WebAnalytics

Stacy Mwende

3/9/2020

# 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.

# Defining the question

The company would like to understand customer’s behavior with more specifications on the characteristics of the different customer groups.

# Data Sourcing

Data Sourcing is the process of extracting data from external or internal front/back office systems comprising an institution’s data Infrastructure for diverse purposes. In our case, the data source was Google Analytics. Google analytics works by the inclusion of a block of JavaScript code on pages in your website. When users to your website view a page, this JavaScript code references a JavaScript file which then executes the tracking operation for Analytics.

# Reading the data

**Get a working directory and load the data**

dir <- "C:/Users/comp5/Downloads/R\_Program"  
my\_data <- file.path(dir,"online\_shoppers\_intention.csv")  
shoppers <- read.csv(my\_data) # Reading the dataset  
head(shoppers)

## Administrative Administrative\_Duration Informational  
## 1 0 0 0  
## 2 0 0 0  
## 3 0 -1 0  
## 4 0 0 0  
## 5 0 0 0  
## 6 0 0 0  
## Informational\_Duration ProductRelated ProductRelated\_Duration  
## 1 0 1 0.000000  
## 2 0 2 64.000000  
## 3 -1 1 -1.000000  
## 4 0 2 2.666667  
## 5 0 10 627.500000  
## 6 0 19 154.216667  
## BounceRates ExitRates PageValues SpecialDay Month OperatingSystems  
## 1 0.20000000 0.2000000 0 0 Feb 1  
## 2 0.00000000 0.1000000 0 0 Feb 2  
## 3 0.20000000 0.2000000 0 0 Feb 4  
## 4 0.05000000 0.1400000 0 0 Feb 3  
## 5 0.02000000 0.0500000 0 0 Feb 3  
## 6 0.01578947 0.0245614 0 0 Feb 2  
## Browser Region TrafficType VisitorType Weekend Revenue  
## 1 1 1 1 Returning\_Visitor FALSE FALSE  
## 2 2 1 2 Returning\_Visitor FALSE FALSE  
## 3 1 9 3 Returning\_Visitor FALSE FALSE  
## 4 2 2 4 Returning\_Visitor FALSE FALSE  
## 5 3 1 4 Returning\_Visitor TRUE FALSE  
## 6 2 1 3 Returning\_Visitor FALSE FALSE

**Get number of rows and columns**

nrow(shoppers)

## [1] 12330

ncol(shoppers)

## [1] 18

The dataset has a total of 12330 entries and 18 columns

**Column Definition** “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

Bounce rate feature for a web page refers to the percentage of visitors who enter the site from that page and then leave (“bounce”) without triggering any other requests to the analytics server during that session.

Exit rate is the percentage of site exits that occurred from a specified page or set of pages.

Page Value feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

Special Day feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine’s Day) in which the sessions are more likely to be finalized with the transaction

Month refers to the month of the web visit

Region means the demographic location

Visitor type refers to either a returning visitor or a new vivitor

##Getting the datatypes

sapply(shoppers, class)

## 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" "factor" "integer"   
## Browser Region TrafficType   
## "integer" "integer" "integer"   
## VisitorType Weekend Revenue   
## "factor" "logical" "logical"

All the variables have the appropriate datatype hence we proceed with our analysis.

## Data Cleaning

*Check for null values*

sum(is.na(shoppers))

## [1] 112

There are 112 missing values

*Delete the missing values*

online\_shop <- na.omit(shoppers)  
sum(is.na(online\_shop))

## [1] 0

We decided to delete the null values since we has a huge dataset and that wouldn’t affect our analysis.

*Check for duplicates*

any(duplicated(online\_shop))

## [1] TRUE

Duplicate records increase computation time and decrease model accuracy, and hence must be removed

*Removing duplicated records*

library(dplyr)

##   
## Attaching package: 'dplyr'

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

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

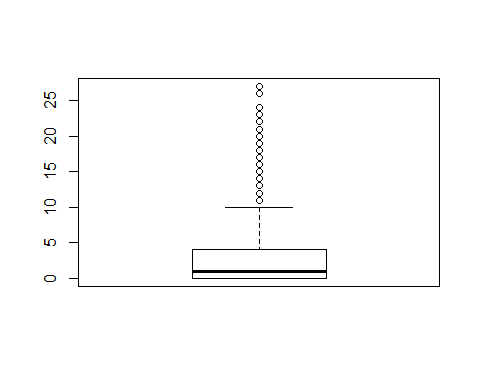
new\_data <- distinct(online\_shop)  
any(duplicated(new\_data))

## [1] FALSE

The duplicates have been removed.

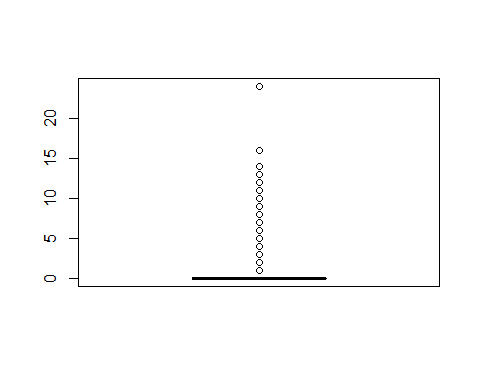
*Check for outliers* This is done on the numerical columns of the dataset

boxplot(new\_data$Administrative, echo=FALSE)



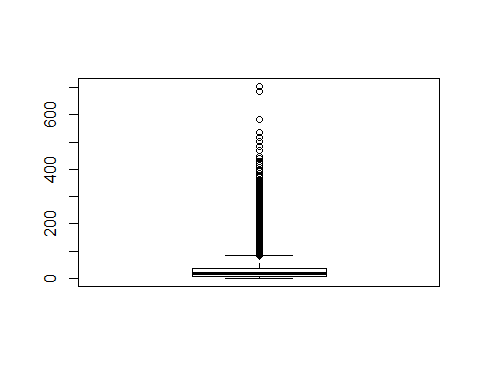
There are outliers in the administrative column but we will not remove them since it is normal for a page to have visits or lacl visits depending on what the customer is intrested in. This also translates to the Administrative\_Duration column

boxplot(new\_data$Informational, echo=FALSE)



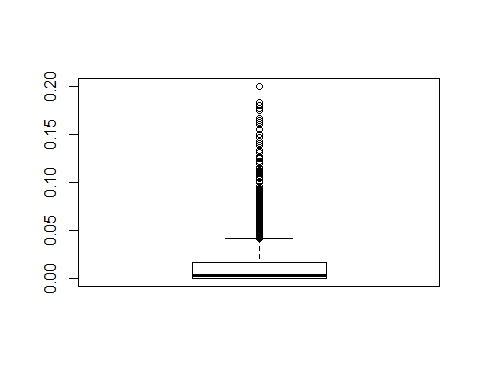
There are also outliers in this particular column as observed above which means that the Informational\_duration column has outliers.

boxplot(new\_data$ProductRelated, echo=FALSE)

