visitor   FALSE   FALSE
## 12330           2       New_Visitor    TRUE   FALSE
```

Checking the structure

```
#check datatypes
str(shoppers)
```

```
## '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                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3
3 3 3 3 3 3 3 3 ...
##  $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3
```

```
3 3 3 3 3 ...
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Checking the shape of the dataset

```
dim(shoppers)

## [1] 12330    18

# the data has 18 variables and 12330 observations
```

Creating a Dataframe

```
# Changing the type of the loaded dataset to a dataframe
df = as.data.frame(shoppers)
# Cleaning column names, by making them uniform
colnames(df) = tolower(colnames(df))
head(df)

##   administrative administrative_duration informational
informational_duration
## 1              0                       0             0
0
## 2              0                       0             0
0
## 3              0                      -1             0
-1
## 4              0                       0             0
0
## 5              0                       0             0
0
## 6              0                       0             0
0
##   productrelated productrelated_duration bouncerates exitrates pagevalues
## 1              1                0.000000  0.20000000 0.2000000          0
## 2              2               64.000000  0.00000000 0.1000000          0
## 3              1               -1.000000  0.20000000 0.2000000          0
## 4              2                2.666667  0.05000000 0.1400000          0
## 5             10              627.500000  0.02000000 0.0500000          0
## 6             19              154.216667  0.01578947 0.0245614          0
##   specialday month operatingsystems browser region traffictype
## 1          0   Feb                1       1      1           1
## 2          0   Feb                2       2      1           2
## 3          0   Feb                4       1      9           3
## 4          0   Feb                3       2      2           4
## 5          0   Feb                3       3      1           4
## 6          0   Feb                2       2      1           3
##         visitortype weekend revenue
## 1 Returning_Visitor   FALSE   FALSE
## 2 Returning_Visitor   FALSE   FALSE
## 3 Returning_Visitor   FALSE   FALSE
## 4 Returning_Visitor   FALSE   FALSE
## 5 Returning_Visitor    TRUE   FALSE
## 6 Returning_Visitor   FALSE   FALSE
```

Statistical summaries of the dataframe

```
# Previewing some statistical summaries of the dataset
summary(df)
```

```
##  administrative   administrative_duration informational  
##  Min.   : 0.000   Min.   :  -1.00         Min.   : 0.000  
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.000  
##  Median : 1.000   Median :   8.00         Median : 0.000  
##  Mean   : 2.318   Mean   :  80.91         Mean   : 0.504  
##  3rd Qu.: 4.000   3rd Qu.:  93.50         3rd Qu.: 0.000  
##  Max.   :27.000   Max.   :3398.75         Max.   :24.000  
##  NA's   :14       NA's   :14              NA's   :14      
##  informational_duration productrelated   productrelated_duration
##  Min.   :  -1.00        Min.   :  0.00   Min.   :   -1.0        
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  185.0        
##  Median :   0.00        Median : 18.00   Median :  599.8        
##  Mean   :  34.51        Mean   : 31.76   Mean   : 1196.0        
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1466.5        
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5        
##  NA's   :14             NA's   :14       NA's   :14             
##   bouncerates         exitrates          pagevalues         specialday     
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000  
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000  
##  Median :0.003119   Median :0.02512   Median :  0.000   Median :0.00000  
##  Mean   :0.022152   Mean   :0.04300   Mean   :  5.889   Mean   :0.06143  
##  3rd Qu.:0.016684   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000  
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000  
##  NA's   :14         NA's   :14                                            
##      month       operatingsystems    browser          region     
##  May    :3364   Min.   :1.000    Min.   : 1.000   Min.   :1.000  
##  Nov    :2998   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000  
##  Mar    :1907   Median :2.000    Median : 2.000   Median :3.000  
##  Dec    :1727   Mean   :2.124    Mean   : 2.357   Mean   :3.147  
##  Oct    : 549   3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000  
##  Sep    : 448   Max.   :8.000    Max.   :13.000   Max.   :9.000  
##  (Other):1337                                                    
##   traffictype                visitortype     weekend          revenue       
##  Min.   : 1.00   New_Visitor      : 1694   Mode :logical   Mode :logical  
##  1st Qu.: 2.00   Other            :   85   FALSE:9462      FALSE:10422    
##  Median : 2.00   Returning_Visitor:10551   TRUE :2868      TRUE :1908     
##  Mean   : 4.07                                                            
##  3rd Qu.: 4.00                                                            
##  Max.   :20.00                                                            
## 
```

DATA CLEANING

CHECKING AND DEALING WITH DUPLICATES

Checking whether the dataset has duplicated values

```
# Checking for duplicated data
anyDuplicated(df)

## [1] 159

#our data has 159 duplicated values
```

Dropping the duplicates

```
#drop duplicates
#install.packages("dplyr")
#library(dplyr)
df1 = distinct(df)
# Ckecking whether the duplicates have been successfully dropped
anyDuplicated(df1)

## [1] 0
```

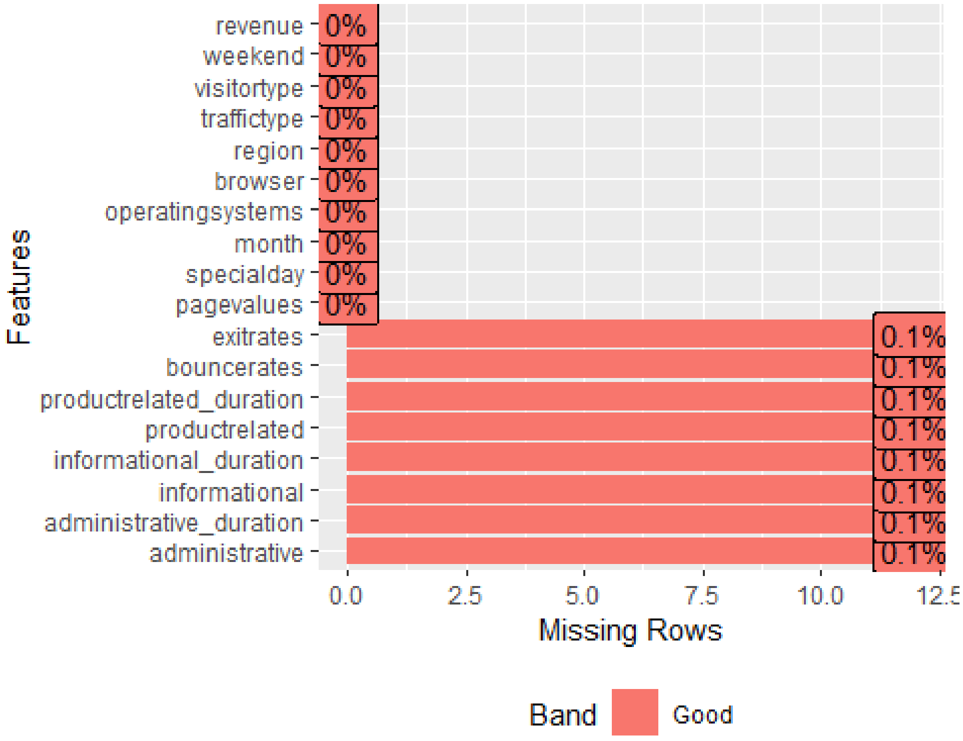
CHECKING AND DEALING WITH MISSING VALUES

Checking for missing values

```
# Checking for missing values
colSums(is.na(df1))

##          administrative administrative_duration           informational 
##                      12                      12                      12 
##  informational_duration          productrelated productrelated_duration 
##                      12                      12                      12 
##             bouncerates               exitrates              pagevalues 
##                      12                      12                       0 
##              specialday                   month        operatingsystems 
##                       0                       0                       0 
##                 browser                  region             traffictype 
##                       0                       0                       0 
##             visitortype                 weekend                 revenue 
##                       0                       0                       0
```

Visualizing the missing values

```
library(DataExplorer)

plot_missing(df1)
```



Drop the missing values from the dataset

```
# Dropping missing values
df_2= na.omit(df1)
```

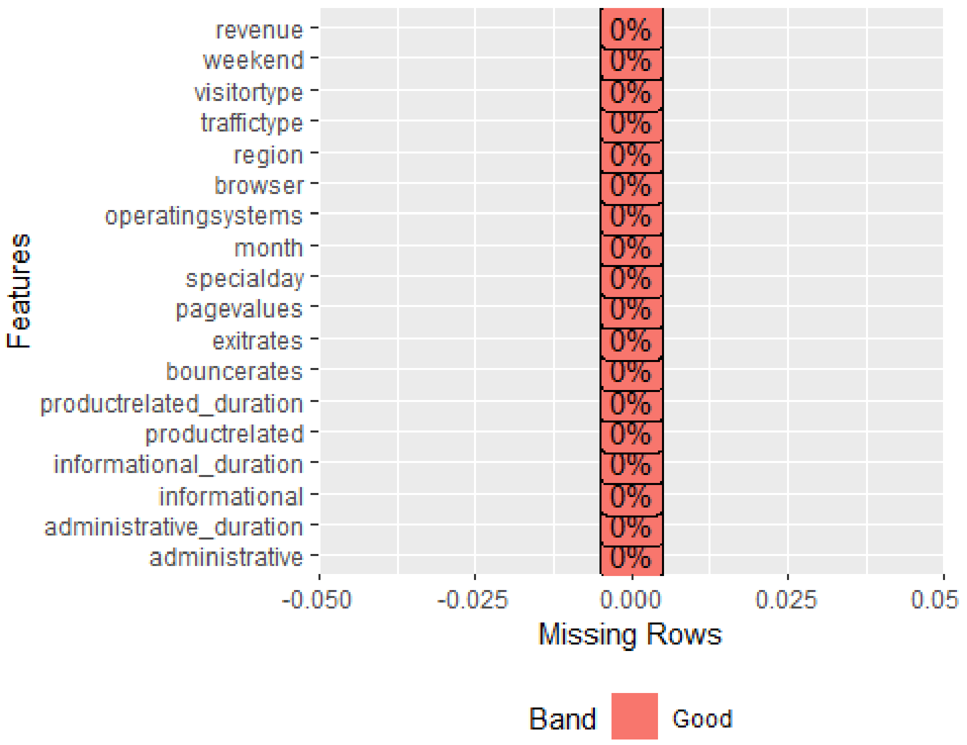
Confriming that the missing values have been deleted.

```
colSums(is.na(df_2))
```

```
##          administrative administrative_duration           informational 
##                       0                       0                       0 
##  informational_duration          productrelated productrelated_duration 
##                       0                       0                       0 
##             bouncerates               exitrates              pagevalues 
##                       0                       0                       0 
##              specialday                   month        operatingsystems 
##                       0                       0                       0 
##                 browser                  region             traffictype 
##                       0                       0                       0 
##             visitortype                 weekend                 revenue 
##                       0                       0                       0
```

Plotting the data after cleaning the data

```
plot_missing(df_2)
```



CONSISTENCY OF THE DATASET

Making sure the the data has the correct datatype. Changing the datatypes of the below variables to factor

```
cat_cols = c('month', 'operatingsystems',   'browser',  'region',
'traffictype', 'visitortype')
# Changing columns to factors
for( i in cat_cols){
   df_2[,i] = as.factor(df_2[,i])
}
```

