Kira Pastinina customer’s behavior model using Unsupervised Learning

Christopher Toromo

2022-06-04

1. Problem Definition

The brand’s Sales and Marketing team of Kira Plastinina would like to understand the customer’s behavior from data that has been collected over the past year. More specifically, to learn the characteristics of customer groups.

2. Data Sourcing

The dataset for this Independent project can be found here http://bit.ly/EcommerceCustomersDataset

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.

3. Checking the Data

i). Reading the Data

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

ii). Previewing the Data

```
# Checking the top records

head(customer)

##   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

# Checking the Bottom records

tail(customer)

##       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
```

```
# Number of Records

cat('Our dataset contains', nrow(customer), 'rows and', ncol(customer),'columns.')
```

```
## Our dataset contains 12330 rows and 18 columns.
```

```
# Checking Datatypes

str(customer)
```

```
## '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"
"Returning_Visitor" ...
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

Although all columns contain the requisite datatypes, the ‘chr’ columns’ data type will be transformed to factors.

```
customer <- as.data.frame(unclass(customer),
                          stringsAsFactors = TRUE)



# Checking the dataset again
str(customer)
```

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

4. Performing Data Cleaning

a) Validation

Checking for unnecessary columns that do not contribute to the study.

```
colnames(customer)
```

```
##  [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"
```

## ***Checking for invalid values***

```
summary(customer)
```

```
##  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
##
```

b). Constistency

```
# Checking for missing values

colSums(is.na(customer))

##          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 shall be dropping the Missing values to avoid inconsistency in our dataset.

```
customer <- na.omit(customer)

# checking to see the missing values are no longer there

colSums(is.na(customer))

##          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
```

c). Completeness

```
# Checking for duplicates values.
sum(duplicated(customer))

## [1] 117
```

We have 117 duplicate records in our Dataset. We shall be dropping them.

```
customer <- unique(customer)

# Checking to confirm they have been removed

sum(duplicated(customer))
```

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

d). Uniformity

```
# Checking the Uniformity of the column names

colnames(customer)
```

```
##  [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"
```

The columns are in good format and uniform hence easy to apply on various models.

e).Checking for outliers

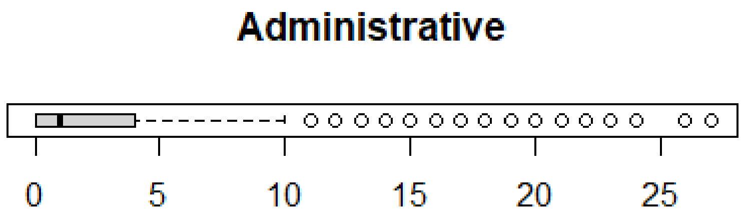
```
library("dplyr")
```

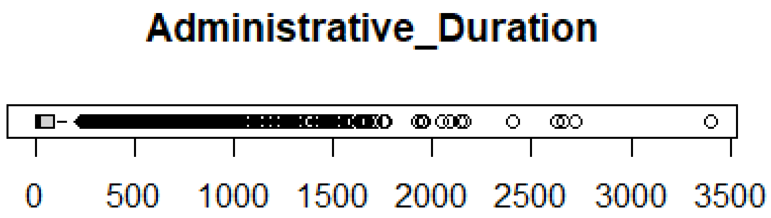
```
## 
## Attaching package: 'dplyr'
```

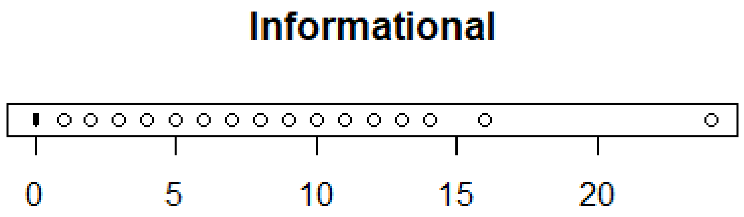
```
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```

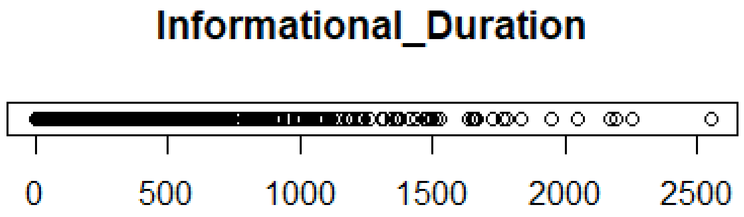
```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

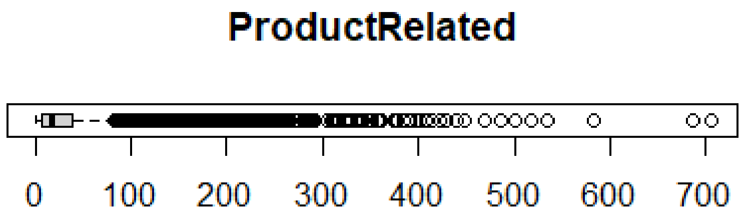
```
num_col <- select_if(customer, is.numeric)

par(mfrow = c(2,1))

for (i in 1:14){
 boxplot(num_col[,i], main = names(num_col)[i], horizontal = TRUE)
}
```