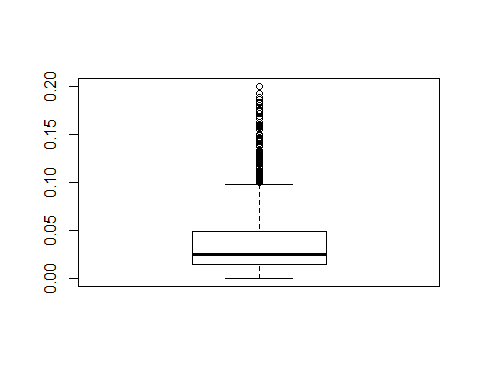


The product related column also has outliers which means that the website has different products projected and hence different total visits on the particular page.

boxplot(new\_data$BounceRates, echo=FALSE)

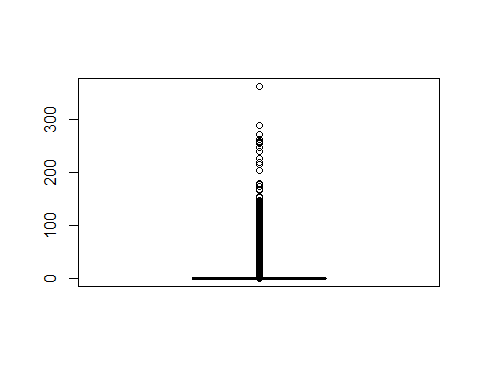
 The bounce rates column has outliers since different pages have different bounce rates which is determined by a customer when he or she exits a page without checking on another page in the website.

boxplot(new\_data$ExitRates, echo=FALSE)



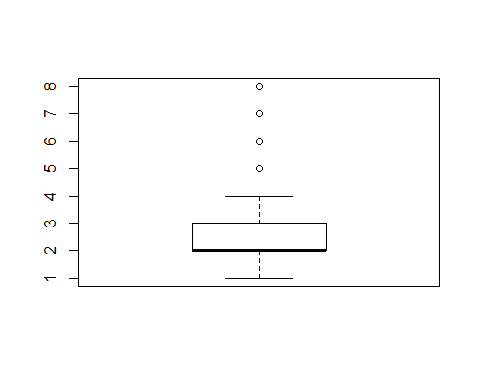
Knowing that exit rate is calculated by considering the page on your site that the visitor left from, that means that different pages have different exit rates hence the outliers.

boxplot(new\_data$PageValues, echo=FALSE)



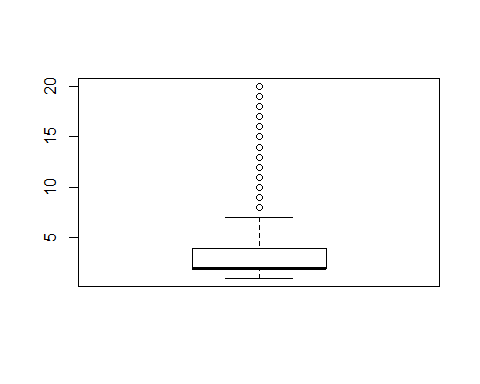
Page Value is the average value for a page that a user visited before landing on the goal page or completing an Ecommerce transaction (or both). This value is intended to give you an idea of which page in your site contributed more to your site’s revenue. If the page wasn’t involved in an ecommerce transaction for your website in any way, then the Page Value for that page will be $0 since the page was never visited in a session where a transaction occurred. Hence the outliers in the plot above.

boxplot(new\_data$OperatingSystems, echo=FALSE)



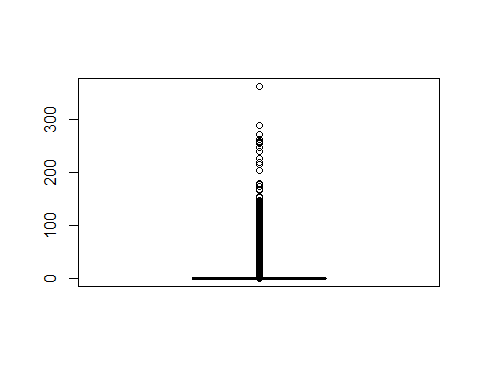
There are outliers in the operating systems column

boxplot(new\_data$TrafficType, echo=FALSE)



The traffic type column has outliers as observed above.

boxplot(new\_data$PageValues, echo=FALSE)



**Get a summary of the dataset provided**

summary(new\_data)

## Administrative Administrative\_Duration Informational   
## Min. : 0.00 Min. : -1.00 Min. : 0.0000   
## 1st Qu.: 0.00 1st Qu.: 0.00 1st Qu.: 0.0000   
## Median : 1.00 Median : 9.00 Median : 0.0000   
## Mean : 2.34 Mean : 81.68 Mean : 0.5088   
## 3rd Qu.: 4.00 3rd Qu.: 94.75 3rd Qu.: 0.0000   
## Max. :27.00 Max. :3398.75 Max. :24.0000   
##   
## Informational\_Duration ProductRelated ProductRelated\_Duration  
## Min. : -1.00 Min. : 0.00 Min. : -1.0   
## 1st Qu.: 0.00 1st Qu.: 8.00 1st Qu.: 193.6   
## Median : 0.00 Median : 18.00 Median : 609.5   
## Mean : 34.84 Mean : 32.06 Mean : 1207.5   
## 3rd Qu.: 0.00 3rd Qu.: 38.00 3rd Qu.: 1477.6   
## Max. :2549.38 Max. :705.00 Max. :63973.5   
##   
## BounceRates ExitRates PageValues SpecialDay   
## Min. :0.00000 Min. :0.00000 Min. : 0.000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.01422 1st Qu.: 0.000 1st Qu.:0.00000   
## Median :0.00293 Median :0.02500 Median : 0.000 Median :0.00000   
## Mean :0.02045 Mean :0.04150 Mean : 5.952 Mean :0.06197   
## 3rd Qu.:0.01667 3rd Qu.:0.04848 3rd Qu.: 0.000 3rd Qu.:0.00000   
## Max. :0.20000 Max. :0.20000 Max. :361.764 Max. :1.00000   
##   
## Month OperatingSystems Browser Region   
## May :3328 Min. :1.000 Min. : 1.000 Min. :1.000   
## Nov :2983 1st Qu.:2.000 1st Qu.: 2.000 1st Qu.:1.000   
## Mar :1853 Median :2.000 Median : 2.000 Median :3.000   
## Dec :1706 Mean :2.124 Mean : 2.358 Mean :3.153   
## Oct : 549 3rd Qu.:3.000 3rd Qu.: 2.000 3rd Qu.:4.000   
## Sep : 448 Max. :8.000 Max. :13.000 Max. :9.000   
## (Other):1332   
## TrafficType VisitorType Weekend   
## Min. : 1.000 New\_Visitor : 1693 Mode :logical   
## 1st Qu.: 2.000 Other : 81 FALSE:9343   
## Median : 2.000 Returning\_Visitor:10425 TRUE :2856   
## Mean : 4.075   
## 3rd Qu.: 4.000   
## Max. :20.000   
##   
## Revenue   
## Mode :logical   
## FALSE:10291   
## TRUE :1908   
##   
##   
##   
##

Comparing all the page views, The Product related page tops with most views(705) and Top duration which means people spent more time on these pages hence the brand’s Sales and Marketing team should consider making easy acces to such pages. We also observe that both the bounce rate and exit rate hold the same percentage of 20%. The page with the highest page value was at 361.76. The month of May topped the list as been the month when people visited to site most. The visitorType column shows that Returning visitors where more on the website which means the clients either loved the products in place hence the company should consider having new trends in the market. The weekend column also tells us that most people visited the website during the weekends.

