Kira Plastinina Customer Insights Project

Kennedy Muriuki

11/09/2020

Defining the question

In this weeks problem, I will be working as a data scientist for a client, Kira Plastinina. Kira Plastinina is a Russian brand that is sold throughout retail stores in Russia, Ukraine, Kazakhstan, Belarus, China, Philippines, and Armenia. It's marketing team would like to better understand their customers behaviors and therefore requested me to draw insights on the characteristics of various customer groups.

Defining the metric of success

The metric of success is to obtain several distinct clusters of the Kira Plastina's customers and their characteristics.

Data Sourcing

The dataset was sourced from the company's database of existing customers. The dataset was availed by the brand's sales and marketing team.

The Experimental Design

For me to be able to obtain the required results, the following steps will be undertaken

1. Problem definition
2. Data sourcing
3. Checking the data
4. Performing data cleaning
5. Perform Exploratory data analysis
6. Implementing the solution
7. Challenging the solution
8. Follow up questions

Checking the data

Loading the dataset

```
# the dataset has been downloaded to a local repository and will be loaded as a csv file
data <- read.csv(file.choose())
head(data)
```

```
##   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 bottom of the dataset
tail(data)
```

```
##       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
```

```
# displaying the structure of the dataset
str(data)
```

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

```
# displaying the dimension of the dataset
dim(data)
```

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

the data contains 12,330 entries and 18 columns

```
# checking the names of the columns and their datatypes in the dataset

columns = colnames(data)
for (column in seq(length(colnames(data)))){
  print(columns[column])
  print(class(data[, column]))
  cat('\n')
}
```

```
## [1] "Administrative"
## [1] "integer"
##
## [1] "Administrative_Duration"
## [1] "numeric"
##
```

```
## [1] "Informational"
## [1] "integer"
##
## [1] "Informational_Duration"
## [1] "numeric"
##
## [1] "ProductRelated"
## [1] "integer"
##
## [1] "ProductRelated_Duration"
## [1] "numeric"
##
## [1] "BounceRates"
## [1] "numeric"
##
## [1] "ExitRates"
## [1] "numeric"
##
## [1] "PageValues"
## [1] "numeric"
##
## [1] "SpecialDay"
## [1] "numeric"
##
## [1] "Month"
## [1] "character"
##
## [1] "OperatingSystems"
## [1] "integer"
##
## [1] "Browser"
## [1] "integer"
##
## [1] "Region"
## [1] "integer"
##
## [1] "TrafficType"
## [1] "integer"
##
## [1] "VisitorType"
## [1] "character"
##
## [1] "Weekend"
## [1] "logical"
##
## [1] "Revenue"
## [1] "logical"
```

Checking for missing values

```
# checking if the dataset contains any missing values
any(is.na(data))
```

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

```
# checking the columns with missing data
colSums(is.na(data))
```

```
##          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
```

```
# since the missing data is not a lot I will skip the missing data
df <- na.omit(data)

# cheking the dimension of the new dataset
dim(df)
```

```
## [1] 12316    18
```

checking for duplicates

```
# checking for duplicated data in the dataset
any(duplicated(df))
```

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

```
# identifying the duplicated data
dup <- df[duplicated(df),]
dup
```

```
##       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
```

data showed above as duplicated did not look like duplicated data. They had a lot of similar entries in some columns but did not have entirely similar column entries. Therefore I will not remove the 177 rows as this might cause inconsistencies within the data set and affect final results

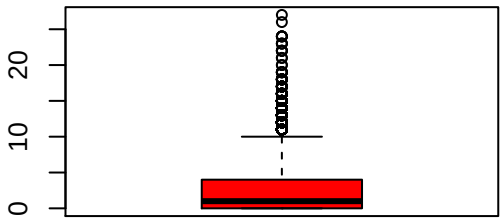
checking for outliers

```
# obtaining the numerical columns
numerical = df[,c(1:10)]
head(numerical)
```

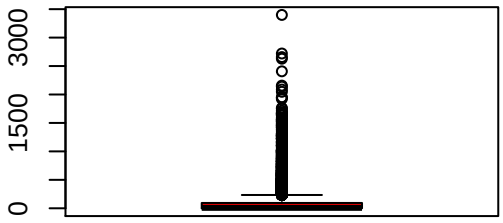
```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                      -1             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
## 1          0
## 2          0
## 3          0
## 4          0
## 5          0
## 6          0
```

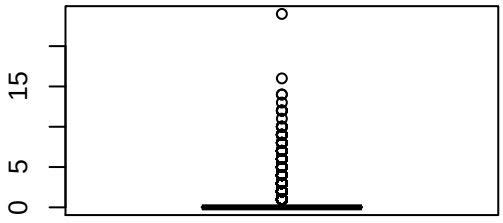
```r
# generating boxplots for the numerical columns
par(mfrow=c(2,2), mar=c(5,4,2,2))

for (i in names(numerical)){
  x <- (numerical)[,i]
  boxplot(x, xlab= i, col="red")
  boxplot.stats(x)$out
}
```



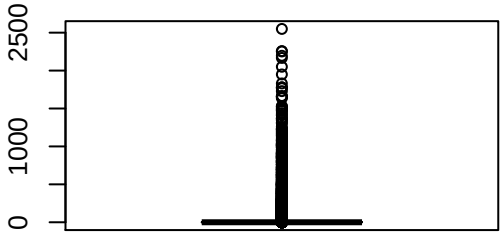
Administrative



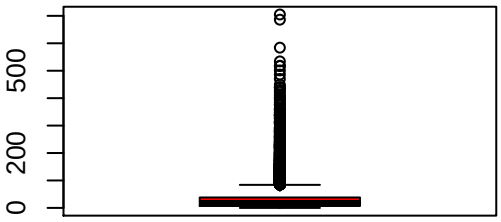
Administrative_Duration



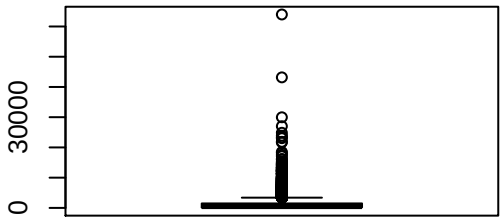
Informational



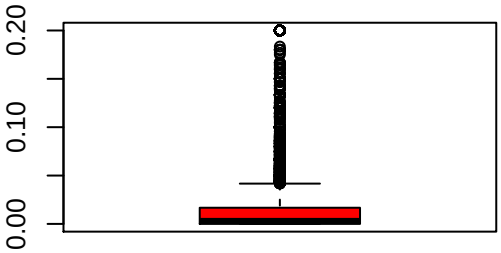
Informational_Duration



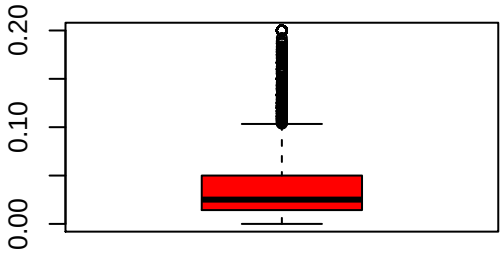
ProductRelated



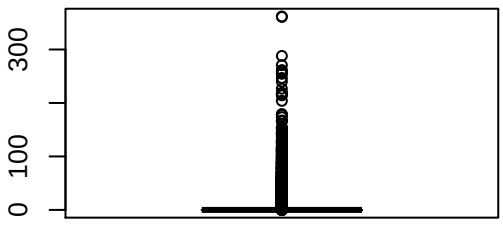
ProductRelated_Duration



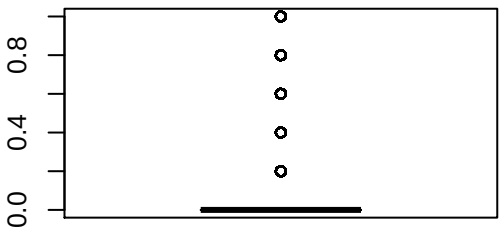
BounceRates



ExitRates



PageValues



SpecialDay

```
# dealing with the outliers
```

```
# checking the summary 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
##  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
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.0000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.0000
##  Median :0.003119   Median :0.02512   Median :  0.000   Median :0.0000
##  Mean   :0.022152   Mean   :0.04300   Mean   :  5.896   Mean   :0.0615
##  3rd Qu.:0.016684   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.0000
```

```
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.0000
##     Month           OperatingSystems    Browser           Region
##  Length:12316       Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mode  :character   Median :2.000    Median : 2.000   Median :3.000
##                     Mean   :2.124    Mean   : 2.358   Mean   :3.148
##                     3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##                     Max.   :8.000    Max.   :13.000   Max.   :9.000
##   TrafficType    VisitorType         Weekend         Revenue
##  Min.   : 1.00   Length:12316       Mode :logical   Mode :logical
##  1st Qu.: 2.00   Class :character   FALSE:9451      FALSE:10408
##  Median : 2.00   Mode  :character   TRUE :2865      TRUE :1908
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
```

Univariate Exploratory Data Analysis

Measures of central tendancies

```
# obtaining the mean of the numerical columns

for (i in names(numerical)){
  x <- numerical[,i]
  mean <- mean(x)
  print(paste("The mean ", i , "is" ,  mean))
  cat('\n')
}
```

```
## [1] "The mean  Administrative is 2.31779798635921"
##
## [1] "The mean  Administrative_Duration is 80.9061763519009"
##
## [1] "The mean  Informational is 0.503978564468983"
##
## [1] "The mean  Informational_Duration is 34.5063873375142"
##
## [1] "The mean  ProductRelated is 31.7638843780448"
##
## [1] "The mean  ProductRelated_Duration is 1196.03705685414"
##
## [1] "The mean  BounceRates is 0.0221524619360182"
##
## [1] "The mean  ExitRates is 0.0430025384157194"
##
## [1] "The mean  PageValues is 5.89595237471679"
##
## [1] "The mean  SpecialDay is 0.0614972393634297"
```

```
# obtaining the median of the numerical columns
```

```r
for (i in names(numerical)){
  x <- numerical[,i]
  mean <- median(x)
  print(paste("The mean ", i , "is" ,  mean))
  cat('\n')
}
```

```
## [1] "The mean  Administrative is 1"
##
## [1] "The mean  Administrative_Duration is 8"
##
## [1] "The mean  Informational is 0"
##
## [1] "The mean  Informational_Duration is 0"
##
## [1] "The mean  ProductRelated is 18"
##
## [1] "The mean  ProductRelated_Duration is 599.76619045"
##
## [1] "The mean  BounceRates is 0.003119412"
##
## [1] "The mean  ExitRates is 0.025124489"
##
## [1] "The mean  PageValues is 0"
##
## [1] "The mean  SpecialDay is 0"
```

```r
# displaying the mode of the numerical columns

getmode <- function(a){
  uniqv <- unique(a)
  uniqv[which.max(tabulate(match(a,uniqv)))]
}

# looping through the columns to get the mode

for (i in names(numerical)){
  x <- numerical[,i]
  mode <- getmode(x)
  print(paste("column", i , ":" ,  mode))
  cat('\n')
}
```

```
## [1] "column Administrative : 0"
##
## [1] "column Administrative_Duration : 0"
##
## [1] "column Informational : 0"
##
## [1] "column Informational_Duration : 0"
##
## [1] "column ProductRelated : 1"
##
```

```
## [1] "column ProductRelated_Duration : 0"
##
## [1] "column BounceRates : 0"
##
## [1] "column ExitRates : 0.2"
##
## [1] "column PageValues : 0"
##
## [1] "column SpecialDay : 0"
```

Measures of dispersion

```
# obtaining the five number summary of the numerical columns

for (i in names(numerical)){
  x <- numerical[,i]
  quantile <- quantile(x)
  print(paste(i))
  cat('\n')
  print(quantile)
  cat('\n')
}
```

```
## [1] "Administrative"
##
##   0%  25%  50%  75% 100%
##    0    0    1    4   27
##
## [1] "Administrative_Duration"
##
##      0%     25%     50%     75%    100%
##   -1.00    0.00    8.00   93.50 3398.75
##
## [1] "Informational"
##
##   0%  25%  50%  75% 100%
##    0    0    0    0   24
##
## [1] "Informational_Duration"
##
##       0%      25%      50%      75%     100%
##   -1.000    0.000    0.000    0.000 2549.375
##
## [1] "ProductRelated"
##
##   0%  25%  50%  75% 100%
##    0    7   18   38  705
##
## [1] "ProductRelated_Duration"
##
##         0%        25%        50%        75%       100%
##    -1.0000   185.0000   599.7662  1466.4799 63973.5222
```

```
##
## [1] "BounceRates"
##
##          0%         25%         50%         75%        100%
## 0.000000000 0.000000000 0.003119412 0.016683674 0.200000000
##
## [1] "ExitRates"
##
##         0%        25%        50%        75%       100%
## 0.00000000 0.01428571 0.02512449 0.05000000 0.20000000
##
## [1] "PageValues"
##
##       0%      25%      50%      75%     100%
##   0.0000   0.0000   0.0000   0.0000 361.7637
##
## [1] "SpecialDay"
##
##   0%  25%  50%  75% 100%
##    0    0    0    0    1
```

```r
# showing the variances and standard deviation of the munerical columns
for (i in names(numerical)){
  x <- numerical[,i]
  Sdev <- sd(x)
  var <- var(x)
  print(paste(i))
  cat('\n')
  print(paste("Variance :", round(var, digits = 2), "Standard deviation :", round(Sdev, digits = 2)))
  cat('\n')
}
```

