R_Programming_week_2_IP_(CLUSTERING)

Ruth Nguli

2022-03-24

Defining the Question

To understand customer's behavior from data that collected over the past year.

Metric for success

The metric of success will be attained on identifying the characteristics of customer groups.

Understanding the business 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.

Experimental Design

Define the question, the metric for success, the context, experimental design taken.

Read and explore the given dataset.

Cleaning Data

Perform Exploratory Data Cleaning (Univariate & Bivariate)

create a Clustering model

Conclusion

Recommendations

Data Source

The dataset source link: http://bit.ly/EcommerceCustomersDataset

Reading Data

```
# Loading and reading data from source link
#
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

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

# Previewing the head of data
#
head(brand)
```

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

```
# Previewing tail of data
#
tail(brand)
```

```
##       Administrative Administrative_Duration Informational
## 12325              0                       0             1
```

```
## 12326              3                    145             0
## 12327              0                      0             0
## 12328              0                      0             0
## 12329              4                     75             0
## 12330              0                      0             0
##       Informational_Duration ProductRelated ProductRelated_Duration BounceRates
## 12325                      0             16                503.000 0.000000000
## 12326                      0             53               1783.792 0.007142857
## 12327                      0              5                465.750 0.000000000
## 12328                      0              6                184.250 0.083333333
## 12329                      0             15                346.000 0.000000000
## 12330                      0              3                 21.250 0.000000000
##        ExitRates PageValues SpecialDay Month OperatingSystems Browser Region
## 12325 0.03764706    0.00000          0   Nov                2       2      1
## 12326 0.02903061   12.24172          0   Dec                4       6      1
## 12327 0.02133333    0.00000          0   Nov                3       2      1
## 12328 0.08666667    0.00000          0   Nov                3       2      1
## 12329 0.02105263    0.00000          0   Nov                2       2      3
## 12330 0.06666667    0.00000          0   Nov                3       2      1
##       TrafficType       VisitorType Weekend Revenue
## 12325           1 Returning_Visitor   FALSE   FALSE
## 12326           1 Returning_Visitor    TRUE   FALSE
## 12327           8 Returning_Visitor    TRUE   FALSE
## 12328          13 Returning_Visitor    TRUE   FALSE
## 12329          11 Returning_Visitor   FALSE   FALSE
## 12330           2       New_Visitor    TRUE   FALSE
```

```
# Checking the data structure
#
str(brand)
```

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 -1 0 0 0 -1 -1 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 1 2 3 ...
##  $ ProductRelated_Duration: num  0 64 -1 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : chr  "Feb" "Feb" "Feb" "Feb" ...
##  $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : chr  "Returning_Visitor" "Returning_Visitor" "Returning_Visitor" "Returni
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

The dataset has 12330 observations (rows) and 18 variables(columns) The data types are appropriately represented

```r
# Looking at the columns
#
library(tidyverse)
colnames(brand)
```

```
##  [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"
```

Data Cleaning

```r
# looking at data summary
#
summary(brand)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :  -1.00         Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.000
##  Median : 1.000   Median :   8.00         Median : 0.000
##  Mean   : 2.318   Mean   :  80.91         Mean   : 0.504
##  3rd Qu.: 4.000   3rd Qu.:  93.50         3rd Qu.: 0.000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.000
##  NA's   :14       NA's   :14              NA's   :14
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :  -1.00        Min.   :  0.00   Min.   :   -1.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  185.0
##  Median :   0.00        Median : 18.00   Median :  599.8
##  Mean   :  34.51        Mean   : 31.76   Mean   : 1196.0
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1466.5
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##  NA's   :14             NA's   :14       NA's   :14
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003119   Median :0.02512   Median :  0.000   Median :0.00000
##  Mean   :0.022152   Mean   :0.04300   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016684   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##  NA's   :14         NA's   :14
##     Month           OperatingSystems    Browser           Region
##  Length:12330       Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mode  :character   Median :2.000    Median : 2.000   Median :3.000
##                     Mean   :2.124    Mean   : 2.357   Mean   :3.147
##                     3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
```

```
##                     Max.   :8.000    Max.   :13.000   Max.   :9.000
##
##   TrafficType    VisitorType         Weekend         Revenue
##  Min.   : 1.00   Length:12330       Mode :logical   Mode :logical
##  1st Qu.: 2.00   Class :character   FALSE:9462      FALSE:10422
##  Median : 2.00   Mode  :character   TRUE :2868      TRUE :1908
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
##
```

- there are some NA values which will need be imputed

```
## Calling Amelia and mice libraries for data imputation
#
library(Amelia)
```

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

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.0, built: 2021-05-26)
## ## Copyright (C) 2005-2022 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(mice)
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(VIM)
```

```
## Loading required package: colorspace
```

```
## Loading required package: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: 'VIM'
```

```
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
## Confirming if the data has missing
#
mapply(anyNA,brand)
```

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

```
# Imputing the missing values by predicting missing values with mice package
#

miss_mod <- mice(brand[, c("Administrative" , "Administrative_Duration", "Informational", "Informational
```

```
##
##  iter imp variable
##   1   1  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   1   2  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   1   3  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   1   4  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   1   5  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   2   1  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   2   2  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   2   3  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   2   4  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   2   5  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   3   1  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   3   2  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   3   3  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   3   4  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   3   5  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   4   1  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   4   2  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   4   3  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   4   4  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   4   5  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   5   1  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
```

```
##   5   2  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   5   3  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   5   4  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
##   5   5  Administrative  Administrative_Duration  Informational  Informational_Duration  ProductRela
```

```
completed <- complete(miss_mod)
```

```
# placing predicted missing values into the main data set
# ---
#
brand$Administrative <- completed$Administrative
brand$Administrative_Duration <- completed$Administrative_Duration
brand$Informational <- completed$Informational
brand$Informational_Duration <- completed$Informational_Duration
brand$ProductRelated <- completed$ProductRelated
brand$ProductRelated_Duration <- completed$ProductRelated_Duration
brand$BounceRates <- completed$BounceRates
brand$ExitRates <- completed$ExitRates
```