## Exploratory Data Analysis

**a) Administrative\_Duration Column** *Mean*

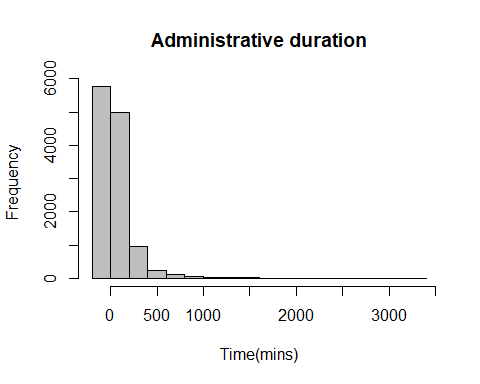
mean(new\_data$Administrative\_Duration)

## [1] 81.68214

The average time spent on the administrative page was 81.68 mins

**Use a histogram to visualize the distribution**

hist(new\_data$Administrative\_Duration, main = "Administrative duration", xlab = "Time(mins)", col = "gray")



The above scale of the timeframe runs from 0-3500 because it is over a period of 1 month. We observe that most people did not concentrate on the administrative column hence the high frequency of zero(0) Another group of customers spent less than 500 mins per month in the administrative page with very few people spending more that 500 mins on the same page.

**b) Informational\_Duration** *Mean*

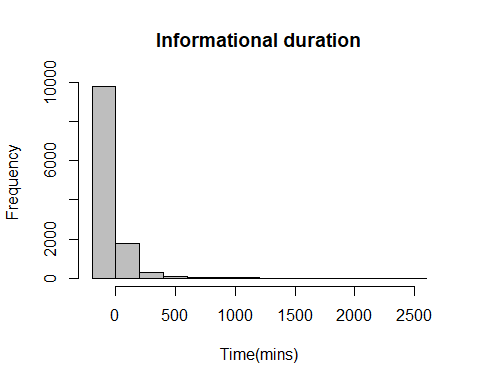
mean(new\_data$Informational\_Duration)

## [1] 34.83734

On average people spent 34.84 mins on the information page.

**A visualization of the Information\_Duration column**

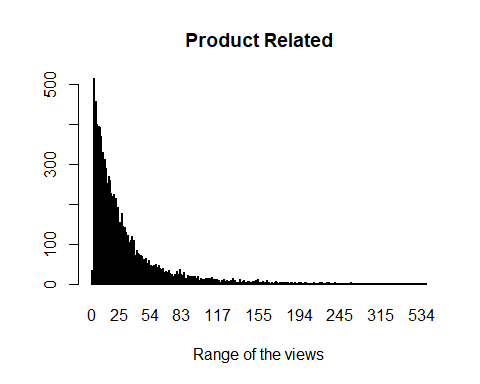
hist(new\_data$Informational\_Duration, main = "Informational duration", xlab = "Time(mins)", col = "gray")



The histogram shows that there are -ve values which is an anomality will will be corrected while doing modelling by scaling the data. All the customers that visited the specific page during that month spent less that 900 mins on the same page.

**c) ProductRelated**

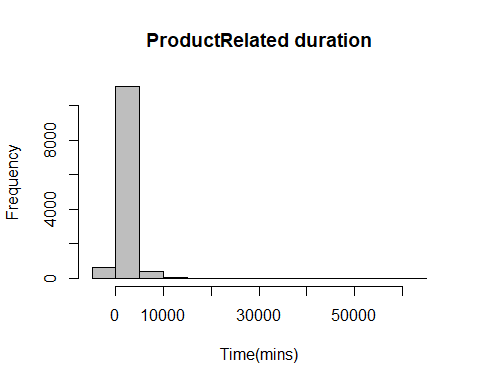
Products <- new\_data$ProductRelated  
Products\_freq <- table(Products)  
barplot(Products\_freq, main = "Product Related", xlab = "Range of the views")



The plot above shows that the views progressively reduced as you move across the graph.

**d) ProductRelated\_Duration**

hist(new\_data$ProductRelated\_Duration, main = "ProductRelated duration", xlab = "Time(mins)", col = "gray")



The data is distributed from -1 to 21857 hence the big scale on the x-axis The bigger group of cuatomers who visited the site spent less that 10,000 mins on the product.

**e) Bounce rates column** Get the highest bounce rate

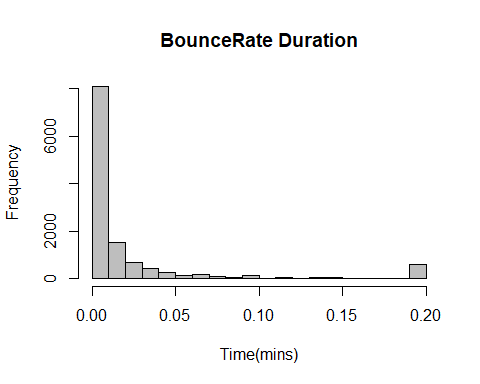
max(new\_data$BounceRates)

## [1] 0.2

The maximum bounce rate was 20%

**A visualization of the Bouncerate Column**

hist(new\_data$BounceRates, main = "BounceRate Duration", xlab = "Time(mins)", col = "gray")



The bounce rate of the website was distributed from 0.00 to 0.20 where the different pages had a bounce rate but not as high.

**f) ExitRates** Check for the highest exit rate

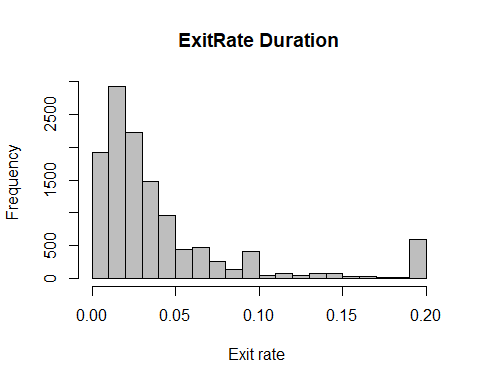
max(new\_data$ExitRates)

## [1] 0.2

A 20% exit rate was the highest

**A visualization of the Exitrates Column**

hist(new\_data$ExitRates, main = "ExitRate Duration", xlab = "Exit rate ", col = "gray")



The exit rate is skewed to the left of the graph which means that the exit rate of different pages was mostly between 0.00-0.06%

**g)PageValues Column** *Mean*

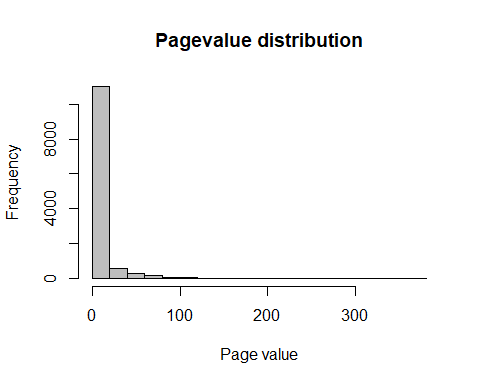
mean(new\_data$PageValues)

## [1] 5.9525

The average page value projected was 5.95

**Plotting the distribution of the Pagevalue column**

hist(new\_data$PageValues, main = "Pagevalue distribution", xlab = "Page value", col = "gray")



Page Value is how you measure the monetary performance of your web pages in Google Analytics. From the distribution above, we observe that the pave value was in a range of 0-362.

