EcommerceCustomers

VICTOR NYARIBO

5/28/2021

EcommerceCustomers

a) Data Analytic Question

The aim of this project is to to understand customer's behavior from a one year data set.

b) Success Metrics

- Successful Loading the data.
- Successful Handling missing data.
- Successful Outliers detection.
- Successful Outlier Visualization.
- Successful Handling outliers.
- Successful Univariate analysis.
- Successful Bivariate analysis.

c) Context

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

d) Data Understanding

Variables

- The dataset consists of 10 numerical and 8 categorical attributes.
- 'Revenue' attribute has been be used as a 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 "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 "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
- 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.

e) Experimental Design

- Formulation of the research question.
- Data Sourcing
- Check the Data
- Perform Data Cleaning
- Perform Exploratory Data Analysis (Univariate, Bivariate & Multivariate)
- Implement the Solution
- Challenging the Solution
- Follow up Questions

Data Importation

```
Ecommerce_data<- read.csv("http://bit.ly/EcommerceCustomersDataset",header =T)
```

converting data.frame data into data.table

```
Ecommerce_data<-as.data.table(Ecommerce_data)
class(Ecommerce_data) #checking class
```

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

Data Columns

```
kable(colnames(Ecommerce_data))
```

x
Administrative
Administrative_Duration
Informational
Informational_Duration
ProductRelated
ProductRelated_Duration
BounceRates
ExitRates
PageValues
SpecialDay
Month
OperatingSystems
Browser
Region
TrafficType
VisitorType

x
Weekend
Revenue

Check for missing values

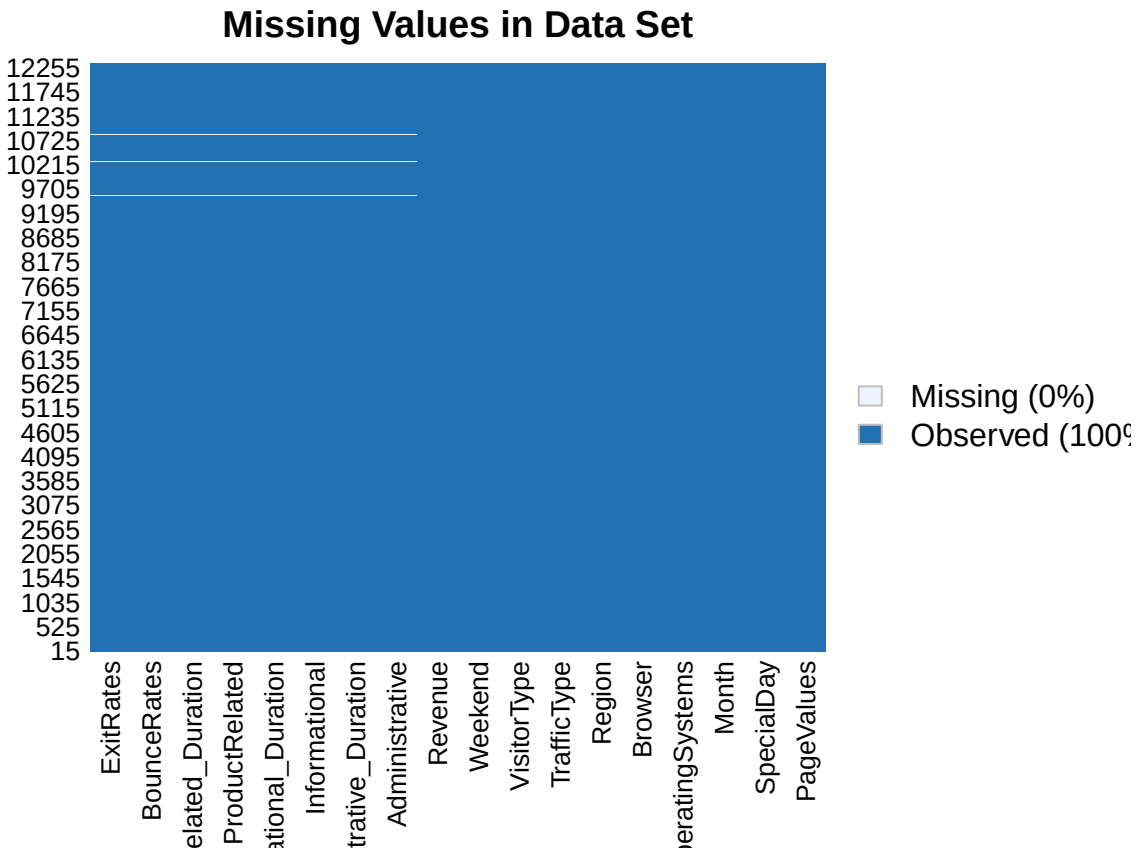
```
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 4.0.5
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.6, built: 2019-11-24)
## ## Copyright (C) 2005-2021 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
missmap(Ecommerce_data,main="Missing Values in Data Set")
```



```
#colSums(is.na(Ecommerce_data))
```

any NAs in data set?

```
colSums(is.na(Ecommerce_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
```

Now lets find the duplicated rows in the dataset df and assign to a variable duplicated_rows below.

```
duplicated_rows <- Ecommerce_data[duplicated(Ecommerce_data),]
#Lets print out the variable duplicated_rows and see these duplicated rows
#kable(duplicated_rows)
```

Removing these duplicated rows in the data set or showing these unique items and assigning to a variable unique_items below

```
unique_items <- Ecommerce_data[!duplicated(Ecommerce_data), ]
```

Encoding Categorical Variables

```
library(encode)
```

```
## Warning: package 'encode' was built under R version 4.0.5
```

