Moringa_Data_Science_Prep_W13_Independent_Project_2020_02_Gideon_Cheruiyot_DataReport

Gideon Cheruiyot

06/03/2020

#Problem Definition: Kira Plastinina is an online retail store available for customers in that wants to know the chararcteristics of their customers in Russia, Ukraine, Kazakhstan, Belarus, China, Philippines, and Armenia. In detail, the sales and marketing team would like to know the characteristics of the different customer groups. #Data Undesrtanding *The dataset has 10 numerical and 8 categorical features. 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 value of the "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. *The value of the "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that was the last in the session.* The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction. *The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day) in which the sessions are more likely to be finalized with the transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina's day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.* The dataset also includes the operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year. #The appropriateness of the data will be addressed by checking:

- whether the data is accurate?
- whether the dataset is enough to sufficiently address the problem at hand?
- whether the dtaset was biased or imbalanced? # Experimental design:
- loading and previewing the dataset, checking datatypes
- Cleaning , checking for outliers
- Conducting exploratory data analysis to find out patterns and relationships
- Comunicating the obsertavtions and findings
- Drawing conclusion

Challenging the solution.

#loading the dataset

```
store_df <- read.csv('online_shoppers_intention.csv')

head(store_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
```

#checking the datasets

```
str(store_df)
```

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 -1 0 0 0 -1 -1 0 0 ...
```

```
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 1 2 3 ...
##  $ ProductRelated_Duration: num  0 64 -1 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3
3 3 3 3 3 3 3 3 ...
##  $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3
3 3 3 3 3 ...
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

#Checking for missing values

```
#missmap(store_df)

colSums(is.na(store_df))

##          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
```

#Checking shape of the dataset

```
dim(store_df)

## [1] 12330    18
```

#deleting missing values

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

#summary statistics

```
summary(store_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
##  May    :3363   Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Nov    :2998   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mar    :1894   Median :2.000    Median : 2.000   Median :3.000
##  Dec    :1727   Mean   :2.124    Mean   : 2.358   Mean   :3.148
##  Oct    : 549   3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##  Sep    : 448   Max.   :8.000    Max.   :13.000   Max.   :9.000
##  (Other):1337
##   TrafficType                VisitorType     Weekend          Revenue
##  Min.   : 1.00   New_Visitor      : 1694   Mode :logical   Mode :logical
##  1st Qu.: 2.00   Other            :   85   FALSE:9451      FALSE:10408
##  Median : 2.00   Returning_Visitor:10537   TRUE :2865      TRUE :1908
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
##
```

#Descriptive statistics

```
library(psych)

## Warning: package 'psych' was built under R version 3.6.3

describe(store_df)

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

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

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

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

##                          vars     n    mean      sd median trimmed    mad
min
## Administrative              1 12316    2.32    3.32   1.00    1.63   1.48
0
## Administrative_Duration     2 12316   80.91  176.86   8.00   42.19  11.86
-1
## Informational               3 12316    0.50    1.27   0.00    0.18   0.00
0
## Informational_Duration      4 12316   34.51  140.83   0.00    3.60   0.00
-1
## ProductRelated              5 12316   31.76   44.49  18.00   22.78  19.27
0
## ProductRelated_Duration     6 12316 1196.04 1914.37 599.77  821.41 743.05
-1
## BounceRates                 7 12316    0.02    0.05   0.00    0.01   0.00
0
## ExitRates                   8 12316    0.04    0.05   0.03    0.03   0.02
0
## PageValues                  9 12316    5.90   18.58   0.00    1.30   0.00
0
## SpecialDay                 10 12316    0.06    0.20   0.00    0.00   0.00
0
## Month*                     11 12316    6.16    2.37   7.00    6.35   1.48
1
## OperatingSystems           12 12316    2.12    0.91   2.00    2.06   0.00
1
## Browser                    13 12316    2.36    1.72   2.00    2.00   0.00
1
## Region                     14 12316    3.15    2.40   3.00    2.79   2.97
1
## TrafficType                15 12316    4.07    4.02   2.00    3.22   1.48
1
## VisitorType*               16 12316    2.72    0.69   3.00    2.90   0.00
1
## Weekend                    17 12316     NaN      NA     NA     NaN     NA
Inf
## Revenue                    18 12316     NaN      NA     NA     NaN     NA
Inf
##                             max    range  skew kurtosis    se
## Administrative            27.00    27.00  1.96     4.69  0.03
## Administrative_Duration 3398.75  3399.75  5.61    50.48  1.59
```

```
## Informational              24.00    24.00  4.03    26.89  0.01
## Informational_Duration   2549.38  2550.38  7.57    76.18  1.27
## ProductRelated            705.00   705.00  4.34    31.17  0.40
## ProductRelated_Duration 63973.52 63974.52  7.26   137.03 17.25
## BounceRates                0.20     0.20  2.95     7.75  0.00
## ExitRates                  0.20     0.20  2.15     4.04  0.00
## PageValues               361.76   361.76  6.38    65.53  0.17
## SpecialDay                 1.00     1.00  3.30     9.89  0.00
## Month*                    10.00     9.00 -0.83    -0.37  0.02
## OperatingSystems           8.00     7.00  2.07    10.45  0.01
## Browser                   13.00    12.00  3.24    12.73  0.02
## Region                     9.00     8.00  0.98    -0.15  0.02
## TrafficType               20.00    19.00  1.96     3.48  0.04
## VisitorType*               3.00     2.00 -2.06     2.28  0.01
## Weekend                    -Inf     -Inf    NA       NA    NA
## Revenue                    -Inf     -Inf    NA       NA    NA
```

column names

```
colnames(store_df)
```

```
##  [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"
```