```
# confirming if there no more missing values
#
anyNA(brand)
```

```
## [1] FALSE
```

```
# checking duplicates
#
anyDuplicated(brand)
```

```
## [1] 159
```

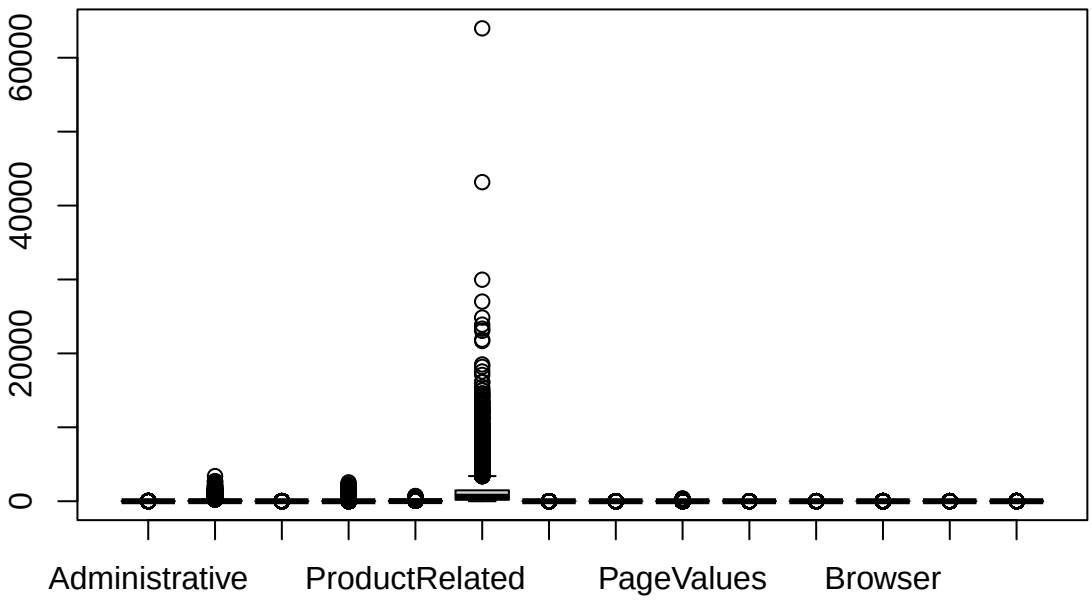
*presence of duplicates

```
# Dropping duplicated
#
brand <- brand %>% distinct()

# confirming if duplicates were successfully dropped
#
anyDuplicated(brand)
```

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

Successfully dropped duplicates in the data

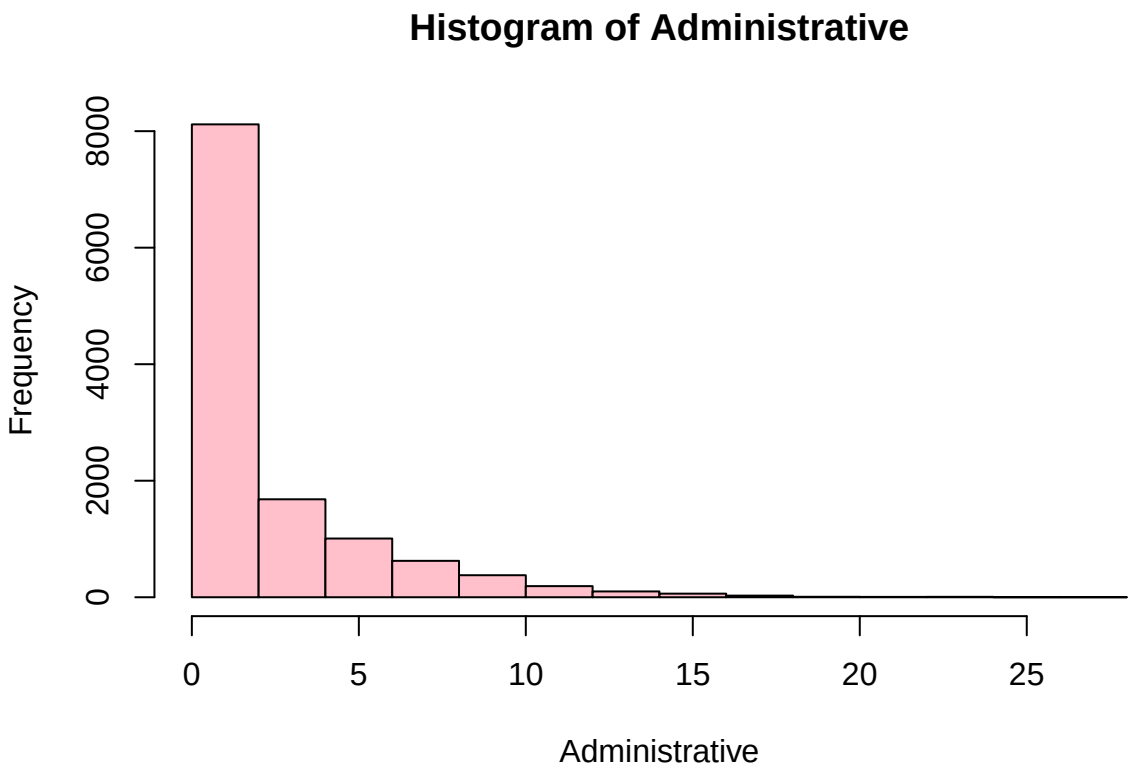
```
# checking outliers
#
non_cat <- brand %>% select("Administrative" , "Administrative_Duration", "Informational", "Informationa
boxplot(non_cat)
```



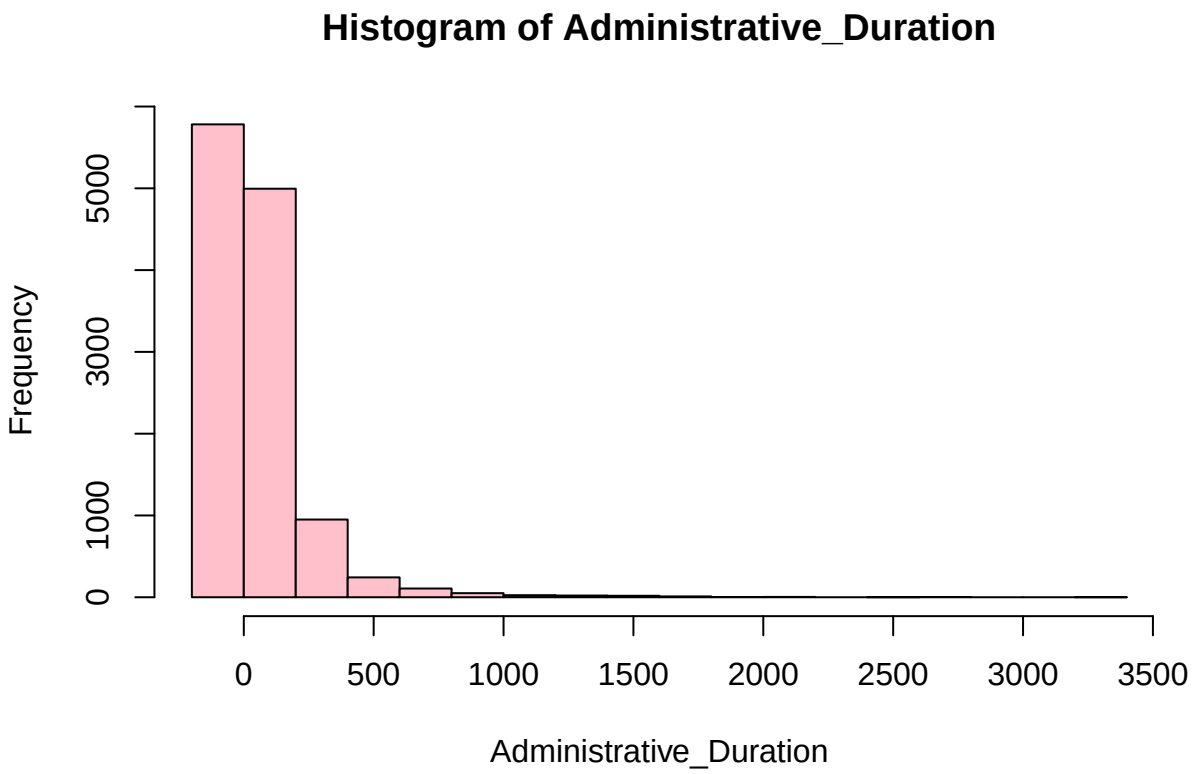
*There are noticeable outliers which represent real time data.

Univariate Analysis