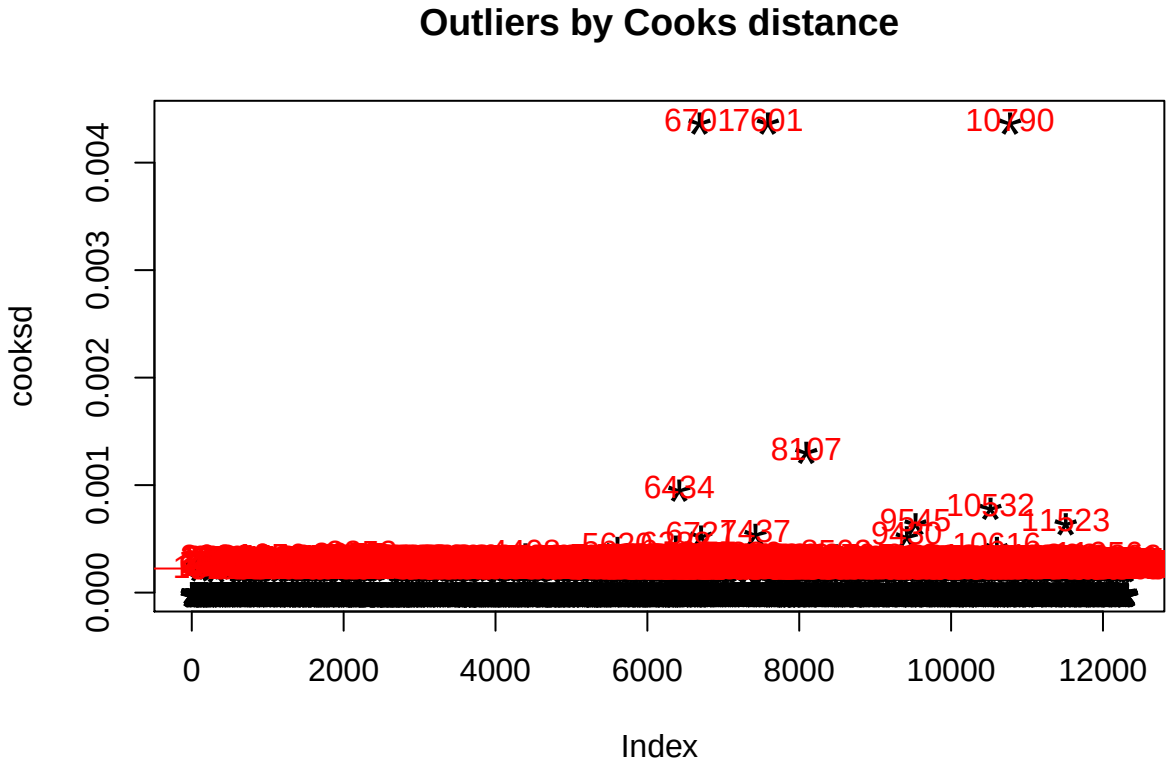
```
##
## Attaching package: 'encode'
```

```
## The following object is masked from 'package:forcats':
##
##     as_factor
```

```
Ecommerce_data$Weekend<-as.factor(Ecommerce_data$Weekend)
Ecommerce_data$Weekend<-unclass(Ecommerce_data$Weekend)         # Convert categorical variables


Ecommerce_data$Revenue<-as.factor(Ecommerce_data$Revenue)
Ecommerce_data$Revenue<-unclass(Ecommerce_data$Revenue)



Ecommerce_data$VisitorType<-as.factor(Ecommerce_data$VisitorType)
Ecommerce_data$VisitorType<-unclass(Ecommerce_data$VisitorType)

Ecommerce_data$Month<-as.factor(Ecommerce_data$Month)
Ecommerce_data$Month<-unclass(Ecommerce_data$Month)
```

Outlier Treatment

```
mod <- lm( Revenue~ExitRates, data=Ecommerce_data)
cooksd <- cooks.distance(mod)

#Influence measures
#In general use, those observations that have a cook's distance greater than 4 times
#the mean may be classified as Outlier


plot(cooksd, pch="*", cex=2, main="Outliers by Cooks distance")  # plot cook's distance
abline(h = 4*mean(cooksd, na.rm=T), col="red")  # add cutoff line
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4*mean(cooksd, na.rm=T),names(cooksd),""), co
```



Tibbles

A tibble is a special kind of data.frame used by dplyr and other packages of the tidyverse. Tidyverse is a set of packages for data science that work in harmony because they share common data representations and API design. When a data.frame is turned into a tibble its class will change.

```
class(Ecommerce_data)
```

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

```
Ecommerce_data<- tbl_df(Ecommerce_data)
```

```
## Warning: `tbl_df()` was deprecated in dplyr 1.0.0.
## Please use `tibble::as_tibble()` instead.
```

```r
class(Ecommerce_data)
```

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

Data Overview

```
## Rows: 12,330
## Columns: 18
## $ Administrative          <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0...
## $ Administrative_Duration <dbl> 0, 0, -1, 0, 0, 0, -1, -1, 0, 0, 0, 0, 0, 0...
## $ Informational           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ Informational_Duration  <dbl> 0, 0, -1, 0, 0, 0, -1, -1, 0, 0, 0, 0, 0, 0...
## $ ProductRelated          <int> 1, 2, 1, 2, 10, 19, 1, 1, 2, 3, 3, 16, 7, 6...
## $ ProductRelated_Duration <dbl> 0.000000000, 64.000000000, -1.000000000, 2....
## $ BounceRates             <dbl> 0.200000000, 0.000000000, 0.200000000, 0.05...
## $ ExitRates               <dbl> 0.200000000, 0.100000000, 0.200000000, 0.14...
## $ PageValues              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ SpecialDay              <dbl> 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.4, 0.0, 0.8...
## $ Month                   <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3...
## $ OperatingSystems        <int> 1, 2, 4, 3, 3, 2, 2, 1, 2, 2, 1, 1, 1, 2, 3...
## $ Browser                 <int> 1, 2, 1, 2, 3, 2, 4, 2, 2, 4, 1, 1, 1, 5, 2...
## $ Region                  <int> 1, 1, 9, 2, 1, 1, 3, 1, 2, 1, 3, 4, 1, 1, 3...
## $ TrafficType             <int> 1, 2, 3, 4, 4, 3, 3, 5, 3, 2, 3, 3, 3, 3, 3...
## $ VisitorType             <int> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3...
## $ Weekend                 <int> 1, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1...
## $ Revenue                 <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
```