#Dropping columns that we do not need. *Browser,OperatingSystems,OperatingSystems features may not be important in helping us address the problem statement. Thus, we have dropped the features.

```
library(dplyr)
```

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

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

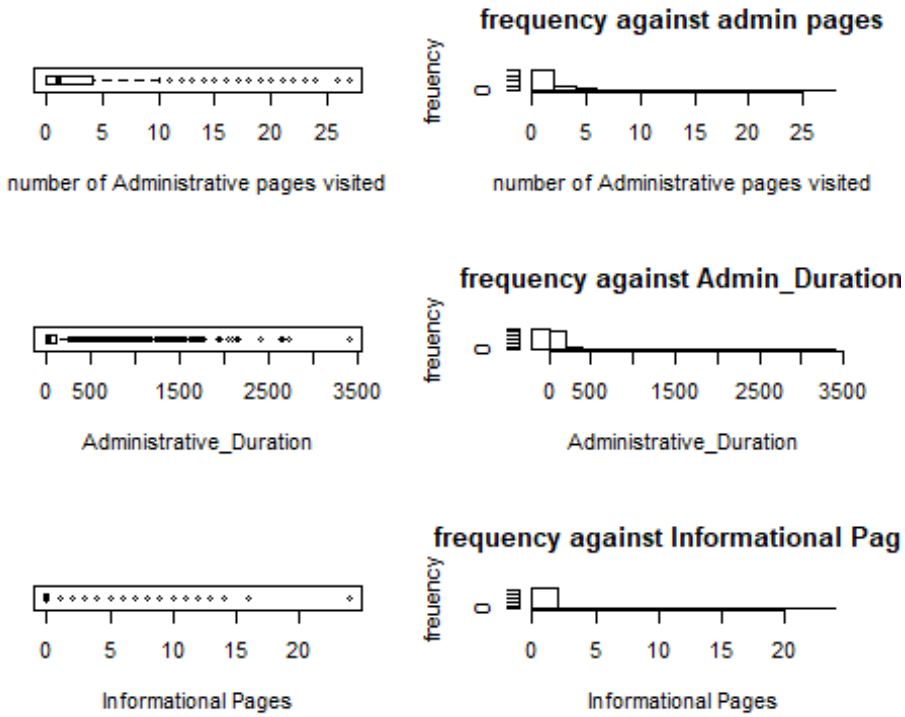
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(base)
library(stats)
 store_df <- select (store_df,-c(Browser,OperatingSystems,OperatingSystems))
dim(store_df)
```

```
## [1] 12316    16
```

#checkin for outliers and Univariate analysis

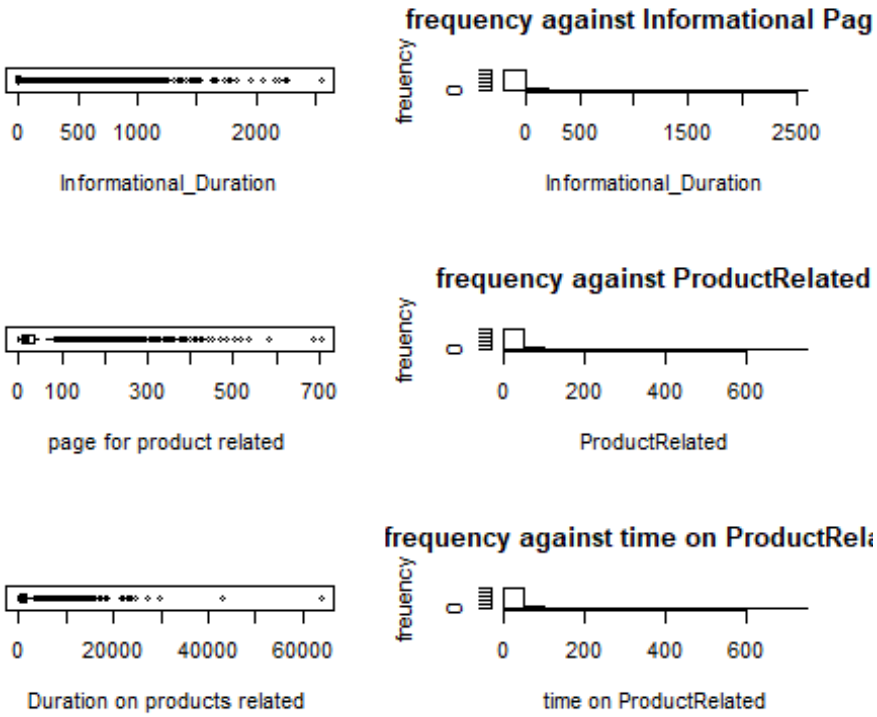
```
par(mfrow=c(3,2))
boxplot(store_df$Administrative ,xlab = 'number of Administrative pages
visited', horizontal=TRUE)
hist(store_df$Administrative , xlab = 'number of Administrative pages
visited', ylab = 'freuency',main = 'frequency against admin pages')
boxplot(store_df$Administrative_Duration ,xlab = 'Administrative_Duration',
horizontal=TRUE)
hist(store_df$Administrative_Duration , xlab = 'Administrative_Duration',
ylab = 'freuency',main = 'frequency against Admin_Duration')
boxplot(store_df$Informational ,xlab = 'Informational Pages',
horizontal=TRUE)
hist(store_df$Informational , xlab = 'Informational Pages', ylab =
'freuency',main = 'frequency against Informational Pages')
```



The most frequented administartive page was between 1 and 2, and users spent an average of 81 seconds were spent o the administrative page. This can help us understand the estimated pages the users visits to login to the the page before navigating to other pages. There are outliers on the both adminidtarive pages and administrative duration. Extreme outliers could be as a results of users who leave there devices unattended to after login in.