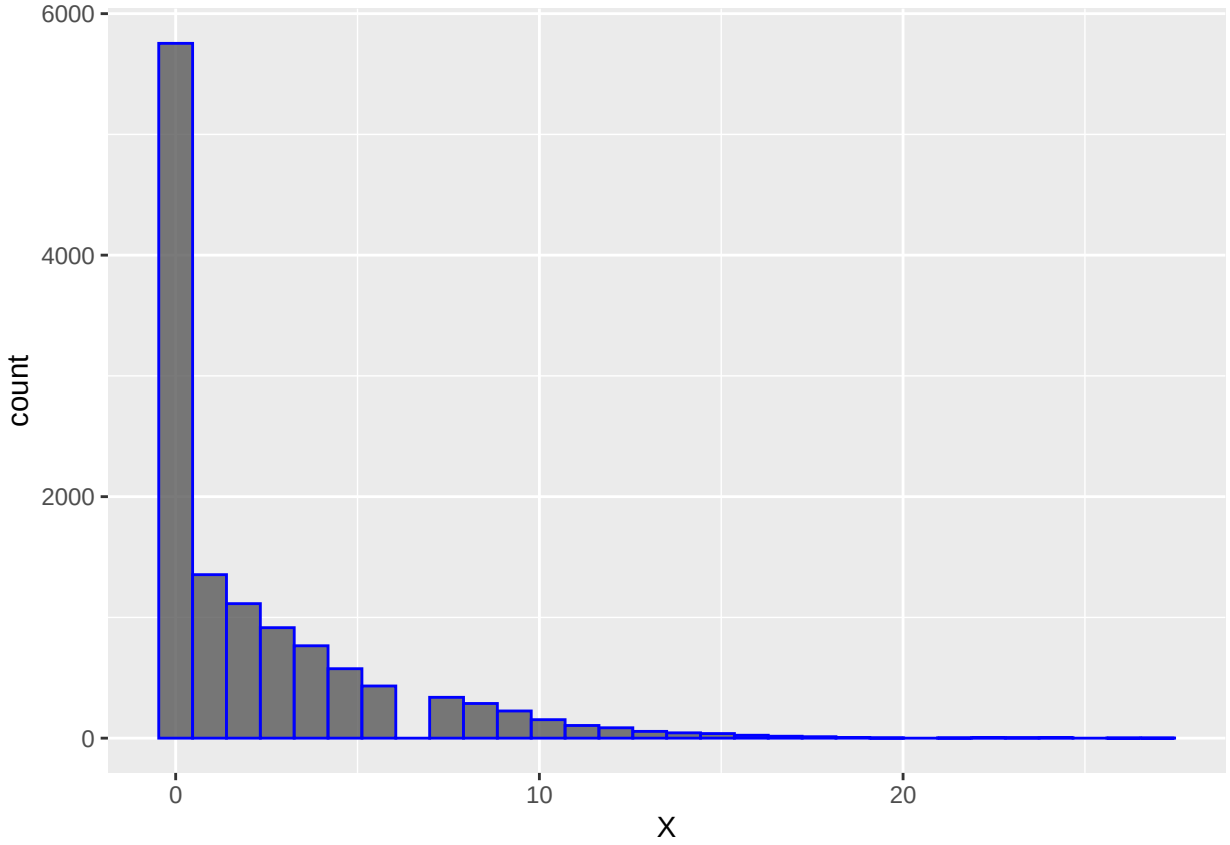
```
## [1] "Administrative"
##
## [1] "Variance : 11.04 Standard deviation : 3.32"
##
## [1] "Administrative_Duration"
##
## [1] "Variance : 31279.61 Standard deviation : 176.86"
##
## [1] "Informational"
##
## [1] "Variance : 1.61 Standard deviation : 1.27"
##
## [1] "Informational_Duration"
##
## [1] "Variance : 19831.82 Standard deviation : 140.83"
##
## [1] "ProductRelated"
##
## [1] "Variance : 1979.39 Standard deviation : 44.49"
##
## [1] "ProductRelated_Duration"
##
```

```
## [1] "Variance : 3664822.11 Standard deviation : 1914.37"
##
## [1] "BounceRates"
##
## [1] "Variance : 0 Standard deviation : 0.05"
##
## [1] "ExitRates"
##
## [1] "Variance : 0 Standard deviation : 0.05"
##
## [1] "PageValues"
##
## [1] "Variance : 345.14 Standard deviation : 18.58"
##
## [1] "SpecialDay"
##
## [1] "Variance : 0.04 Standard deviation : 0.2"
```

Histograms

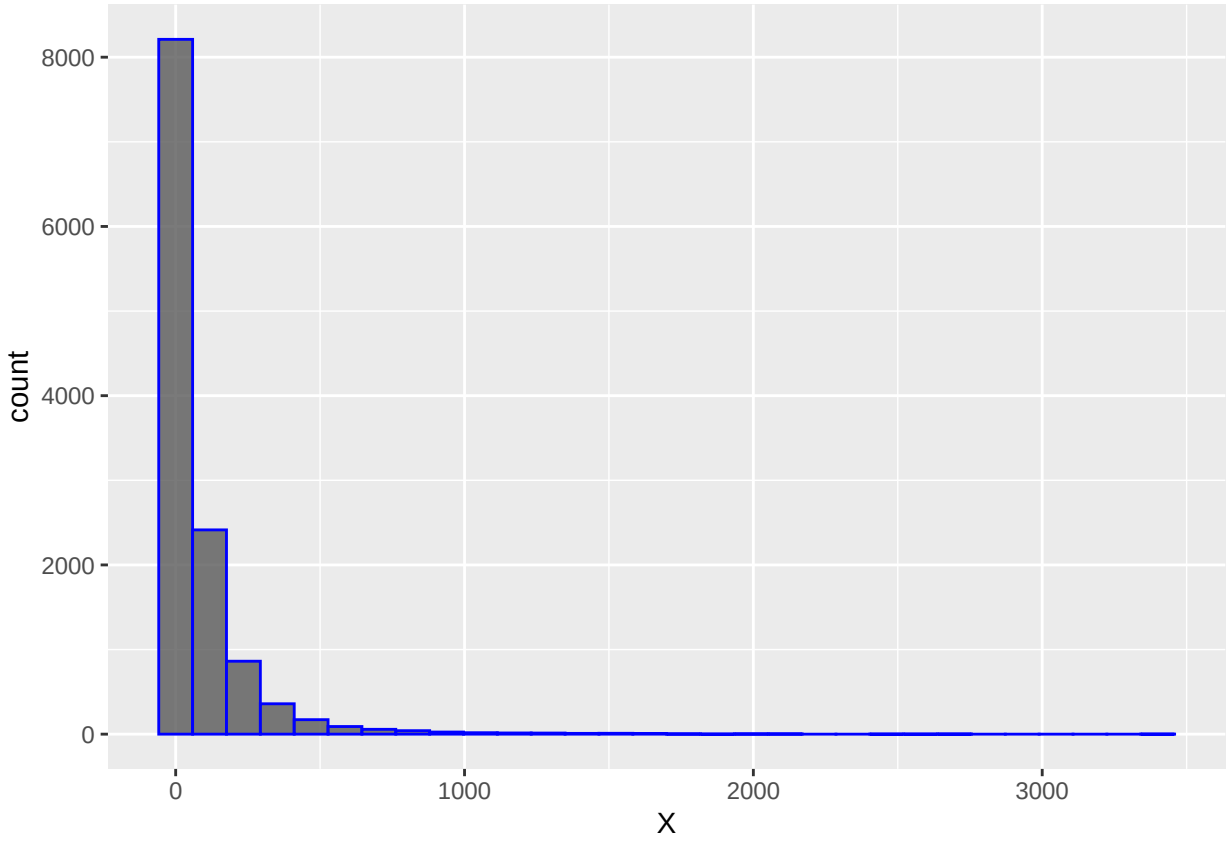
```
# creating a function to generate histogram plots for the numerical columns
library(ggplot2)
dens_fun = function(X) {
     ggplot(numerical, aes(x= X)) +
      geom_histogram(color="blue", alpha=0.8)
}

# generating histogram for Administrative
dens_fun(df$Administrative)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



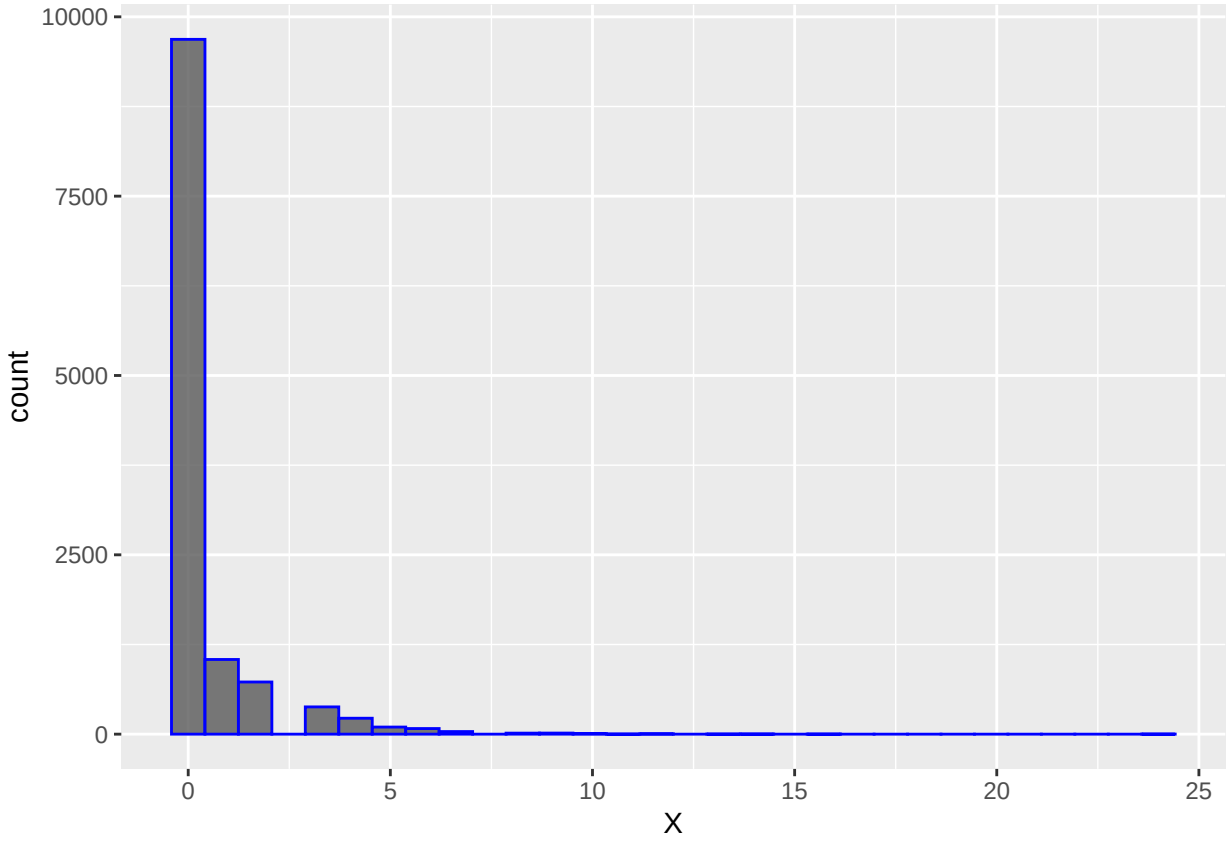

```
dens_fun(numerical$Administrative_Duration)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