Number of columns

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

Dimesion

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

Columnames

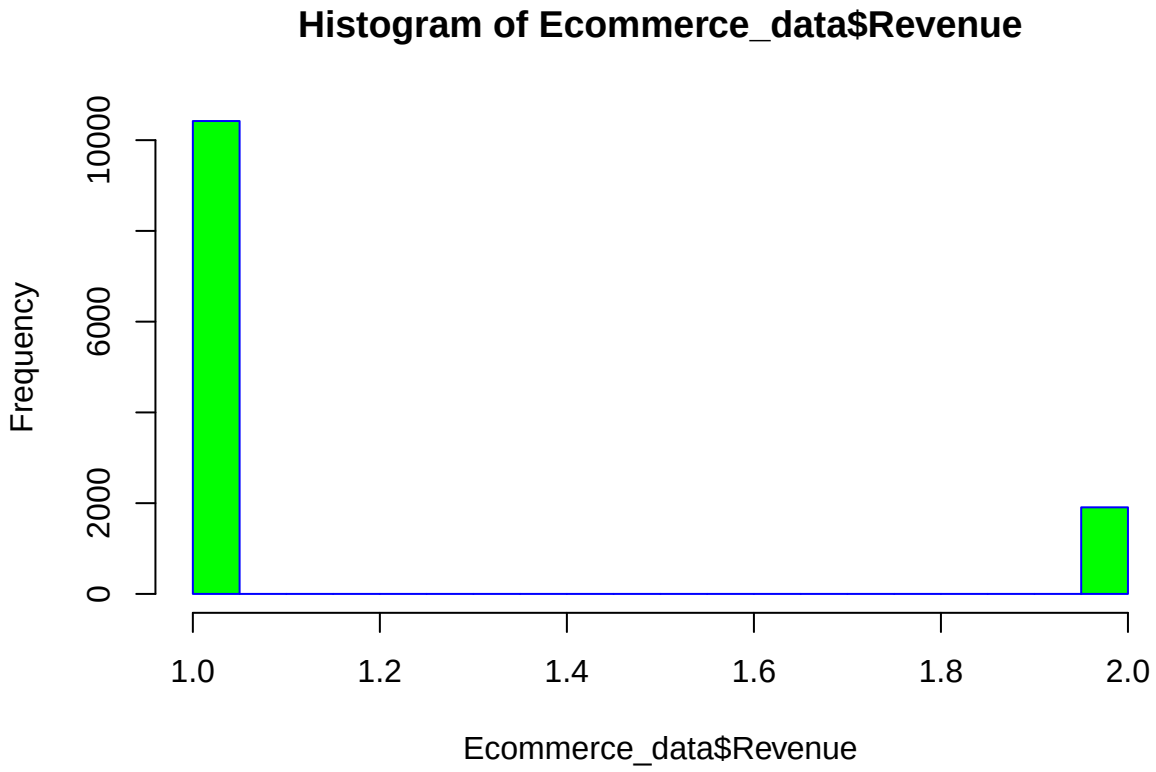
```
##  [1] "Administrative"          "Administrative_Duration"
##  [3] "Informational"           "Informational_Duration"
##  [5] "ProductRelated"          "ProductRelated_Duration"
##  [7] "BounceRates"             "ExitRates"
##  [9] "PageValues"              "SpecialDay"
## [11] "Month"         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

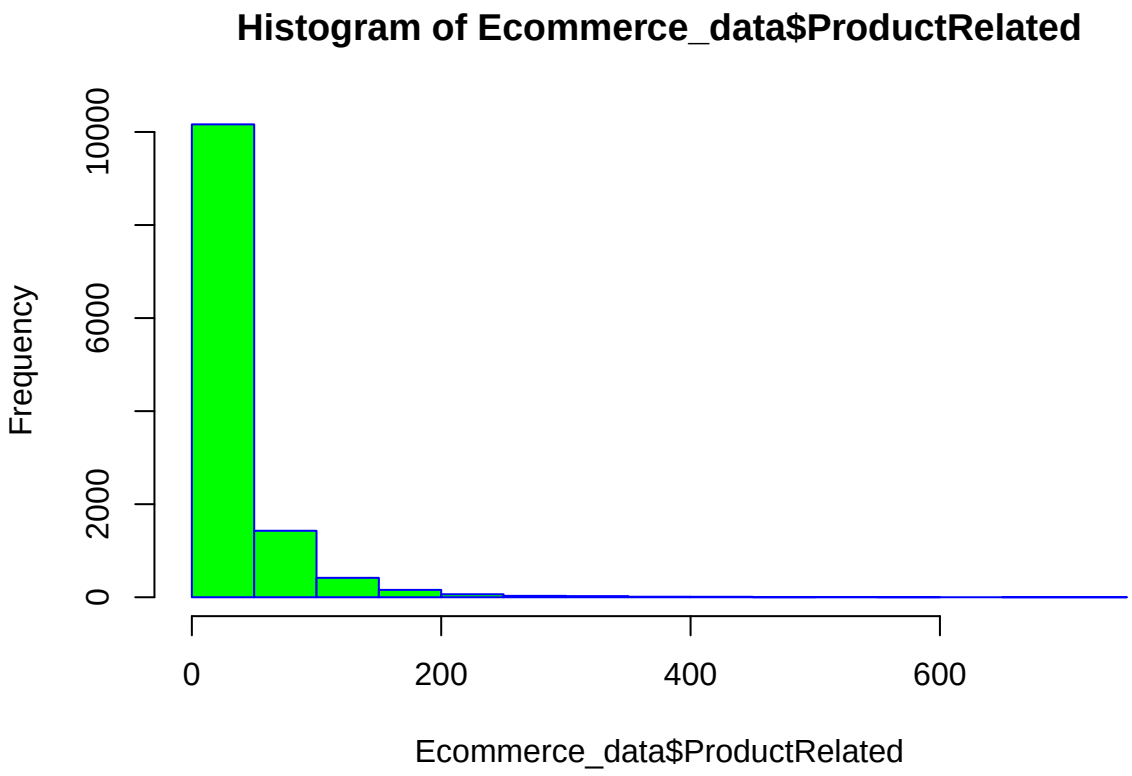
Column data types