Administrative
0
5
10
15
20
25

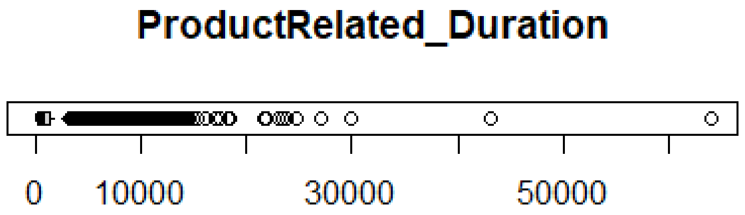

Administrative_Duration
0
500
1000
1500
2000
2500
3000
3500


Informational
0
5
10
15
20

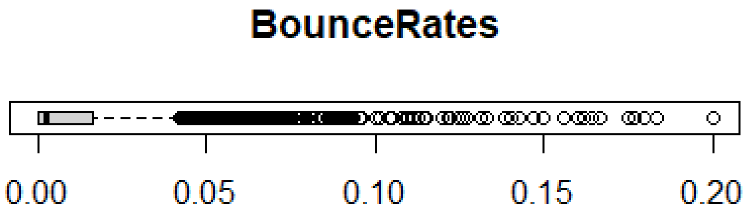

Informational_Duration
0
500
1000
1500
2000
2500



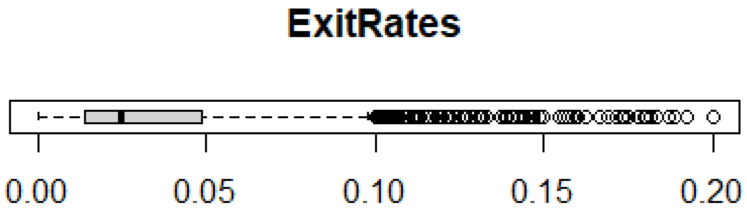
ProductRelated



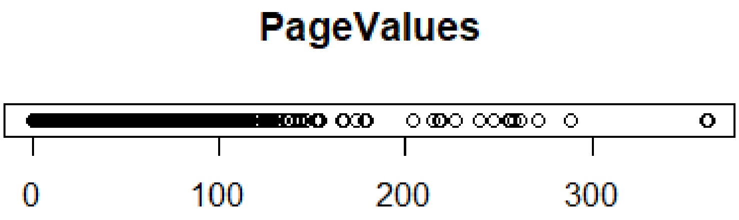
ProductRelated_Duration

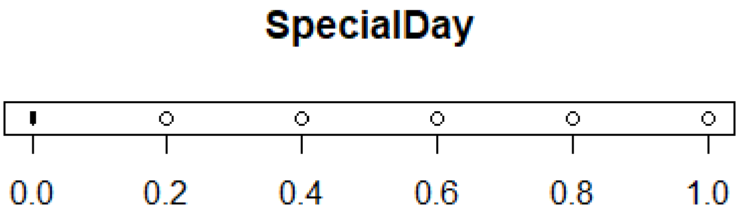


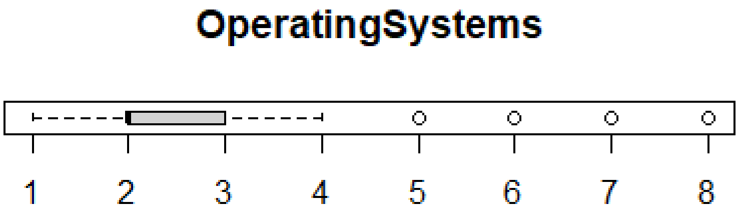
BounceRates

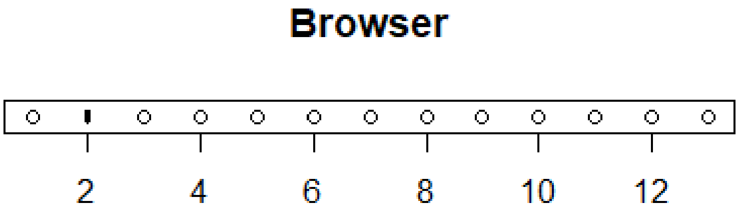


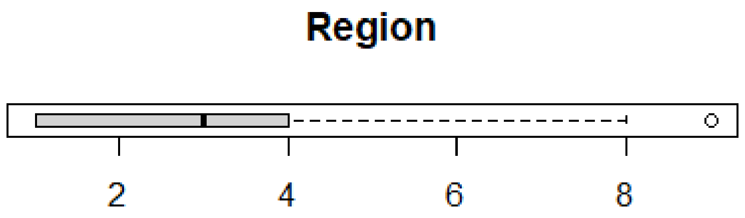
ExitRates

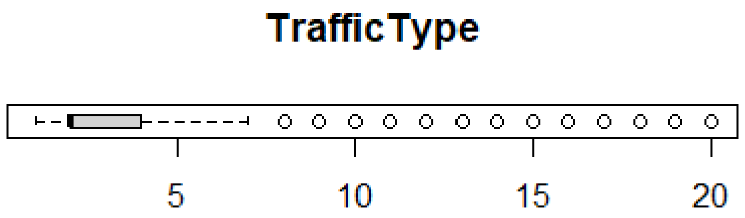

PageValues
0
100
200
300


SpecialDay
0.0
0.2
0.4
0.6
0.8
1.0


OperatingSystems
1
2
3
4
5
6
7
8


Browser
2
4
6
8
10
12





We shall not be removing the outliers because they will be essential in our study. They represent actual behavior of customers.