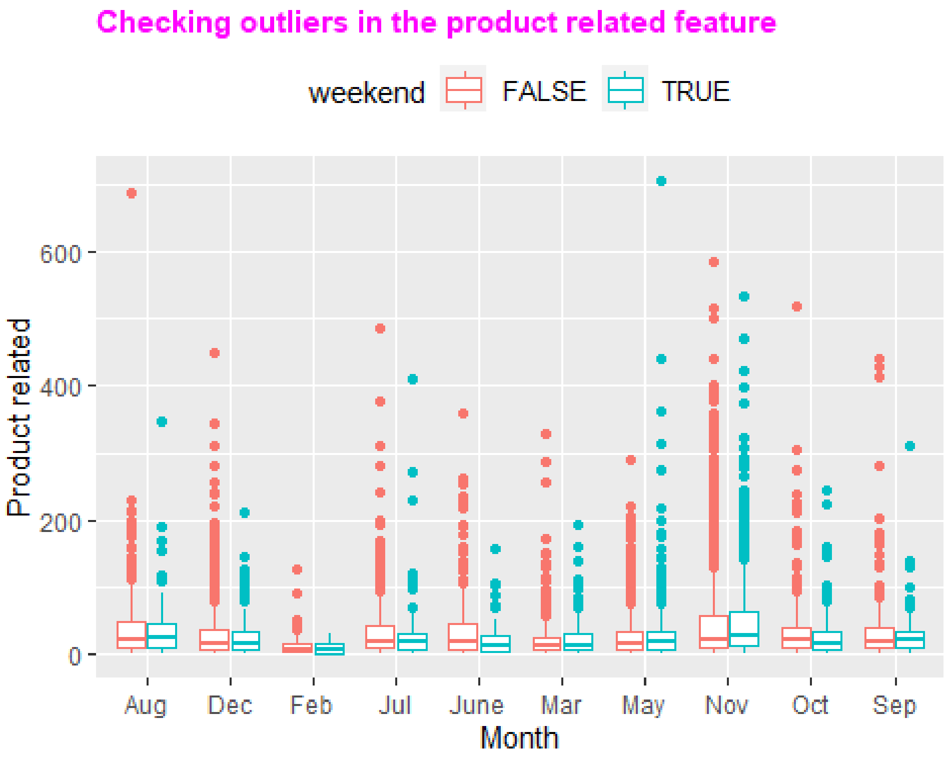
CHECKING AND DEALING WITH OUTLIERS

Checking for Outliers

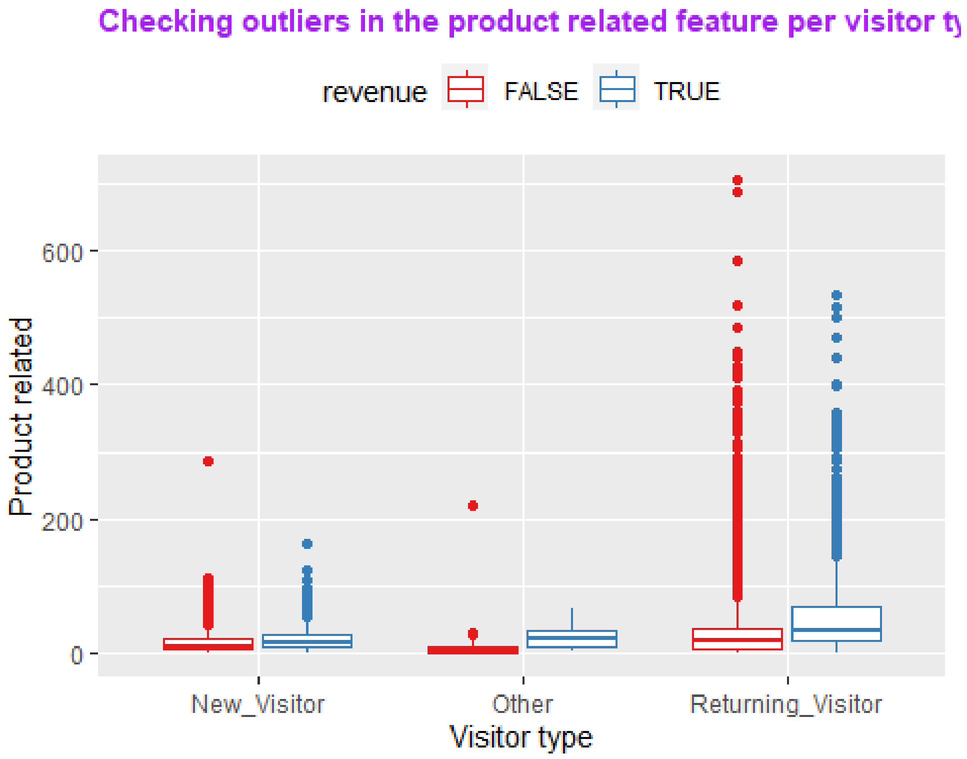
```
#Loading the necessary packages
library("ggplot2")
```

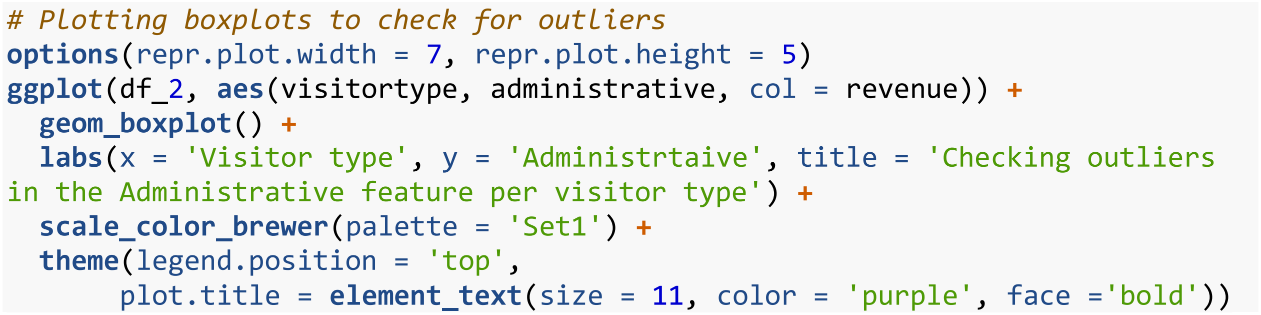
Boxplot to check for outliers

```
options(repr.plot.width = 11, repr.plot.height = 5)
ggplot(df_2, aes(month, productrelated, col = weekend)) +
  geom_boxplot() +
  labs(x = 'Month', y = 'Product related', title = 'Checking outliers in the
product related feature') +
  theme(legend.position = 'top', legend.text = element_text(size = 10),
       plot.title = element_text(size = 11, color = 'magenta', face =
'bold'))
```

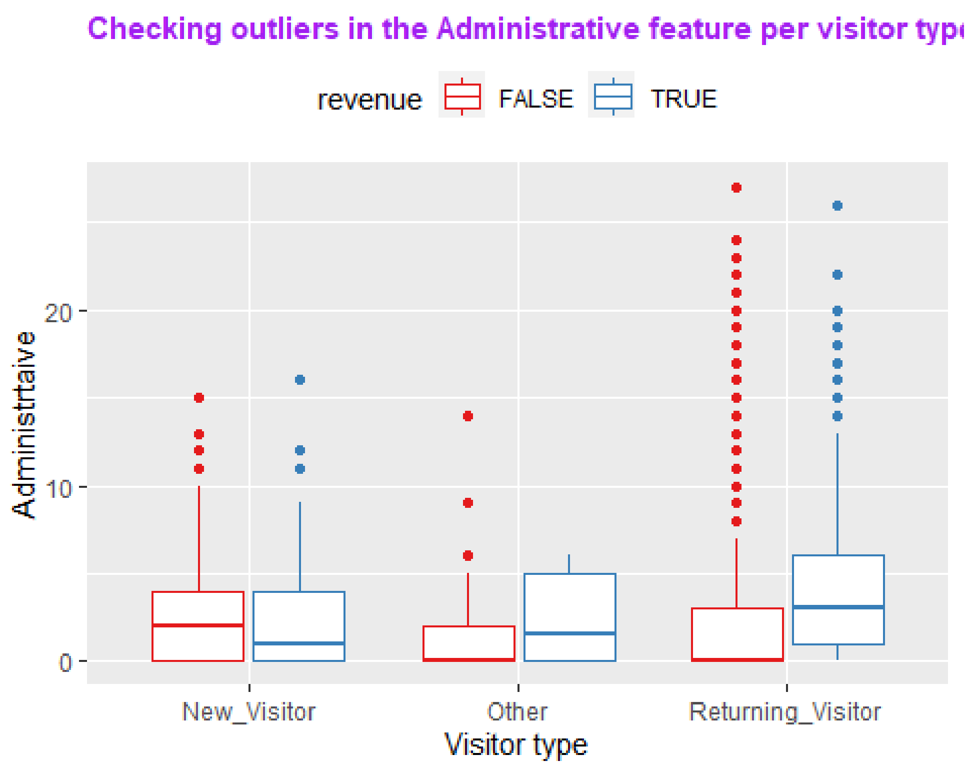


```
# Plotting boxplots to check for outliers
options(repr.plot.width = 7, repr.plot.height = 5)
ggplot(df_2, aes(visitortype, productrelated, col = revenue)) +
  geom_boxplot() +
  labs(x = 'Visitor type', y = 'Product related', title = 'Checking outliers
in the product related feature per visitor type') +
  scale_color_brewer(palette = 'Set1') +
  theme(legend.position = 'top',
       plot.title = element_text(size = 11, color = 'purple', face ='bold'))
```