```
##          Administrative Administrative_Duration           Informational
##               "integer"               "numeric"               "integer"
##  Informational_Duration          ProductRelated ProductRelated_Duration
##               "numeric"               "integer"               "numeric"
##             BounceRates               ExitRates              PageValues
```

```
##              "numeric"              "numeric"              "numeric"
##            SpecialDay                  Month       OperatingSystems
##              "numeric"            "integer"              "integer"
##                Browser                 Region            TrafficType
##              "integer"            "integer"              "integer"
##            VisitorType                Weekend                Revenue
##              "integer"            "integer"              "integer"
```

UNIVARIATE ANALYSIS

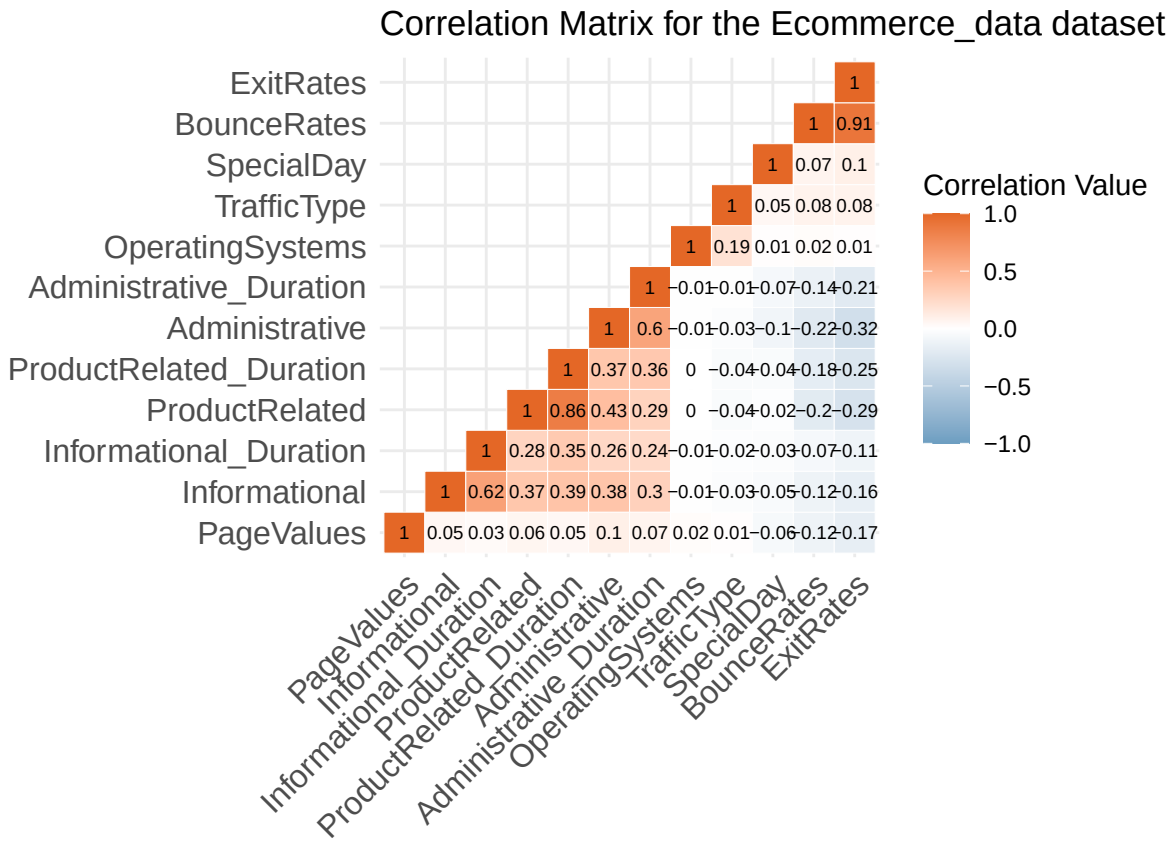





```
## [1] 31.7638843780448
## [1] 18
## [1] 1
```

Correlation Matrix for the Ecommerce_data dataset

```
## Warning: package 'ggcorrplot' was built under R version 4.0.5
```



Variables are not strongly correlated.

DATA SCALING

Scaling

At this point we fit data to a a range of between 0 and 1.