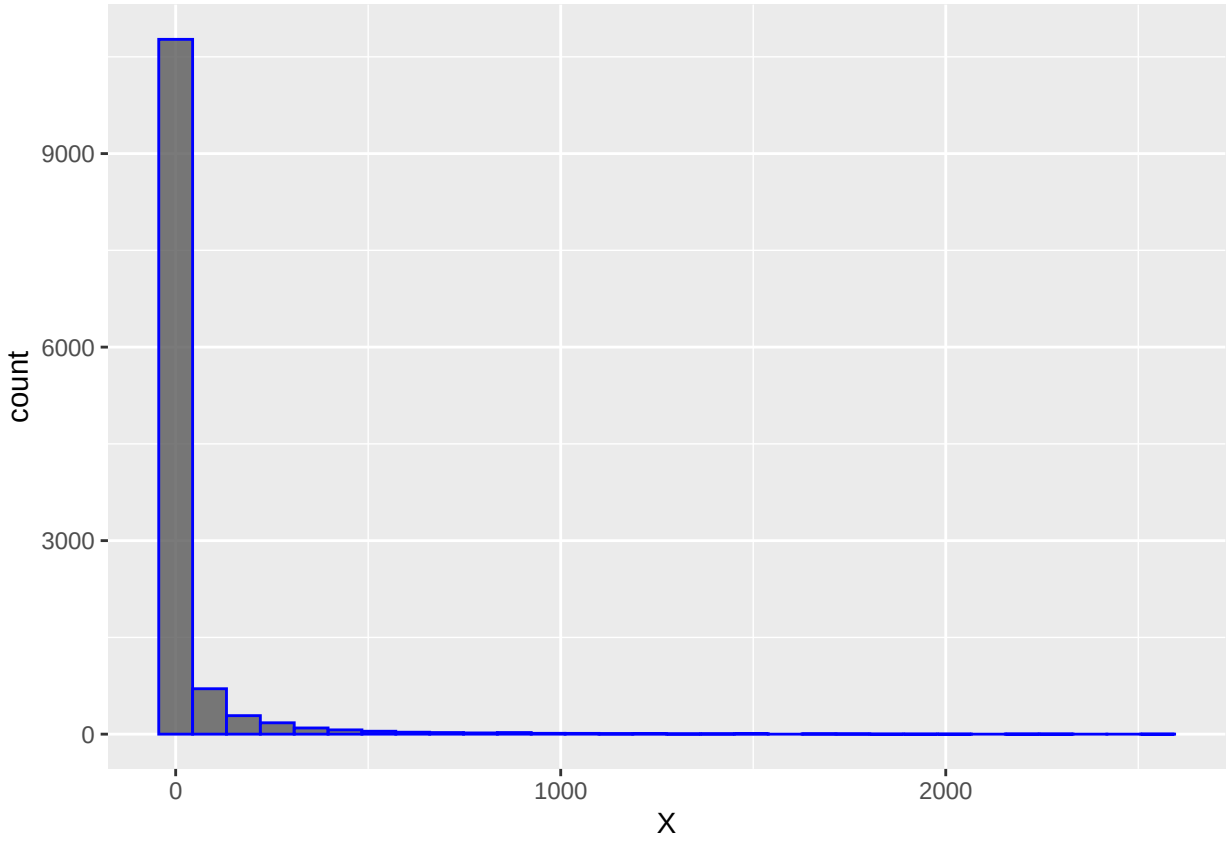

```
dens_fun(numerical$Informational)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



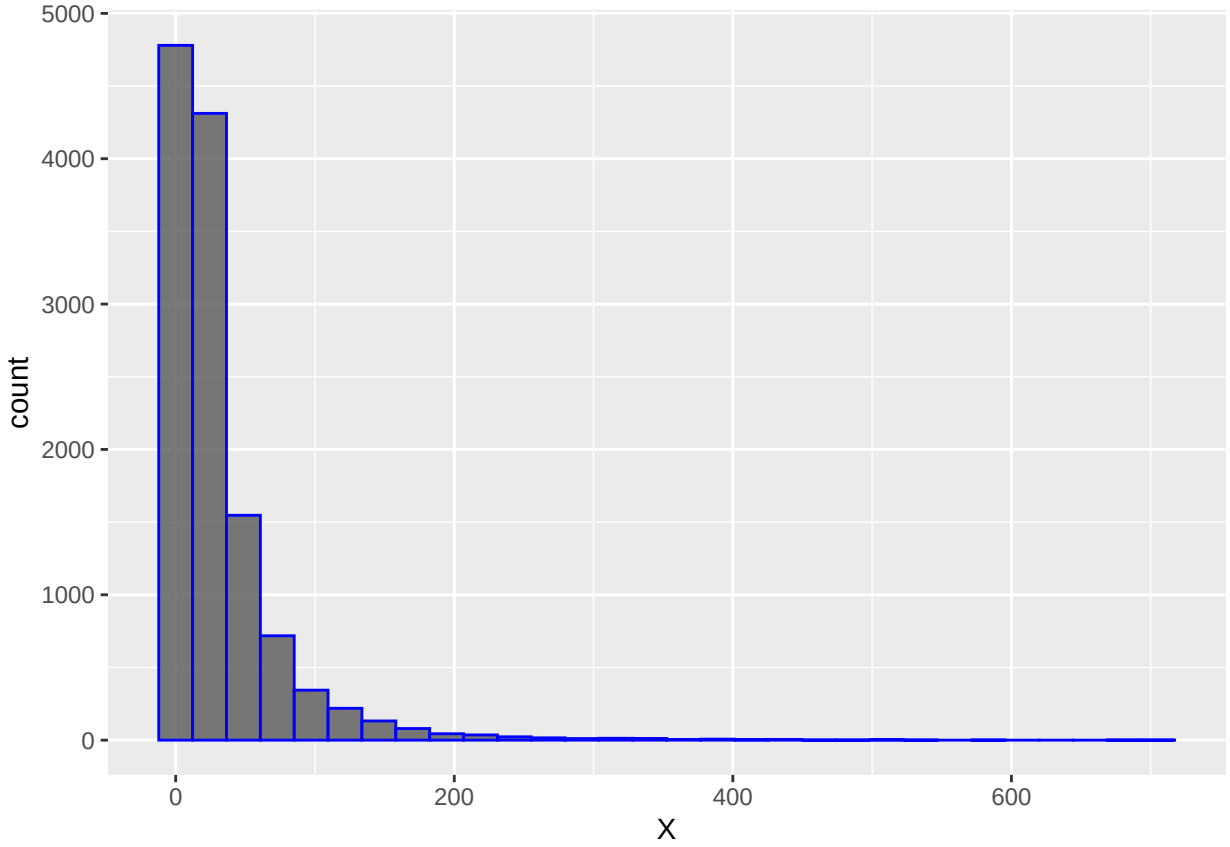

```
dens_fun(numerical$Informational_Duration)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



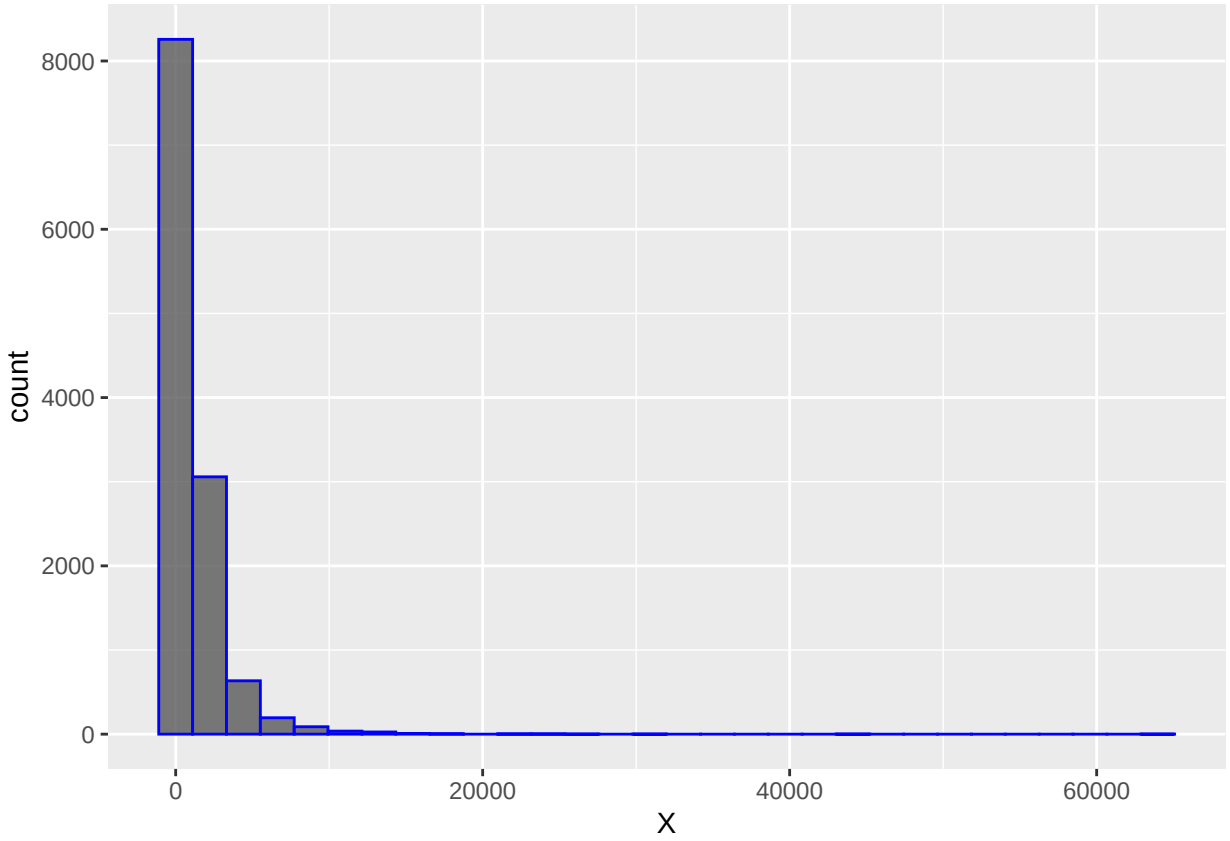

```
dens_fun(numerical$ProductRelated )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



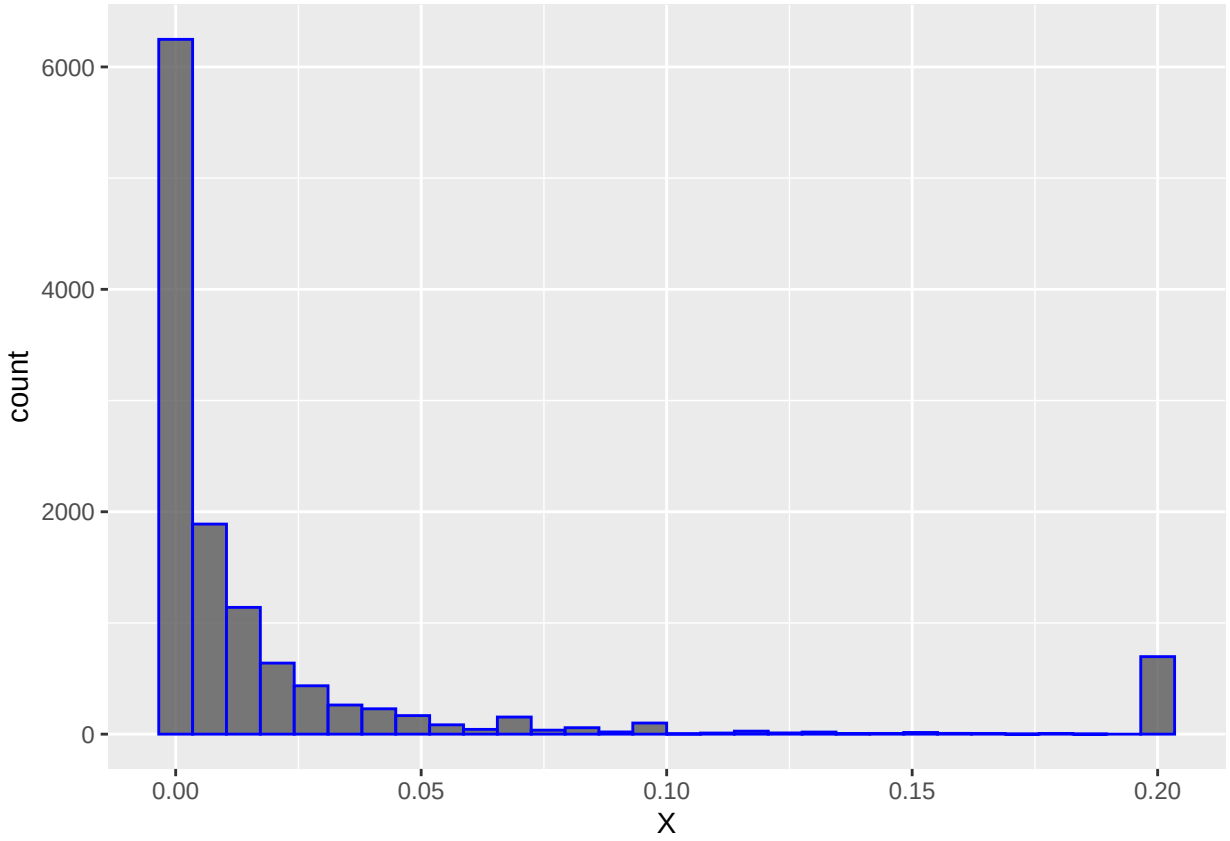