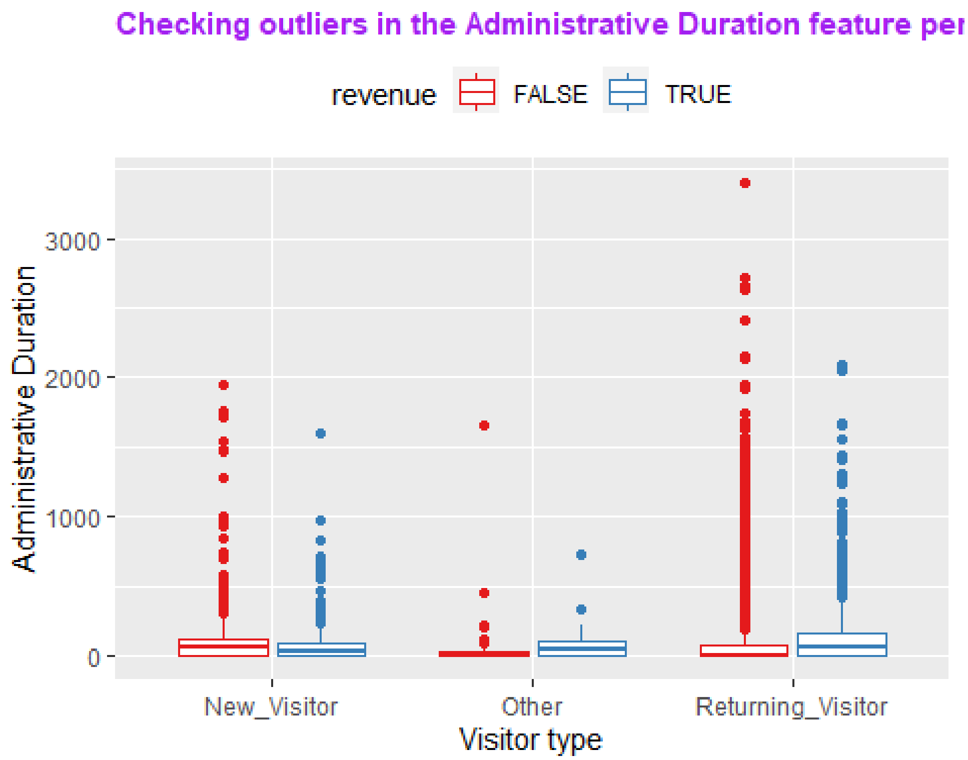




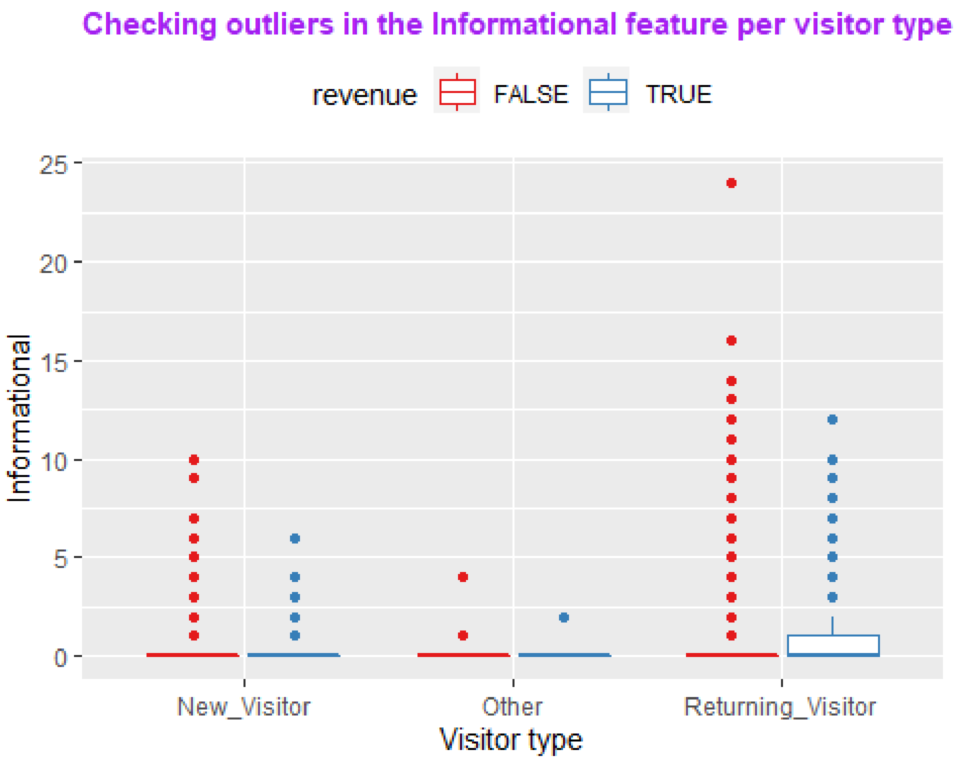
```
# Plotting boxplots to check for outliers
options(repr.plot.width = 7, repr.plot.height = 5)
ggplot(df_2, aes(visitortype, administrative, col = revenue)) +
  geom_boxplot() +
  labs(x = 'Visitor type', y = 'Administrtaive', title = 'Checking outliers
in the Administrative feature per visitor type') +
  scale_color_brewer(palette = 'Set1') +
  theme(legend.position = 'top',
       plot.title = element_text(size = 11, color = 'purple', face ='bold'))
```



```
options(repr.plot.width = 7, repr.plot.height = 5)
ggplot(df_2, aes(visitortype, administrative_duration, col = revenue)) +
  geom_boxplot() +
  labs(x = 'Visitor type', y = 'Administrative Duration', title = 'Checking
outliers in the Administrative Duration feature per visitor type') +
  scale_color_brewer(palette = 'Set1') +
  theme(legend.position = 'top',
       plot.title = element_text(size = 11, color = 'purple', face ='bold'))
```



```
# Plotting boxplots to check for outliers
options(repr.plot.width = 7, repr.plot.height = 5)
ggplot(df_2, aes(visitortype, informational, col = revenue)) +
  geom_boxplot() +
  labs(x = 'Visitor type', y = 'Informational', title = 'Checking outliers in
the Informational feature per visitor type') +
  scale_color_brewer(palette = 'Set1') +
  theme(legend.position = 'top',
       plot.title = element_text(size = 11, color = 'purple', face ='bold'))
```



The boxplots show that the data has a lot of outliers. We will however analyse them more in the subsequent analyses. The outliers will not be omitted.

EXPLORATORY DATA ANALYSIS

UNIVARIATE ANALYSIS

This is the analysis of individual variables.

The analysis includes:

Measures of central tendancy: Mean, Median, Mode

Measures of dispersion: Min, Max, Range, Quartiles, Variance, Standard deviation

Other measures include: Skewness, Kurtosis

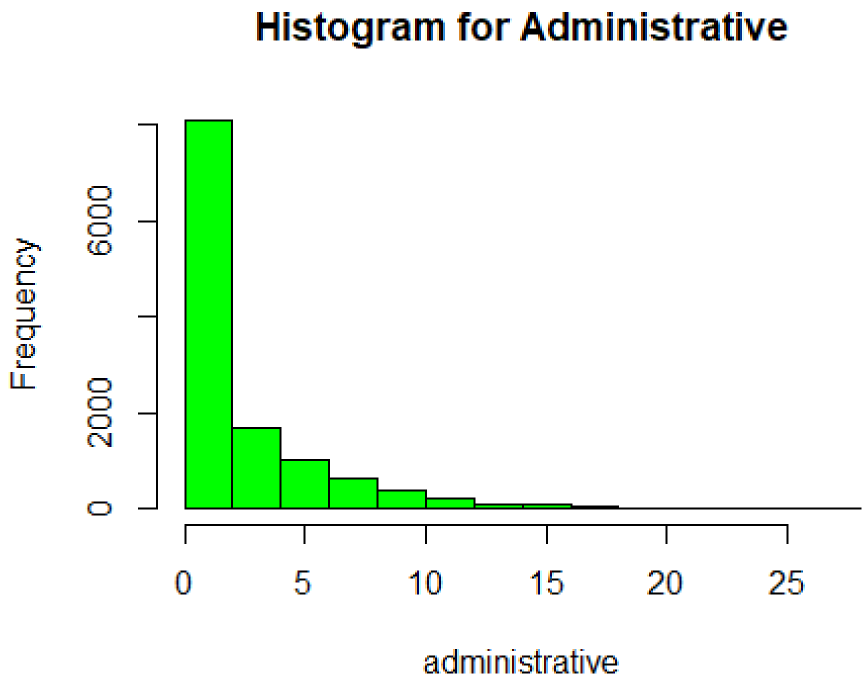
Univariate Graphs: Histogram, Box plots, Bar plots, Kernel density plots

HISTOGRAMS AND SKEWNESS

Importing the necessary packages

```r
# install.packages("moments")
library(moments)
```

```
# histogram of the admistrative feature
# Find the measures of skewness and kurtosis

hist(df_2$administrative,
     main="Histogram for Administrative",
     xlab="administrative",
     border="black",
     col="green")
```

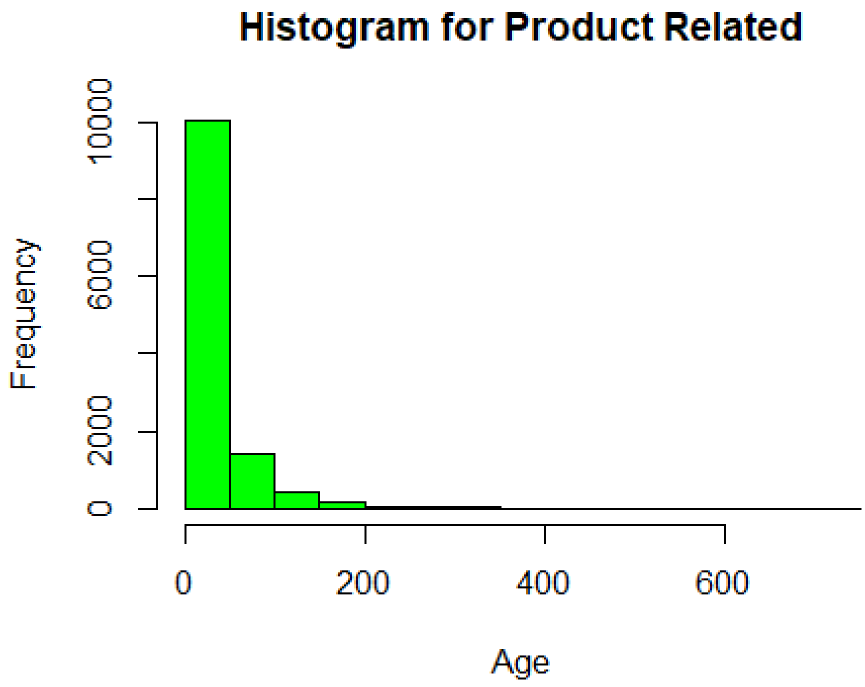


```
skewness(df_2$administrative)

## [1] 1.946248

kurtosis(df_2$administrative)

## [1] 7.636106
```

The administrative column is skewed to the right It has a skewness of 1.946248 which means that the data is positively skewed.

The kurtosis of the variable is 7.636106. This means that the data is not flat but it is peaky.