```
par(mfrow=c(3,2))
boxplot(store_df$Informational_Duration ,xlab = 'Informational_Duration',
horizontal=TRUE)
hist(store_df$Informational_Duration , xlab = 'Informational_Duration', ylab
```

```
= 'freuency',main = 'frequency against Informational Pages')
boxplot(store_df$ProductRelated ,xlab = 'page for product related ',
horizontal=TRUE)
hist(store_df$ProductRelated , xlab = 'ProductRelated', ylab =
'freuency',main = 'frequency against ProductRelated')
boxplot(store_df$ProductRelated_Duration ,xlab = 'Duration on products
related ', horizontal=TRUE)
hist(store_df$ProductRelated , xlab = 'time on ProductRelated', ylab =
'freuency',main = 'frequency against time on ProductRelated')
```



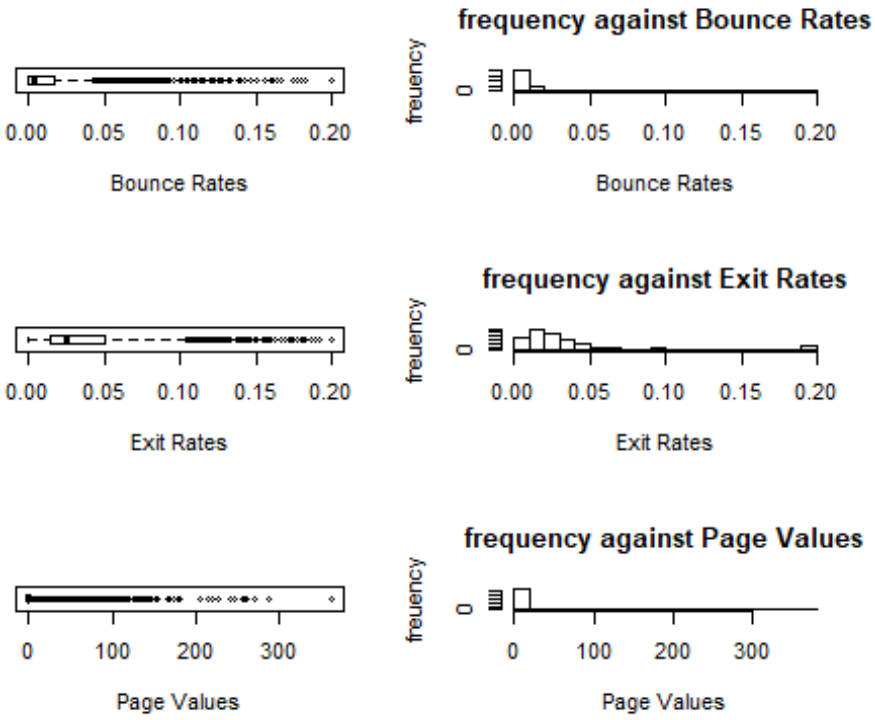
users spent and average of 35 seconds on the informational pages, this could suggest that users are less interested in information and are more aware of what they want or looking for. recommendation* Curate the information to be more appealing to the users.

```
colnames(store_df)
```

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

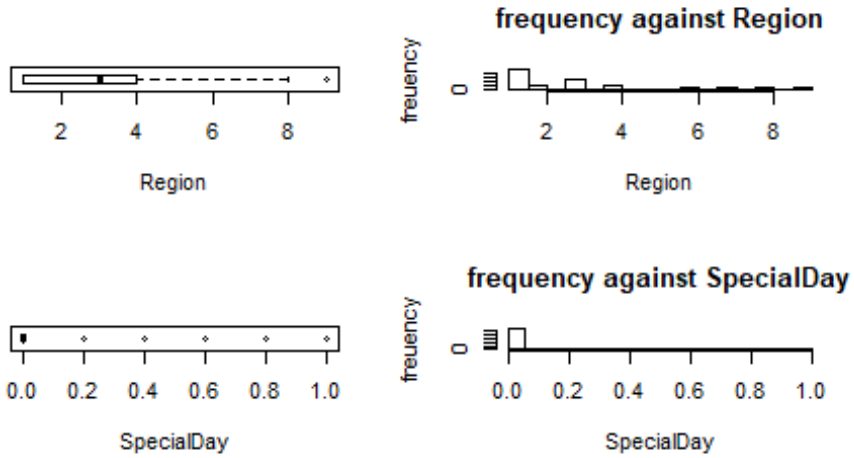
```
par(mfrow=c(3,2))
boxplot(store_df$BounceRates ,xlab = 'Bounce Rates', horizontal=TRUE)
hist(store_df$BounceRates , xlab = 'Bounce Rates', ylab = 'freuency',main =
```