```
##          Administrative Administrative_Duration           Informational
##               "numeric"              "numeric"               "numeric"
##  Informational_Duration         ProductRelated ProductRelated_Duration
##               "numeric"              "numeric"               "numeric"
##             BounceRates              ExitRates              PageValues
##               "numeric"              "numeric"               "numeric"
##              SpecialDay                  Month        OperatingSystems
##               "numeric"              "numeric"               "numeric"
##                 Browser                 Region             TrafficType
##               "numeric"              "numeric"               "numeric"
##             VisitorType                Weekend                 Revenue
##               "numeric"              "numeric"               "numeric"
```

```
##  Administrative            Administrative_Duration    Informational
##  Min.   :-0.697553315445   Min.   :-0.4631119318990   Min.   :-0.3966145153
##  1st Qu.:-0.697553315445   1st Qu.:-0.4574577559560   1st Qu.:-0.3966145153
##  Median :-0.396598133177   Median :-0.4122243484130   Median :-0.3966145153
##  Mean   : 0.000000000000   Mean   : 0.0000000000000   Mean   : 0.0000000000
##  3rd Qu.: 0.506267413627   3rd Qu.: 0.0712076946994   3rd Qu.:-0.3966145153
```

```
##  Max.   : 7.428236605790   Max.   :18.7596727298000   Max.   :18.4905942636
##  Informational_Duration    ProductRelated            ProductRelated_Duration
##  Min.   :-0.252130420513   Min.   :-0.713950149982   Min.   :-0.625289513962
##  1st Qu.:-0.245029432792   1st Qu.:-0.556612621914   1st Qu.:-0.528129740227
##  Median :-0.245029432792   Median :-0.309367934951   Median :-0.311470658414
##  Mean   : 0.000000000000   Mean   : 0.000000000000   Mean   : 0.000000000000
##  3rd Qu.:-0.245029432792   3rd Qu.: 0.140167859529   3rd Qu.: 0.141269707626
##  Max.   :17.858051139300   Max.   :15.132186605400   Max.   :32.792721801700
##   BounceRates                ExitRates                 PageValues
##  Min.   :-0.457439096937   Min.   :-0.886151476923   Min.   :-0.317363329018
##  1st Qu.:-0.457439096937   1st Qu.:-0.591766377196   1st Qu.:-0.317363329018
##  Median :-0.393024540561   Median :-0.368412202579   Median :-0.317363329018
##  Mean   : 0.000000000000   Mean   : 0.000000000000   Mean   : 0.000000000000
##  3rd Qu.:-0.112928194454   3rd Qu.: 0.144196392726   3rd Qu.:-0.317363329018
##  Max.   : 3.672477462750   Max.   : 3.235240001670   Max.   :19.155410219800
##    SpecialDay                    Month                  OperatingSystems
##  Min.   :-0.309001044651   Min.   :-2.1775390650800   Min.   :-1.233204815900
##  1st Qu.:-0.309001044651   1st Qu.:-0.0691939820213   1st Qu.:-0.136191416650
##  Median :-0.309001044651   Median : 0.3524750345910   Median :-0.136191416650
##  Mean   : 0.000000000000   Mean   : 0.0000000000000   Mean   : 0.000000000000
##  3rd Qu.:-0.309001044651   3rd Qu.: 0.7741440512040   3rd Qu.: 0.960821982605
##  Max.   : 4.715631733160   Max.   : 1.6174820844300   Max.   : 6.445888978880
##     Browser                    Region                 TrafficType
##  Min.   :-0.790198793691   Min.   :-0.894184051425   Min.   :-0.7629277724290
##  1st Qu.:-0.208136093745   1st Qu.:-0.894184051425   1st Qu.:-0.5144557382310
##  Median :-0.208136093745   Median :-0.061617748091   Median :-0.5144557382310
##  Mean   : 0.000000000000   Mean   : 0.000000000000   Mean   : 0.0000000000000
##  3rd Qu.:-0.208136093745   3rd Qu.: 0.354665403576   3rd Qu.:-0.0175116698347
##  Max.   : 6.194553605660   Max.   : 2.436081161910   Max.   : 3.9580408773300
##   VisitorType                Weekend
##  Min.   :-2.485954873610   Min.   :-0.550561450598
##  1st Qu.: 0.408040137815   1st Qu.:-0.550561450598
##  Median : 0.408040137815   Median :-0.550561450598
##  Mean   : 0.000000000000   Mean   : 0.000000000000
##  3rd Qu.: 0.408040137815   3rd Qu.:-0.550561450598
##  Max.   : 0.408040137815   Max.   : 1.816180198810
```

Normalizing

Data normalization is a process in which data attributes within a data model are organized to increase the cohesion of entity types.