**h) SpecialDay Column**

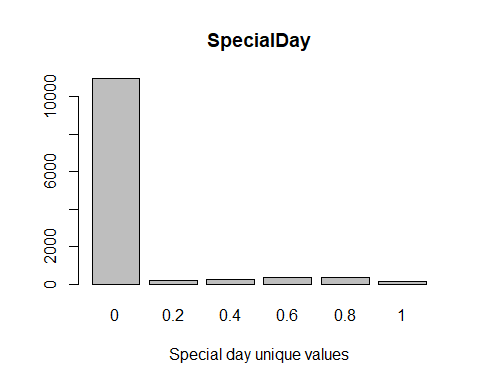
unique(new\_data$SpecialDay)

## [1] 0.0 0.4 0.8 1.0 0.2 0.6

SpecialDay indicates the closeness of the site visiting time to a specific special day.

**A barplot to represent the distribution of the data**

specialday <- new\_data$SpecialDay  
special\_freq <- table(specialday)  
barplot(special\_freq, main = "SpecialDay", xlab = "Special day unique values")



We observe that the special day values were in a range of 0-10. The value was 0 before and after the special day and in accordance to the data, most visits where at off special days.

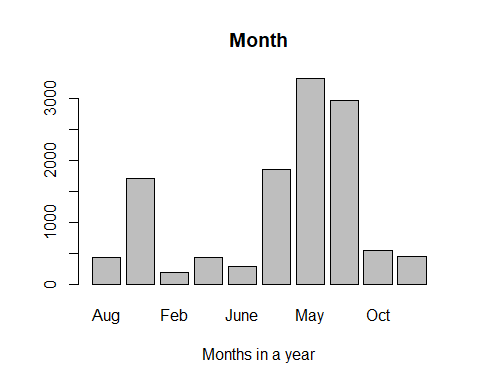
**i) Month Column** Get the unique values represented in the dataset

unique(new\_data$Month)

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

**Use a barplot to visualize**

months <- new\_data$Month  
months\_freq <- table(months)  
barplot(months\_freq, main = "Month", xlab = "Months in a year")



May led the list with most visits in the website followed by November, March and December respectively. With may being psrt of the Spring season, the brand’s Sales and Marketing team should consider having more products at such a time and also have enticing brands represented. The rest of the months had few visits.

**j) Browser Column**

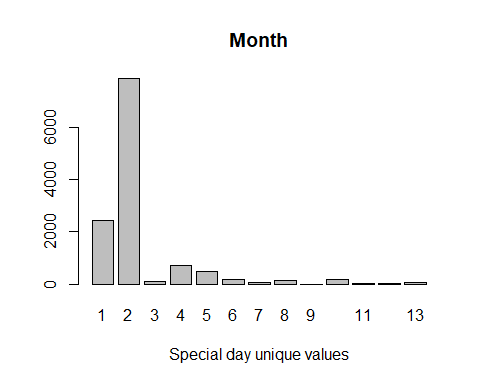
unique(new\_data$Browser)

## [1] 1 2 3 4 5 6 7 10 8 9 12 13 11

There were 13 different browsers represented in the data

**Use a barplot to visualize**

browsers <- new\_data$Browser  
browsers\_freq <- table(browsers)  
barplot(browsers\_freq, main = "Month", xlab = "Special day unique values")



A browser is an application used to access and view websites. From the different types of websites provided above, close to 7,000 people who visited the site used website type 2 followed by type one. Few people accessed the wibesite using the other types of browsers.

**k) Region Column**

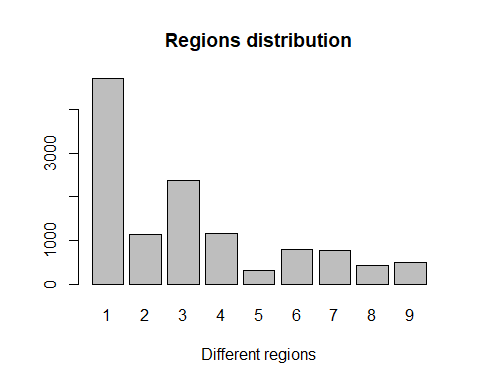
unique(new\_data$Region)

## [1] 1 9 2 3 4 5 6 7 8

From the dataset provided, the website visitors were from 9 different regions

**Use a barplot to visualize**

regions <- new\_data$Region  
region\_freq <- table(regions)  
barplot(region\_freq, main = "Regions distribution", xlab = "Different regions")



Most website visitors over 4000 were from region 1 followed by region 3 with a balance observed in regions 2 and 4. Regions 5-9 had less than 1000 people visit the website.

**l) TrafficType Column**

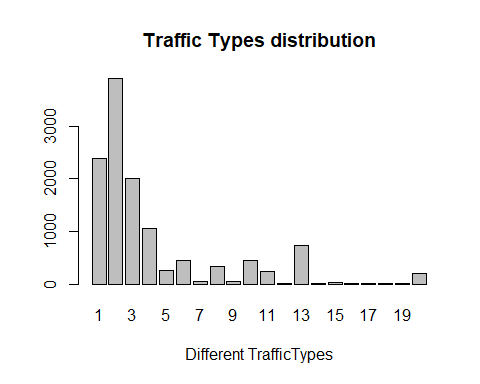
unique(new\_data$TrafficType)

## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 18 19 16 17 20

A traffic type is a particular identifier type for any hierarchy of your customer base. Traffic types in Split are completely customizable and can be any database key you choose to send to Split hence the range of 1-20

**Use a barplot to visualize**

Traffictype <- new\_data$TrafficType  
traffictypes\_freq <- table(Traffictype )  
barplot(traffictypes\_freq, main = "Traffic Types distribution", xlab = "Different TrafficTypes")



Traffic type 2 division had most visitors under it followed by type 1,3,and 4 respectively.

**m) VisitorType Column**

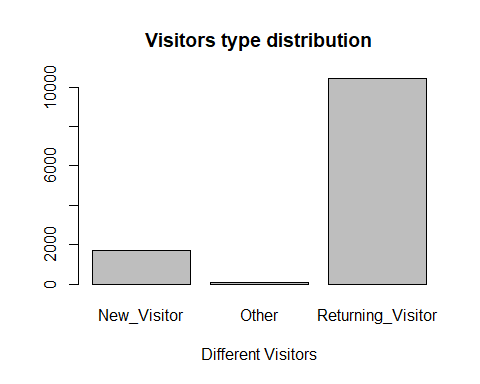
unique(new\_data$VisitorType)

## [1] Returning\_Visitor New\_Visitor Other   
## Levels: New\_Visitor Other Returning\_Visitor

There were 3 unique classes of visitors. New Visitors are those navigating to your site for the first time on a specific device while Returning Visitors have visited your site before and are back for more.

**Use a barplot to visualize**

Visitors <- new\_data$VisitorType  
visitorstypes\_freq <- table(Visitors)  
barplot(visitorstypes\_freq, main = "Visitors type distribution", xlab = "Different Visitors")



Over 10,000 peope were returning visitors which is a good thing for the company showing the effectiveness of the inbound digital marketing techniques across the web. The company should purpose to improve the number of new visitors either by offering better products with better prices to improve their sales.