5. Performing Exploratory Data Analysis (Univariate, Bivariate & Multivariate)

Univariate Analysis

Descriptive statistics

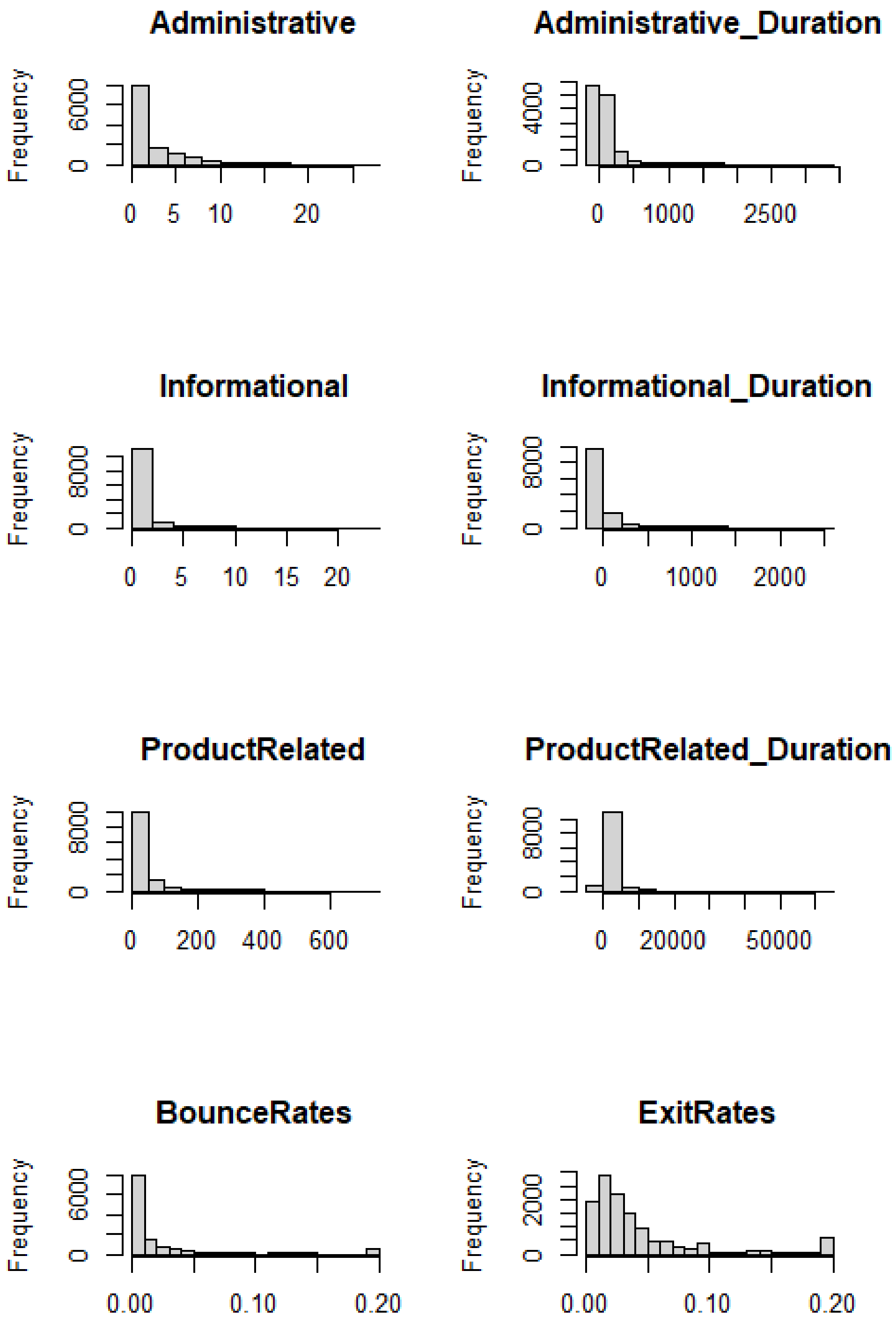
```
library("psych")

describe(customer)

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to min; returning Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf

## Warning in FUN(newX[, i], ...): no non-missing arguments to max; returning -Inf

##                         vars     n   mean     sd median trimmed   mad min
## Administrative             1 12199   2.34   3.33   1.00    1.66  1.48   0
## Administrative_Duration    2 12199  81.68 177.53   9.00   42.87 13.34  -1
## Informational              3 12199   0.51   1.28   0.00    0.18  0.00   0
## Informational_Duration     4 12199  34.84 141.46   0.00    3.73  0.00  -1
## ProductRelated             5 12199   32.06   44.60  18.00   23.06  19.27   0
## ProductRelated_Duration    6 12199 1207.51 1919.93 609.54  832.36 745.12  -1
## BounceRates                7 12199    0.02    0.05   0.00    0.01   0.00   0
## ExitRates                  8 12199    0.04    0.05   0.03    0.03   0.02   0
## PageValues                 9 12199    5.95   18.66   0.00    1.33   0.00   0
## SpecialDay                10 12199    0.06    0.20   0.00    0.00   0.00   0
## Month*                    11 12199    6.17    2.37   7.00    6.36   1.48   1
## OperatingSystems          12 12199    2.12    0.91   2.00    2.06   0.00   1
## Browser                   13 12199    2.36    1.71   2.00    2.00   0.00   1
## Region                    14 12199    3.15    2.40   3.00    2.79   2.97   1
## TrafficType               15 12199    4.07    4.02   2.00    3.22   1.48   1
## VisitorType*              16 12199    2.72    0.69   3.00    2.89   0.00   1
## Weekend                   17 12199     NaN      NA     NA     NaN     NA Inf
## Revenue                   18 12199     NaN      NA     NA     NaN     NA Inf
##                              max    range  skew kurtosis    se
## Administrative             27.00    27.00  1.95     4.63  0.03
## Administrative_Duration  3398.75  3399.75  5.59    50.09  1.61
## Informational              24.00    24.00  4.01    26.64  0.01
## Informational_Duration   2549.38  2550.38  7.54    75.45  1.28
## ProductRelated            705.00   705.00  4.33    31.04  0.40
## ProductRelated_Duration 63973.52 