```
dens_fun(numerical$ProductRelated_Duration)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



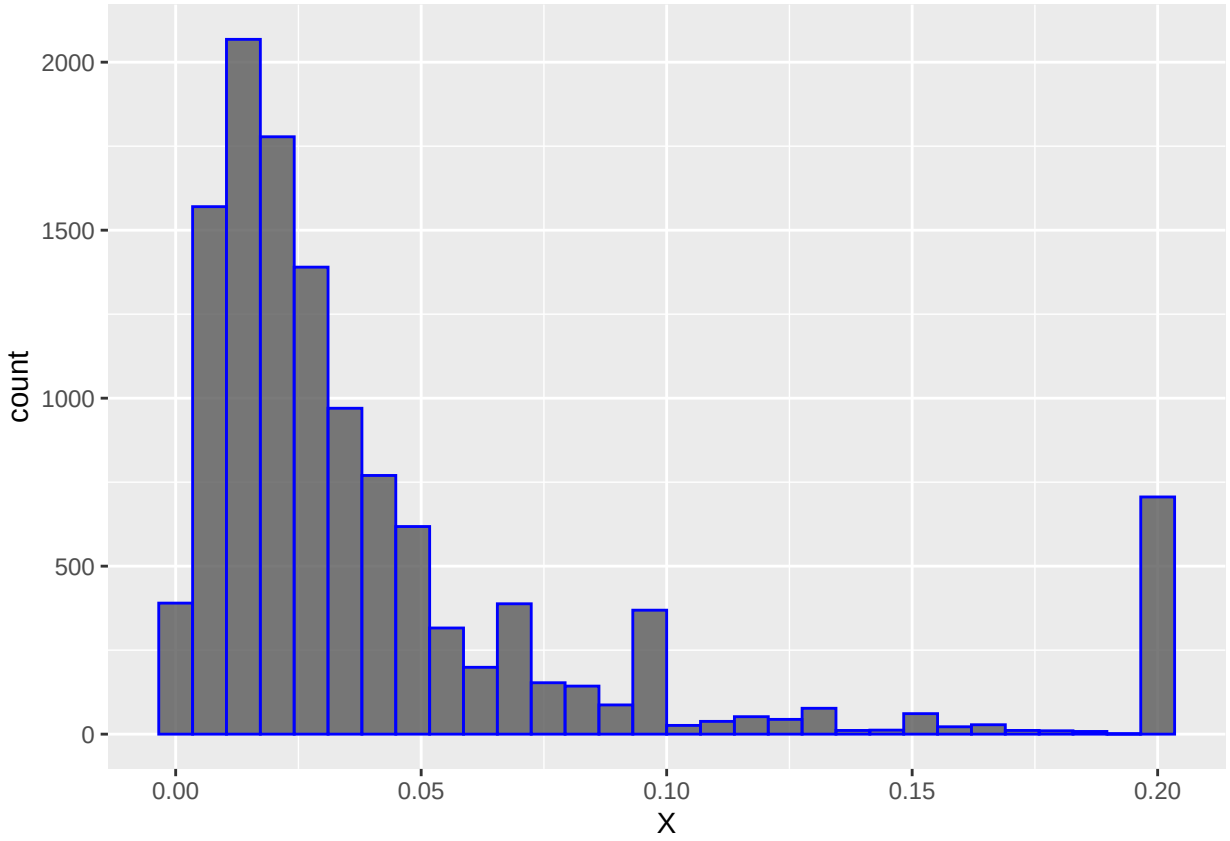

```
dens_fun(numerical$BounceRates)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



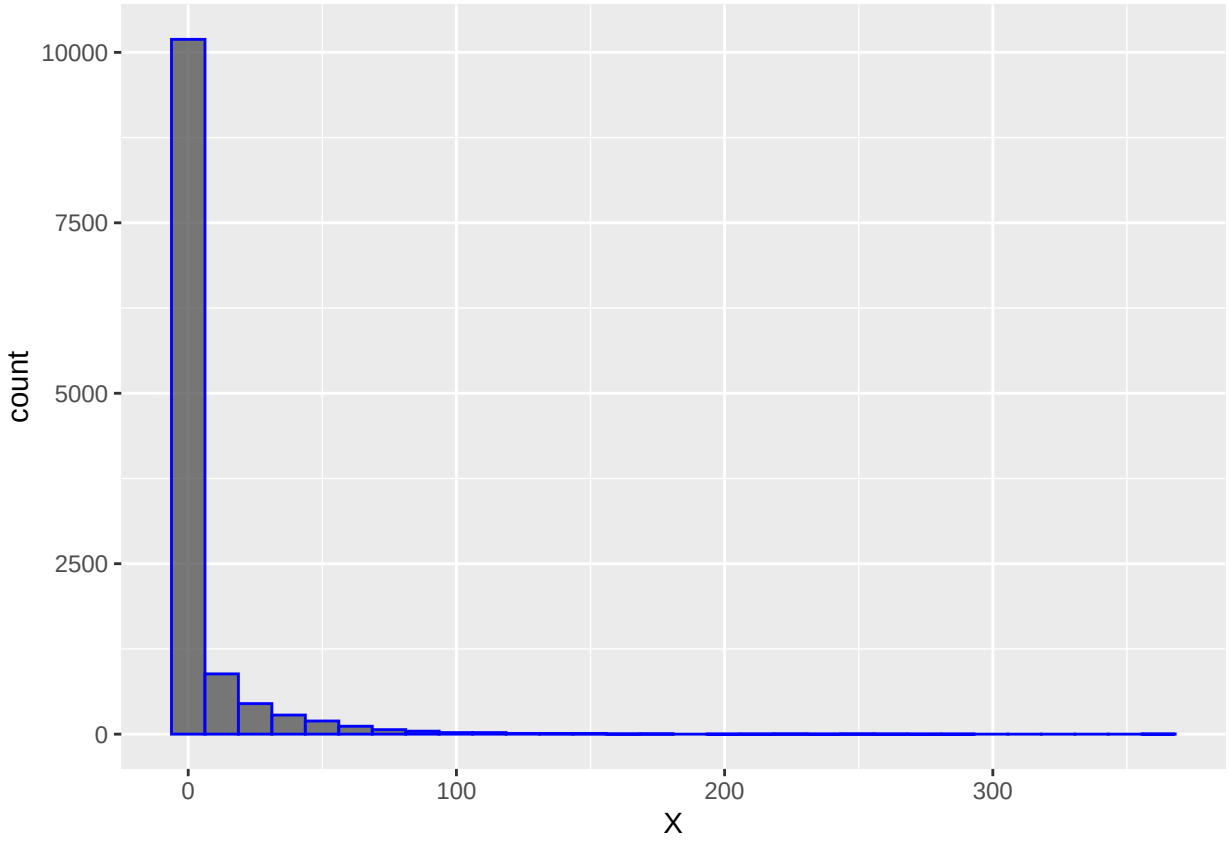

```
dens_fun(numerical$ExitRates)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

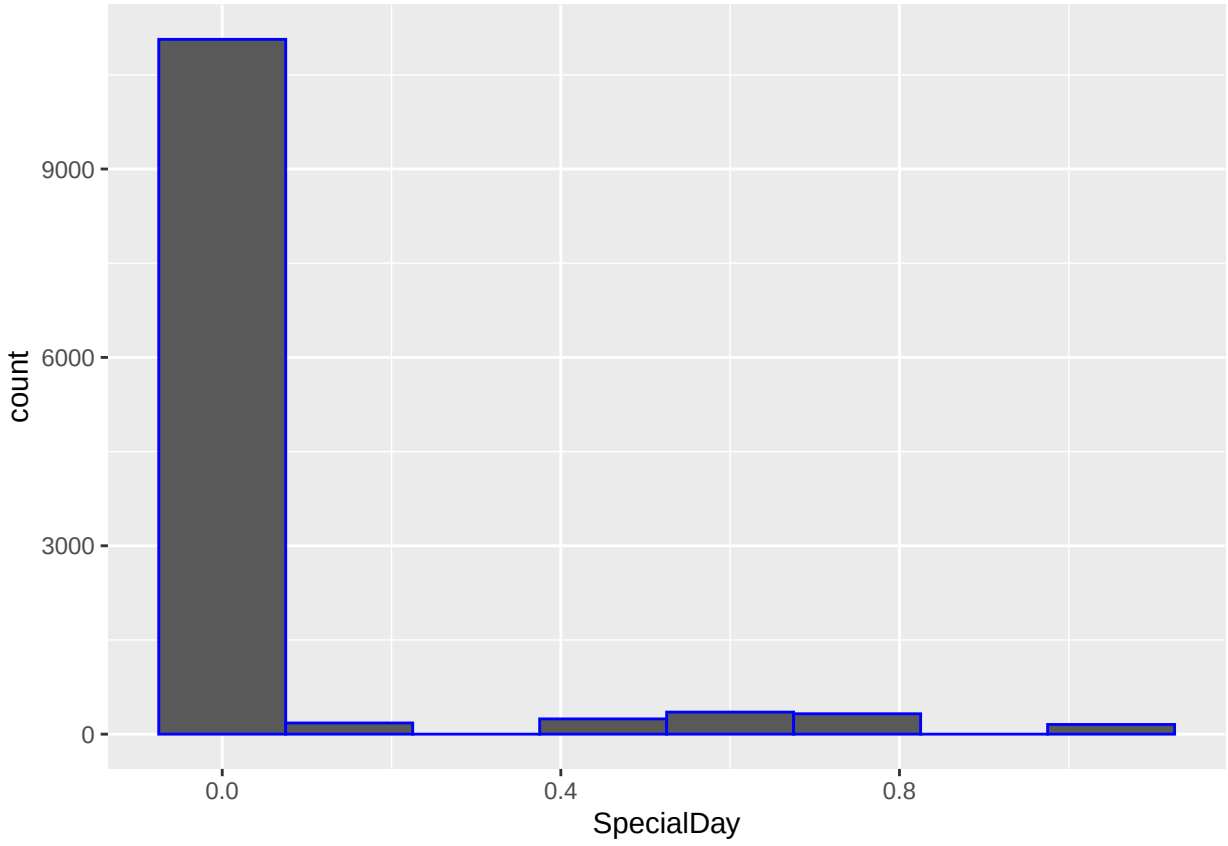



```
dens_fun(numerical$PageValues)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#dens_fun(numerical$SpecialDay)
ggplot(numerical, aes(x= SpecialDay)) +
  geom_histogram(binwidth= 0.15, color="blue")
```