```
##  Administrative           Administrative_Duration   Informational
##  Min.   : 0.00000000000   Min.   :  -1.0000000000   Min.   : 0.000000000000
##  1st Qu.: 0.00000000000   1st Qu.:   0.0000000000   1st Qu.: 0.000000000000
##  Median : 1.00000000000   Median :   8.0000000000   Median : 0.000000000000
##  Mean   : 2.31779798636   Mean   :  80.9061763519   Mean   : 0.503978564469
##  3rd Qu.: 4.00000000000   3rd Qu.:  93.5000000000   3rd Qu.: 0.000000000000
##  Max.   :27.00000000000   Max.   :3398.7500000000   Max.   :24.000000000000
##  Informational_Duration    ProductRelated          ProductRelated_Duration
##  Min.   :  -1.0000000000   Min.   :  0.000000000   Min.   :   -1.00000000
##  1st Qu.:   0.0000000000   1st Qu.:  7.000000000   1st Qu.:  185.00000000
##  Median :   0.0000000000   Median : 18.000000000   Median :  599.76619045
##  Mean   :  34.5063873375   Mean   : 31.763884378   Mean   : 1196.03705685
##  3rd Qu.:   0.0000000000   3rd Qu.: 38.000000000   3rd Qu.: 1466.47990175
```

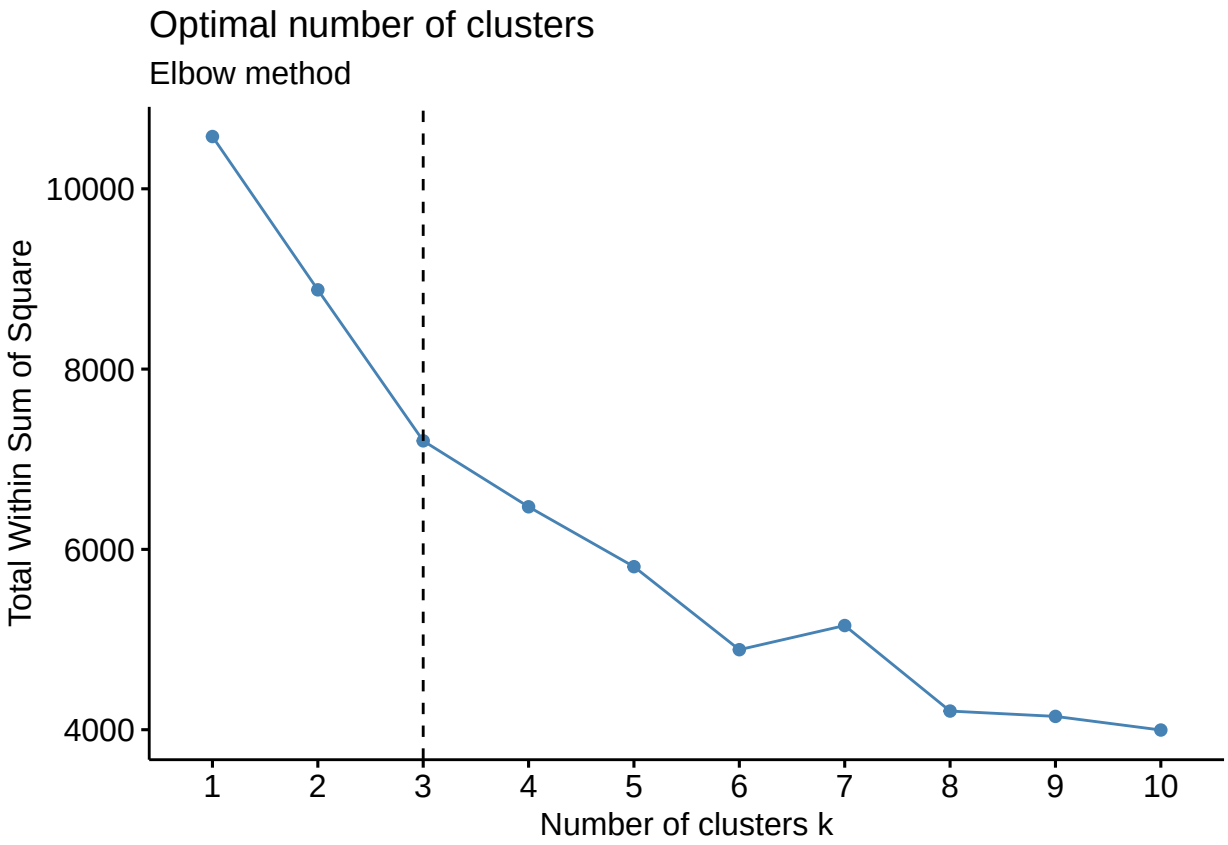
```
##  Max.   :2549.3750000000   Max.   :705.000000000   Max.   :63973.52223000
##   BounceRates               ExitRates                  PageValues
##  Min.   :0.000000000000   Min.   :0.0000000000000   Min.   :  0.00000000000
##  1st Qu.:0.000000000000   1st Qu.:0.0142857140000   1st Qu.:  0.00000000000
##  Median :0.003119412000   Median :0.0251244890000   Median :  0.00000000000
##  Mean   :0.022152461936   Mean   :0.0430025384157   Mean   :  5.89595237472
##  3rd Qu.:0.016683673750   3rd Qu.:0.0500000000000   3rd Qu.:  0.00000000000
##  Max.   :0.200000000000   Max.   :0.2000000000000   Max.   :361.76374190000
##    SpecialDay                    Month                 OperatingSystems
##  Min.   :0.0000000000000   Min.   : 1.00000000000   Min.   :1.00000000000
##  1st Qu.:0.0000000000000   1st Qu.: 6.00000000000   1st Qu.:2.00000000000
##  Median :0.0000000000000   Median : 7.00000000000   Median :2.00000000000
##  Mean   :0.0614972393634   Mean   : 6.16409548555   Mean   :2.12414745047
##  3rd Qu.:0.0000000000000   3rd Qu.: 8.00000000000   3rd Qu.:3.00000000000
##  Max.   :1.0000000000000   Max.   :10.00000000000   Max.   :8.00000000000
##     Browser                    Region                TrafficType
##  Min.   : 1.00000000000   Min.   :1.00000000000   Min.   : 1.00000000000
##  1st Qu.: 2.00000000000   1st Qu.:1.00000000000   1st Qu.: 2.00000000000
##  Median : 2.00000000000   Median :3.00000000000   Median : 2.00000000000
##  Mean   : 2.35758363105   Mean   :3.14801883728   Mean   : 4.07047742774
##  3rd Qu.: 2.00000000000   3rd Qu.:4.00000000000   3rd Qu.: 4.00000000000
##  Max.   :13.00000000000   Max.   :9.00000000000   Max.   :20.00000000000
##   VisitorType               Weekend                  Revenue
##  Min.   :1.00000000000   Min.   :1.00000000000   Min.   :1.00000000000
##  1st Qu.:3.00000000000   1st Qu.:1.00000000000   1st Qu.:1.00000000000
##  Median :3.00000000000   Median :1.00000000000   Median :1.00000000000
##  Mean   :2.71800909386   Mean   :1.23262422865   Mean   :1.15492042871
##  3rd Qu.:3.00000000000   3rd Qu.:1.00000000000   3rd Qu.:1.00000000000
##  Max.   :3.00000000000   Max.   :2.00000000000   Max.   :2.00000000000
```

Finding optimal number of clusters

Method 1:Elbow

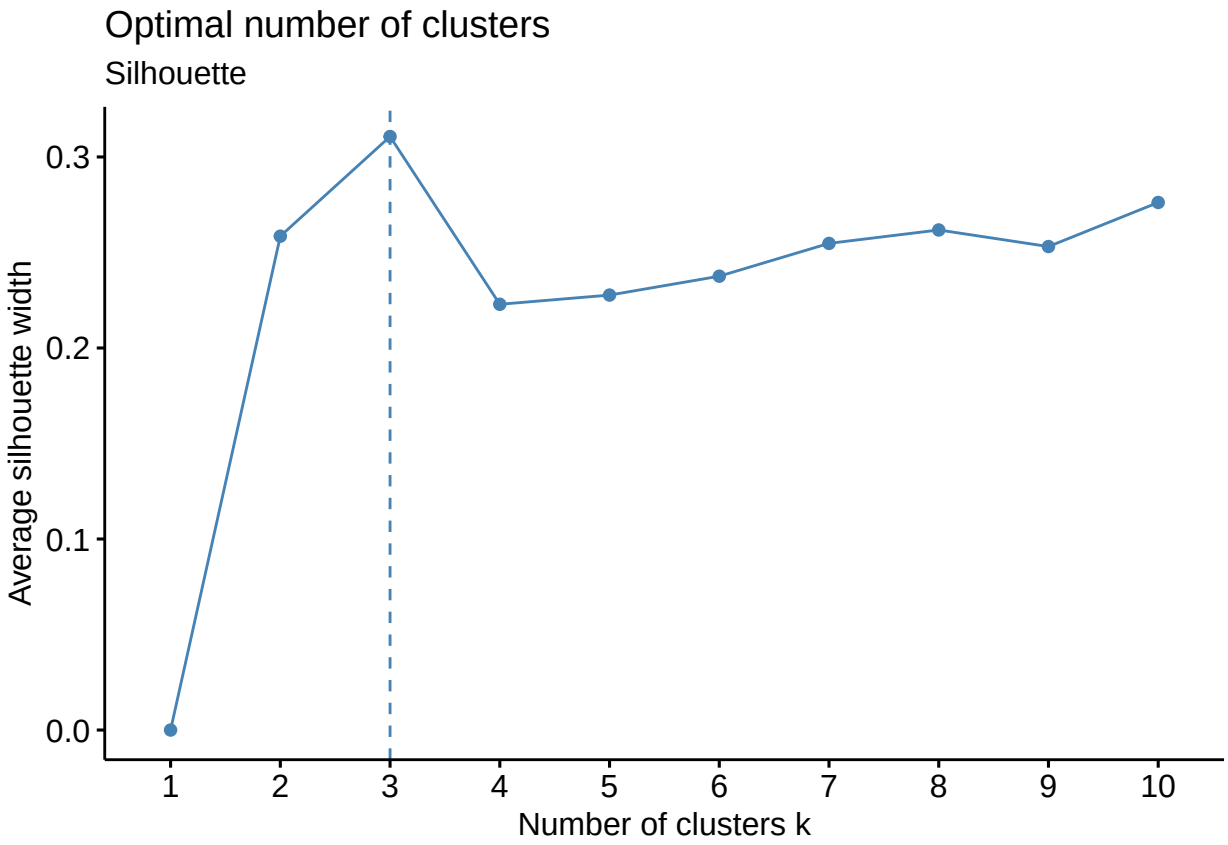
```
## Warning: package 'factoextra' was built under R version 4.0.5
```

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



According to these observations, it's possible to define k = 3 as the optimal number of clusters in the data.
Method 2:Silhouette