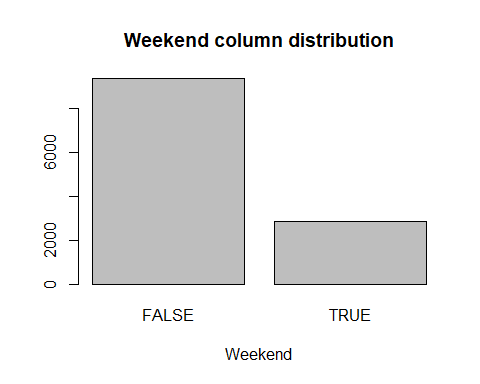
**n) Weekend Column**

unique(new\_data$Weekend)

## [1] FALSE TRUE

This shows that the visits to the website was either during weekends or otherdays. **Use a barplot to visualize**

weekends <- new\_data$Weekend  
weekend\_freq <- table(weekends)  
barplot(weekend\_freq, main = "Weekend column distribution", xlab = "Weekend")



Most visits to the website happened off weekends according to the visulaization above. About 3,000 people visisted the site over the weekend.

**O) Revenue Column**

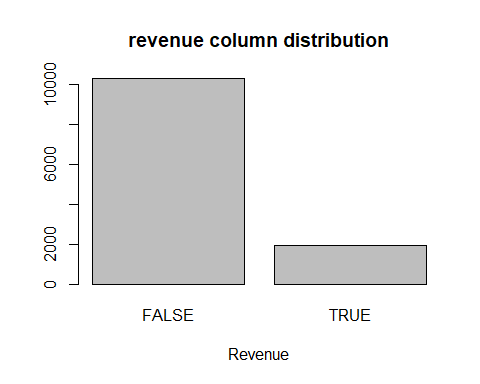
unique(new\_data$Revenue)

## [1] FALSE TRUE

This shows that a page either mad revenue or not

**Use a barplot to visualize**

revenue <- new\_data$Revenue  
revenue\_freq <- table(revenue)  
barplot(revenue\_freq, main = "revenue column distribution", xlab = "Revenue")



Most visits did not go to the point of making reveunue in accordance to the above visualization. About 2000 visits to the different pages made money for the company.

## Bivariate Analysis

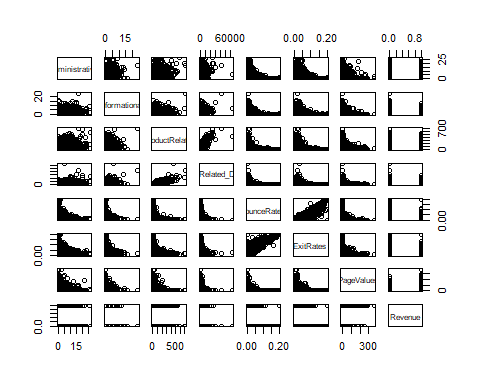
**Get an overview of the correlation between the numerical variables**

cov(new\_data[,-c(0,2,4,10,11,12,13,14,15,16,17)])

## Administrative Informational ProductRelated  
## Administrative 11.09456996 1.594806280 63.6117036  
## Informational 1.59480628 1.627709681 21.2021821  
## ProductRelated 63.61170357 21.202182071 1989.2412959  
## ProductRelated\_Duration 2372.71642208 945.703033133 73668.6330189  
## BounceRates -0.03231259 -0.006343127 -0.3918681  
## ExitRates -0.04794942 -0.009414909 -0.5902590  
## PageValues 6.02328225 1.128072018 45.0324519  
## Revenue 0.16479971 0.043383040 2.5264373  
## ProductRelated\_Duration BounceRates  
## Administrative 2372.71642 -0.032312588  
## Informational 945.70303 -0.006343127  
## ProductRelated 73668.63302 -0.391868084  
## ProductRelated\_Duration 3686121.49674 -15.200227238  
## BounceRates -15.20023 0.002061387  
## ExitRates -21.78350 0.001896814  
## PageValues 1821.19283 -0.098258013  
## Revenue 104.59767 -0.002397839  
## ExitRates PageValues Revenue  
## Administrative -0.047949418 6.02328225 0.164799707  
## Informational -0.009414909 1.12807202 0.043383040  
## ProductRelated -0.590258984 45.03245187 2.526437312  
## ProductRelated\_Duration -21.783499809 1821.19282970 104.597665309  
## BounceRates 0.001896814 -0.09825801 -0.002397839  
## ExitRates 0.002138800 -0.14976966 -0.003432087  
## PageValues -0.149769655 348.11318376 3.333606363  
## Revenue -0.003432087 3.33360636 0.131954161

The above matrix shows the correlation coefficient between the different numerical variables in the dataset

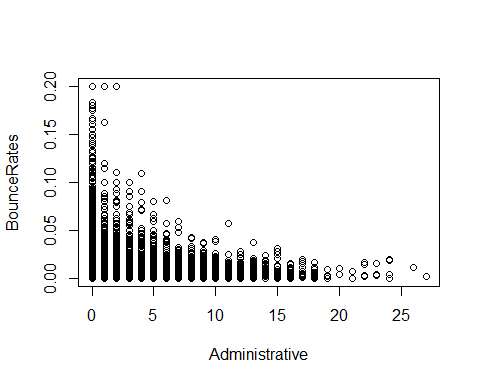
pairs(new\_data[,-c(0,2,4,10,11,12,13,14,15,16,17)])



A plot of the correlation coeffients matrix visualizes the same above. There is a positive correlation between bounce rates Vs exit rates and ProductRelated Vs Product\_duration.

**Administrative Vs BounceRates**

attach(new\_data)  
plot(Administrative, BounceRates)



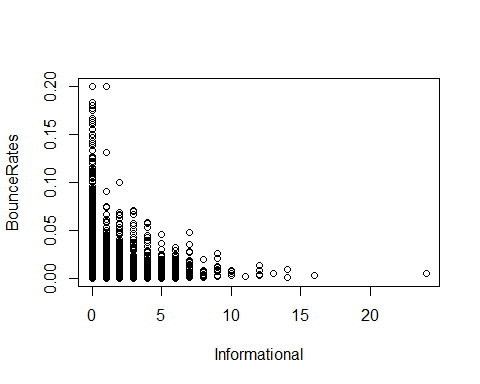
There is a negative correlation between the 2 variables as shown above. As the independent variable(Administrative) increases, the dependent variable(bounce rates) decreases. Increase in the independent variable means increase in the number of views which means a visitor is likely to move to another page in the website which will reduce the bounce rate respectively.

**Informational Vs BounceRates**

attach(new\_data)

## The following objects are masked from new\_data (pos = 3):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

plot(Informational , BounceRates)



There is a negative correlation between the 2 variables which is a good thing for the company in that increase in the Informational column would mean that clients have more intrest about what the company offers leading to naviagtion to another pages hence reducing the bounce rate.