63974.52  7.25   136.57 17.38
## BounceRates                 0.20     0.20  3.15     9.25  0.00
## ExitRates                   0.20     0.20  2.23     4.62  0.00
## PageValues                361.76   361.76  6.35    64.93  0.17
## SpecialDay                  1.00     1.00  3.28     9.78  0.00
## Month*                     10.00     9.00 -0.83    -0.37  0.02
## OperatingSystems            8.00     7.00  2.03    10.27  0.01
## Browser                    13.00    12.00  3.22    12.53  0.02
## Region                      9.00     8.00  0.98    -0.16  0.02
## TrafficType                20.00    19.00  1.96     3.47  0.04
## VisitorType*                3.00     2.00 -2.05     2.23  0.01
## Weekend                     -Inf     -Inf    NA       NA    NA
## Revenue                     -Inf     -Inf    NA       NA    NA
```

```
#Making a dataframe of numeric data descriptive statistics

library("moments")

num_col <- Filter(is.numeric, customer)
desc_stats <- data.frame(
  min = apply(num_col, 2, min),
  median = apply(num_col, 2, median),
  mean_df = apply(num_col, 2, mean),
  SD = apply(num_col, 2, sd),
  max = apply(num_col, 2, max),
  skew = apply(num_col,2, skewness),
  Kurt = apply(num_col,2, kurtosis)
)
stats <- round(desc_stats,1)
stats
```

```
##                         min median mean_df     SD     max skew  Kurt
## Administrative            0    1.0     2.3    3.3    27.0  1.9   7.6
## Administrative_Duration  -1    9.0    81.7  177.5  3398.8  5.6  53.1
## Informational             0    0.0     0.5    1.3    24.0  4.0  29.6
## Informational_Duration   -1    0.0    34.8  141.5  2549.4  7.5  78.5
## ProductRelated            0   18.0    32.1   44.6   705.0  4.3  34.0
## ProductRelated_Duration  -1  609.5  1207.5 1919.9 63973.5  7.3 139.6
## BounceRates               0    0.0     0.0    0.0     0.2  3.2  12.3
## ExitRates                 0    0.0     0.0    0.0     0.2  2.2   7.6
## PageValues                0    0.0     6.0   18.7   361.8  6.3  67.9
## SpecialDay                0    0.0     0.1    0.2     1.0  3.3  12.8
## OperatingSystems          1    2.0     2.1    0.9     8.0  2.0  13.3
## Browser                   1    2.0     2.4    1.7    13.0  3.2  15.5
## Region                    1    3.0     3.2    2.4     9.0  1.0   2.8
## TrafficType               1    2.0     4.1    4.0    20.0  2.0   6.5
```