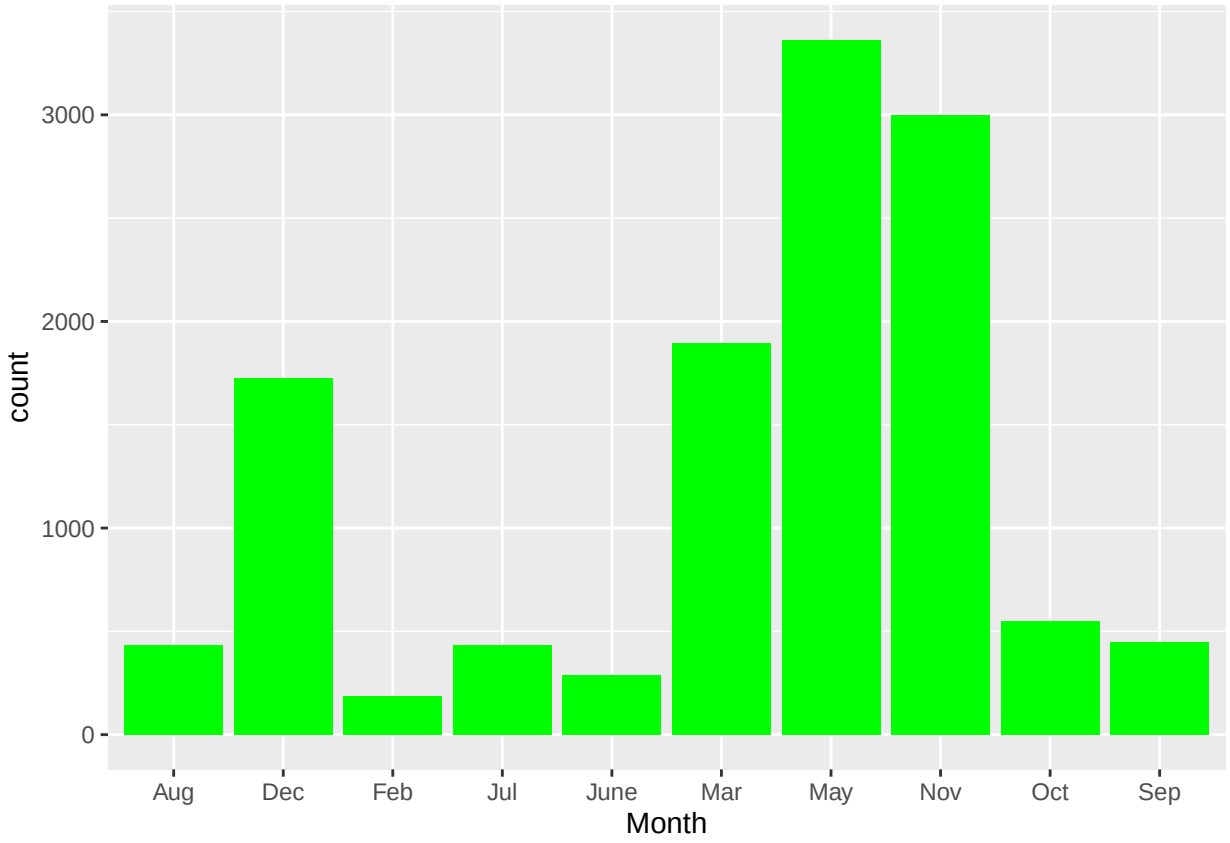
Bar graphs

```
head(df)
```

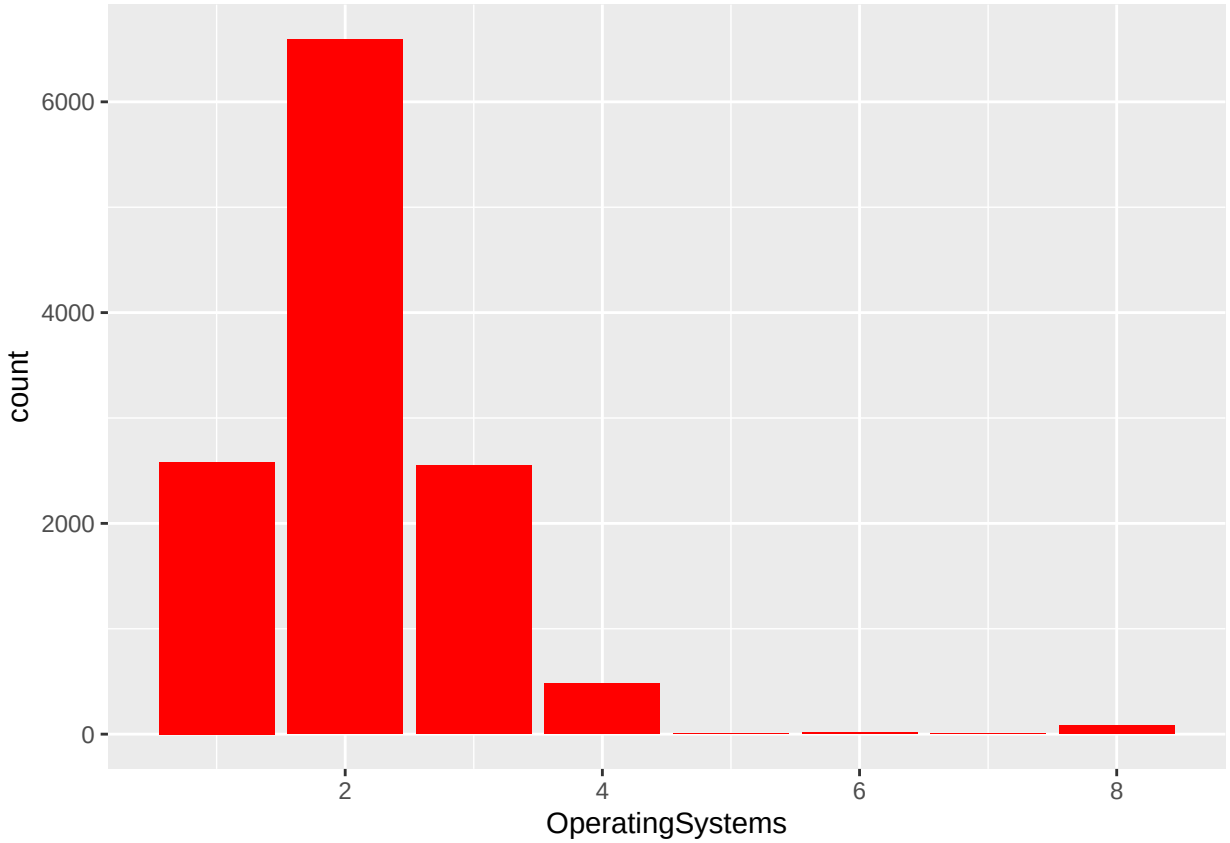
```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                      -1          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

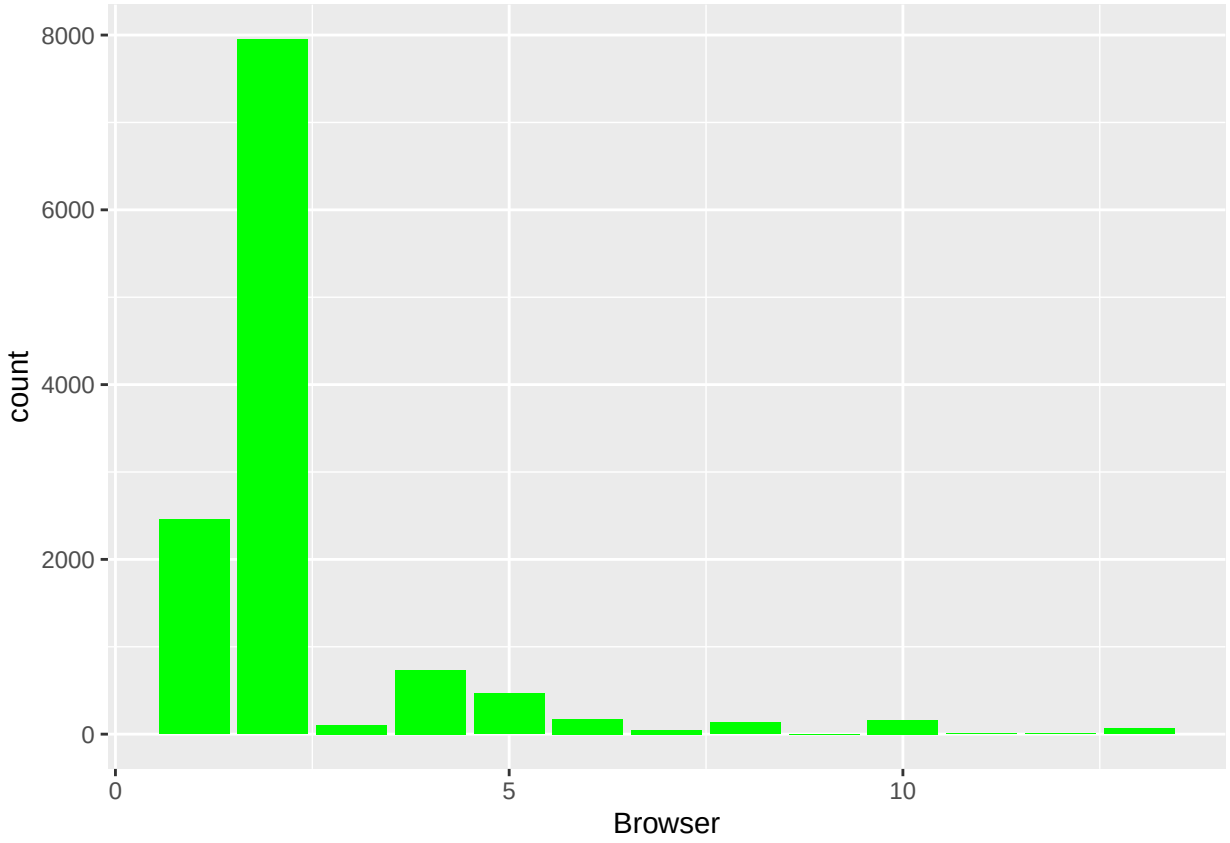
```r
# visualizing the column months

ggplot(df, aes(Month)) +
  geom_bar(fill = "green")
```



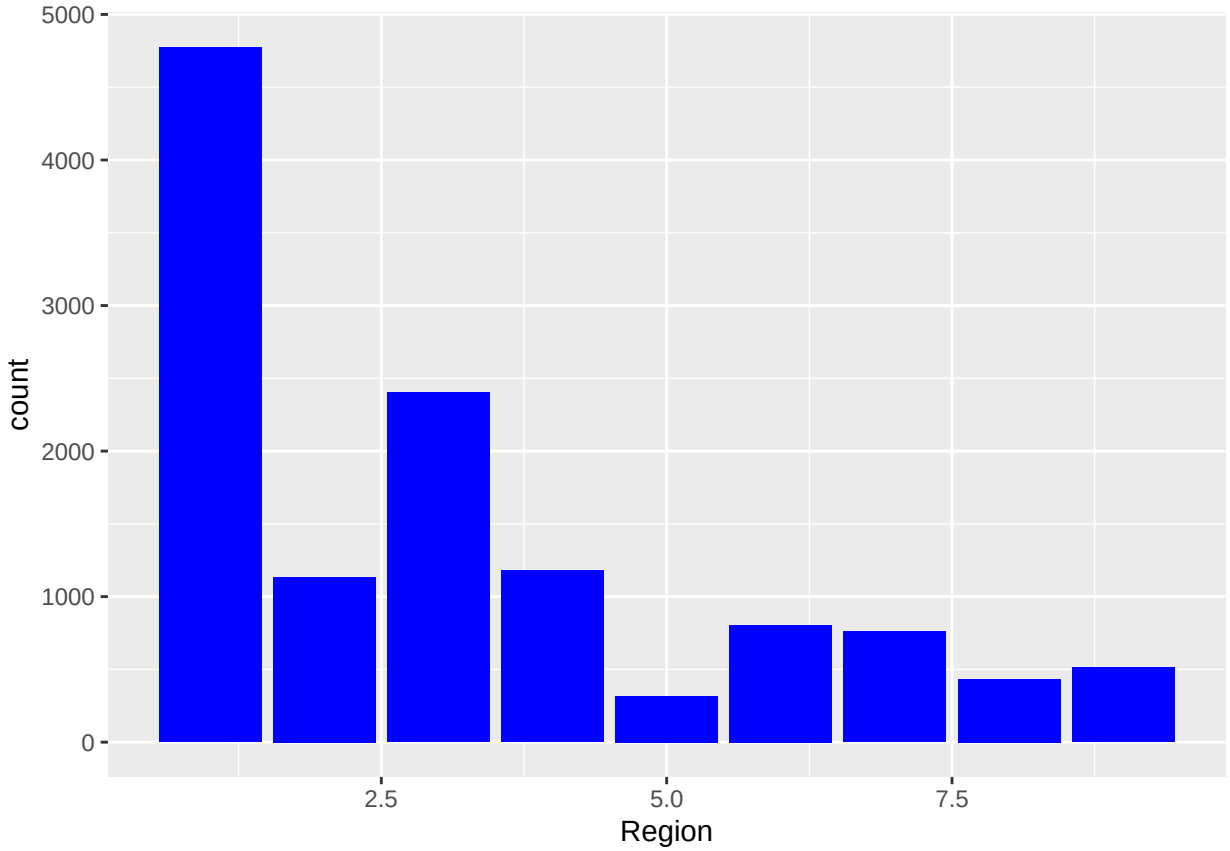