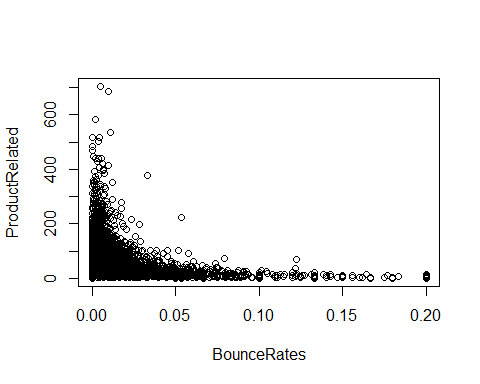
**ProductRelated Vs BounceRates**

attach(new\_data)

## The following objects are masked from new\_data (pos = 3):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 4):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

plot(BounceRates, ProductRelated )



As the product related pages got more views, the bouncerate reduced significantly as seen in the above graph.

**Exitrates Vs Bouncerates**

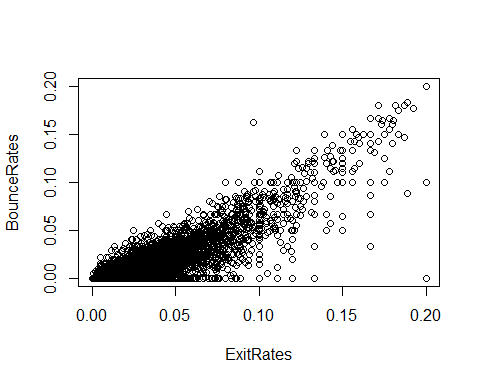
attach(new\_data)

## The following objects are masked from new\_data (pos = 3):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 4):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 5):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

plot(ExitRates , BounceRates)



Bounce Rate: the percentage of single-engagement sessions Exit Rate: the percentage of exits on a page There is a positive linear relationship between the 2 variables. The company should purpose to keep such rates low.

**Exitrates Vs Bouncerates**

attach(new\_data)

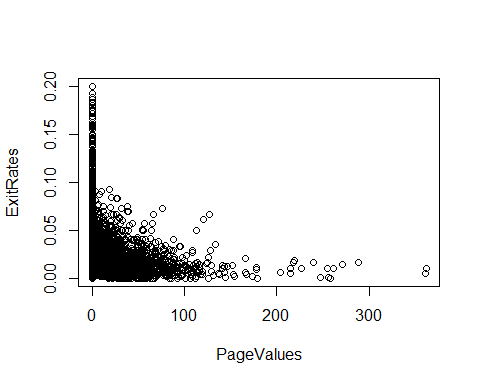
## The following objects are masked from new\_data (pos = 3):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 4):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 5):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 6):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

plot(PageValues , ExitRates)



There is a negative relationship between the 2 variables in that an increase in one variables affects the other variable negatively. Exit Rate is the percentage of exits on a page. This means that increase in the exit rate affects the pagevalue negatively in that it affects the monetary peformance of the different pages.

**Exitrates Vs Bouncerates**

attach(new\_data)

## The following objects are masked from new\_data (pos = 3):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

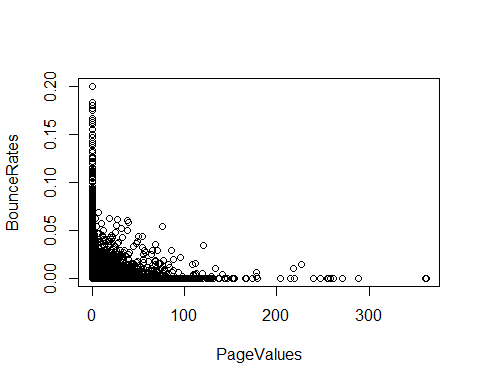
## The following objects are masked from new\_data (pos = 4):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 5):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 6):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

## The following objects are masked from new\_data (pos = 7):  
##   
## Administrative, Administrative\_Duration, BounceRates, Browser,  
## ExitRates, Informational, Informational\_Duration, Month,  
## OperatingSystems, PageValues, ProductRelated,  
## ProductRelated\_Duration, Region, Revenue, SpecialDay,  
## TrafficType, VisitorType, Weekend

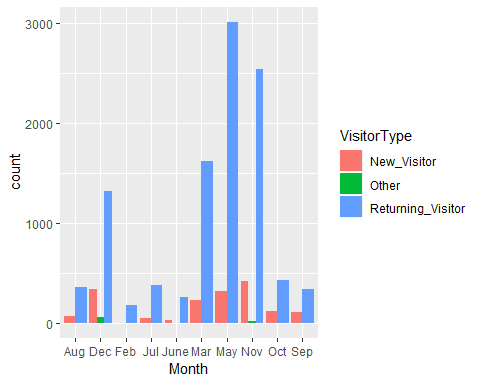
plot(PageValues , BounceRates)



A negative relatioship is observed in the above plot which means that an incease in the bounce rates affects the page value negatively. The company should therefore make sure that the bounce rate is kept low by checking the factors that lead to the incease in the rate.

**Month Vs VisitorType**

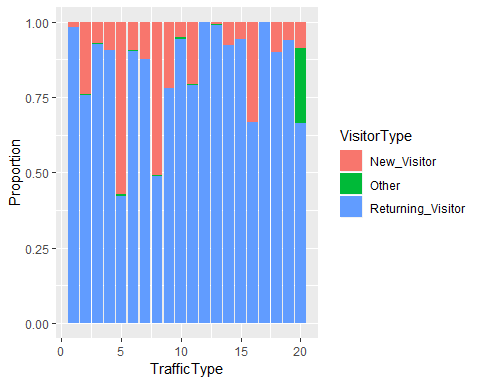
library(ggplot2)  
  
# grouped bar plot  
ggplot(new\_data,   
 aes(x = Month,   
 fill = VisitorType)) +   
 geom\_bar(position = "dodge")



We observe that in all the months, returning visitors were more compared to new visitors. November had abit more new visitors compared to the rest of the months. The brand’s Sales and Marketing team should aim to keep the returning visitors to avoid any churning at any point.

**TrafficType Vs VisitorsType**

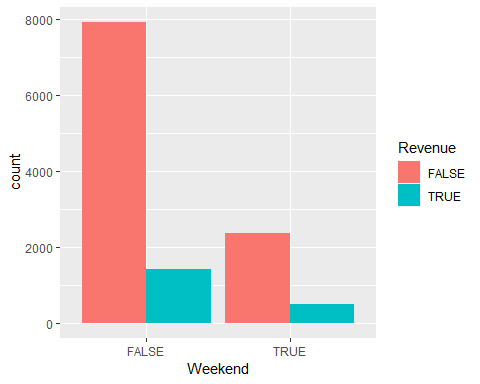
library(ggplot2)  
  
ggplot(new\_data,   
 aes(x = TrafficType,   
 fill = VisitorType)) +   
 geom\_bar(position = "fill") +  
 labs(y = "Proportion")



In the range(0-20) of the traffic types, a bigger prortion of the traffic was brought by the returning visitors. Thi means that the company is offering good services in the sense of the different web applications concerned. For traffic type 5, new visitors occupied a bigger proportion of the traffic.