```
# Plotting a Histograms of Administrative and duration
#
attach(brand)
hist(Administrative, col="pink")
```

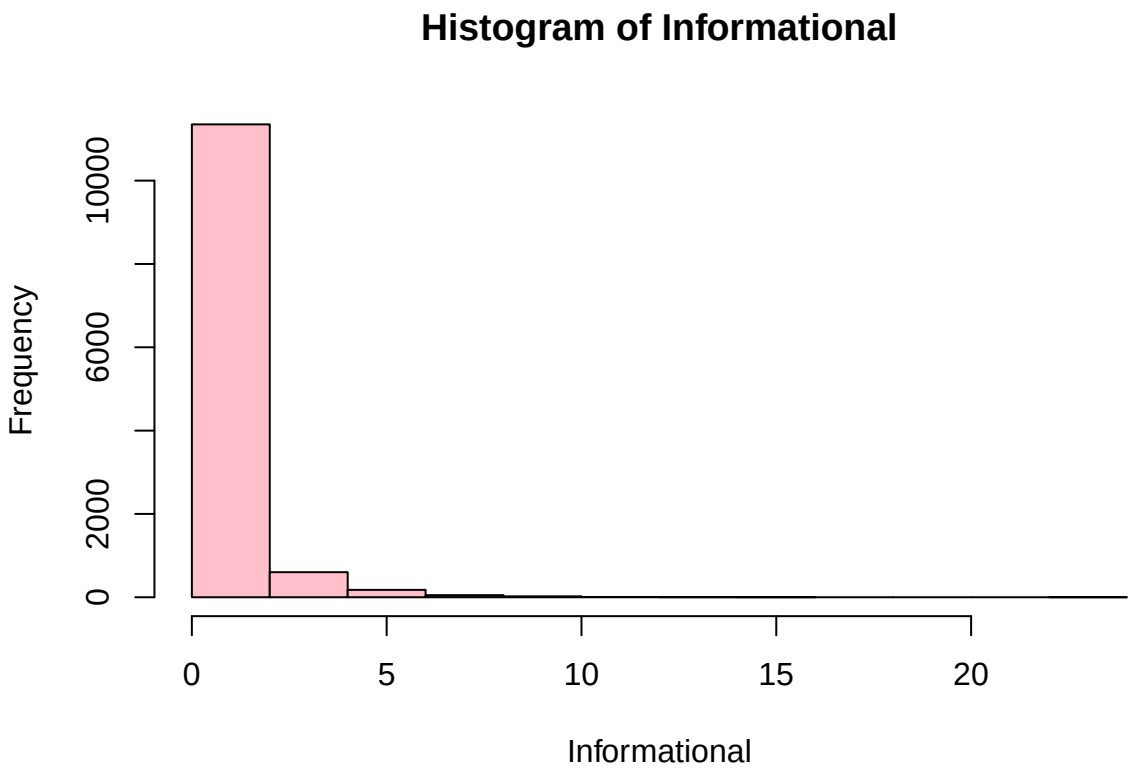