```r
# visualizing column OperatingSystems

ggplot(df, aes(OperatingSystems)) +
  geom_bar(fill = "red")
```



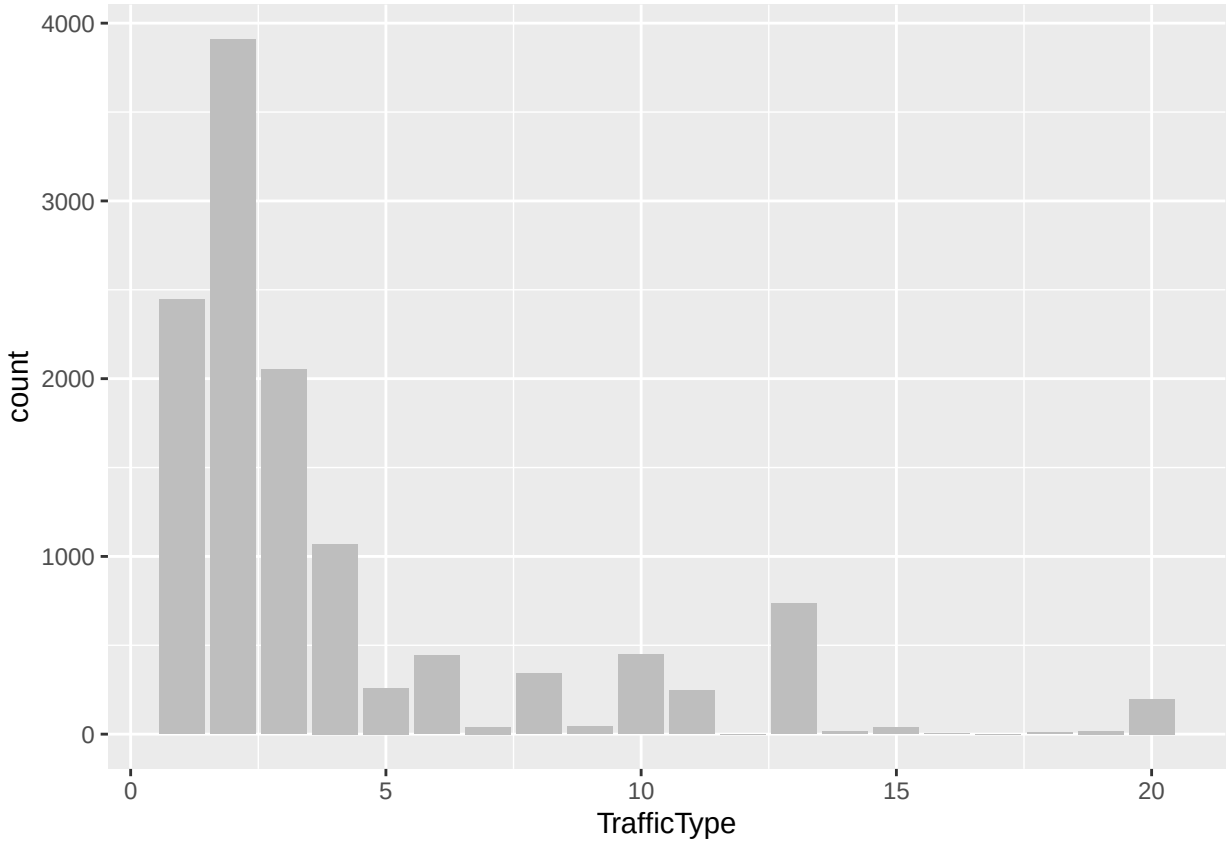

```
# visualizing browser column
ggplot(df, aes(Browser)) +
  geom_bar(stat="count", fill = "green")
```



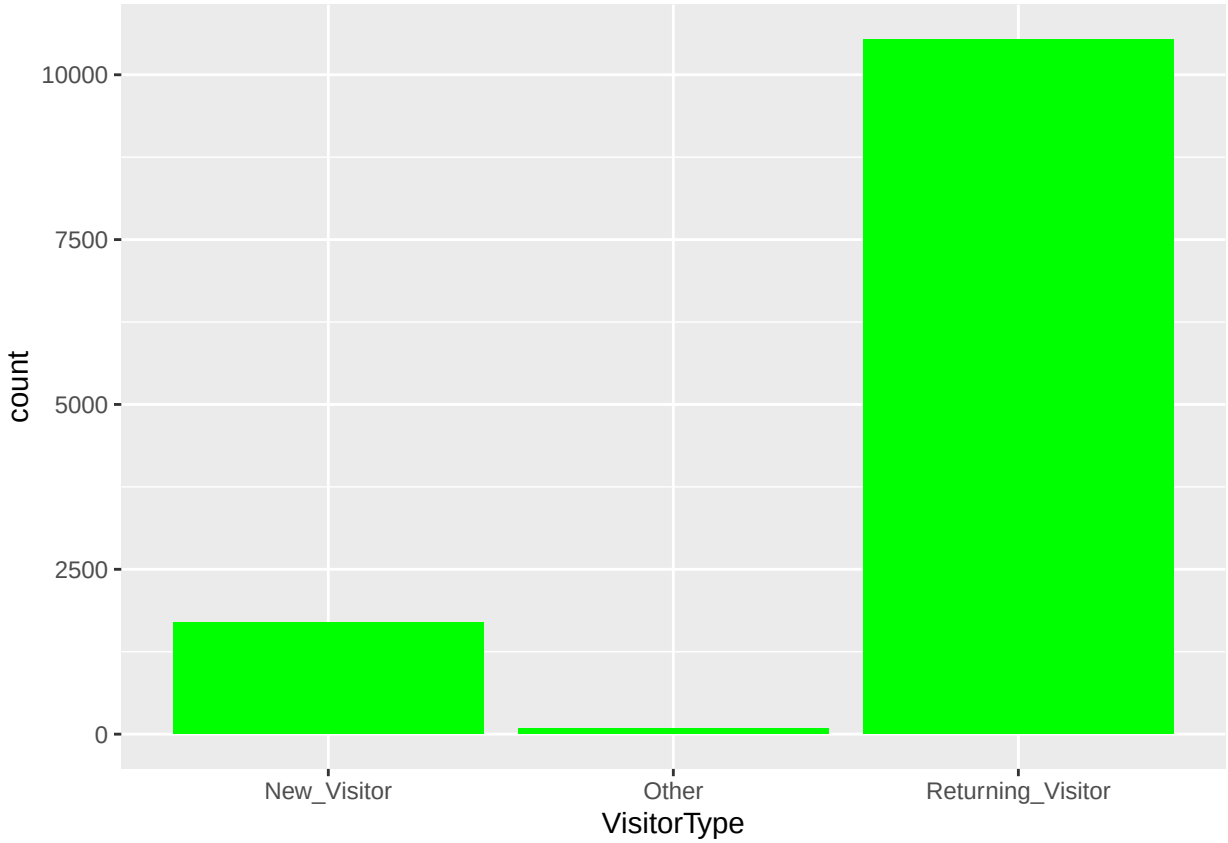

```
# visualizing column OperatingSystems

ggplot(df, aes(Region)) +
  geom_bar(fill = "blue")
```



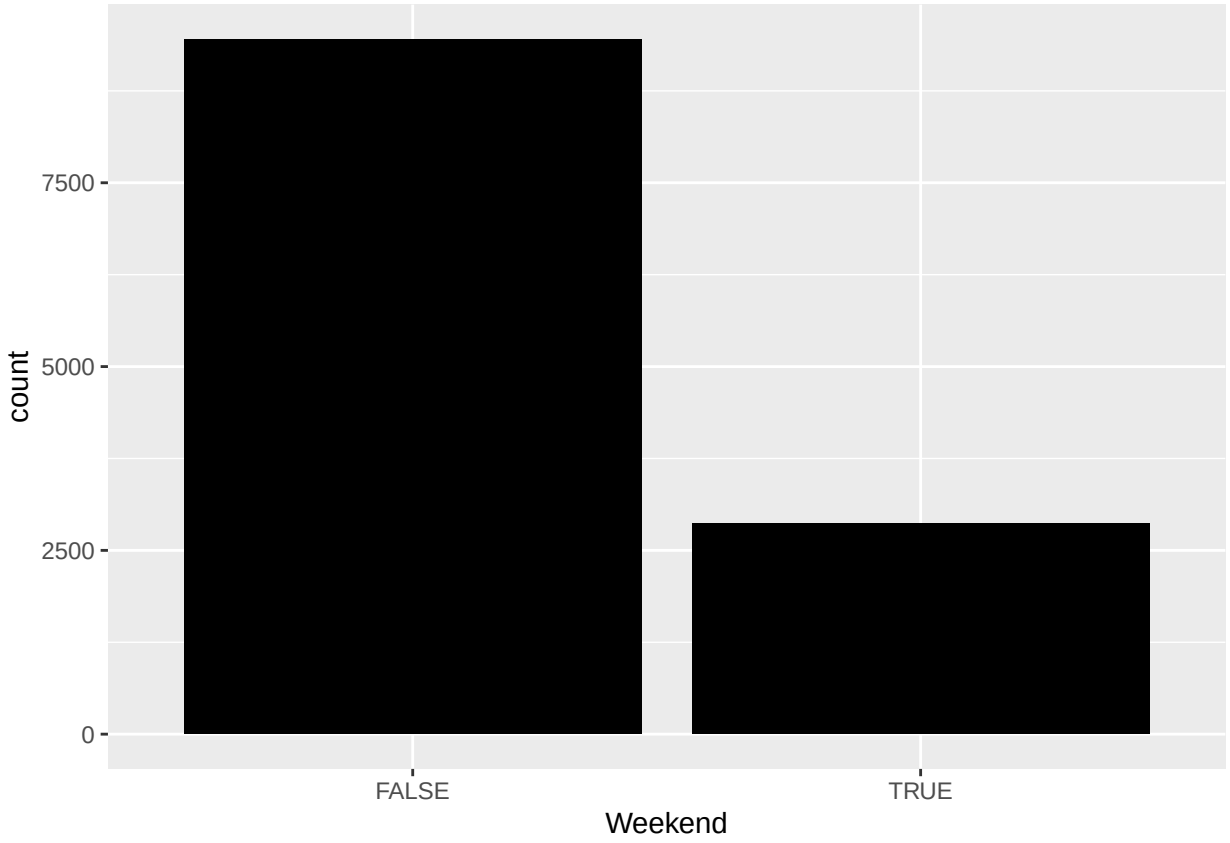

```
# visualizing column OperatingSystems

ggplot(df, aes(TrafficType)) +
  geom_bar(fill = "grey")
```



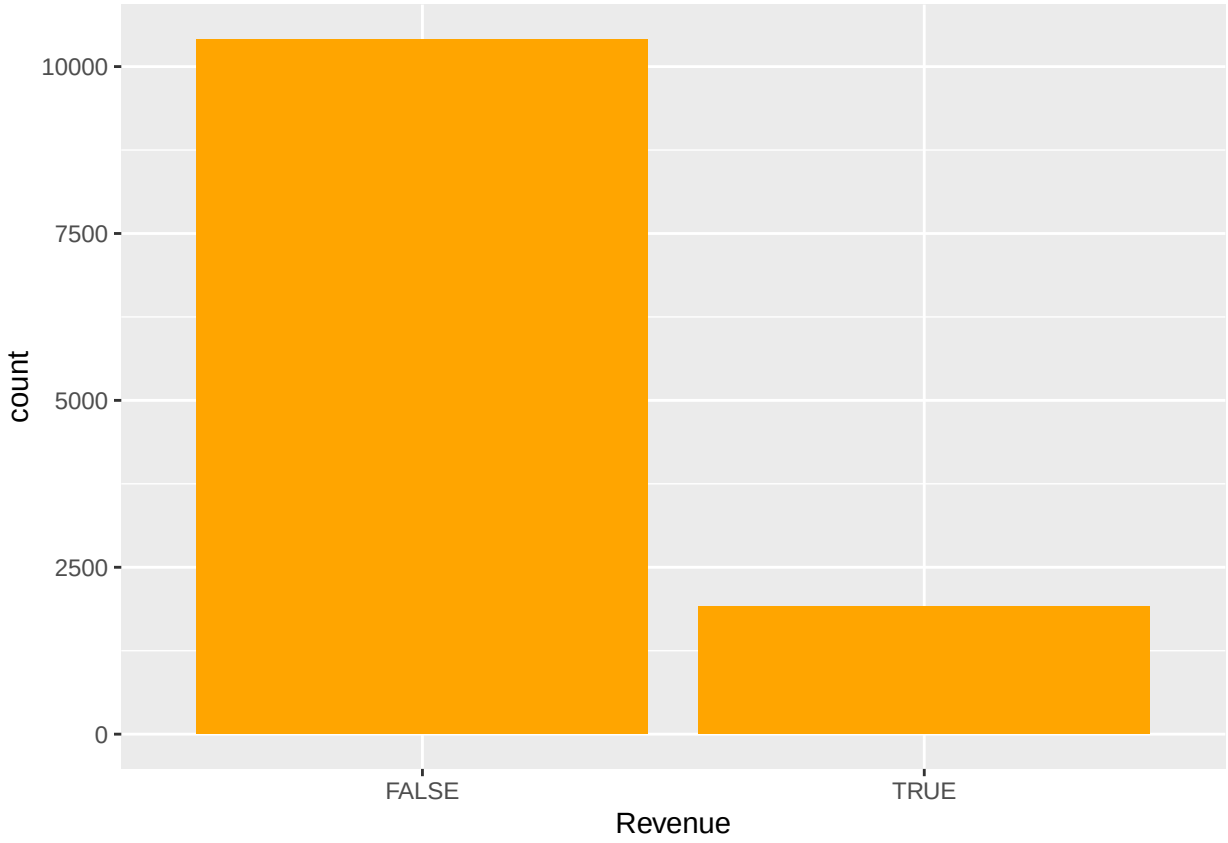