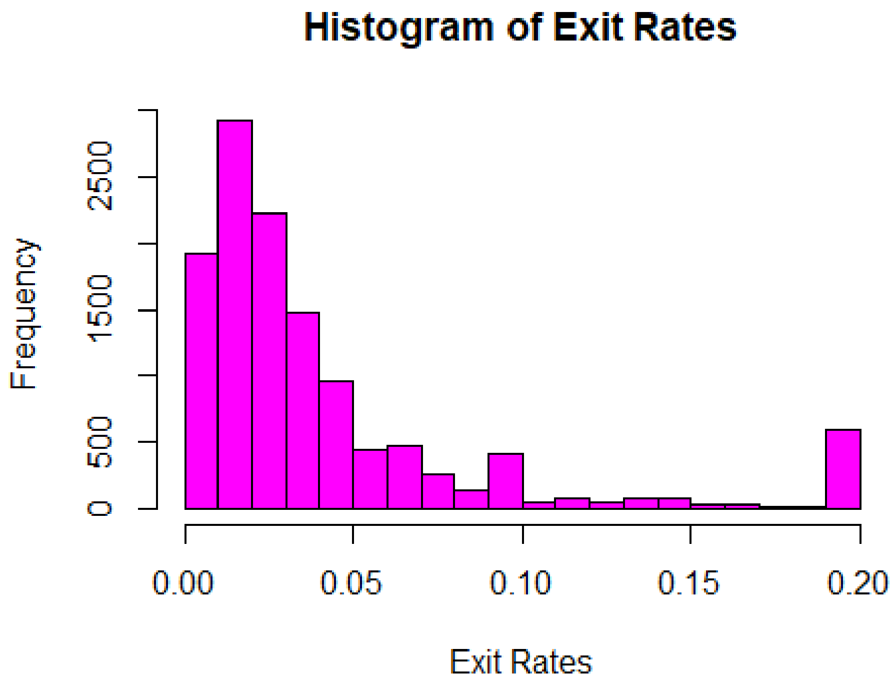
```
# Histogram of the product related variable
hist(df_2$productrelated,
     main="Histogram for Product Related",
     xlab="Age",
     border="black",
     col="green")
```



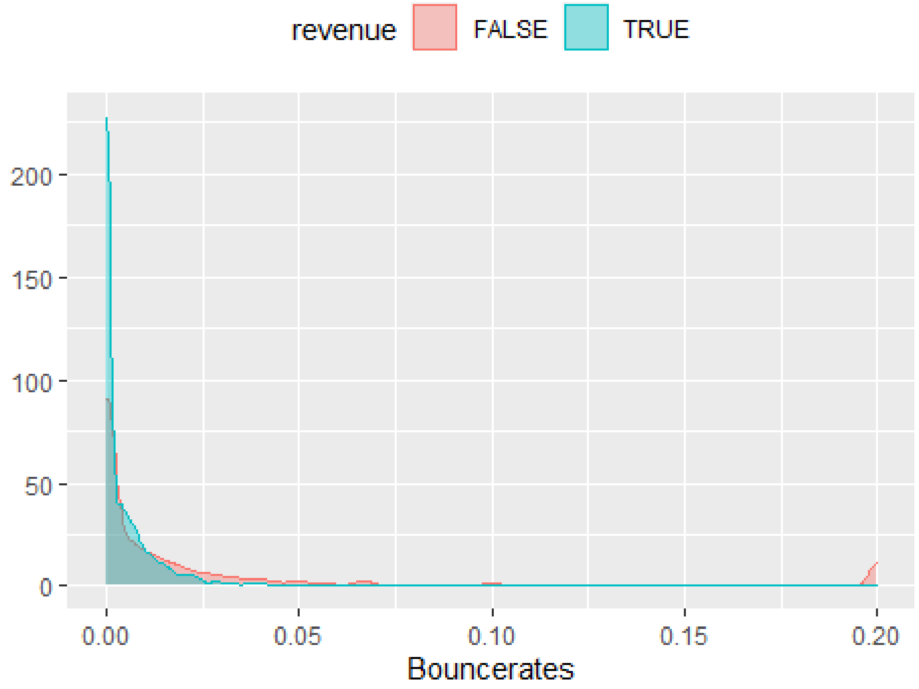
```
skewness(df_2$productrelated)

## [1] 4.332134

kurtosis(df_2$productrelated)

## [1] 34.04903
```

The product related column is also skewed to the right. The measure of skewness is 4.33, which shows a positive skewness. The kurtosis is 34.04903

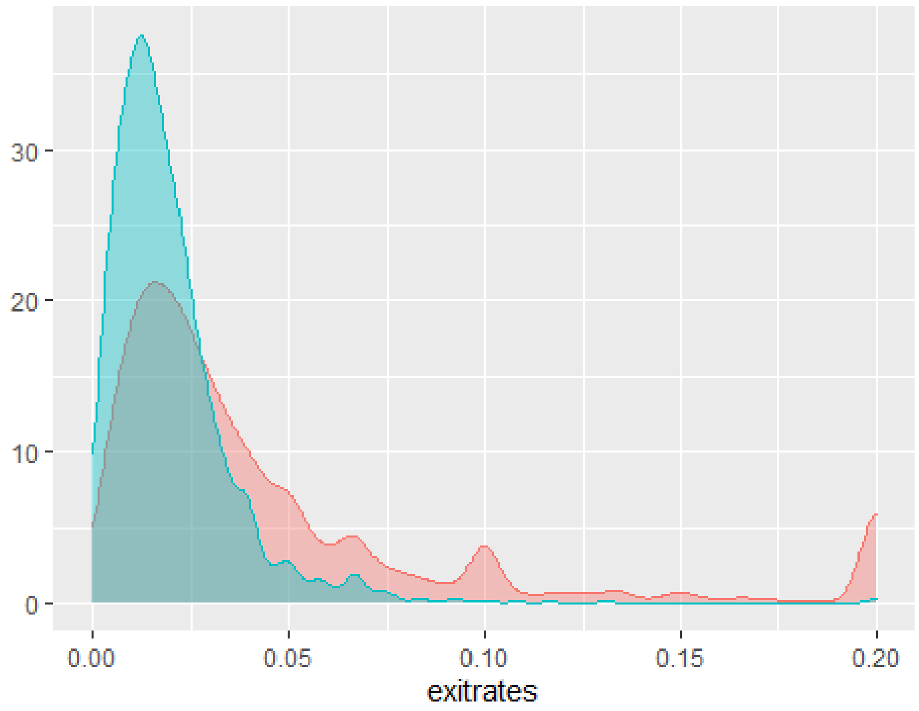
```
# plotting a histogram of Exit Rates
hist(df_2$exitrates,
     main = "Histogram of Exit Rates",
     xlab = "Exit Rates",
     col = "magenta")
```



```
skewness(df_2$exitrates)

## [1] 2.233125

kurtosis(df_2$exitrates)

## [1] 7.624252

p2 = ggplot(df_2, aes(bouncerates, col = revenue)) +
  geom_density(aes(fill = revenue), alpha = 0.4) +
  labs(x = 'Bouncerates', y = '', title = '') +
  theme(legend.position = 'top')
p2
```



```
p3 = ggplot(df_2, aes(exitrates, col = revenue)) +
  geom_density(aes(fill = revenue), alpha = 0.4) +
  labs(x = 'exitrates', y = '', title = '') +
  theme(legend.position = 'none',
        plot.title = element_text(size = 12))
p3
```