Showing distribution using Histogram

```
par(mfrow = c(2,2))
for (i in 1:13){
  hist(num_col[,i],main = names(num_col)[i], xlab = NULL)
}
```



Administrative

Administrative_Duration

Informational

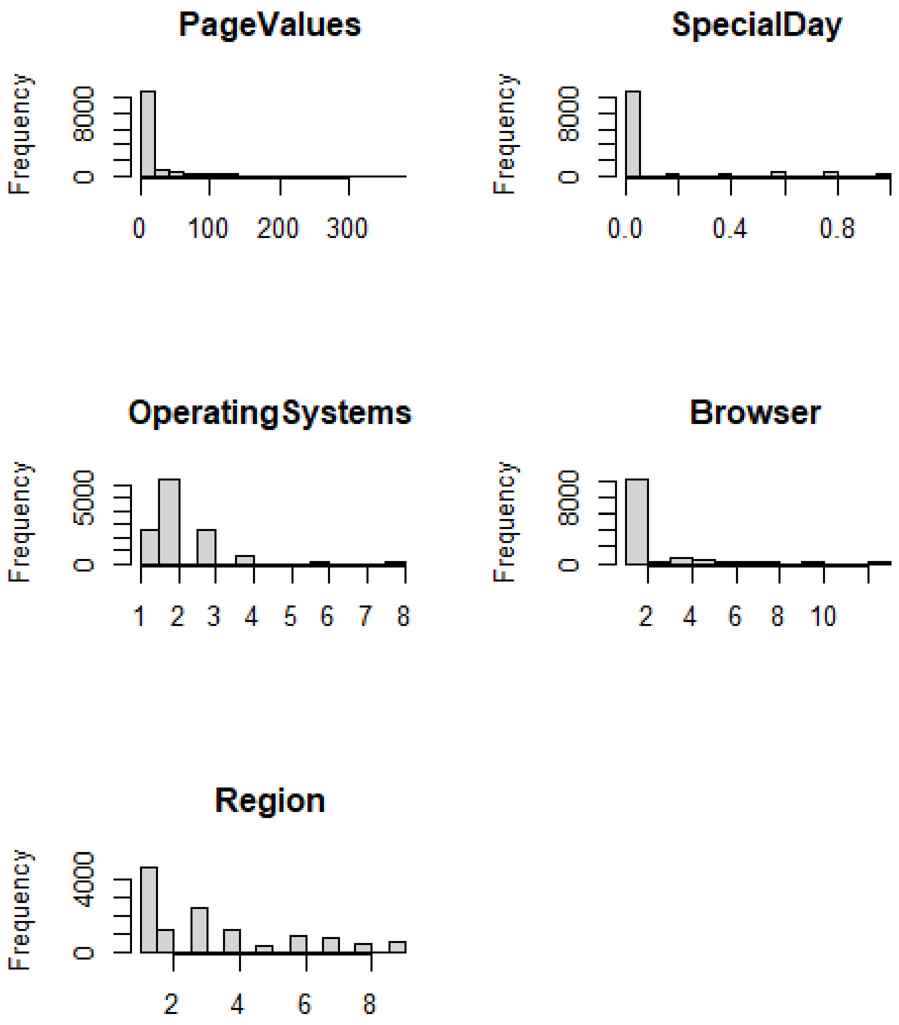
Informational_Duration

ProductRelated

ProductRelated_Duration

BounceRates

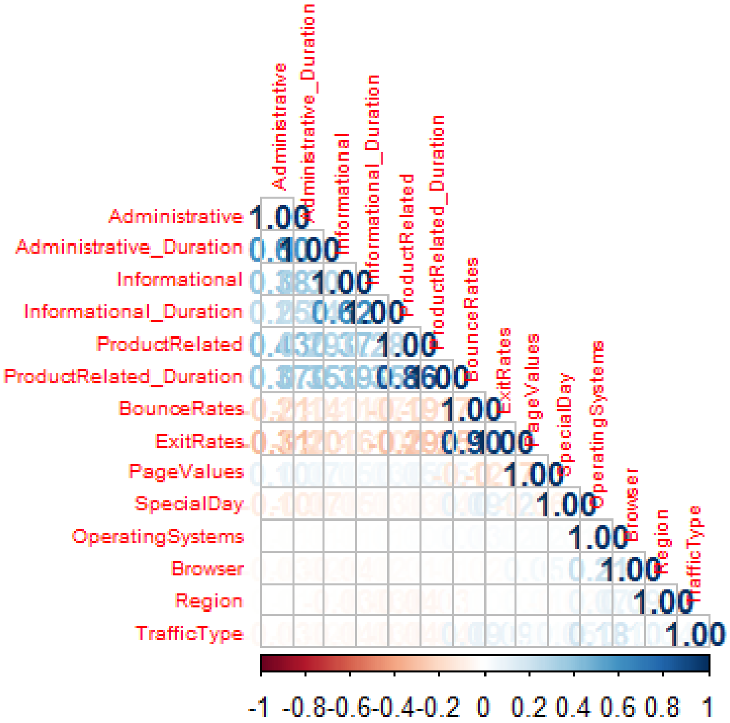
ExitRates



### Bivariate Analysis

```
#Correlation Matrix

library("corrplot")
```

```
## corrplot 0.92 loaded

corr_matrix <- cor(num_col)

corrplot(corr_matrix, method='number',type = 'lower',tl.cex = 0.6)
```



Multivariate Analysis

Dimensionality Reduction

```
# Scaling the dataset

customer_sc <- scale(num_col)

head(customer_sc)

##   Administrative Administrative_Duration Informational Informational_Duration
## 1     -0.7025315              -0.4601081    -0.3988128             -0.2462725
## 2     -0.7025315              -0.4601081    -0.3988128             -0.2462725
## 3     -0.7025315              -0.4657410    -0.3988128             -0.2533417
## 4     -0.7025315              -0.4601081    -0.3988128             -0.2462725
## 5     -0.7025315              -0.4601081    -0.3988128             -0.2462725
## 6     -0.7025315              -0.4601081    -0.3988128             -0.2462725
##   ProductRelated ProductRelated_Duration  BounceRates  ExitRates PageValues
## 1     -0.6963635              -0.6289343  3.954699721  3.4273070 -0.3190356
## 2     -0.6739424              -0.5955997 -0.450343788  1.2650121 -0.3190356
## 3     -0.6963635              -0.6294551  3.954699721  3.4273070 -0.3190356
## 4     -0.6739424              -0.6275453  0.650917089  2.1299300 -0.3190356
## 5     -0.4945739              -0.3020990 -0.009839437  0.1838646 -0.3190356
## 6     -0.2927843              -0.5486101 -0.102577188 -0.3661929 -0.3190356
##   SpecialDay OperatingSystems    Browser     Region TrafficType
## 1 -0.3103105       -1.2396607 -0.7939682 -0.8962939 -0.76562243
## 2 -0.3103105       -0.1371074 -0.2093703 -0.8962939 -0.51660683
## 3 -0.3103105        2.0679992 -0.7939682  2.4336556 -0.26759123
## 4 -0.3103105        0.9654459 -0.2093703 -0.4800502 -0.01857564
## 5 -0.3103105        0.9654459  0.3752276 -0.8962939 -0.01857564
## 6 -0.3103105       -0.1371074 -0.2093703 -0.8962939 -0.26759123
```

```
summary(customer_sc)
```

```
##  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
##  OperatingSystems     Browser            Region          TrafficType
##  Min.   :-1.2397   Min.   :-0.7940   Min.   :-0.89629   Min.   :-0.76562
##  1st Qu.:-0.1371   1st Qu.:-0.2094   1st Qu.:-0.89629   1st Qu.:-0.51661
##  Median :-0.1371   Median :-0.2094   Median :-0.06381   Median :-0.51661
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.9654   3rd Qu.:-0.2094   3rd Qu.: 0.35244   3rd Qu.:-0.01858
##  Max.   : 6.4782   Max.   : 6.2212   Max.   : 2.43366   Max.   : 3.96567
```