```
'frequency against Bounce Rates')
boxplot(store_df$ExitRates ,xlab = 'Exit Rates', horizontal=TRUE)
hist(store_df$ExitRates , xlab = 'Exit Rates', ylab = 'freuency',main =
'frequency against Exit Rates')
boxplot(store_df$PageValues ,xlab = 'Page Values', horizontal=TRUE)
hist(store_df$PageValues , xlab = 'Page Values', ylab = 'freuency',main =
'frequency against Page Values')
```



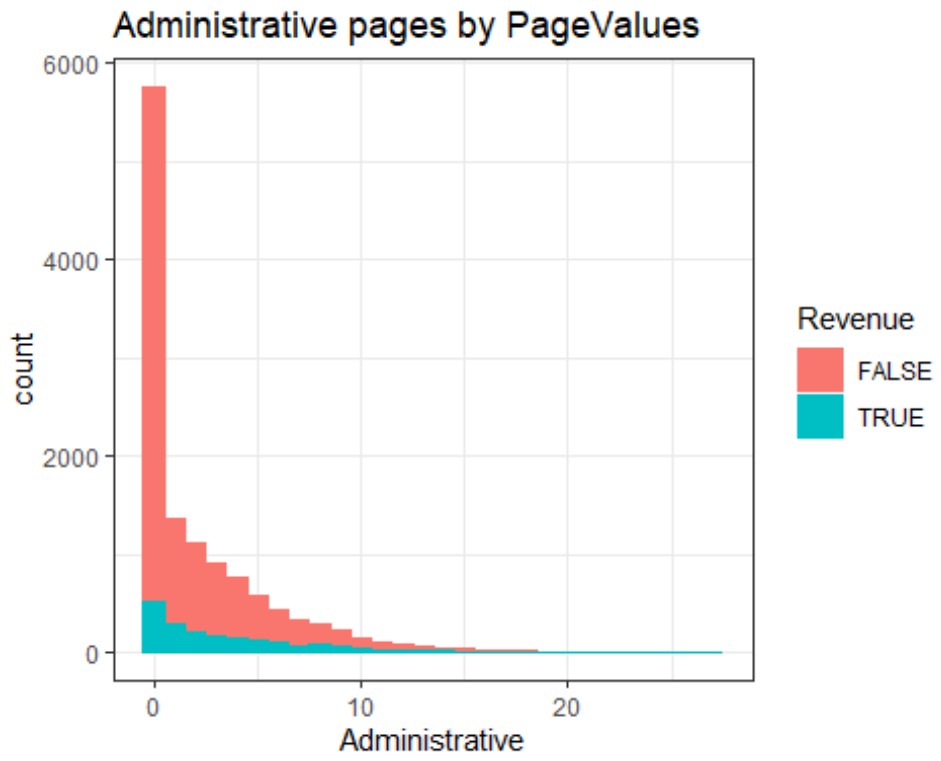
The average bounce rate was below 1%, suggesting that users were interested in the page and did not leave without exploring other pages. On the other hand, the mean of the page value was 6, indicating the users make visit but they do not end up making transcations or purchases. Recommendation* finding out what is hindering the users from making purchases.Reason may range form poor user intereface, less secure payment systems.