```
# plotting the graph of visitor type

ggplot(df, aes(VisitorType)) +
  geom_bar(stat="count", fill = "green")
```




```
# visualizing the weekend column
ggplot(df, aes(Weekend)) +
  geom_bar(stat="count", fill = "black")
```




```
# visualizing the revenue column
ggplot(df, aes(Revenue)) +
  geom_bar(stat="count", fill = "orange")
```



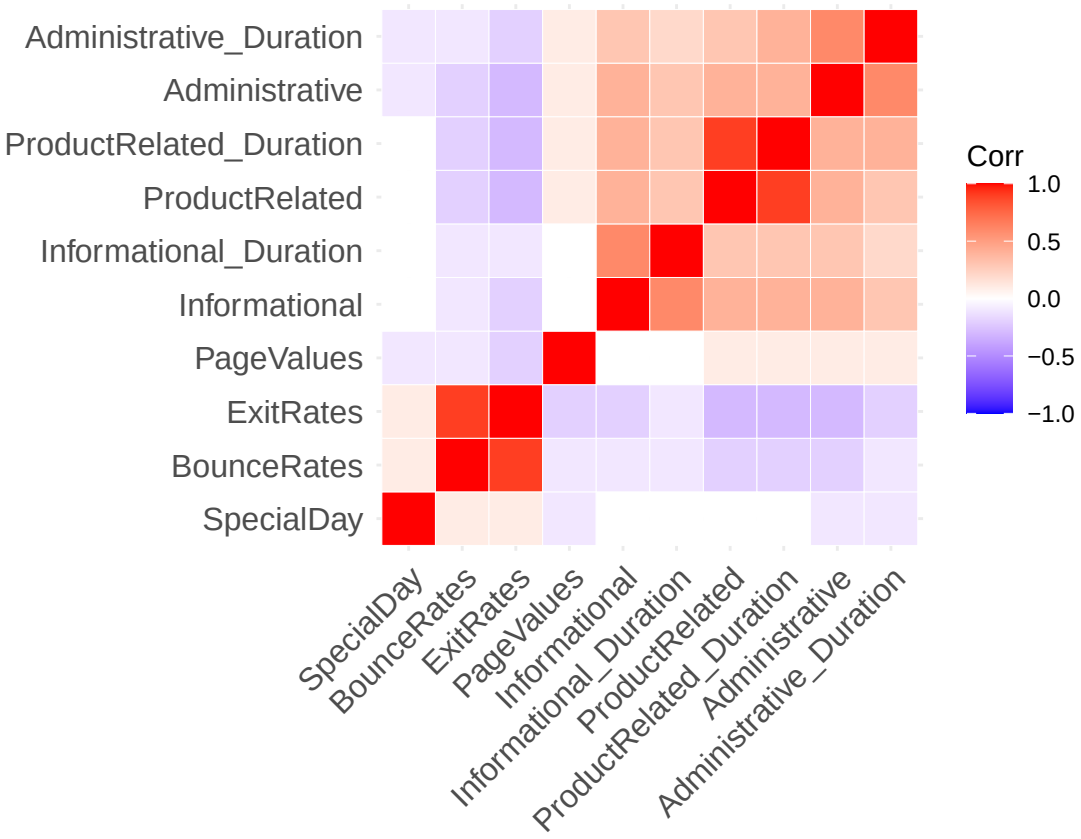
Bivariate analysis

```
# performing a correlation plot for the numerical variables

library(ggcorrplot)

corr <- round(cor(numerical), 1)
head(corr,4)
```

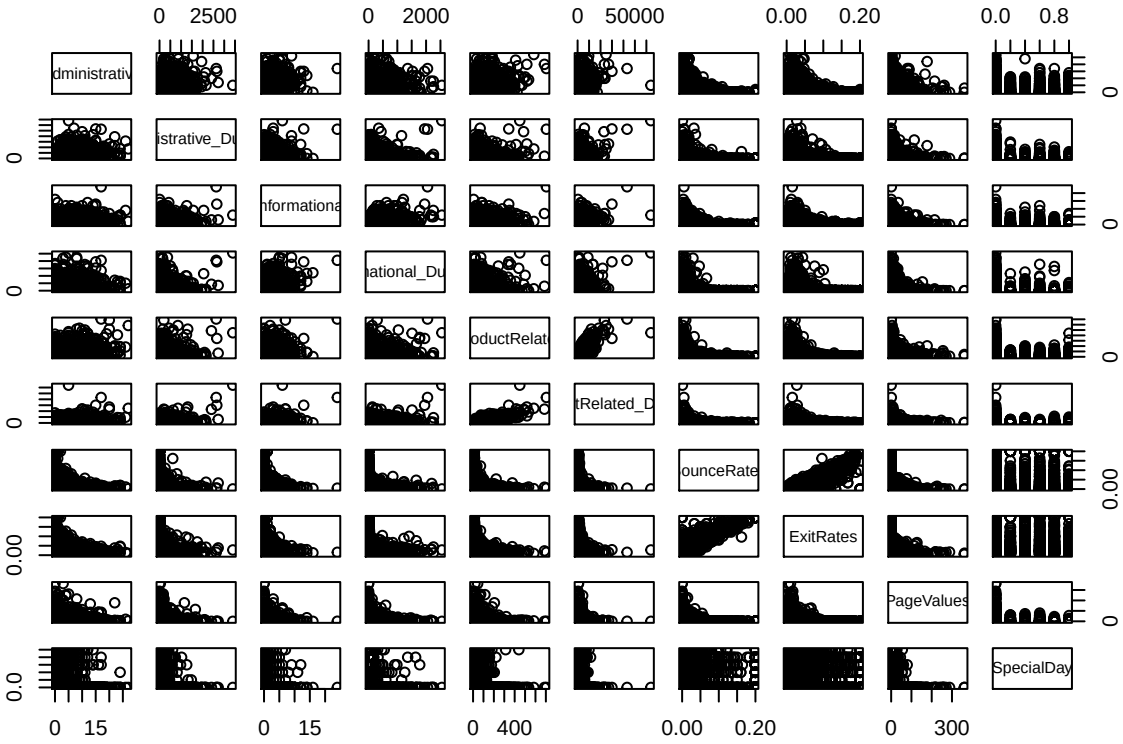
```
##                          Administrative Administrative_Duration Informational
## Administrative                      1.0                     0.6           0.4
## Administrative_Duration             0.6                     1.0           0.3
## Informational                       0.4                     0.3           1.0
## Informational_Duration              0.3                     0.2           0.6
##                          Informational_Duration ProductRelated
## Administrative                              0.3             0.4
## Administrative_Duration                     0.2             0.3
## Informational                               0.6             0.4
## Informational_Duration                      1.0             0.3
##                          ProductRelated_Duration BounceRates ExitRates
## Administrative                               0.4        -0.2      -0.3
## Administrative_Duration                      0.4        -0.1      -0.2
## Informational                                0.4        -0.1      -0.2
## Informational_Duration                       0.3        -0.1      -0.1
```

```
##                           PageValues SpecialDay
## Administrative                   0.1       -0.1
## Administrative_Duration          0.1       -0.1
## Informational                    0.0        0.0
## Informational_Duration           0.0        0.0
```

```r
ggcorrplot(corr, hc.order = TRUE, outline.col = "white")
```




```r
# perfoming pairplots for the numerical variables
col <- df[,c(1:10)]
pairs(col)
```



Implimenting the solution

K-Means Clustering

This method involves partitioning the data set into clusters or k groups.

```
# Since clustering is unsupervised learning I will remove the class attribute which is revenue
df_new <- df[,c(1:9)]
df_class <- df[,c(10)]
```

```
# looking at the predictor data
head(df_new)
```

```
##   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
```

```r
# normalizing the dataset
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
```

```r
# printing unique values for the categorical columns
for (i in names(df[,c(11:18)])){
  print(unique(df[i]))
}
```

```
##      Month
## 1      Feb
## 185    Mar
## 2092   May
## 5456   Oct
## 5457  June
## 5461   Jul
## 5463   Aug
## 5464   Nov
## 5469   Sep
## 7983   Dec
##      OperatingSystems
```

```
## 1                   1
## 2                   2
## 3                   4
## 4                   3
## 261                 7
## 2415                6
## 3365                8
## 6317                5
##      Browser
## 1          1
## 2          2
## 5          3
## 7          4
## 14         5
## 29         6
## 69         7
## 245       10
## 267        8
## 285        9
## 3051      12
## 5680      13
## 6317      11
##    Region
## 1       1
## 3       9
## 4       2
## 7       3
## 12      4
## 21      5
## 31      6
## 36      7
## 38      8
##      TrafficType
## 1              1
## 2              2
## 3              3
## 4              4
## 8              5
## 32             6
## 99             7
## 185            8
## 202            9
## 205           10
## 213           11
## 220           12
## 311           13
## 463           14
## 688           15
## 2124          18
## 2425          19
## 2576          16
## 2641          17
## 2768          20
##            VisitorType
```

```
## 1    Returning_Visitor
## 94         New_Visitor
## 5680             Other
##   Weekend
## 1   FALSE
## 5    TRUE
##    Revenue
## 1    FALSE
## 66    TRUE
```

One hot encoding the categorical columns

```
library(caret)
```

```
## Loading required package: lattice
```

```
dummy <- dummyVars("~.", data=df, fullRank=T)
df_enc <- data.frame(predict(dummy, newdata=df))
str(df_enc)
```

```
## 'data.frame':    12316 obs. of  27 variables:
##  $ Administrative              : num  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration     : num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ Informational               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration      : num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ ProductRelated              : num  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 ...
##  $ MonthDec                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MonthFeb                    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ MonthJul                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MonthJune                   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MonthMar                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MonthMay                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MonthNov                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MonthOct                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ MonthSep                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ OperatingSystems            : num  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                     : num  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                      : num  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType                 : num  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorTypeOther            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ VisitorTypeReturning_Visitor: num  1 1 1 1 1 1 1 1 1 1 ...
##  $ WeekendTRUE                 : num  0 0 0 0 1 0 0 1 0 0 ...
##  $ RevenueTRUE                 : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
# scaling the numerical columns
df_scale <- scale(df_enc[,1:10])
head(df_scale)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 2     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 3     -0.6975533              -0.4631119    -0.3966145             -0.2521304
## 4     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 5     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 6     -0.6975533              -0.4574578    -0.3966145             -0.2450294
##   ProductRelated ProductRelated_Duration BounceRates  ExitRates PageValues
## 1     -0.6914734              -0.6247671  3.67247746  3.2352400 -0.3173633
## 2     -0.6689966              -0.5913358 -0.45743910  1.1745443 -0.3173633
## 3     -0.6914734              -0.6252895  3.67247746  3.2352400 -0.3173633
## 4     -0.6689966              -0.6233742  0.57504004  1.9988226 -0.3173633
## 5     -0.4891823              -0.2969835 -0.04444744  0.1441964 -0.3173633
## 6     -0.2868911              -0.5442099 -0.13139305 -0.3800157 -0.3173633
##   SpecialDay
## 1  -0.309001
## 2  -0.309001
## 3  -0.309001
## 4  -0.309001
## 5  -0.309001
## 6  -0.309001
```