```
hist(Administrative_Duration, col="pink")
```

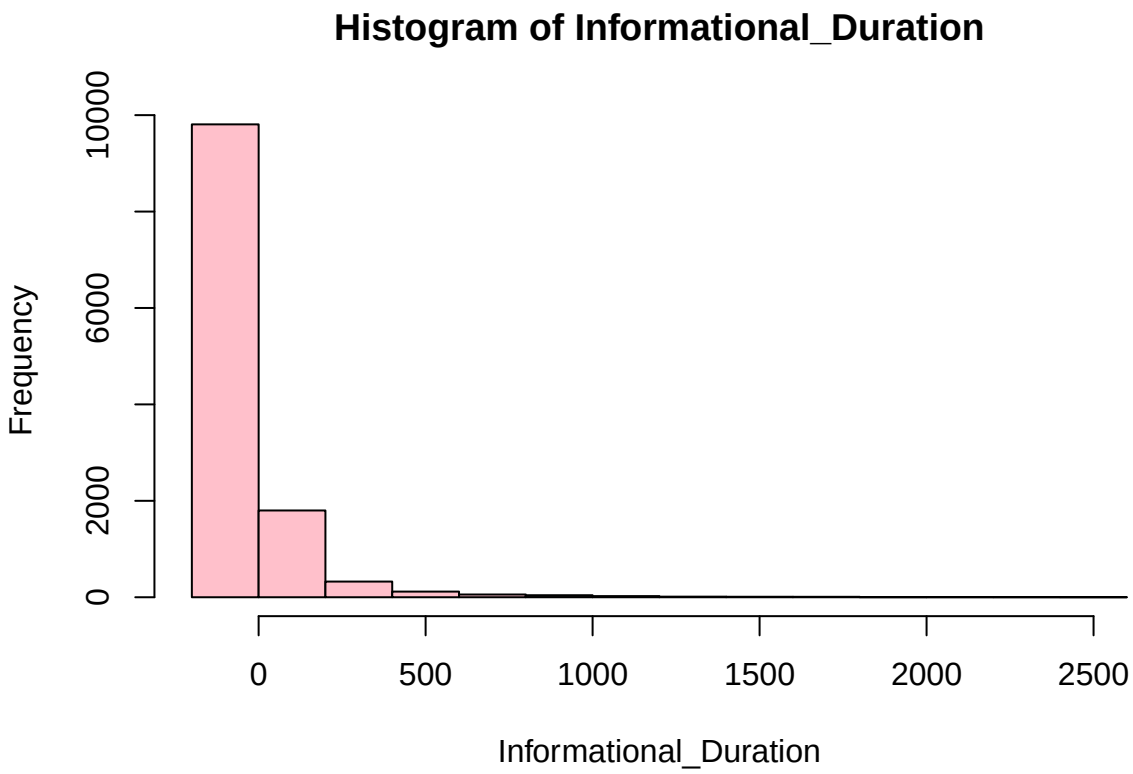


Administrative is skewed to the right. Majority of the visitors did not visit the Administrative page. Majority of the visitors spend less time on the page.

```
# Plotting a Histograms of Informational and duration
#
hist(Informational, col="pink")
```

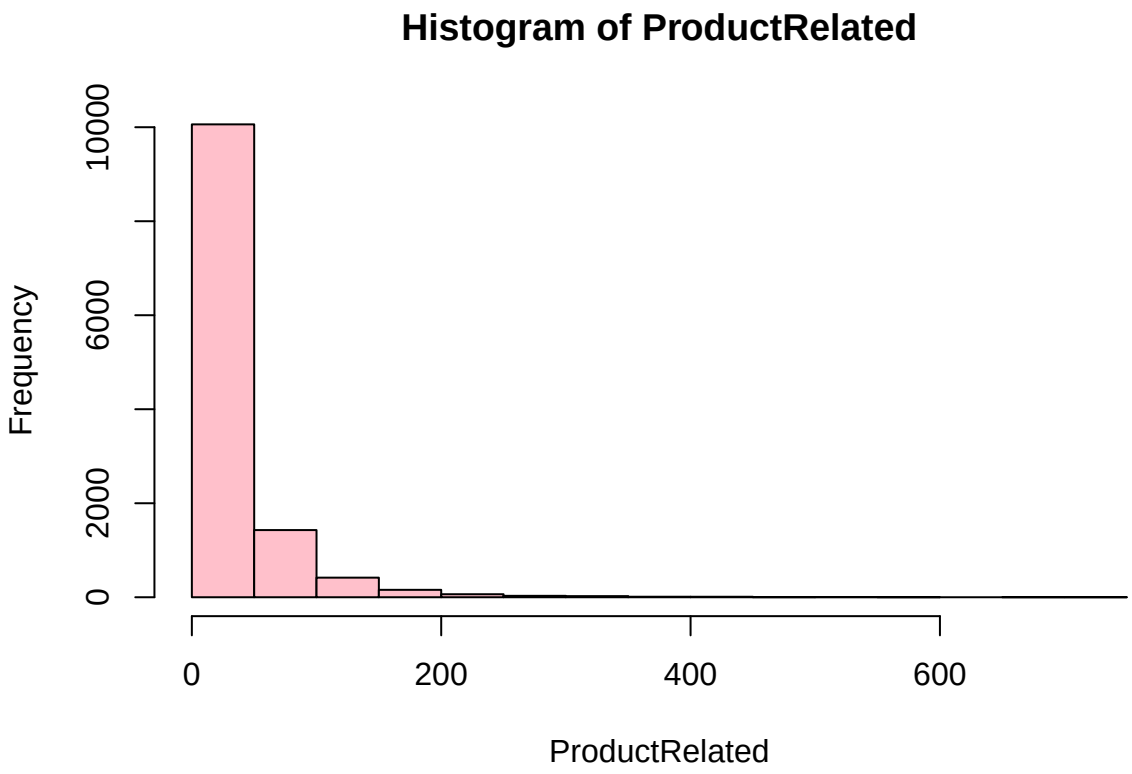