```
par(mfrow=c(3,2))
boxplot(store_df$Region ,xlab = 'Region', horizontal=TRUE)
hist(store_df$Region , xlab = 'Region', ylab = 'freuency',main = 'frequency
against Region')
boxplot(store_df$SpecialDay ,xlab = 'SpecialDay', horizontal=TRUE)
hist(store_df$SpecialDay , xlab = 'SpecialDay', ylab = 'freuency',main =
'frequency against SpecialDay')
```

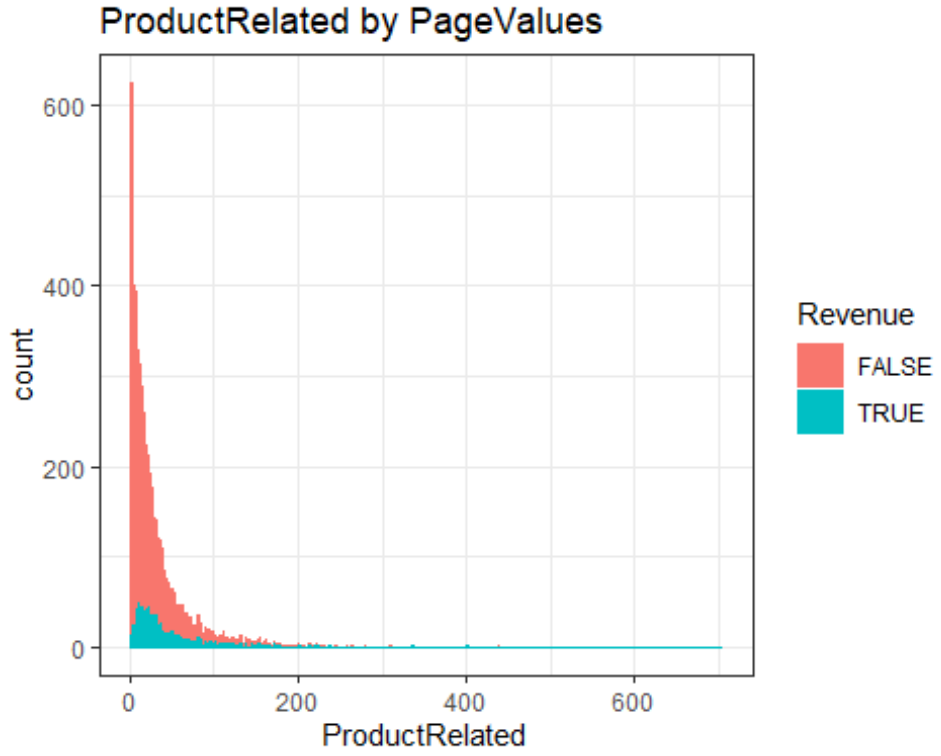


#Bivariate analysis

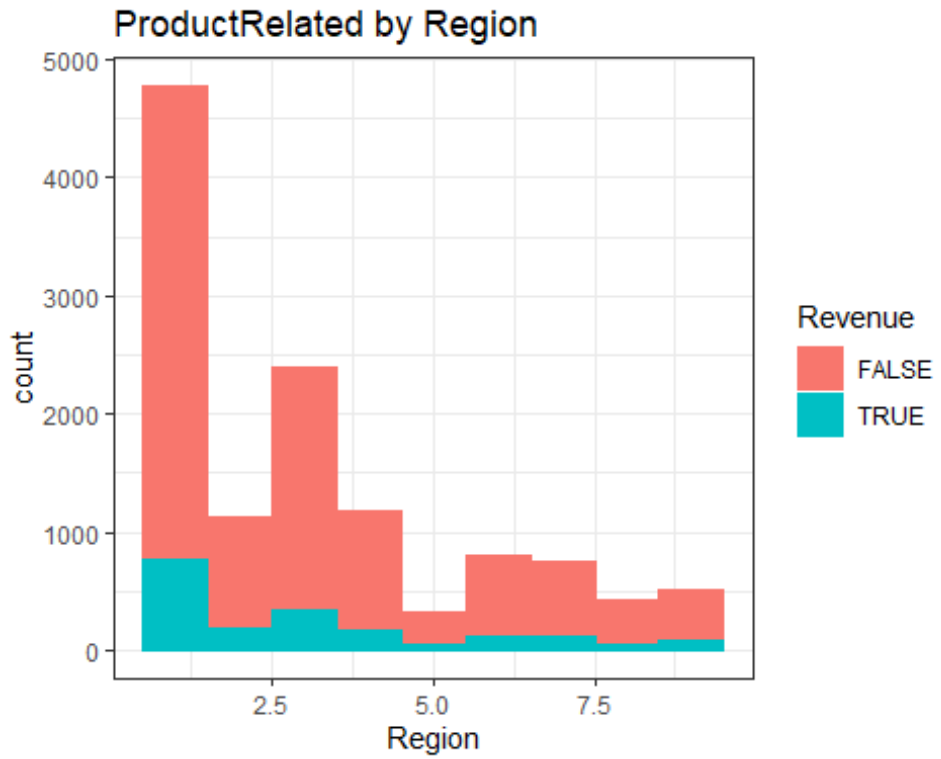
```
library(ggplot2)

##
## Attaching package: 'ggplot2'

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

library(psych)
par(mfrow=c(3,2))
c <- ggplot(store_df, aes(x=Administrative, fill=Revenue, color=Revenue)) +
geom_histogram(binwidth = 1) + labs(title="Administrative pages by
PageValues")
c + theme_bw()
```



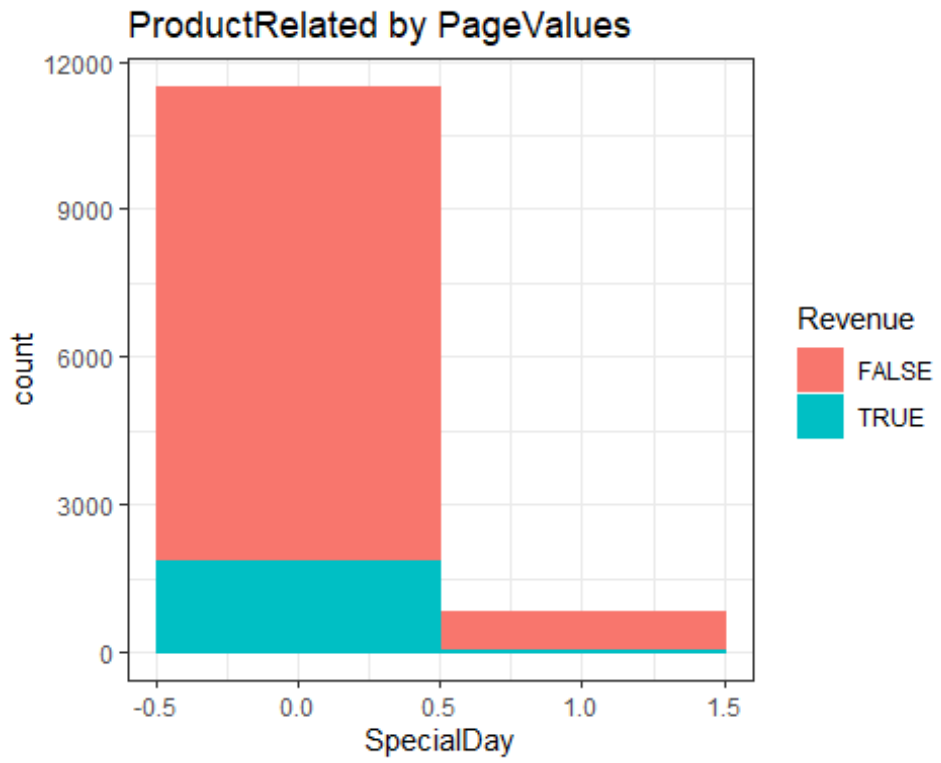