```
## Warning: package 'cluster' was built under R version 4.0.5
```



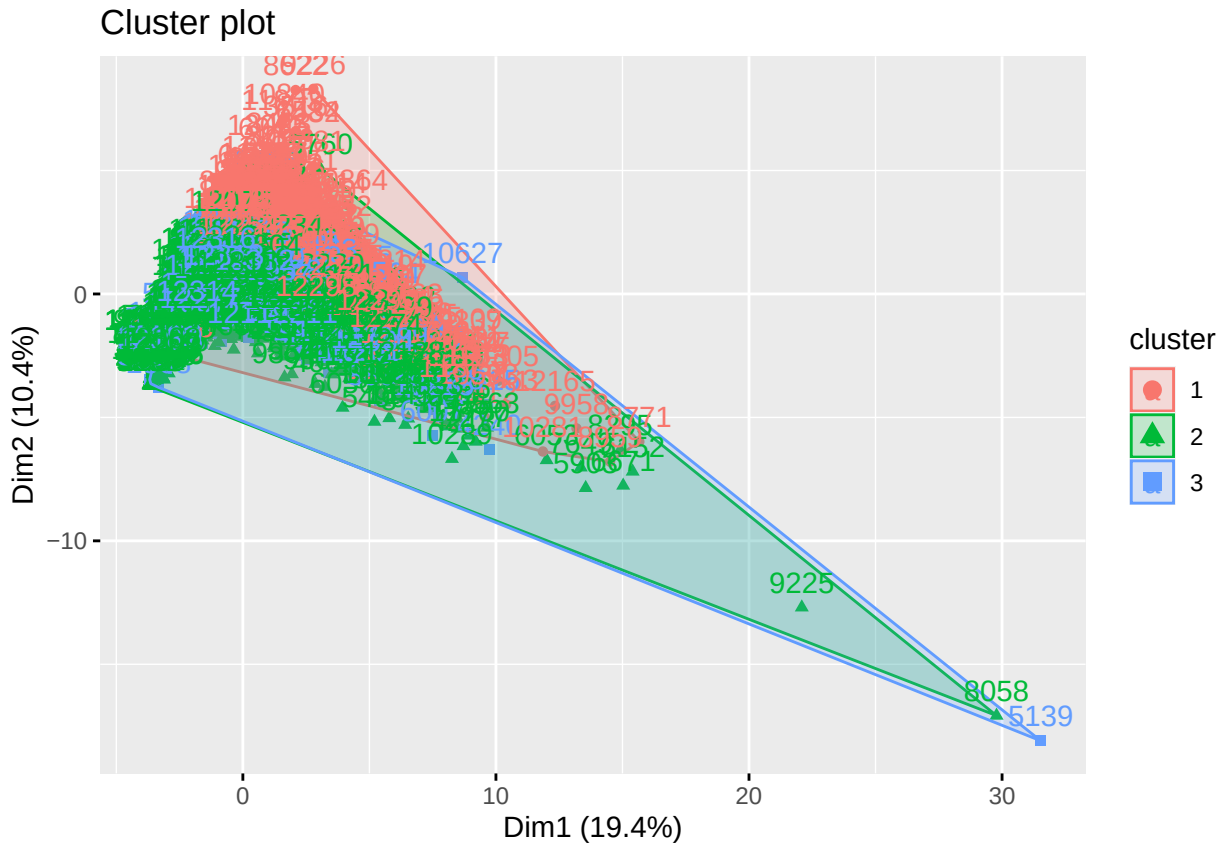
Solution Implimentation

K-MEANS CLUSTERING