```
hist(Informational_Duration, col="pink")
```

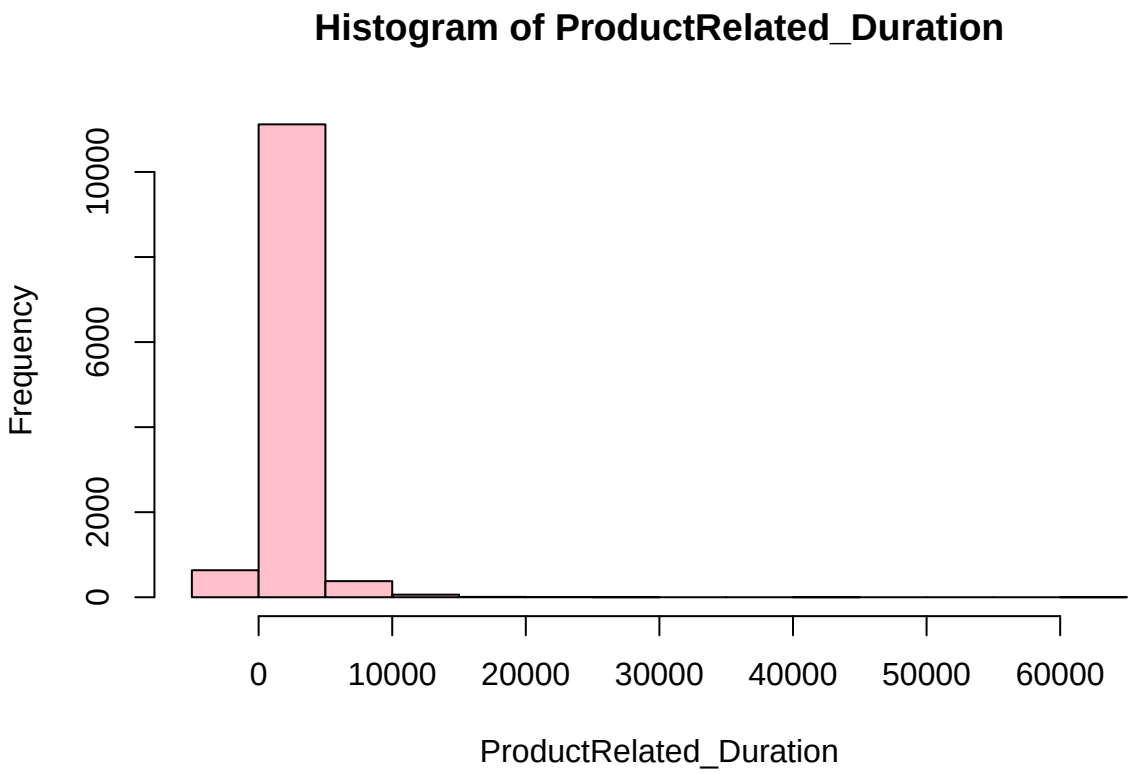


- Informational and Informational_Duration are skewed to the right.
- most visitors did not visit the Informational page.
- most visitors who visited the page spend less time on the page.

```
## Plotting  Histograms of Product related and duration
#
hist(ProductRelated, col="pink")
```

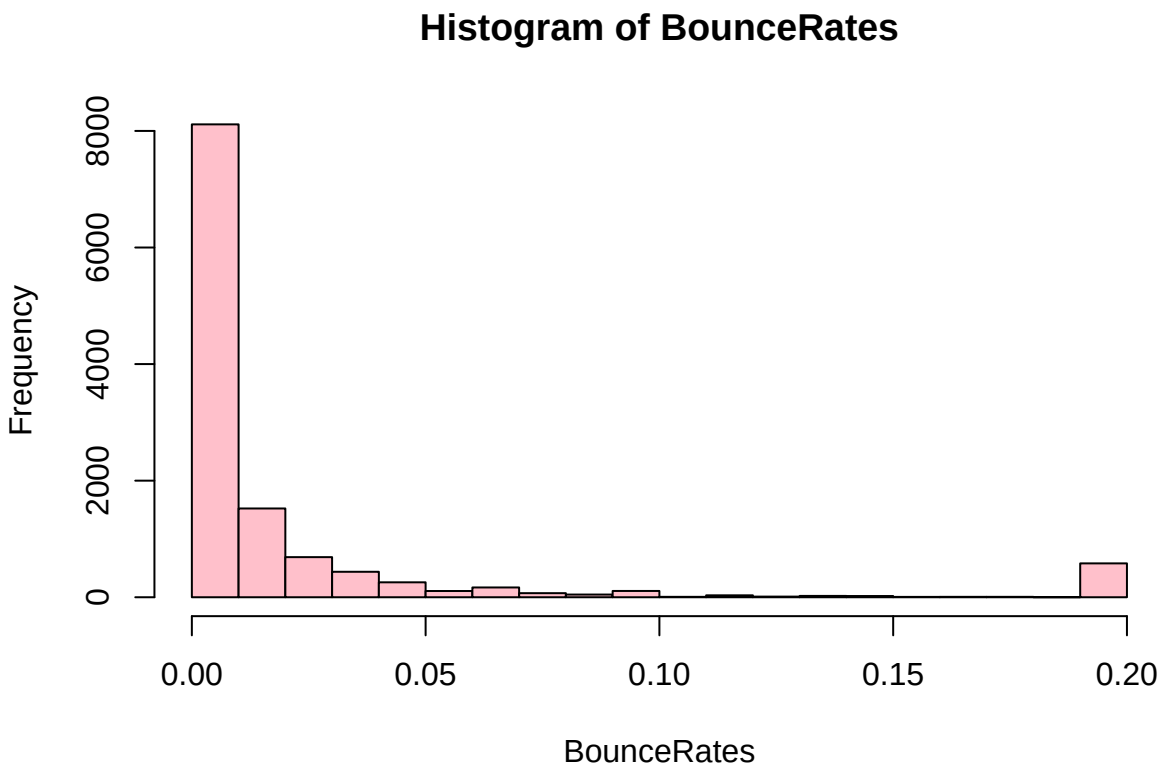