```
c <- ggplot(store_df, aes(x=ProductRelated, fill=Revenue, color=Revenue)) +
geom_histogram(binwidth = 1) + labs(title="ProductRelated by PageValues")
c + theme_bw()
```



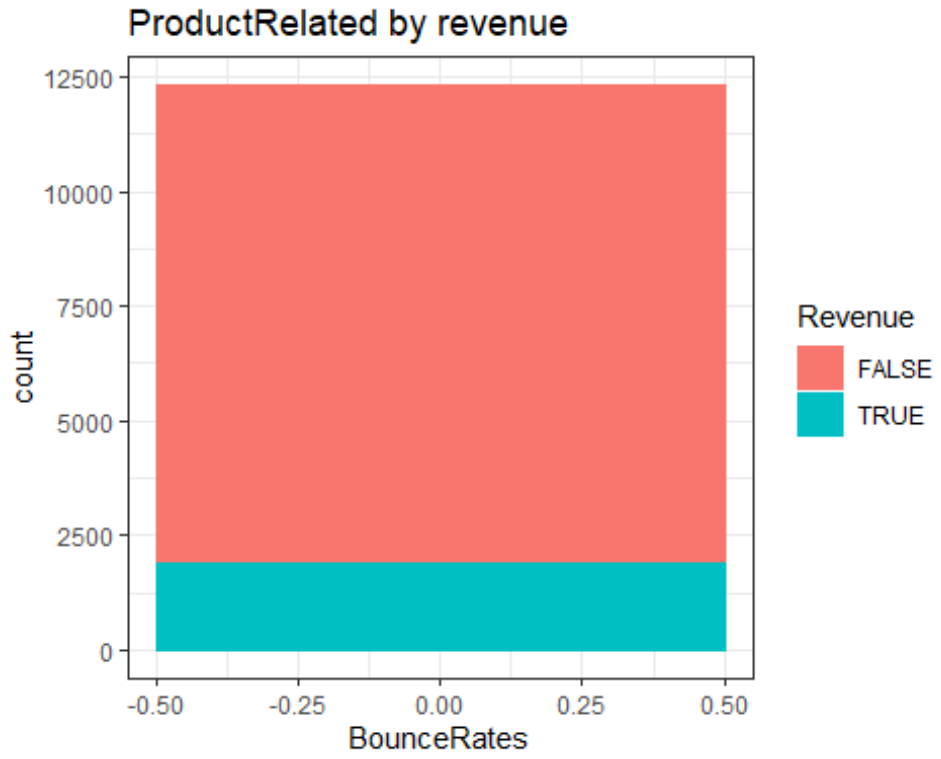
```
p <- ggplot(store_df, aes(x=Region, fill=Revenue, color=Revenue)) +
geom_histogram(binwidth = 1) + labs(title="ProductRelated by Region")
p + theme_bw()
```




```
p <- ggplot(store_df, aes(x=SpecialDay, fill=Revenue, color=Revenue)) +
geom_histogram(binwidth = 1) + labs(title="ProductRelated by PageValues")
p + theme_bw()
```

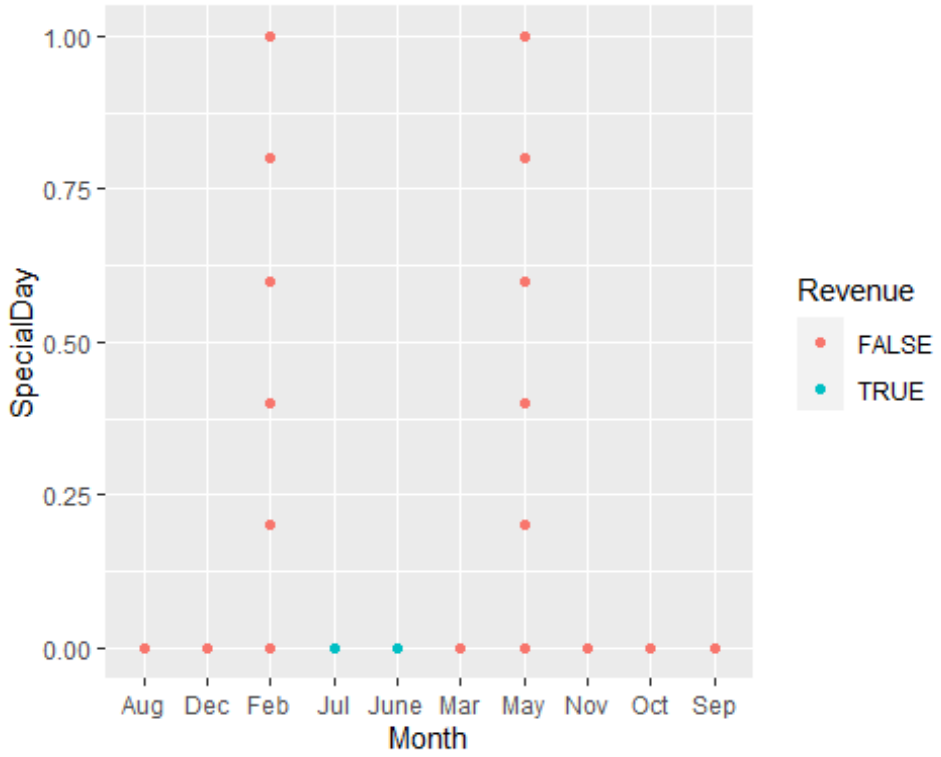