```
## [1] 1908 8042 2366
```

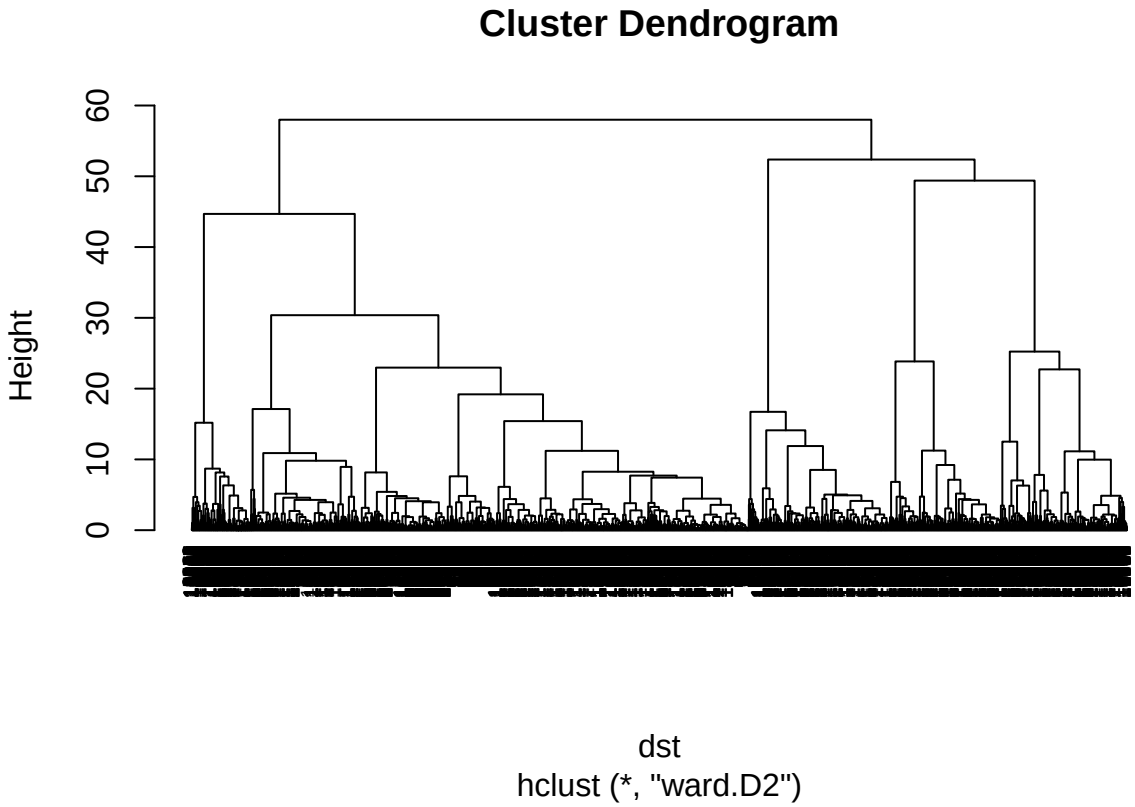
```
##       Administrative Administrative_Duration      Informational
## 1 0.1256891062970746      0.0354388540022979 0.0327568134171907
## 2 0.0766347048366429      0.0213562199162860 0.0176210312525906
## 3 0.0850161234776630      0.0242395782237748 0.0229994364609749
##   Informational_Duration     ProductRelated ProductRelated_Duration
## 1     0.0229814935650393 0.0683832166168575      0.0293430814245083
## 2     0.0114896632008332 0.0400193311356514      0.0165549916060556
## 3     0.0148838175714926 0.0433595319029034      0.0174660673205600
##          BounceRates         ExitRates         PageValues         SpecialDay
## 1 0.0255857632023061 0.097775841284067 0.0753655356711580 0.0231656184486373
## 2 0.1327516052499383 0.245099538176448 0.0055622120335342 0.0701815468788861
## 3 0.1047094507227387 0.207290585145816 0.0051542385631132 0.0628909551986475
##               Month  OperatingSystems           Browser            Region
## 1 0.623107384113692 0.156109613656783 0.121112858141162 0.260285639412998
## 2 0.562076874188299 0.161136177923039 0.115000414490600 0.270579457846307
## 3 0.573823612285161 0.162359618403575 0.100345167652862 0.268068469991547
##         TrafficType       VisitorType           Weekend Revenue
## 1 0.159025708926404 0.774633123689727 0.261530398322851       1
## 2 0.162600295815383 0.884419298681920 0.000000000000000       0
## 3 0.160297192685854 0.840659340659341 1.000000000000000       0
```

Cluster visualization



As visualized the data is clustered into two distinct clusters with one overlapping the two.

HIERACHICAL CLUSTERING



Challenging the solution