Applying the Principle Compenent Analysis

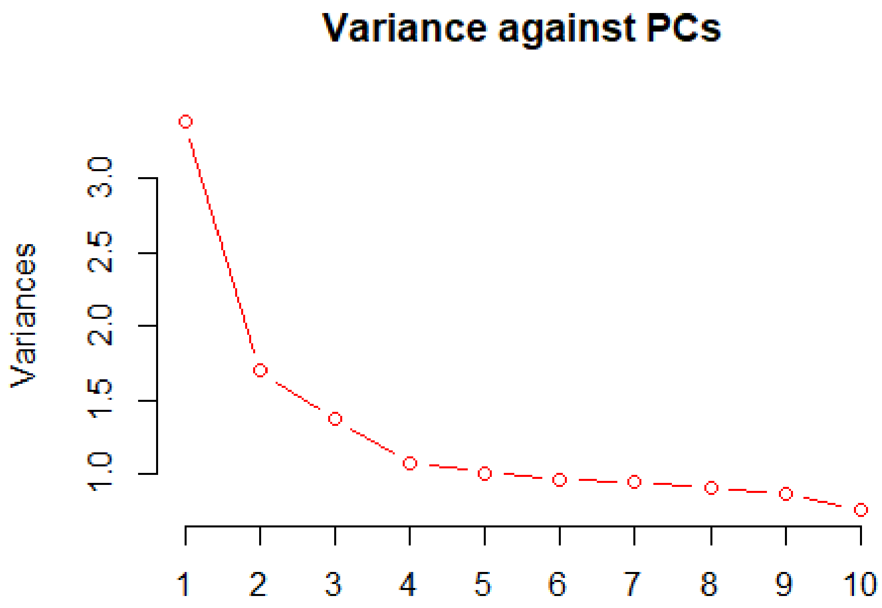
```
customer_sc.pca <- prcomp(customer_sc, center = TRUE, scale = TRUE)
print(customer_sc.pca)
```

```
## Standard deviations (1, .., p=14):
##  [1] 1.8401010 1.3030684 1.1743779 1.0377092 1.0059577 0.9856398 0.9736821
##  [8] 0.9576303 0.9298795 0.8729808 0.6502062 0.5935555 0.3519072 0.2929192
## 
## Rotation (n x k) = (14 x 14):
##                                  PC1         PC2          PC3         PC4
## Administrative           0.38174831 -0.05389571  0.034330189 -0.25483540
## Administrative_Duration  0.32880068 -0.10688051  0.040028236 -0.32113386
## Informational            0.34868758 -0.27428680 -0.031715069 -0.17278982
## Informational_Duration   0.29716046 -0.29468954 -0.030178618 -0.16620112
## ProductRelated           0.41138032 -0.15246032  0.031450889  0.40153735
## ProductRelated_Duration  0.41341349 -0.19218143  0.034406884  0.36600557
## BounceRates             -0.27252341 -0.60563878 -0.006891667 -0.12543730
## ExitRates               -0.32133883 -0.57439219 -0.020420692 -0.09487117
## PageValues               0.09128055  0.18120380  0.144494992 -0.29646434
## SpecialDay              -0.07744055 -0.13106791  0.029171630  0.55300044
## OperatingSystems        -0.01521708 -0.03823080  0.598590850  0.06704353
## Browser                 -0.01886564  0.03946340  0.551687097  0.02761936
## Region                  -0.02413005  0.04449186  0.299485640 -0.23034894
## TrafficType             -0.05567707 -0.10808315  0.467518982  0.05006552
##                                 PC5          PC6         PC7          PC8
## Administrative          -0.33548530 -0.093624614 -0.27514185  0.010881536
## Administrative_Duration -0.39539550 -0.118330393 -0.35730724  0.021429673
## Informational            0.46401823  0.010281210 -0.05953477  0.008275883
## Informational_Duration   0.59385784  0.026992696 -0.03482989  0.009724208
## ProductRelated          -0.21525638  0.103215220  0.28736923 -0.045568845
## ProductRelated_Duration -0.18237976  0.108437356  0.28724283 -0.039040092
## BounceRates             -0.18586853 -0.020254333  0.14957627 -0.082412434
## ExitRates               -0.12945291  0.002411283  0.11432435 -0.048848648
## PageValues               0.02282527 -0.434564511  0.43439273 -0.678372112
## SpecialDay               0.13021229 -0.124488654 -0.52617878 -0.522649405
## OperatingSystems         0.06115479 -0.081747082  0.02277972  0.263466751
## Browser                  0.08773987  0.103576367  0.19544260  0.101423808
## Region                  -0.05600937  0.773046363 -0.14673946 -0.413293138
## TrafficType             -0.01332820 -0.366353490 -0.25598440  0.051154709
##                                  PC9          PC10         PC11          PC12
## Administrative          -0.148423551 -0.0216837717 -0.581039341  0.4459814969
## Administrative_Duration -0.209230312  0.0090995860  0.564320475 -0.3050229330
## Informational            0.010189907 -0.0081230047 -0.391745425 -0.6315161654
## Informational_Duration   0.013675715  0.0270082421  0.360362305  0.5519809657
## ProductRelated           0.117233445  0.0105090305 -0.106591057  0.0340588806
## ProductRelated_Duration  0.114891620  0.0278295208  0.204962782 -0.0441297695
## BounceRates             -0.044816062 -0.0237803024 -0.069816899  0.0356549641
## ExitRates               -0.055931553 -0.0006497547 -0.005463743 -0.0029626397
## PageValues              -0.005223884 -0.0977300950  0.023122365 -0.0077459387
## SpecialDay              -0.286049152 -0.0720185436 -0.014605927  0.0037694621
## OperatingSystems        -0.061985699 -0.7422159237  0.018248164  0.0020656298
## Browser                 -0.585019019  0.5314668040 -0.033401220 -0.0094935347
## Region                   0.244421092 -0.0423243224  0.004150026 -0.0031163072
## TrafficType              0.643536423  0.3838360389 -0.010987911  0.0002231436
##                                 PC13         PC14
## Administrative           0.167736543 -0.031063530
## Administrative_Duration -0.145890070 -0.025088993
## Informational            0.028725269  0.004237148
## Informational_Duration  -0.077827901 -0.009956400
## ProductRelated          -0.667734985 -0.177224718
## ProductRelated_Duration  0.672816489  0.131697721
## BounceRates           -0.151391960  0.668871622
## ExitRates              0.148368300 -0.707104492
## PageValues             0.006174431 -0.039985387
## SpecialDay             0.010426029  0.018370927
## OperatingSystems       0.004091795 -0.008009905
## Browser               -0.005462687  0.010699285
## Region                -0.003988623 -0.005353012
## TrafficType           -0.002044921 -0.002450879
```

```
summary(customer_sc.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.8401 1.3031 1.17438 1.03771 1.00596 0.98564 0.97368
## Proportion of Variance 0.2419 0.1213 0.09851 0.07692 0.07228 0.06939 0.06772
## Cumulative Proportion  0.2419 0.3631 0.46165 0.53857 0.61085 0.68024 0.74796
##                           PC8     PC9    PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.9576 0.92988 0.87298 0.6502 0.59356 0.35191 0.29292
## Proportion of Variance 0.0655 0.06176 0.05444 0.0302 0.02516 0.00885 0.00613
## Cumulative Proportion  0.8135 0.87523 0.92966 0.9599 0.98503 0.99387 1.00000
```