```
p <- ggplot(store_df, aes(x=BounceRates, fill=Revenue, color=Revenue),
StatBin = 10) +
geom_histogram(binwidth = 1) + labs(title="ProductRelated by revenue")
p + theme_bw()
```



Findings There is no relationship between the bounce rates and Revenue. Revenue increases with decrease in the administartive pages and productedRelated pages.Aditional most frequented administrative and ProductRelated pages visited did not generate any revenue *most frequent users were from region 1 and are the users that generated most revenue.

```
ggplot(data = store_df) +
  geom_point(mapping = aes(x =Month  ,  y= SpecialDay , fill = Revenue,
colour = Revenue))
```



* Special days were noticed on the month of Feburary and May. This could be due to valentines and Mothers day on the respective months.

```
colnames(store_df)
```

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

#Multivariate analysis

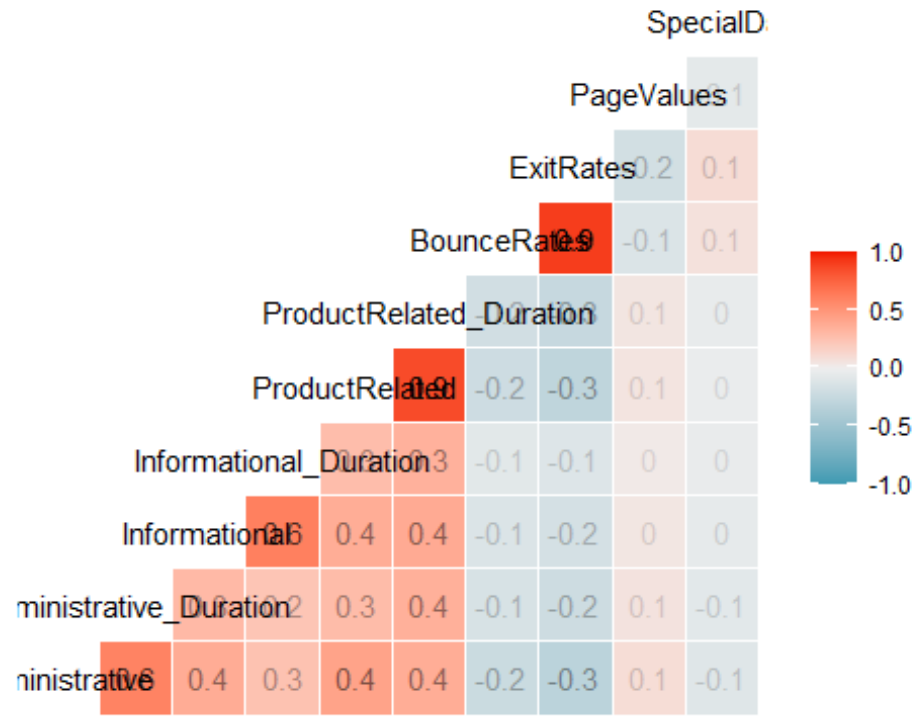
```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa