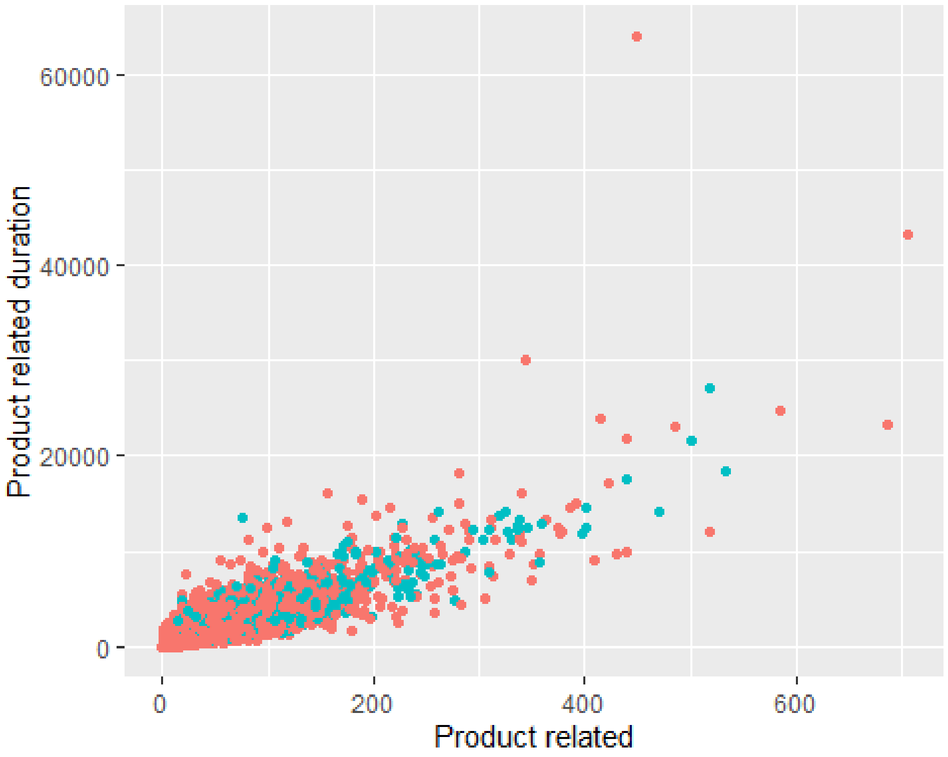


MEASURES OF CENTRAL TENDENCIES AND DISPERSION

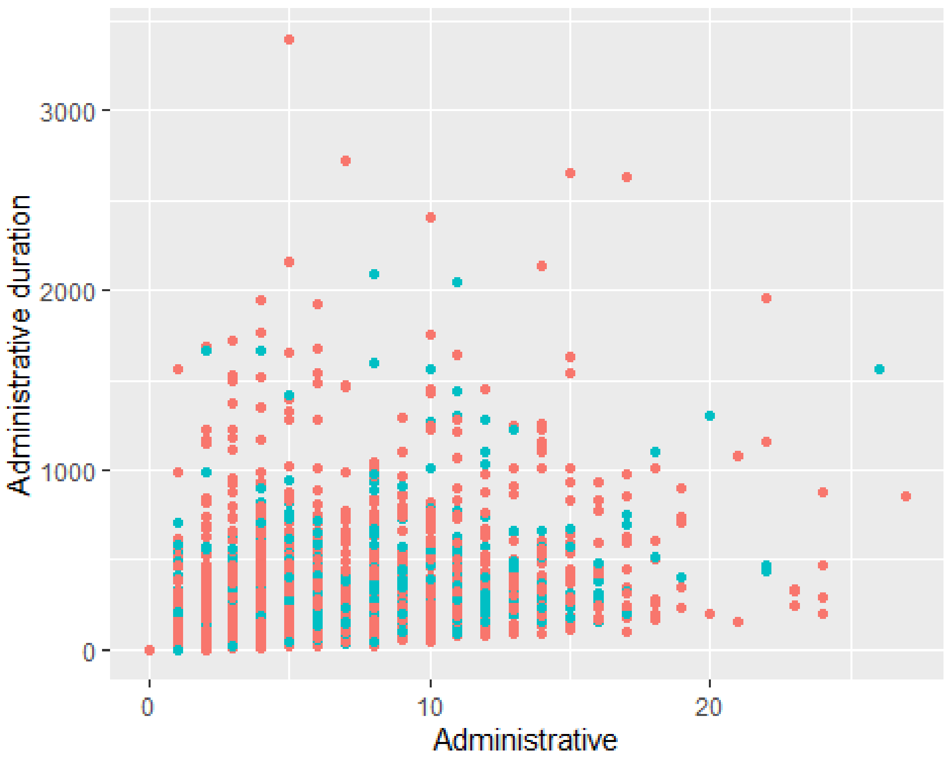
BIVARIATE ANALYSIS

SCATTER PLOTS

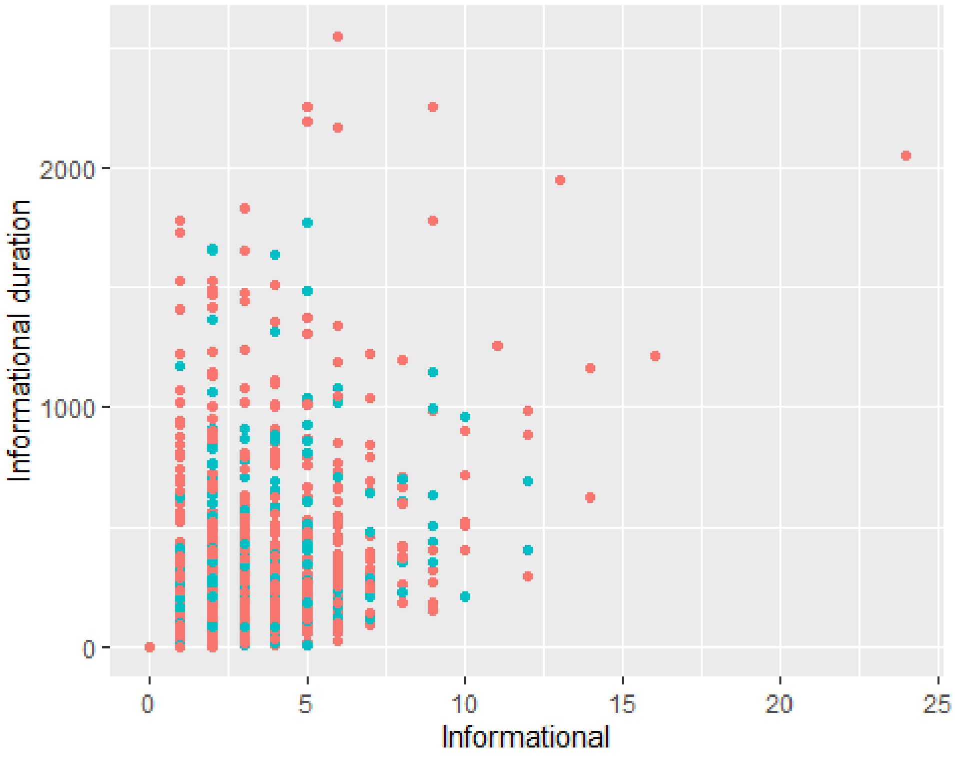
```
options(repr.plot.width = 11, repr.plot.height = 5)
p1 = ggplot(df_2, aes(productrelated, productrelated_duration, col =
revenue)) +
    geom_point() + theme(legend.position = 'none') +
    labs(x='Product related', y ='Product related duration')
p1
```



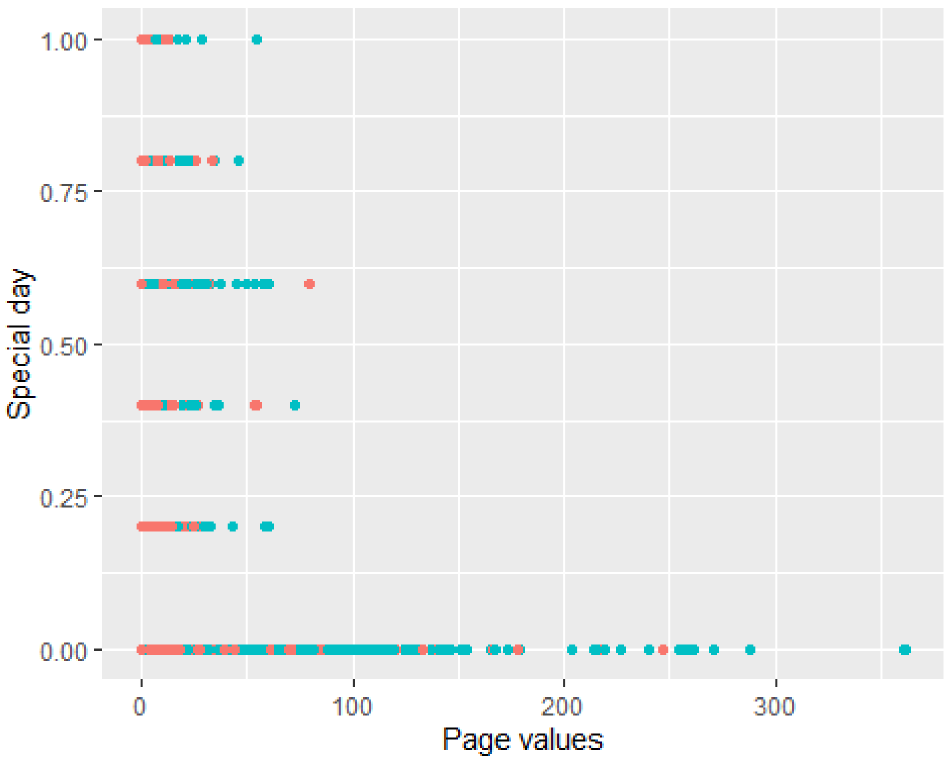
```
p2 = ggplot(df_2, aes(administrative, administrative_duration, col =
revenue)) +
    geom_point() + theme(legend.position = 'none') +
    labs(x = 'Administrative', y = 'Administrative duration')
p2
```



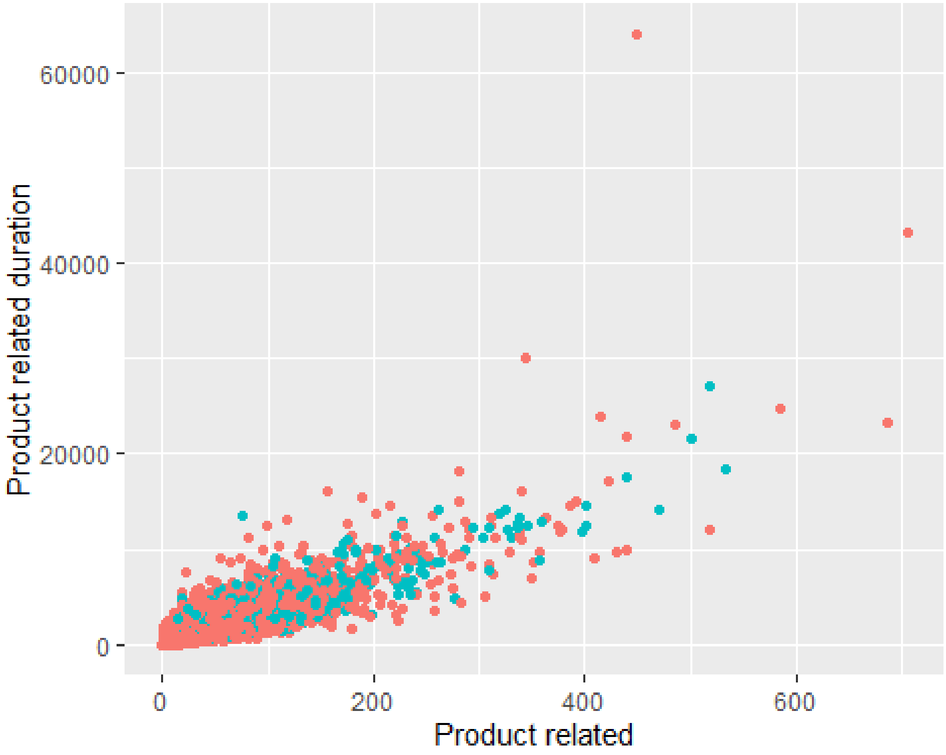
```
p3 = ggplot(df_2, aes(informational, informational_duration, col = revenue))
+
    geom_point() + theme(legend.position = 'none') +
    labs(x = 'Informational', y = 'Informational duration')
p3
```



```
p4 = ggplot(df_2, aes(pagevalues,   specialday  , col = revenue)) +
geom_point() + theme(legend.position = 'none') +
    labs(x = 'Page values', y = 'Special day')
p4
```



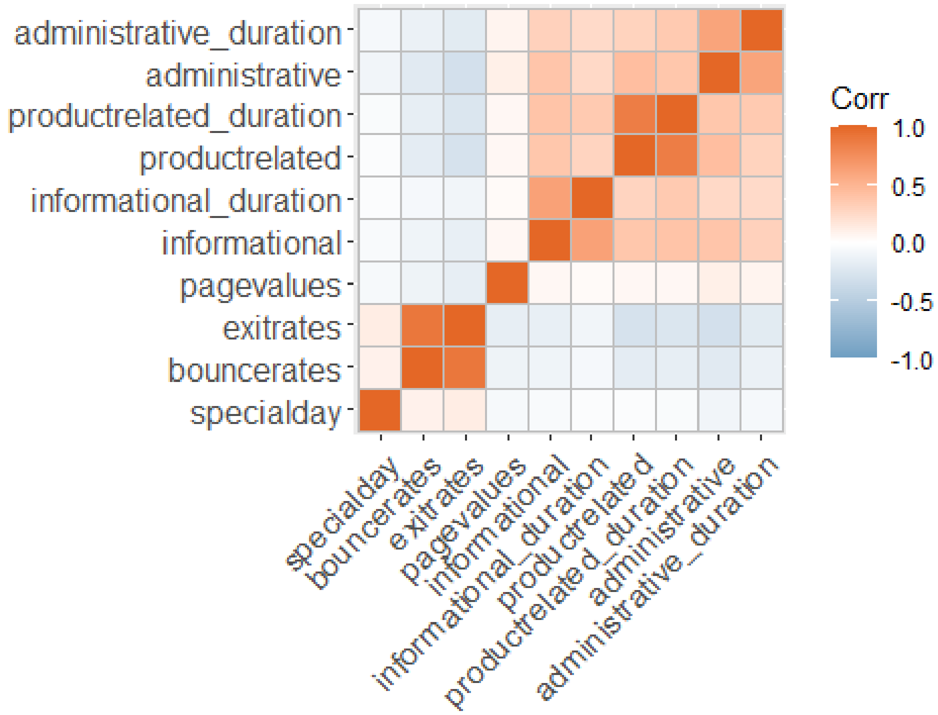
```
p1 = ggplot(df_2, aes(productrelated, productrelated_duration, col =
revenue)) +
    geom_point() + theme(legend.position = 'none') +
    labs(x='Product related', y ='Product related duration')
p1
```



MULTIVARIATE ANALYSIS

#HEATMAP

```
# Plotting a correlogram to check for correlations
library(dplyr)
library(ggcorrplot)
options(repr.plot.width = 4, repr.plot.height = 5)
corr = round(cor(select_if(df_2, is.numeric)), 2)
ggcorrplot(corr, hc.order = T, ggtheme = ggplot2::theme_gray,
   colors = c("#6D9EC1", "white", "#E46726"), lab = F)
```