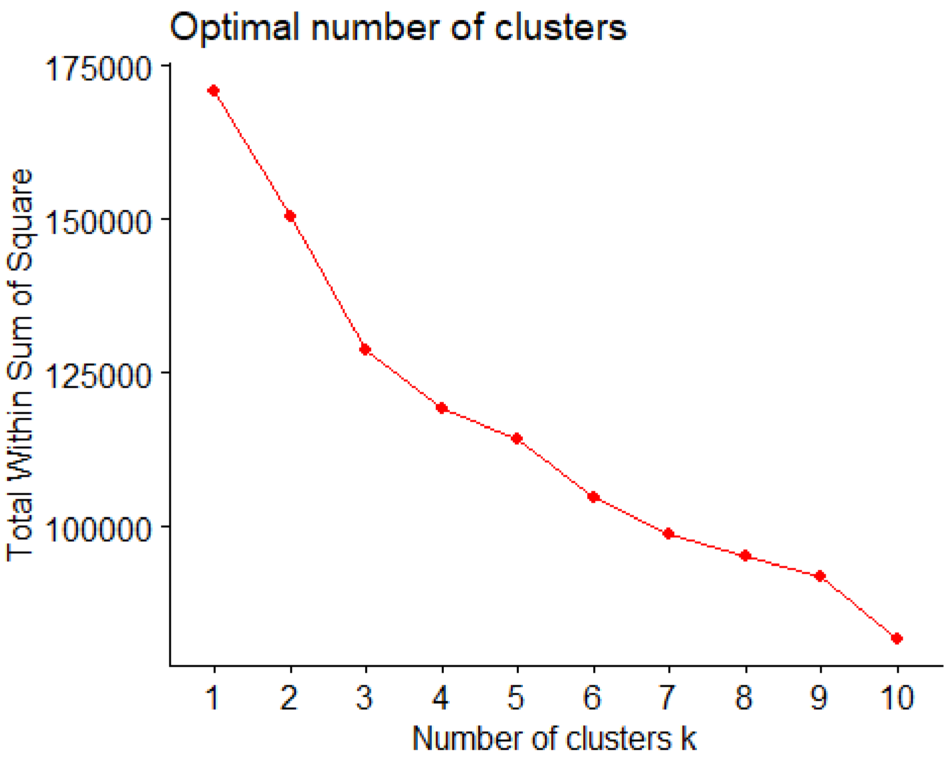
```
# Plotting Variance vs PCs

plot(customer_sc.pca, type = "l" ,col = "red",main = "Variance against PCs")
```



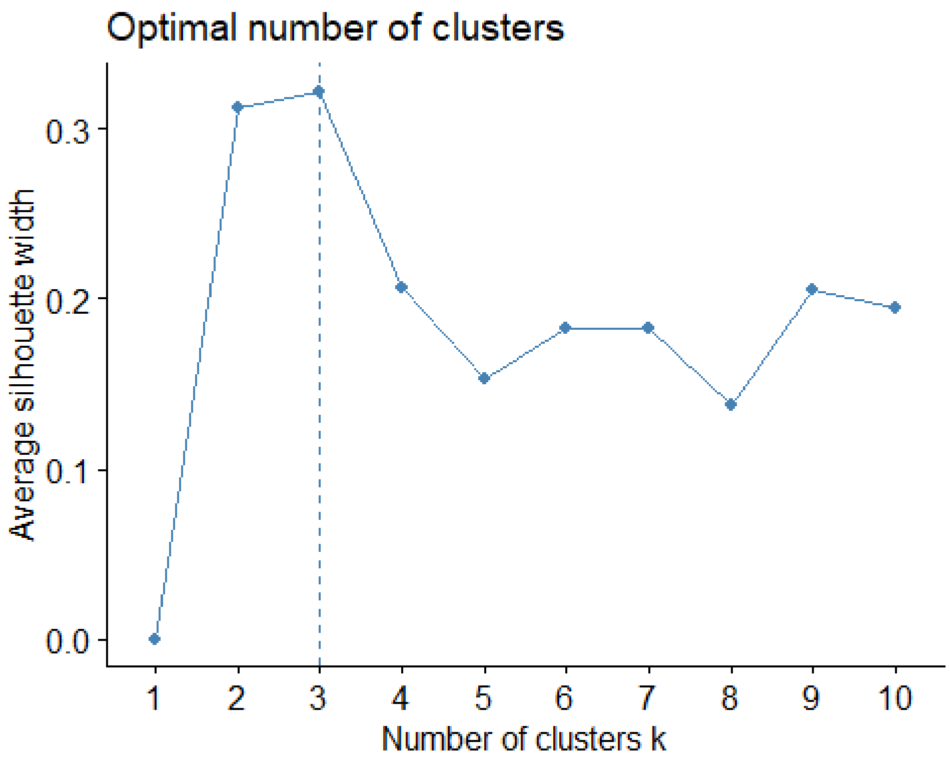
6. Modeling

K-Means Clustering

```
# Determining Optimal clusters (k) Using Elbow method

library("factoextra")

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

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

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

fviz_nbclust(x = customer_sc,FUNcluster = kmeans, method = 'wss',linecolor = "red")
```

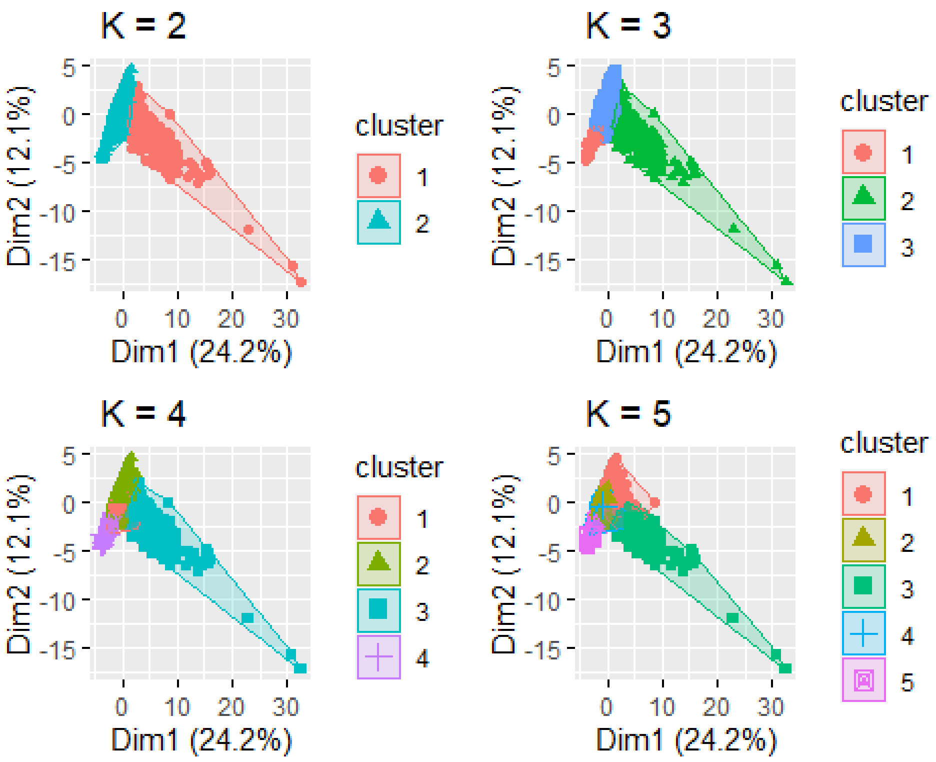


```
# Determining the Optimal clusters (k) Using the average silhouette method

fviz_nbclust(x = customer_sc,FUNcluster = kmeans, method = 'silhouette' )
```