store_cor <- store_df[, c(1,2,3,4,5,6,7,8,9,10)]
ggcorr(store_cor,
       label = TRUE,
       label_alpha = TRUE)
```

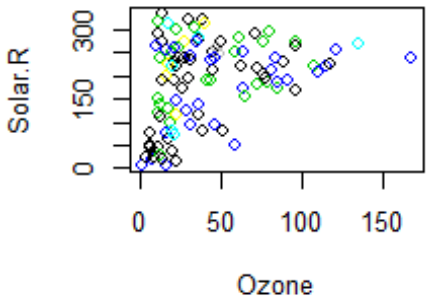


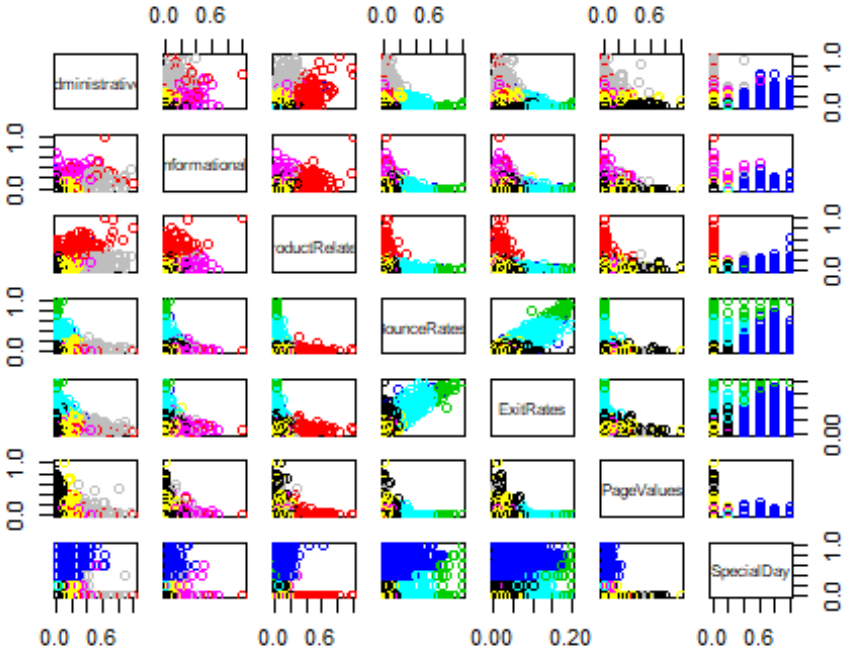
Informational and administrative duration are higly correlated, this could suggest redundancy. special days and exit rate have weak positive correlation to page value. Thus the more and exit rates the more likelihood that user will identify what they like or don't like and be able to intiate transcations on the page. * There is a weak negative correlation between page value and the bounce rates.This suggest that the lesser the bounce rates the higher the likelihood of users making purchases. *Administration and administration duration are have a weak negative corelation to the page value, suggesting that the less time spend on the administration pages the more likely they will transact.* There is no correlation between page value and either of informational and informational_duration. *Recommendation See how informational page can be merged with the administartive page.* Makes the administrative pages more user-friendly and easy to navigate.

#Applying KMeans in clustering

```
#Normalizing the data first
normalize <- function(x){
  return ((x-min(x)) / (max(x)-min(x)))
}
```

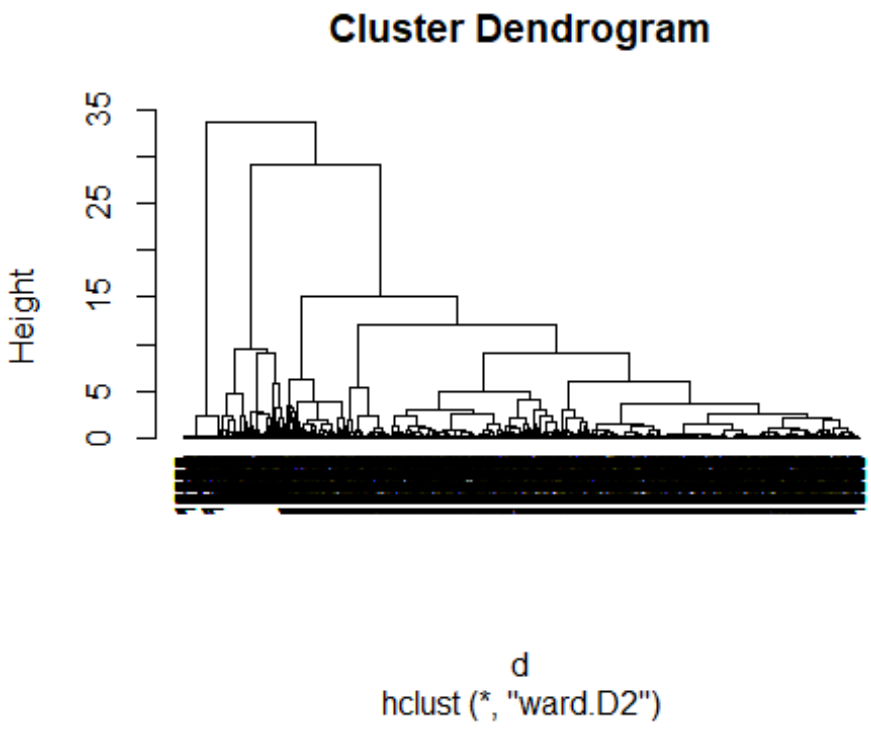
```r
store_cor$Administrative <- normalize(store_cor$Administrative)
store_cor$Administrative_Duration <-
normalize(store_cor$Administrative_Duration)
store_cor$Informational <- normalize(store_cor$Informational)
store_cor$Informational_Duration<-
normalize(store_cor$Informational_Duration)
store_cor$ProductRelated<- normalize(store_cor$ProductRelated)
store_cor$ProductRelated_Duration<-
normalize(store_cor$ProductRelated_Duration)
store_cor$BounceRates<- normalize(store_cor$BounceRates)
store_cor$PageValues<- normalize(store_cor$PageValues)
store_cor$PageValues<- normalize(store_cor$PageValues)
store_cor$SpecialDay<- normalize(store_cor$SpecialDay)
results<- kmeans(store_cor,8)

results$size

## [1] 5694  189  755  970  891  322 2676  819

#results$cluster

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

plot(airquality[,1:2], col = results$cluster)
plot(store_cor[c(1,3,5,7,8,9,10)], col = results$cluster)
```





#Clustering with Hierarchical clustering

```
d <- dist(store_cor, method = "euclidean")
res.hc <- hclust(d, method = "ward.D2" )
plot(res.hc, cex = 0.6, hang = -1)
```


Cluster Dendrogram
Height
0
5
15
25
35
d
hclust (*, "ward.D2")