```
hist(ProductRelated_Duration, col="pink")
```

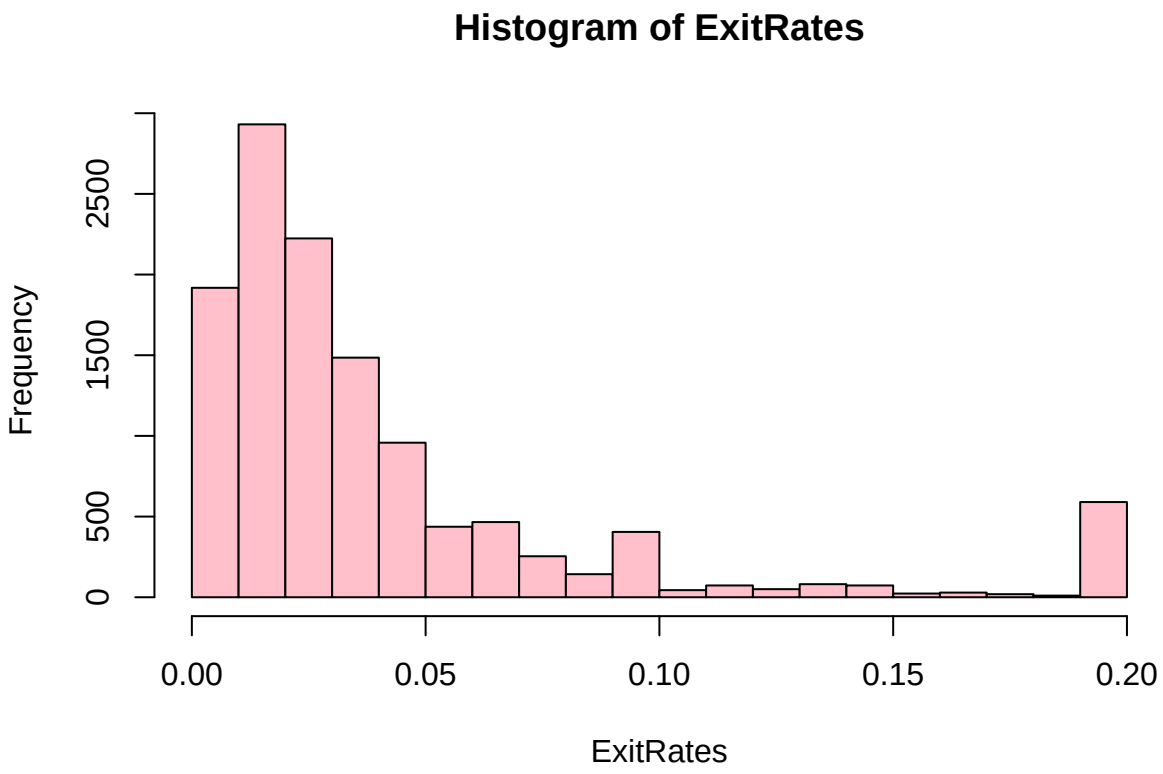


- ProductRelated and ProductRelated_Duration are skewed to the right.
- most visitors did not visit the ProductRelated page.
- most visitors who visited the page spend less time on the page.

```
## Plotting  Histograms of BounceRate and ExitRate
#
hist(BounceRates, col="pink")
```

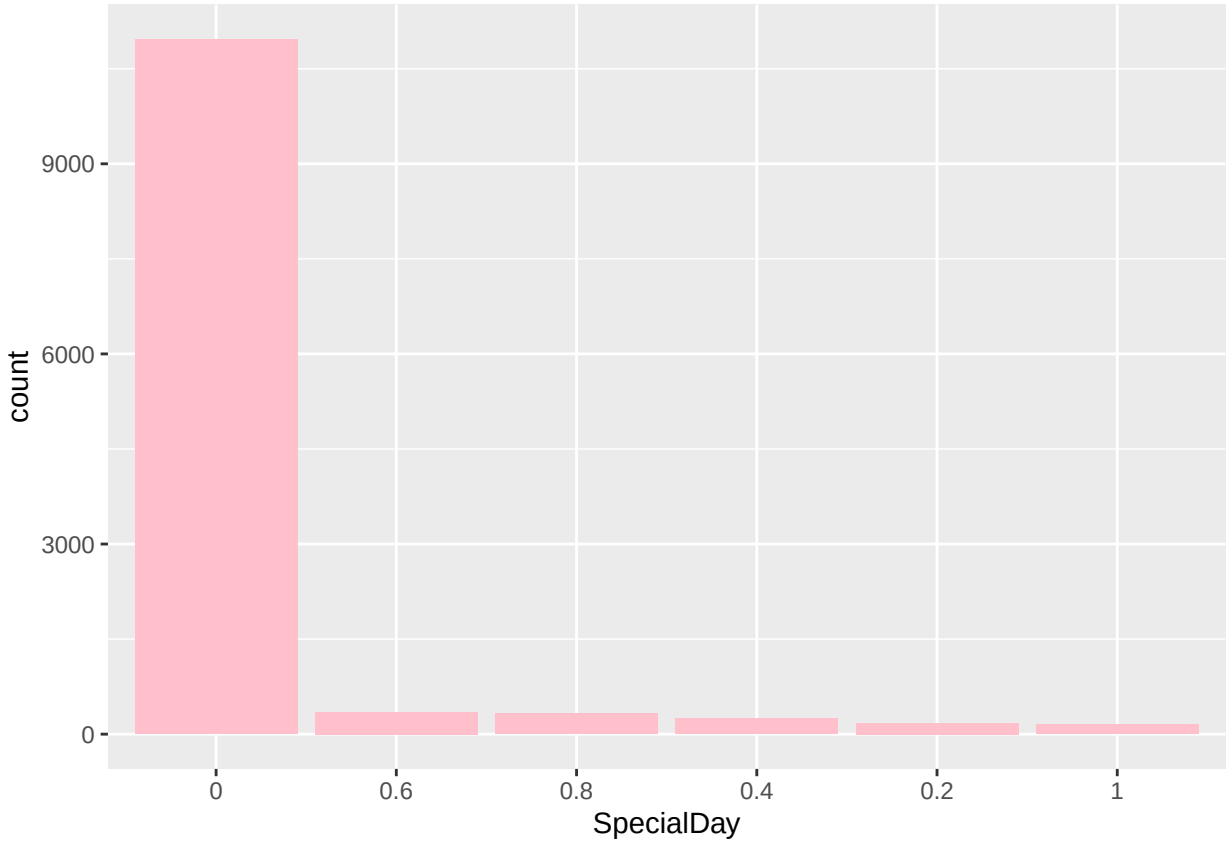