**Revenue Vs Weekend**

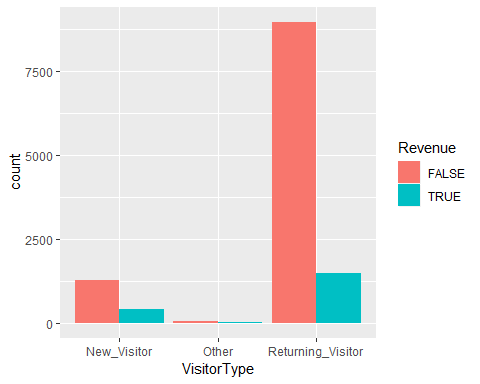
ggplot(new\_data,   
 aes(x = Weekend,   
 fill = Revenue)) +   
 geom\_bar(position = "dodge")



We observe that for the visits that happened off weekends, the pages where not income generating. Less than 2000 visits led to meetingthe company goal of earning revenue by buying a product. For the visits that happened during the weekend, pages viewed did not also bring in any revenue. Since for both visits on both weekends and off weekends did not bring in any revenue, the company should make sure bounce rate and exit rates are managed by making sure that the Url runs well and that there are no unnecessary exits from the different paes before getting to either payment section or even putting an item in the cart.

**VisitorType Vs Revenue**

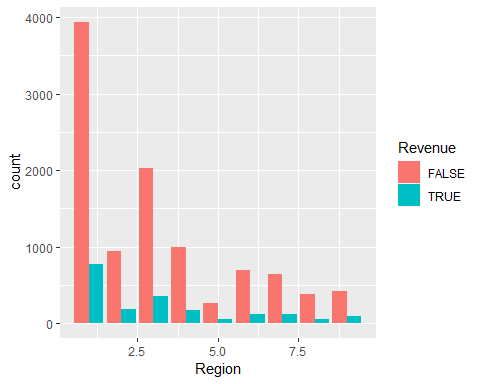
ggplot(new\_data,   
 aes(x = VisitorType,   
 fill = Revenue)) +   
 geom\_bar(position = "dodge")



We observe that the returning visitors but still the revenue acquired was still low. This could mean that the visitors were not exploring on new products hence no purcases. On the new viisitors side, the revenue making views where not optimized which means that the company should have insentives to attract the new customers to making new purchases hence increasing the revenue.

**Region Vs Revenue**

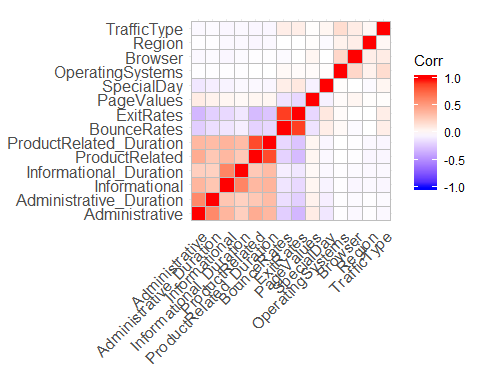
ggplot(new\_data,   
 aes(x = Region,   
 fill = Revenue)) +   
 geom\_bar(position = "dodge")



Across all the regions in the dataset, different site visits where not income generating. This could mean that it’s either there where high bounce and exits rates. This means that the customers did not make to the final goal page.

## Multivariate Analysis

library(ggcorrplot)  
library(dplyr)  
new\_data %>%  
 select\_if(is.numeric) %>%  
 cor %>%   
 ggcorrplot()



There is a positive correlation betwen the various website pages. We also observe a postitive relationship between productRelated Vs ProductRelated\_Duration A positive correlation is observed between the counce rates and exit rates

##Implementing the solution ### K-Means Clustering K-Means is a clustering algorithm used for partitioning a dataset into a set of k-clusters, where k represents the number of gropus pre-specified. The clusters(groups) of observations share similar characteristics The algorithm finds the groups by minimizing the distance between the observations(optimal solutions)

\*Preview the dataset again\*\*

head(new\_data)

## Administrative Administrative\_Duration Informational  
## 1 0 0 0  
## 2 0 0 0  
## 3 0 -1 0  
## 4 0 0 0  
## 5 0 0 0  
## 6 0 0 0  
## Informational\_Duration ProductRelated ProductRelated\_Duration  
## 1 0 1 0.000000  
## 2 0 2 64.000000  
## 3 -1 1 -1.000000  
## 4 0 2 2.666667  
## 5 0 10 627.500000  
## 6 0 19 154.216667  
## BounceRates ExitRates PageValues SpecialDay Month OperatingSystems  
## 1 0.20000000 0.2000000 0 0 Feb 1  
## 2 0.00000000 0.1000000 0 0 Feb 2  
## 3 0.20000000 0.2000000 0 0 Feb 4  
## 4 0.05000000 0.1400000 0 0 Feb 3  
## 5 0.02000000 0.0500000 0 0 Feb 3  
## 6 0.01578947 0.0245614 0 0 Feb 2  
## Browser Region TrafficType VisitorType Weekend Revenue  
## 1 1 1 1 Returning\_Visitor FALSE FALSE  
## 2 2 1 2 Returning\_Visitor FALSE FALSE  
## 3 1 9 3 Returning\_Visitor FALSE FALSE  
## 4 2 2 4 Returning\_Visitor FALSE FALSE  
## 5 3 1 4 Returning\_Visitor TRUE FALSE  
## 6 2 1 3 Returning\_Visitor FALSE FALSE

##Normalization of data Normalization is the process of reorganizing data in a database so that There is no redundancy of data(all data is stored in only one place) and to make sure data dependencies are logical. We start by choosing the features we will use for analysis We will leave out the “Revenue column” since it will be used as the class label

Get the class label in another variable

class\_lbl <- new\_data["Revenue"]  
colnames(class\_lbl)

## [1] "Revenue"

**Convert the categorical variables to numeric**

new\_data$Month <- as.numeric(new\_data$Month)   
new\_data$VisitorType <- as.numeric(new\_data$VisitorType)   
new\_data$Weekend <- as.numeric(new\_data$Weekend)

**Normalize the data**

**Apply the K-Means Algorithm**

# Applying the K-means clustering algorithm with no. of centroids(k)=3  
  
k\_result<- kmeans(new\_data,3)   
  
# Previewing the no. of records in each cluster  
#   
k\_result$size

## [1] 1973 207 10019

There are 3 clusters which means that the customers have been classified into groups based on their characteristics.

# Get the value of cluster center datapoint value(3 centers for k=3)  
  
k\_result$centers

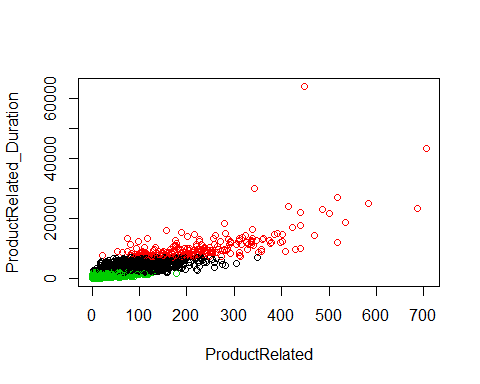
## Administrative Administrative\_Duration Informational  
## 1 4.324886 154.85382 1.1743538  
## 2 7.439614 295.36489 2.7487923  
## 3 1.843797 62.85791 0.3314702  
## Informational\_Duration ProductRelated ProductRelated\_Duration  
## 1 84.73057 77.77293 3356.5135  
## 2 266.72309 236.10628 10886.1744  
## 3 20.22114 18.84030 584.3451  
## BounceRates ExitRates PageValues SpecialDay Month OperatingSystems  
## 1 0.007126288 0.02167382 8.020426 0.05078561 6.498733 2.136341  
## 2 0.005939756 0.01968831 4.521370 0.03091787 6.782609 2.149758  
## 3 0.023369600 0.04585102 5.574840 0.06481685 6.090129 2.121469  
## Browser Region TrafficType VisitorType Weekend Revenue  
## 1 2.311201 3.092752 3.706031 2.902686 0.2184491 0.2508870  
## 2 2.309179 2.584541 3.618357 2.985507 0.2512077 0.3381643  
## 3 2.368400 3.176964 4.156602 2.673421 0.2368500 0.1340453

We see how the the different customers have been clustered to the 3 different clusters.