```
# joining the two dataset
final <- cbind(df_scale, df_enc[,11:27])
head(final)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 2     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 3     -0.6975533              -0.4631119    -0.3966145             -0.2521304
## 4     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 5     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 6     -0.6975533              -0.4574578    -0.3966145             -0.2450294
##   ProductRelated ProductRelated_Duration BounceRates  ExitRates PageValues
## 1     -0.6914734              -0.6247671  3.67247746  3.2352400 -0.3173633
## 2     -0.6689966              -0.5913358 -0.45743910  1.1745443 -0.3173633
## 3     -0.6914734              -0.6252895  3.67247746  3.2352400 -0.3173633
## 4     -0.6689966              -0.6233742  0.57504004  1.9988226 -0.3173633
## 5     -0.4891823              -0.2969835 -0.04444744  0.1441964 -0.3173633
## 6     -0.2868911              -0.5442099 -0.13139305 -0.3800157 -0.3173633
##   SpecialDay MonthDec MonthFeb MonthJul MonthJune MonthMar MonthMay MonthNov
## 1  -0.309001        0        1        0         0        0        0        0
## 2  -0.309001        0        1        0         0        0        0        0
## 3  -0.309001        0        1        0         0        0        0        0
## 4  -0.309001        0        1        0         0        0        0        0
## 5  -0.309001        0        1        0         0        0        0        0
## 6  -0.309001        0        1        0         0        0        0        0
##   MonthOct MonthSep OperatingSystems Browser Region TrafficType
## 1        0        0                1       1      1           1
## 2        0        0                2       2      1           2
## 3        0        0                4       1      9           3
## 4        0        0                3       2      2           4
## 5        0        0                3       3      1           4
## 6        0        0                2       2      1           3
```

```
##   VisitorTypeOther VisitorTypeReturning_Visitor WeekendTRUE RevenueTRUE
## 1                0                            1           0           0
## 2                0                            1           0           0
## 3                0                            1           0           0
## 4                0                            1           0           0
## 5                0                            1           1           0
## 6                0                            1           0           0
```

```
# removing the class attribute
final_att <- final[,c(1:26)]
final_class <- final$RevenueTRUE
head(final_att)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 2     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 3     -0.6975533              -0.4631119    -0.3966145             -0.2521304
## 4     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 5     -0.6975533              -0.4574578    -0.3966145             -0.2450294
## 6     -0.6975533              -0.4574578    -0.3966145             -0.2450294
##   ProductRelated ProductRelated_Duration BounceRates  ExitRates PageValues
## 1     -0.6914734              -0.6247671  3.67247746  3.2352400 -0.3173633
## 2     -0.6689966              -0.5913358 -0.45743910  1.1745443 -0.3173633
## 3     -0.6914734              -0.6252895  3.67247746  3.2352400 -0.3173633
## 4     -0.6689966              -0.6233742  0.57504004  1.9988226 -0.3173633
## 5     -0.4891823              -0.2969835 -0.04444744  0.1441964 -0.3173633
## 6     -0.2868911              -0.5442099 -0.13139305 -0.3800157 -0.3173633
##   SpecialDay MonthDec MonthFeb MonthJul MonthJune MonthMar MonthMay MonthNov
## 1  -0.309001        0        1        0         0        0        0        0
## 2  -0.309001        0        1        0         0        0        0        0
## 3  -0.309001        0        1        0         0        0        0        0
## 4  -0.309001        0        1        0         0        0        0        0
## 5  -0.309001        0        1        0         0        0        0        0
## 6  -0.309001        0        1        0         0        0        0        0
##   MonthOct MonthSep OperatingSystems Browser Region TrafficType
## 1        0        0                1       1      1           1
## 2        0        0                2       2      1           2
## 3        0        0                4       1      9           3
## 4        0        0                3       2      2           4
## 5        0        0                3       3      1           4
## 6        0        0                2       2      1           3
##   VisitorTypeOther VisitorTypeReturning_Visitor WeekendTRUE
## 1                0                            1           0
## 2                0                            1           0
## 3                0                            1           0
## 4                0                            1           0
## 5                0                            1           1
## 6                0                            1           0
```

```
# applying the k-means clustering
result <- kmeans(final_att, 2)
result$size
```

```
## [1]  2099 10217
```

```
result$centers
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1    -0.06338239            -0.019121644   -0.05880732           -0.039914964
## 2     0.01302140             0.003928387    0.01208149            0.008200206
##   ProductRelated ProductRelated_Duration BounceRates   ExitRates   PageValues
## 1    -0.05624437            -0.034528047  0.16064694  0.15748040  0.005700066
## 2     0.01155495             0.007093508 -0.03300361 -0.03235307 -0.001171032
##      SpecialDay  MonthDec  MonthFeb   MonthJul  MonthJune  MonthMar  MonthMay
## 1 -0.0011555110 0.1667461 0.0000000 0.02191520 0.02096236 0.1481658 0.2220105
## 2  0.0002373904 0.1347754 0.0180092 0.03778017 0.02388177 0.1549378 0.2835470
##    MonthNov   MonthOct   MonthSep OperatingSystems  Browser   Region
## 1 0.3573130 0.03239638 0.01476894         2.398761 2.555026 3.279657
## 2 0.2200254 0.04707840 0.04081433         2.067730 2.317021 3.120975
##   TrafficType VisitorTypeOther VisitorTypeReturning_Visitor WeekendTRUE
## 1    12.00429      0.027632206                    0.8361124   0.2472606
## 2     2.44054      0.002642654                    0.8595478   0.2296173
```

```
# showing the table of the clusters
table(result$cluster, final_class)
```

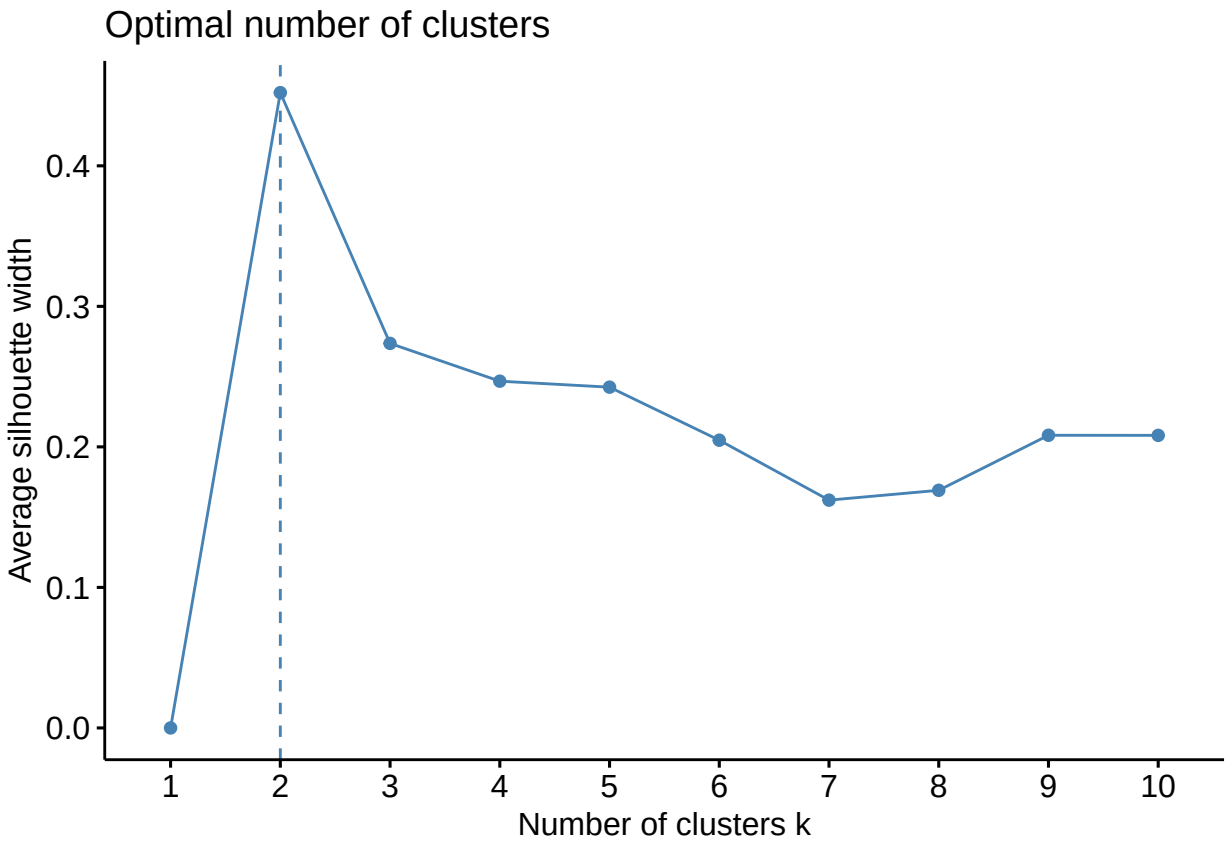
```
##    final_class
##        0    1
##   1 1766  333
##   2 8642 1575
```

```
# visualizing the clusters
# installing the package
# install.packages("factoextra")
png("C:\\plot1.png", width = 480, height = 480, units = "px", bg = "white")
library(factoextra)
```

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

```
fviz_cluster(result, data = final_att)
```

```
# obtaining the optimal number of clusters
fviz_nbclust(x = final_att, FUNcluster = kmeans, method = 'silhouette' )
```



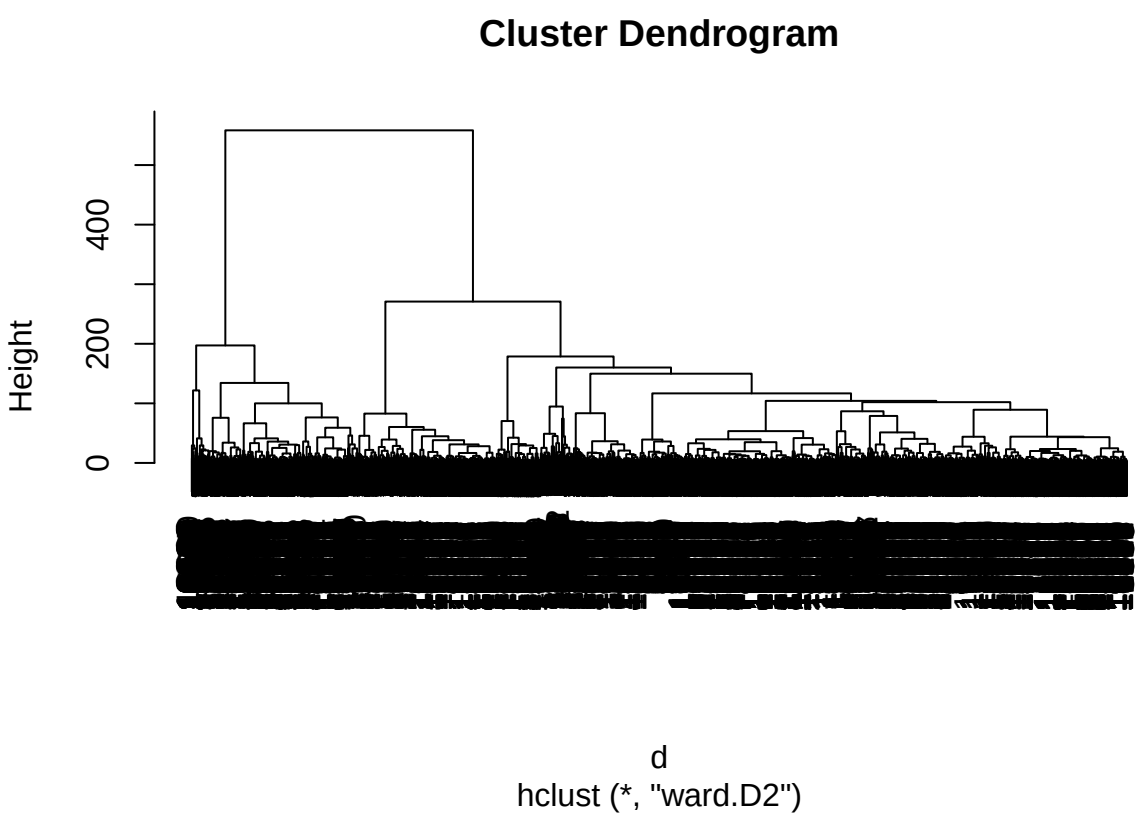
Hierachical clustering

```
# computing the eucledian distance between observations

d <- dist(final_att, method = "euclidean")
```

```
# performing hierachical clustering usig hclust
result_hc <- hclust(d, method = "ward.D2")
```

```
# plotting the dendograms
plot(result_hc, cev=0.6)
```

```
## Warning in graphics:::plotHclust(n1, merge, height, order(x$order), hang, :
## "cev" is not a graphical parameter

## Warning in graphics:::plotHclust(n1, merge, height, order(x$order), hang, :
## "cev" is not a graphical parameter

## Warning in axis(2, at = pretty(range(height)), ...): "cev" is not a graphical
## parameter

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...): "cev"
## is not a graphical parameter
```


Cluster Dendrogram
Height
0
200
400
d
hclust (*, "ward.D2")