K-MEANS CLUSTERING

This is an unsupervised learning technique

Encode categorical variables

```
# Creating a copy of the cleaned dataframe
library(dplyr)
non_dummy_df = data.table::copy(df_2)
# Encoding categorical variables
month = data.frame(model.matrix(~0+df_2$month))
opr = data.frame(model.matrix(~0+df_2$operatingsystems))
brw = data.frame(model.matrix(~0+df_2$browser))
reg = data.frame(model.matrix(~0+df_2$region))
trf = data.frame(model.matrix(~0+df_2$traffictype))
vis = data.frame(model.matrix(~0+df_2$visitortype))
wkn = data.frame(model.matrix(~0+df_2$weekend))
rev = data.frame(model.matrix(~0+df_2$revenue))
# Dropping columns which have already encoded
drop_cols = c('month', 'operatingsystems',  'browser',  'region',
'traffictype', 'visitortype', 'weekend', 'revenue')
df_2= select(data.frame(cbind(df_2, month, opr, brw, reg, trf, vis, wkn,
rev)), -drop_cols)

## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(drop_cols)` instead of `drop_cols` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

Scaling the data

```
df2_scaled <- scale(df_2)
```

Checking the summary of the scaled data

```
summary(df2_scaled)
```

```
##  administrative    administrative_duration informational
##  Min.   :-0.7025   Min.   :-0.46574        Min.   :-0.3988
##  1st Qu.:-0.7025   1st Qu.:-0.46011        1st Qu.:-0.3988
##  Median :-0.4023   Median :-0.40941        Median :-0.3988
##  Mean   : 0.0000   Mean   : 0.00000        Mean   : 0.0000
##  3rd Qu.: 0.4984   3rd Qu.: 0.07361        3rd Qu.:-0.3988
##  Max.   : 7.4035   Max.   :18.68474        Max.   :18.4127
##  informational_duration productrelated    productrelated_duration
##  Min.   :-0.2533        Min.   :-0.7188   Min.   :-0.6295
##  1st Qu.:-0.2463        1st Qu.:-0.5394   1st Qu.:-0.5281
##  Median :-0.2463        Median :-0.3152   Median :-0.3115
##  Mean   : 0.0000        Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.2463        3rd Qu.: 0.1332   3rd Qu.: 0.1407
##  Max.   :17.7758        Max.   :15.0881   Max.   :32.6919
##   bouncerates         exitrates          pagevalues        specialday
##  Min.   :-0.45034   Min.   :-0.8973   Min.   :-0.319   Min.   :-0.3103
##  1st Qu.:-0.45034   1st Qu.:-0.5897   1st Qu.:-0.319   1st Qu.:-0.3103
##  Median :-0.38580   Median :-0.3567   Median :-0.319   Median :-0.3103
##  Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.000   Mean   : 0.0000
##  3rd Qu.:-0.08326   3rd Qu.: 0.1511   3rd Qu.:-0.319   3rd Qu.:-0.3103
##  Max.   : 3.95470   Max.   : 3.4273   Max.   :19.070   Max.   : 4.6969
##  df_2.monthAug     df_2.monthDec     df_2.monthFeb     df_2.monthJul
##  Min.   :-0.1918   Min.   :-0.4032   Min.   :-0.1231   Min.   :-0.1916
##  1st Qu.:-0.1918   1st Qu.:-0.4032   1st Qu.:-0.1231   1st Qu.:-0.1916
##  Median :-0.1918   Median :-0.4032   Median :-0.1231   Median :-0.1916
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.1918   3rd Qu.:-0.4032   3rd Qu.:-0.1231   3rd Qu.:-0.1916
##  Max.   : 5.2126   Max.   : 2.4799   Max.   : 8.1254   Max.   : 5.2188
##  df_2.monthJune    df_2.monthMar     df_2.monthMay     df_2.monthNov
##  Min.   :-0.1547   Min.   :-0.4232   Min.   :-0.6125   Min.   :-0.5689
##  1st Qu.:-0.1547   1st Qu.:-0.4232   1st Qu.:-0.6125   1st Qu.:-0.5689
##  Median :-0.1547   Median :-0.4232   Median :-0.6125   Median :-0.5689
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.1547   3rd Qu.:-0.4232   3rd Qu.: 1.6326   3rd Qu.:-0.5689
##  Max.   : 6.4653   Max.   : 2.3628   Max.   : 1.6326   Max.   : 1.7576
##  df_2.monthOct     df_2.monthSep     df_2.operatingsystems1
##  Min.   :-0.2171   Min.   :-0.1952   Min.   :-0.5138
##  1st Qu.:-0.2171   1st Qu.:-0.1952   1st Qu.:-0.5138
##  Median :-0.2171   Median :-0.1952   Median :-0.5138
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.2171   3rd Qu.:-0.1952   3rd Qu.:-0.5138  
##  Max.   : 4.6064   Max.   : 5.1213   Max.   : 1.9461  
##  df_2.operatingsystems2 df_2.operatingsystems3 df_2.operatingsystems4
##  Min.   :-1.0743        Min.   :-0.5115        Min.   :-0.2019       
##  1st Qu.:-1.0743        1st Qu.:-0.5115        1st Qu.:-0.2019       
##  Median : 0.9308        Median :-0.5115        Median :-0.2019       
##  Mean   : 0.0000        Mean   : 0.0000        Mean   : 0.0000       
##  3rd Qu.: 0.9308        3rd Qu.:-0.5115        3rd Qu.:-0.2019       
##  Max.   : 0.9308        Max.   : 1.9548        Max.   : 4.9517       
##  df_2.operatingsystems5 df_2.operatingsystems6 df_2.operatingsystems7
##  Min.   :-0.02218       Min.   :-0.03949       Min.   :-0.02396      
##  1st Qu.:-0.02218       1st Qu.:-0.03949       1st Qu.:-0.02396      
##  Median :-0.02218       Median :-0.03949       Median :-0.02396      
##  Mean   : 0.00000       Mean   : 0.00000       Mean   : 0.00000      
##  3rd Qu.:-0.02218       3rd Qu.:-0.03949       3rd Qu.:-0.02396      
##  Max.   :45.07771       Max.   :25.31798       Max.   :41.73214      
##  df_2.operatingsystems8 df_2.browser1     df_2.browser2     df_2.browser3    
##  Min.   :-0.07865       Min.   :-0.4982   Min.   :-1.3502   Min.   :-
0.09317  
##  1st Qu.:-0.07865       1st Qu.:-0.4982   1st Qu.:-1.3502   1st Qu.:-
0.09317  
##  Median :-0.07865       Median :-0.4982   Median : 0.7406   Median :-
0.09317  
##  Mean   : 0.00000       Mean   : 0.0000   Mean   : 0.0000   Mean   : 
0.00000  
##  3rd Qu.:-0.07865       3rd Qu.:-0.4982   3rd Qu.: 0.7406   3rd Qu.:-
0.09317  
##  Max.   :12.71378       Max.   : 2.0070   Max.   : 0.7406   Max.   
:10.73180  
##  df_2.browser4     df_2.browser5     df_2.browser6     df_2.browser7    
##  Min.   :-0.2523   Min.   :-0.1993   Min.   :-0.1203   Min.   :-0.0635  
##  1st Qu.:-0.2523   1st Qu.:-0.1993   1st Qu.:-0.1203   1st Qu.:-0.0635  
##  Median :-0.2523   Median :-0.1993   Median :-0.1203   Median :-0.0635  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.:-0.2523   3rd Qu.:-0.1993   3rd Qu.:-0.1203   3rd Qu.:-0.0635  
##  Max.   : 3.9635   Max.   : 5.0176   Max.   : 8.3129   Max.   :15.7461  
##  df_2.browser8     df_2.browser9       df_2.browser10    df_2.browser11    
##  Min.   :-0.1058   Min.   : -0.00905   Min.   :-0.1164   Min.   :-0.02218  
##  1st Qu.:-0.1058   1st Qu.: -0.00905   1st Qu.:-0.1164   1st Qu.:-0.02218  
##  Median :-0.1058   Median : -0.00905   Median :-0.1164   Median :-0.02218  
##  Mean   : 0.0000   Mean   :  0.00000   Mean   : 0.0000   Mean   : 0.00000  
##  3rd Qu.:-0.1058   3rd Qu.: -0.00905   3rd Qu.:-0.1164   3rd Qu.:-0.02218  
##  Max.   : 9.4528   Max.   :110.44003   Max.   : 8.5927   Max.   :45.07771  
##  df_2.browser12     df_2.browser13      df_2.region1      df_2.region2    
##  Min.   :-0.02864   Min.   :-0.06791   Min.   :-0.7932   Min.   :-0.319  
##  1st Qu.:-0.02864   1st Qu.:-0.06791   1st Qu.:-0.7932   1st Qu.:-0.319  
##  Median :-0.02864   Median :-0.06791   Median :-0.7932   Median :-0.319  
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.000  
##  3rd Qu.:-0.02864   3rd Qu.:-0.06791   3rd Qu.: 1.2607   3rd Qu.:-0.319  
##  Max.   :34.91132   Max.   :14.72486   Max.   : 1.2607   Max.   : 3.134  
##   df_2.region3      df_2.region4      df_2.region5      df_2.region6    
##  Min.   :-0.4926   Min.   :-0.3254   Min.   :-0.1633   Min.   :-0.2649  
##  1st Qu.:-0.4926   1st Qu.:-0.3254   1st Qu.:-0.1633   1st Qu.:-0.2649  
##  Median :-0.4926   Median :-0.3254   Median :-0.1633   Median :-0.2649  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.:-0.4926   3rd Qu.:-0.3254   3rd Qu.:-0.1633   3rd Qu.:-0.2649  
##  Max.   : 2.0300   Max.   : 3.0730   Max.   : 6.1221   Max.   : 3.7746  
##   df_2.region7      df_2.region8      df_2.region9     df_2.trafficype1
##  Min.   :-0.2574   Min.   :-0.1914   Min.   :-0.2078   Min.   :-0.4927  
##  1st Qu.:-0.2574   1st Qu.:-0.1914   1st Qu.:-0.2078   1st Qu.:-0.4927  
##  Median :-0.2574   Median :-0.1914   Median :-0.2078   Median :-0.4927  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.:-0.2574   3rd Qu.:-0.1914   3rd Qu.:-0.2078   3rd Qu.:-0.4927  
##  Max.   : 3.8849   Max.   : 5.2251   Max.   : 4.8119   Max.   : 2.0295  
##  df_2.traffictype2 df_2.traffictype3 df_2.traffictype4 df_2.traffictype5
##  Min.   :-0.6864   Min.   :-0.4451   Min.   :-0.3094   Min.   :-0.1476  
##  1st Qu.:-0.6864   1st Qu.:-0.4451   1st Qu.:-0.3094   1st Qu.:-0.1476  
##  Median :-0.6864   Median :-0.4451   Median :-0.3094   Median :-0.1476  
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 1.4568   3rd Qu.:-0.4451   3rd Qu.:-0.3094   3rd Qu.:-0.1476  
##  Max.   : 1.4568   Max.   : 2.2467   Max.   : 3.2315   Max.   : 6.7761  
##  df_2.traffictype6 df_2.traffictype7  df_2.traffictype8 df_2.traffictype9
##  Min.   :-0.1941   Min.   :-0.05735   Min.   :-0.1701   Min.   :-0.05807  
##  1st Qu.:-0.1941   1st Qu.:-0.05735   1st Qu.:-0.1701   1st Qu.:-0.05807  
##  Median :-0.1941   Median :-0.05735   Median :-0.1701   Median :-0.05807  
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000   Mean   : 0.00000  
##  3rd Qu.:-0.1941   3rd Qu.:-0.05735   3rd Qu.:-0.1701   3rd Qu.:-0.05807  
##  Max.   : 5.1512   Max.   :17.43416   Max.   : 5.8790   Max.   :17.21953  
##  df_2.traffictype10 df_2.traffictype11 df_2.traffictype12
df_2.traffictype13
##  Min.   :-0.1957    Min.   :-0.1438    Min.   : -0.00905   Min.   :-0.2519  
##  1st Qu.:-0.1957    1st Qu.:-0.1438    1st Qu.: -0.00905   1st Qu.:-0.2519  
##  Median :-0.1957    Median :-0.1438    Median : -0.00905   Median :-0.2519  
##  Mean   : 0.0000    Mean   : 0.0000    Mean   :  0.00000   Mean   : 0.0000  
##  3rd Qu.:-0.1957    3rd Qu.:-0.1438    3rd Qu.: -0.00905   3rd Qu.:-0.2519  
##  Max.   : 5.1095    Max.   : 6.9559    Max.   :110.44003   Max.   : 3.9693  
##  df_2.traffictype14 df_2.traffictype15 df_2.traffictype16
df_2.traffictype17
##  Min.   :-0.03266   Min.   :-0.0544    Min.   :-0.01568   Min.   : -
0.00905  
##  1st Qu.:-0.03266   1st Qu.:-0.0544    1st Qu.:-0.01568   1st Qu.: -
0.00905  
##  Median :-0.03266   Median :-0.0544    Median :-0.01568   Median : -
0.00905  
##  Mean   : 0.00000   Mean   : 0.0000    Mean   : 0.00000   Mean   :  
0.00000  
##  3rd Qu.:-0.03266   3rd Qu.:-0.0544    3rd Qu.:-0.01568   3rd Qu.: -
0.00905  
##  Max.   :30.61548   Max.   :18.3802    Max.   :63.75735   Max.   
:110.44003
```