```
hist(ExitRates, col="pink")
```

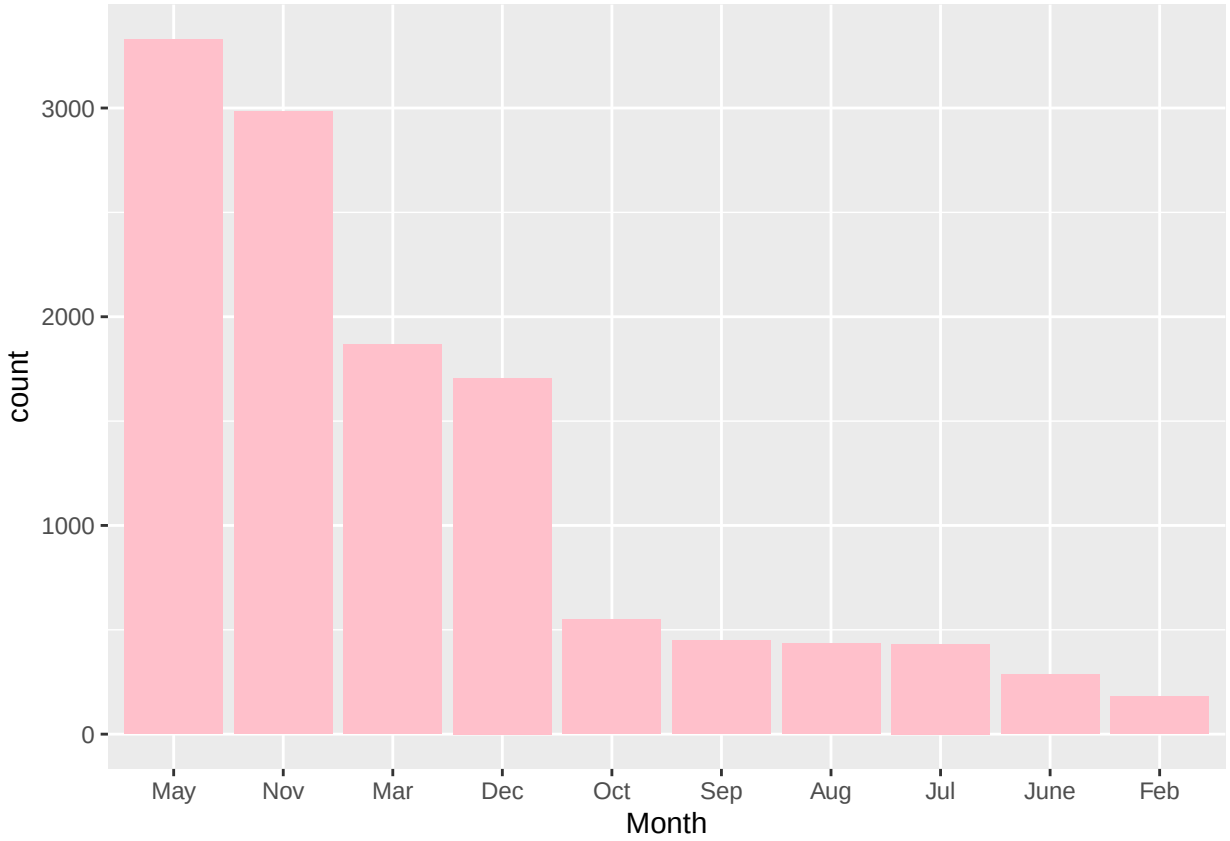


* Bounce rate at 0.00.has the highest frequency. * ExitRate was happening between 0.00 and 0.05 rate.