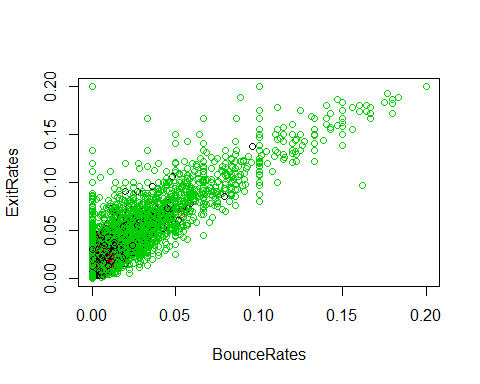
**Plot the distribution of the the clusters**

par(mfrow= c(2,2), mar= c(5,4,2,2))

#Plot the ProductRelated Vs ProductRelated\_Duration Page to see the distribution  
plot(new\_data[c(5,6)], col = k\_result$cluster)



#Plot the BounceRate Vs Exitrate  
plot(new\_data[c(7,8)], col = k\_result$cluster)



## Hierarchical clustering

# We start by computing some descriptive statistics  
desc\_stats <- data.frame(  
 Min = apply(new\_data, 2, min), # minimum  
 Med = apply(new\_data, 2, median), # median  
 Mean = apply(new\_data, 2, mean), # mean  
 SD = apply(new\_data, 2, sd), # Standard deviation  
 Max = apply(new\_data, 2, max) # Maximum  
)  
desc\_stats <- round(desc\_stats, 1)  
head(desc\_stats)

## Min Med Mean SD Max  
## Administrative 0 1.0 2.3 3.3 27.0  
## Administrative\_Duration -1 9.0 81.7 177.5 3398.8  
## Informational 0 0.0 0.5 1.3 24.0  
## Informational\_Duration -1 0.0 34.8 141.5 2549.4  
## ProductRelated 0 18.0 32.1 44.6 705.0  
## ProductRelated\_Duration -1 609.5 1207.5 1919.9 63973.5

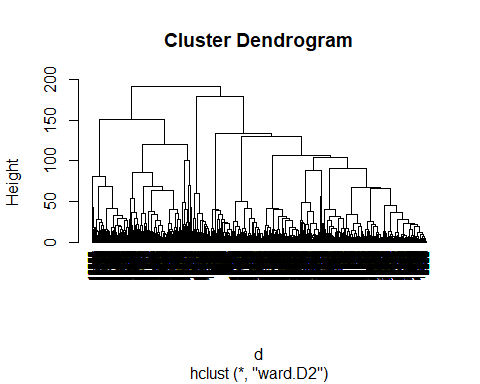
# Scale the data to make it comparable  
#  
new\_data<- scale(new\_data)  
head(new\_data)

## Administrative Administrative\_Duration Informational  
## [1,] -0.7025315 -0.4601081 -0.3988128  
## [2,] -0.7025315 -0.4601081 -0.3988128  
## [3,] -0.7025315 -0.4657410 -0.3988128  
## [4,] -0.7025315 -0.4601081 -0.3988128  
## [5,] -0.7025315 -0.4601081 -0.3988128  
## [6,] -0.7025315 -0.4601081 -0.3988128  
## Informational\_Duration ProductRelated ProductRelated\_Duration  
## [1,] -0.2462725 -0.6963635 -0.6289343  
## [2,] -0.2462725 -0.6739424 -0.5955997  
## [3,] -0.2533417 -0.6963635 -0.6294551  
## [4,] -0.2462725 -0.6739424 -0.6275453  
## [5,] -0.2462725 -0.4945739 -0.3020990  
## [6,] -0.2462725 -0.2927843 -0.5486101  
## BounceRates ExitRates PageValues SpecialDay Month  
## [1,] 3.954699721 3.4273070 -0.3190356 -0.3103105 -1.333953  
## [2,] -0.450343788 1.2650121 -0.3190356 -0.3103105 -1.333953  
## [3,] 3.954699721 3.4273070 -0.3190356 -0.3103105 -1.333953  
## [4,] 0.650917089 2.1299300 -0.3190356 -0.3103105 -1.333953  
## [5,] -0.009839437 0.1838646 -0.3190356 -0.3103105 -1.333953  
## [6,] -0.102577188 -0.3661929 -0.3190356 -0.3103105 -1.333953  
## OperatingSystems Browser Region TrafficType VisitorType  
## [1,] -1.2396607 -0.7939682 -0.8962939 -0.76562243 0.409771  
## [2,] -0.1371074 -0.2093703 -0.8962939 -0.51660683 0.409771  
## [3,] 2.0679992 -0.7939682 2.4336556 -0.26759123 0.409771  
## [4,] 0.9654459 -0.2093703 -0.4800502 -0.01857564 0.409771  
## [5,] 0.9654459 0.3752276 -0.8962939 -0.01857564 0.409771  
## [6,] -0.1371074 -0.2093703 -0.8962939 -0.26759123 0.409771  
## Weekend Revenue  
## [1,] -0.5528638 -0.4305688  
## [2,] -0.5528638 -0.4305688  
## [3,] -0.5528638 -0.4305688  
## [4,] -0.5528638 -0.4305688  
## [5,] 1.8086156 -0.4305688  
## [6,] -0.5528638 -0.4305688

#Find the Euclidean distance between observations using the dist() function  
d <- dist(new\_data, method = "euclidean")

# We then perform hierarchical clustering using the Ward's method  
  
res.hc <- hclust(d, method = "ward.D2" )

# We plot the obtained dendrogram  
  
plot(res.hc, cex = 0.6, hang = -1)



# Conclusion

An overview of the company’s performance is that the company is growing well on the marketting side which is a good thing in the market place. This is as per the low exit and bounce rates(20%) projected by the different pages compared to the Google Analytics figure of above 70% which transaltes to poor performance of the company. The company’s money in is quite low and they should change on the different tactics of handling the website to make sure revenue flows improve.

# Recommendation

The brand’s Sales and Marketing team should make sure that both ExitRates and BounceRates are low enough. This would transalte to customers helping the company meet it’s main goal which is maximising the revenue. With increased pagevalues,that would mean that the company is likely to meet it main focus of making revenue from the online shop in place. The concerned teams should make easy access for customers to the website by reducing the unnecessary ad plays to different pages and also make sure that the loading speed is good to enable a customer move from one page to the other with ease. Based on the 3 clusters from the data, the company should study further each cluster for better understanding and also get to know what incentives to give just incase the company is more money oriented.