```
##  df_2.traffictype18 df_2.traffictype19 df_2.traffictype20
##  Min.   :-0.02864   Min.   :-0.03735   Min.   :-0.1268
##  1st Qu.:-0.02864   1st Qu.:-0.03735   1st Qu.:-0.1268
##  Median :-0.02864   Median :-0.03735   Median :-0.1268
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.:-0.02864   3rd Qu.:-0.03735   3rd Qu.:-0.1268
##  Max.   :34.91132   Max.   :26.76807   Max.   : 7.8868
##  df_2.visitortypeNew_Visitor df_2.visitortypeOther
##  Min.   :-0.4014             Min.   :-0.08175
##  1st Qu.:-0.4014             1st Qu.:-0.08175
##  Median :-0.4014             Median :-0.08175
##  Mean   : 0.0000             Mean   : 0.00000
##  3rd Qu.:-0.4014             3rd Qu.:-0.08175
##  Max.   : 2.4910             Max.   :12.23081
##  df_2.visitortypeReturning_Visitor df_2.weekendFALSE df_2.weekendTRUE
##  Min.   :-2.4241                   Min.   :-1.8086   Min.   :-0.5529
##  1st Qu.: 0.4125                   1st Qu.: 0.5529   1st Qu.:-0.5529
##  Median : 0.4125                   Median : 0.5529   Median :-0.5529
##  Mean   : 0.0000                   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4125                   3rd Qu.: 0.5529   3rd Qu.:-0.5529
##  Max.   : 0.4125                   Max.   : 0.5529   Max.   : 1.8086
##  df_2.revenueFALSE df_2.revenueTRUE
##  Min.   :-2.3223   Min.   :-0.4306
##  1st Qu.: 0.4306   1st Qu.:-0.4306
##  Median : 0.4306   Median :-0.4306
##  Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4306   3rd Qu.:-0.4306
##  Max.   : 0.4306   Max.   : 2.3223
```

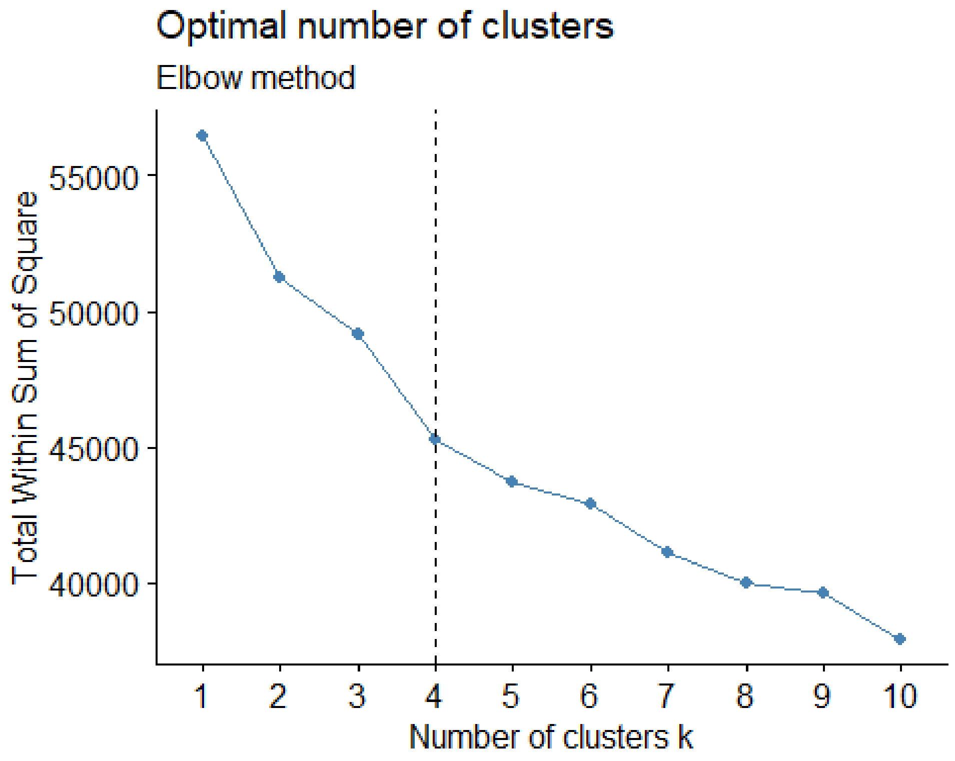
The data still has very varying values and negative values for minimum. Scaling has not fixed this. We should normalize the data.

Normalizing the data

```
main_df = data.table::copy(df_2)

# Normalising the data
df_2 = as.data.frame(apply(df_2, 2,  function(x) (x - min(x))/max(x) -
min(x)))
```

Getting the optimum number of clusters using a dendogram Use the elbow method

```
# installing and loading the necessary packages
#install.packages("factoextra")
library(factoextra)

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

# Using the Elbow method
fviz_nbclust(df_2, kmeans, method = "wss") +
  geom_vline(xintercept = 4, linetype = 2)+
 labs(subtitle = "Elbow method")
```



From the dendogram, the optimum number of clusters is 4 Clustering

```
# Performing clustering with the optimal number of clusters
kmeans_res = kmeans(df_2, 4)
# Checking the cluster centers of each variable
kmeans_res$centers

##   administrative administrative_duration informational
informational_duration
## 1     0.07885088                1.019908    0.01781524
1.012742
## 2     0.12568911                1.035449    0.03275681
1.022991
## 3     0.07621446                1.023821    0.01791785
1.010278
## 4     0.08534075                1.024338    0.02308726
1.014945
##   productrelated productrelated_duration bouncerates  exitrates
pagevalues
## 1     0.04657146                1.019627  0.09102102 0.21289293
0.005980868
## 2     0.06838322                1.029344  0.02558576 0.09777584
0.075365536
## 3     0.03302353                1.013229  0.15828098 0.26218414
```