From the above we clearly see the best value for k is 3

```
# Clustering using the K-means

#We will test out a few values of k
library("gridExtra")

##
## Attaching package: 'gridExtra'

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

K2 <- kmeans(customer_sc, centers = 2, nstart = 50)
K3 <- kmeans(customer_sc, centers = 3, nstart = 50)
K4 <- kmeans(customer_sc, centers = 4, nstart = 50)
K5 <- kmeans(customer_sc, centers = 5, nstart = 50)

#plot these clusters for different K value to compare.

p1 <- fviz_cluster(K2, geom = "point", data = customer_sc) + ggtitle(" K = 2")
p2 <- fviz_cluster(K3, geom = "point", data = customer_sc) + ggtitle(" K = 3")
p3 <- fviz_cluster(K4, geom = "point", data = customer_sc) + ggtitle(" K = 4")
p4 <- fviz_cluster(K5, geom = "point", data = customer_sc) + ggtitle(" K = 5")

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

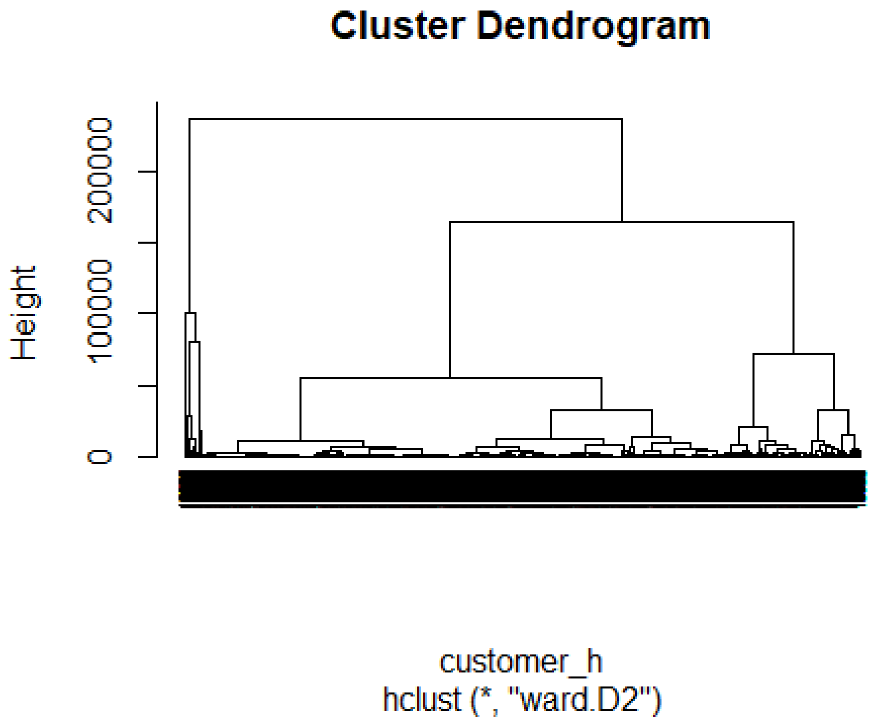


```
set.seed(100)

# Calculating the Kmeans clusters
data_kmeans <- kmeans(num_col, centers = 2, nstart = 25)
summary(data_kmeans)

##              Length Class  Mode
## cluster      12199  -none- numeric
## centers         28  -none- numeric
## totss            1  -none- numeric
## withinss         2  -none- numeric
## tot.withinss     1  -none- numeric
## betweenss        1  -none- numeric
## size             2  -none- numeric
## iter             1  -none- numeric
## ifault           1  -none- numeric

#Comparing the revenue and the model cluster to see if the clusters match

# Adding the clusters as a column to our original dataset

calculated <- customer %>%
  mutate(cluster = data_kmeans$cluster) %>%
  select(Revenue, cluster)

calculated$cluster[calculated$cluster == 1] <- 'FALSE'
```

```
table(calculated$cluster == calculated$Revenue)

##
## FALSE  TRUE
##  2556  9643
```

Hierarchical Clustering

```
# First we use the dist() function to compute the Euclidean distance between observations,
# ---
#

customer_h <- suppressWarnings(dist(customer, method = "euclidean"))

# We then hierarchical clustering using the Ward's method
# ---
#

customer_hc <- hclust(customer_h, method = "ward.D2" )

# We plot the obtained dendrogram


plot(customer_hc, cex = 0.5, hang = -1)
```



8. Conclusion

Region 1 accounted for the majority of traffic and revenue. More regions visit the site over the holidays, contributing significantly to total revenue. The holiday of Mother’s Day generated more income than Valentine’s Day. The majority of the Wednesday visits occurred in May, yet November generated more money than May. Most visitors were attracted by traffic type 2. For the whole 10-month period under consideration, some traffic kinds did not bring in any visitors. When evaluating advertising, it should be deleted or re-evaluated to determine the problem. Return visitors accounted for the majority of revenue and visits. This is an excellent indication of consumer satisfaction.

9. Recommendation

We recommend more supply of the product during the holidays to cover the demand of the product.

10. Follow up Questions

a) Did we have the right data?

Yes, the dataset available for this analysis was relevant to the research problem.

b) Do we need other data to answer the research question?

Yes, to improve the accuracy, we need more data to add more relevant information for the research question.