```
0.005210076
## 4     0.04351968                1.017533  0.10129086 0.20426369
0.005173920
##   specialday df_2.monthAug df_2.monthDec df_2.monthFeb df_2.monthJul
## 1 0.07465940    0.03451408     0.1400999   0.016802906    0.03156222
## 2 0.02316562    0.03983229     0.1132075   0.001572327    0.03459119
## 3 0.06634561    0.03427762     0.1560907   0.022096317    0.03937677
## 4 0.06313110    0.03563852     0.1366143   0.011455240    0.03733560
##   df_2.monthJune df_2.monthMar df_2.monthMay df_2.monthNov df_2.monthOct
## 1     0.02770209     0.1693915     0.3108538     0.1911898    0.04041780
## 2     0.01519916     0.1006289     0.1912998     0.3983229    0.06027254
## 3     0.02606232     0.1430595     0.2773371     0.2266289    0.04022663
## 4     0.01781926     0.1739499     0.2609249     0.2464998    0.04836657
##   df_2.monthSep df_2.operatingsystems1 df_2.operatingsystems2
## 1    0.03746594              0.0000000              1.0000000
## 2    0.04507338              0.1986373              0.6053459
## 3    0.03484419              0.4427762              0.0000000
## 4    0.03139584              0.2571065              0.4145100
##   df_2.operatingsystems3 df_2.operatingsystems4 df_2.operatingsystems5
## 1              0.0000000             0.00000000           0.0000000000
## 2              0.1404612             0.04454927           0.0005241090
## 3              0.4524079             0.08498584           0.0011331445
## 4              0.2821383             0.03945694           0.0004242681
##   df_2.operatingsystems6 df_2.operatingsystems7 df_2.operatingsystems8
## 1            0.000000000           0.0000000000            0.000000000
## 2            0.001048218           0.0005241090            0.008909853
## 3            0.003966006           0.0011331445            0.013597734
## 4            0.001272804           0.0008485363            0.004242681
##   df_2.browser1 df_2.browser2 df_2.browser3 df_2.browser4 df_2.browser5
## 1  0.0006811989     0.7654405   0.000000000  0.1146684832   0.063578565
## 2  0.1912997904     0.6409853   0.002620545  0.0681341719   0.045073375
## 3  0.4186968839     0.5133144   0.019263456  0.0002832861   0.007365439
## 4  0.2460755197     0.6245227   0.013576580  0.0398812049   0.031395842
##   df_2.browser6 df_2.browser7 df_2.browser8 df_2.browser9 df_2.browser10
## 1    0.02588556   0.004995459    0.00000000  0.0000000000   0.0231607629
## 2    0.01048218   0.003144654    0.01100629  0.0000000000   0.0167714885
## 3    0.00368272   0.002266289    0.02237960  0.0002832861   0.0005665722
## 4    0.01145524   0.005515486    0.01484938  0.0000000000   0.0114552397
##   df_2.browser11 df_2.browser12 df_2.browser13 df_2.region1 df_2.region2
## 1   0.0000000000   0.0009082652   0.0006811989    0.3785195   0.09536785
## 2   0.0005241090   0.0015723270   0.0083857442    0.4040881   0.09853249
## 3   0.0011331445   0.0005665722   0.0101983003    0.3920680   0.08555241
## 4   0.0004242681   0.0004242681   0.0004242681    0.3771744   0.09206619
##   df_2.region3 df_2.region4 df_2.region5 df_2.region6 df_2.region7
df_2.region8
## 1    0.1898274   0.08878292   0.02974569   0.06630336   0.07947321
0.02588556
## 2    0.1829140   0.09171908   0.02725367   0.05870021   0.06236897
0.02935010
## 3    0.2070822   0.09830028   0.02124646   0.06713881   0.04050992
```

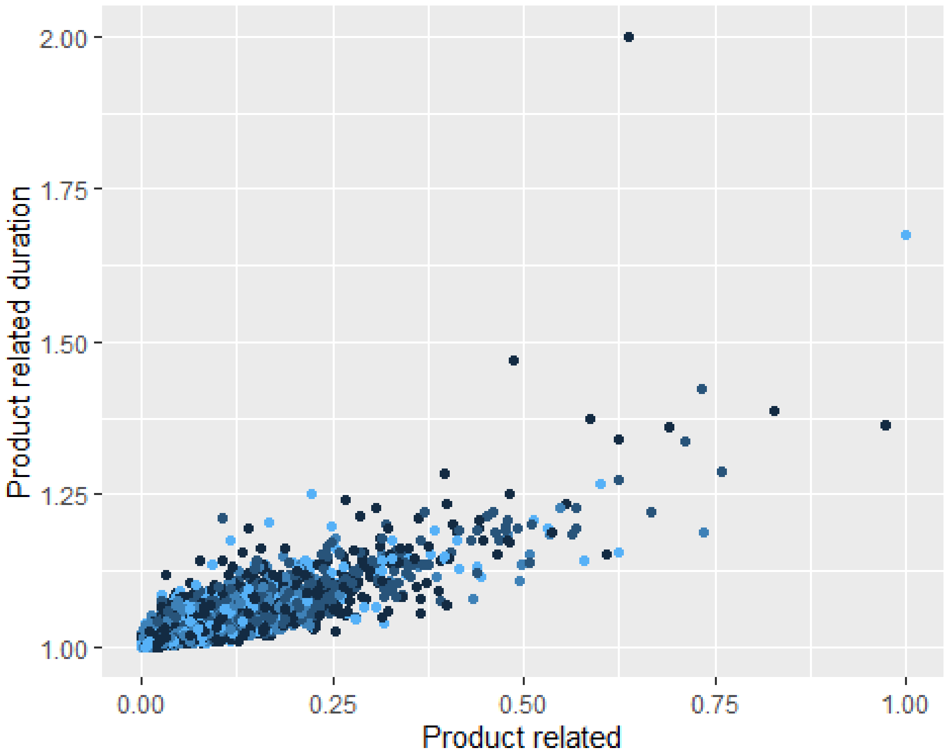
```
0.04900850
## 4    0.1977090   0.10818838   0.02503182   0.06745863   0.06194315
0.03733560
##   df_2.region9 df_2.traffictype1 df_2.traffictype2 df_2.traffictype3
## 1   0.04609446         0.2452316         0.3024523        0.16348774
## 2   0.04507338         0.1373166         0.4439203        0.09433962
## 3   0.03909348         0.1691218         0.2682720        0.22549575
## 4   0.03309291         0.1883751         0.3313534        0.13619007
##   df_2.traffictype4 df_2.traffictype5 df_2.traffictype6 df_2.traffictype7
## 1        0.08492280        0.01998183        0.04382380       0.002724796
## 2        0.08647799        0.02935010        0.02777778       0.006289308
## 3        0.07960340        0.01869688        0.03087819       0.002266289
## 4        0.10436996        0.02121341        0.03733560       0.003394145
##   df_2.traffictype8 df_2.traffictype9 df_2.traffictype10
df_2.traffictype11
## 1        0.01930064      0.0000000000         0.02929155
0.01226158
## 2        0.04979036      0.0020964361         0.04716981
0.02463312
## 3        0.02577904      0.0008498584         0.04135977
0.01983003
## 4        0.03054731      0.0144251167         0.03606279
0.03224438
##   df_2.traffictype12 df_2.traffictype13 df_2.traffictype14
df_2.traffictype15
## 1       0.0002270663         0.06221617       0.0004541326
0.000000000
## 2       0.0000000000         0.02253669       0.0010482180
0.000000000
## 3       0.0000000000         0.08753541       0.0008498584
0.005665722
## 4       0.0000000000         0.04327535       0.0025456088
0.006788290
##   df_2.traffictype16 df_2.traffictype17 df_2.traffictype18
df_2.traffictype19
## 1       0.0004541326       0.0000000000       0.0006811989
0.001589464
## 2       0.0005241090       0.0000000000       0.0000000000
0.000524109
## 3       0.0000000000       0.0000000000       0.0008498584
0.001416431
## 4       0.0000000000       0.0004242681       0.0016970725
0.001697073
##   df_2.traffictype20 df_2.visitortypeNew_Visitor df_2.visitortypeOther
## 1        0.010899183                   0.1146685           0.002497729
## 2        0.026205451                   0.2211740           0.008385744
## 3        0.021529745                   0.1110482           0.013597734
## 4        0.008061095                   0.1586763           0.002545609
##   df_2.visitortypeReturning_Visitor df_2.weekendFALSE df_2.weekendTRUE
## 1                         0.8828338         1.0000000        0.0000000
## 2                         0.7704403         0.7384696        0.2615304
## 3                         0.8753541         1.0000000        0.0000000
## 4                         0.8387781         0.0000000        1.0000000
##   df_2.revenueFALSE df_2.revenueTRUE
## 1                 1                0
## 2                 0                1
## 3                 1                0
## 4                 1                0
```

```
# Previewing the size of observations in each cluster
kmeans_res$size
```

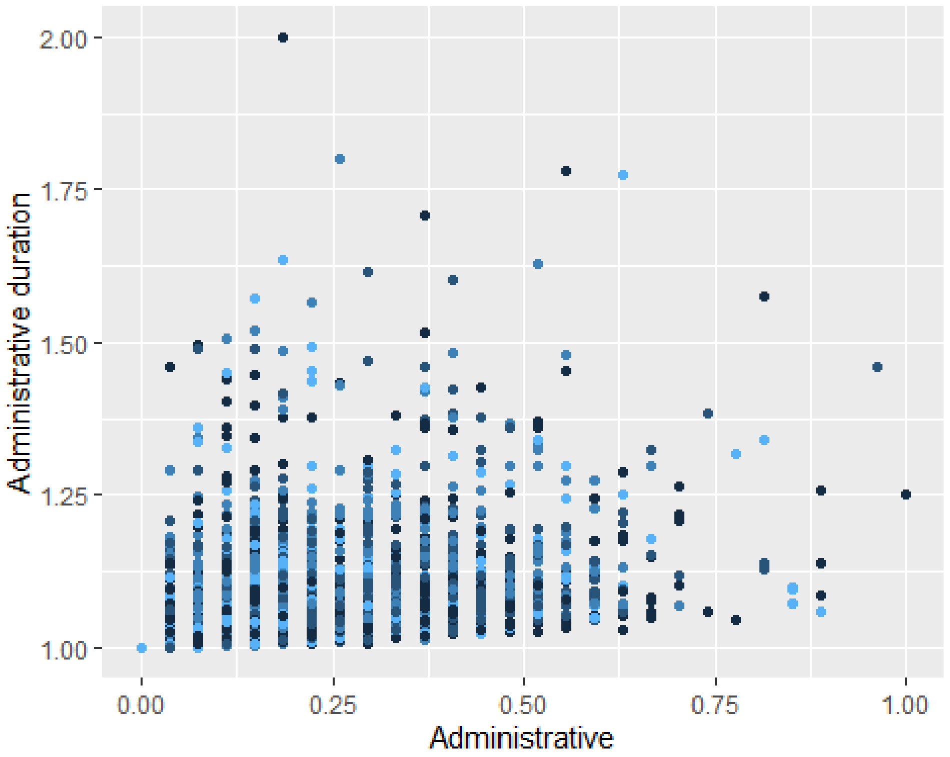
```
## [1] 4404 1908 3530 2357
```

Plotting the clusters to see how some of the features are clustered.

```
# Checking how some features have been clustered
options(repr.plot.width = 11, repr.plot.height = 6)
p1 = ggplot(df_2, aes(productrelated, productrelated_duration, col =
kmeans_res$cluster)) +
    geom_point() + theme(legend.position = 'none') +
    labs(x='Product related', y ='Product related duration')
p1
```



```
p2 = ggplot(df_2, aes(administrative, administrative_duration, col =
kmeans_res$cluster)) +
    geom_point() + theme(legend.position = 'none') +
    labs(x = 'Administrative', y = 'Administrative duration')
p2
```

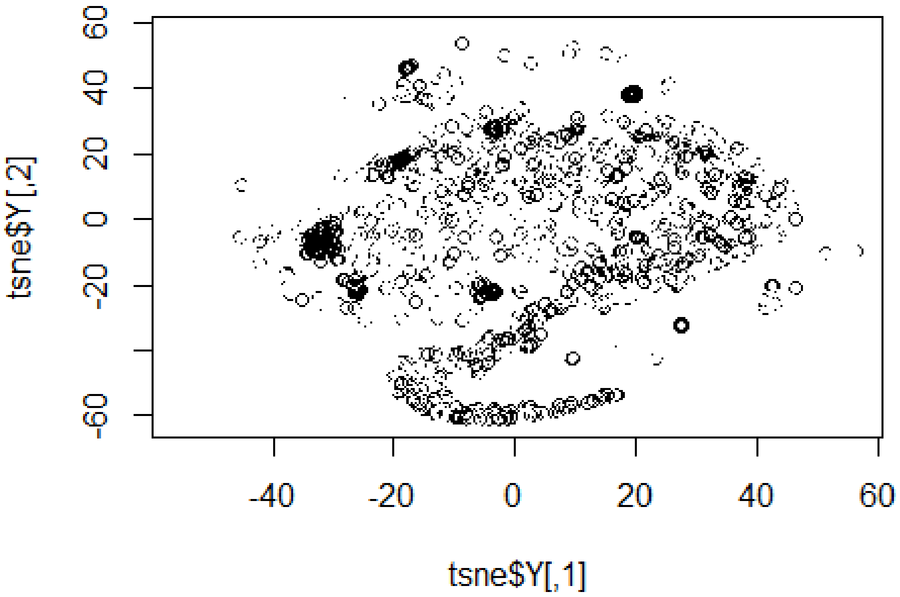


HIERCHICAL CLUSTERING

```
# The euclidean distance and the ward2 method has been used to perform
hierachical clustering
hierachical_res = hclust(dist(df_2, method = 'euclidean'), method =
'ward.D2')

# Reducing the dimensionality of the dataset
pca_res = prcomp(main_df, scale = T, center = T)
```

t-SNE modelling

```
# installing and loading the necessary package
#install.packages("Rtsne")
library(Rtsne)

# modelling
unique_df = unique(non_dummy_df[, 1:18])
tsne = Rtsne(unique_df[, 1:17], epoch=1000)
plot(tsne$Y, col= non_dummy_df$revenue)
```


tsne$Y[,2]
60
40
20
0
-20
-60
-40
-20
0
20
40
60
tsne$Y[,1]