```
## Plotting  Histograms of Pagevalue and Special day
#
#
ggplot(brand, aes(x=reorder(SpecialDay, SpecialDay, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='SpecialDay')
```

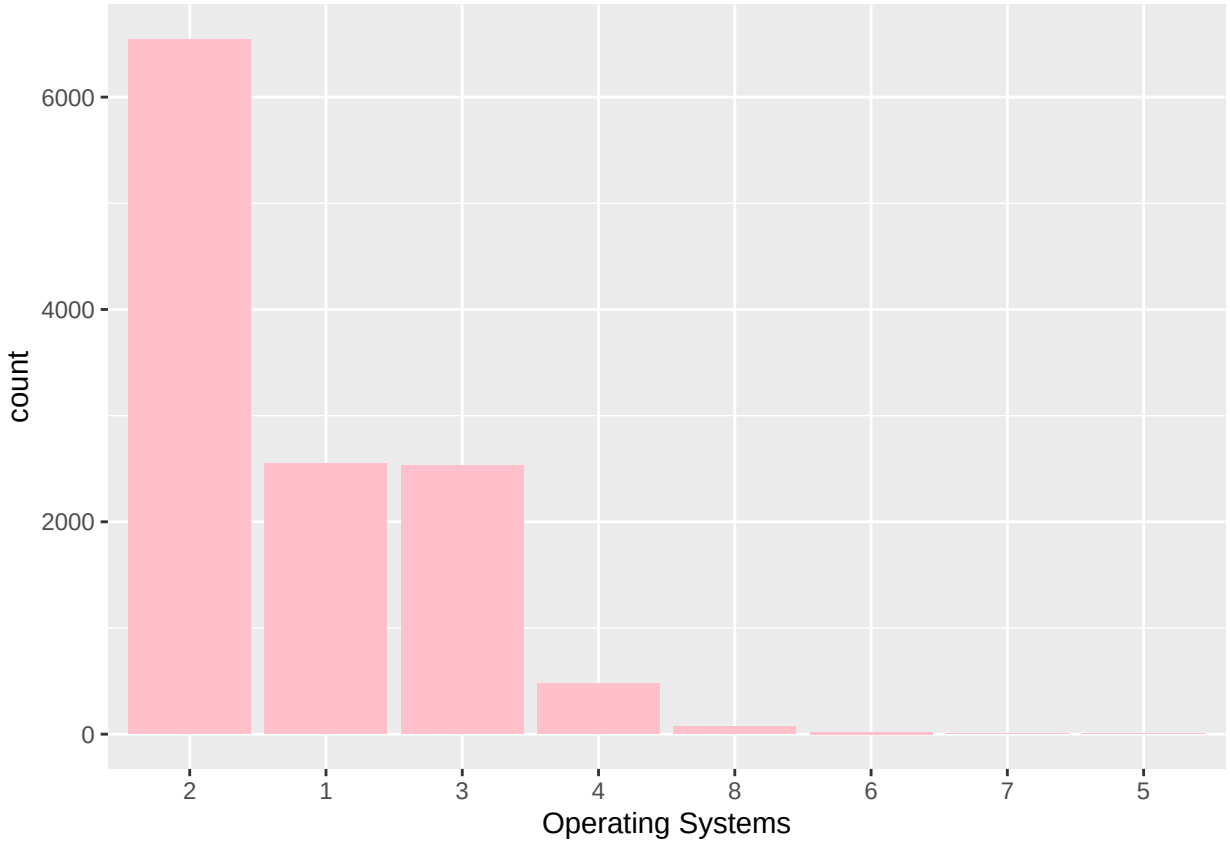



```
## Plotting  Bar plot of Month
#
ggplot(brand, aes(x=reorder(Month, Month, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='Month')
```



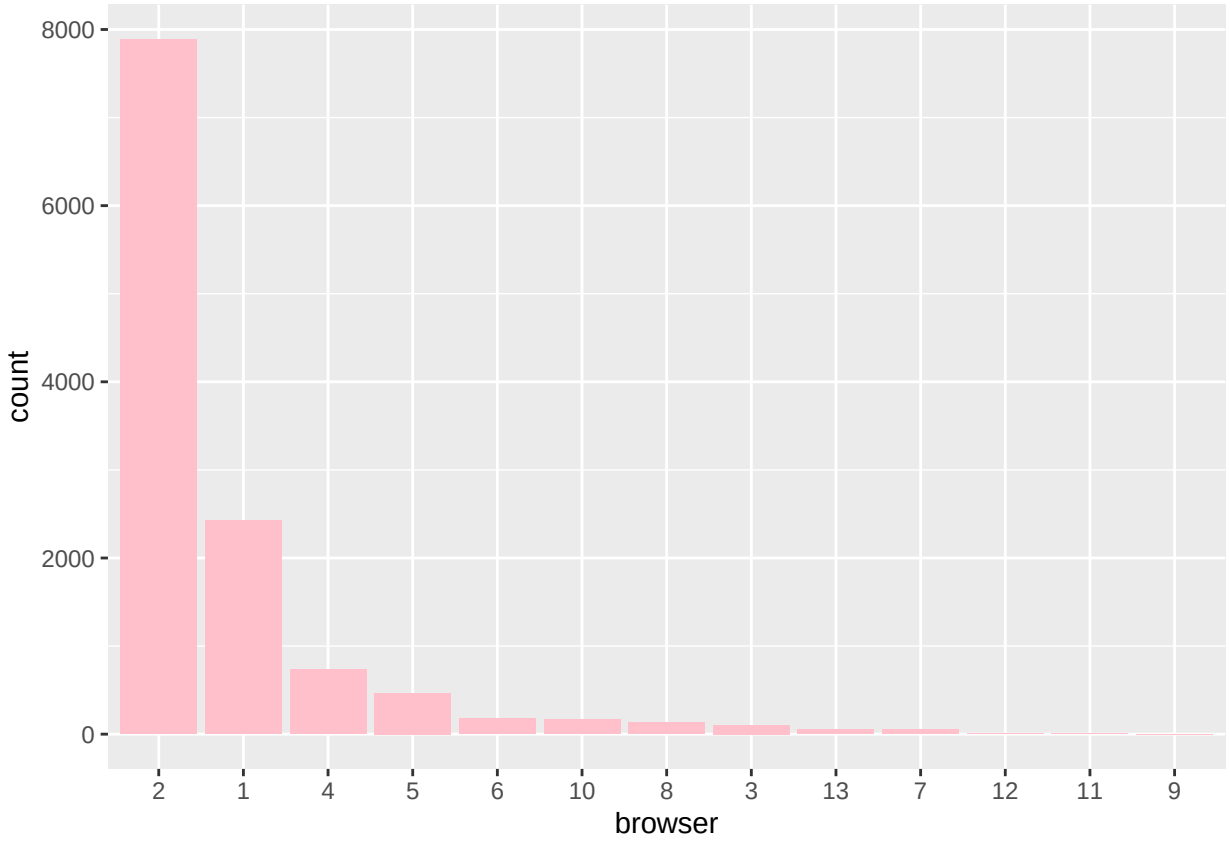
* most customers visited the pages in the months of May,Nov,March and Dec.

```
## Plotting  Bar plot of OperatingSystems
#
ggplot(brand, aes(x=reorder(OperatingSystems, OperatingSystems, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='Operating Systems')
```



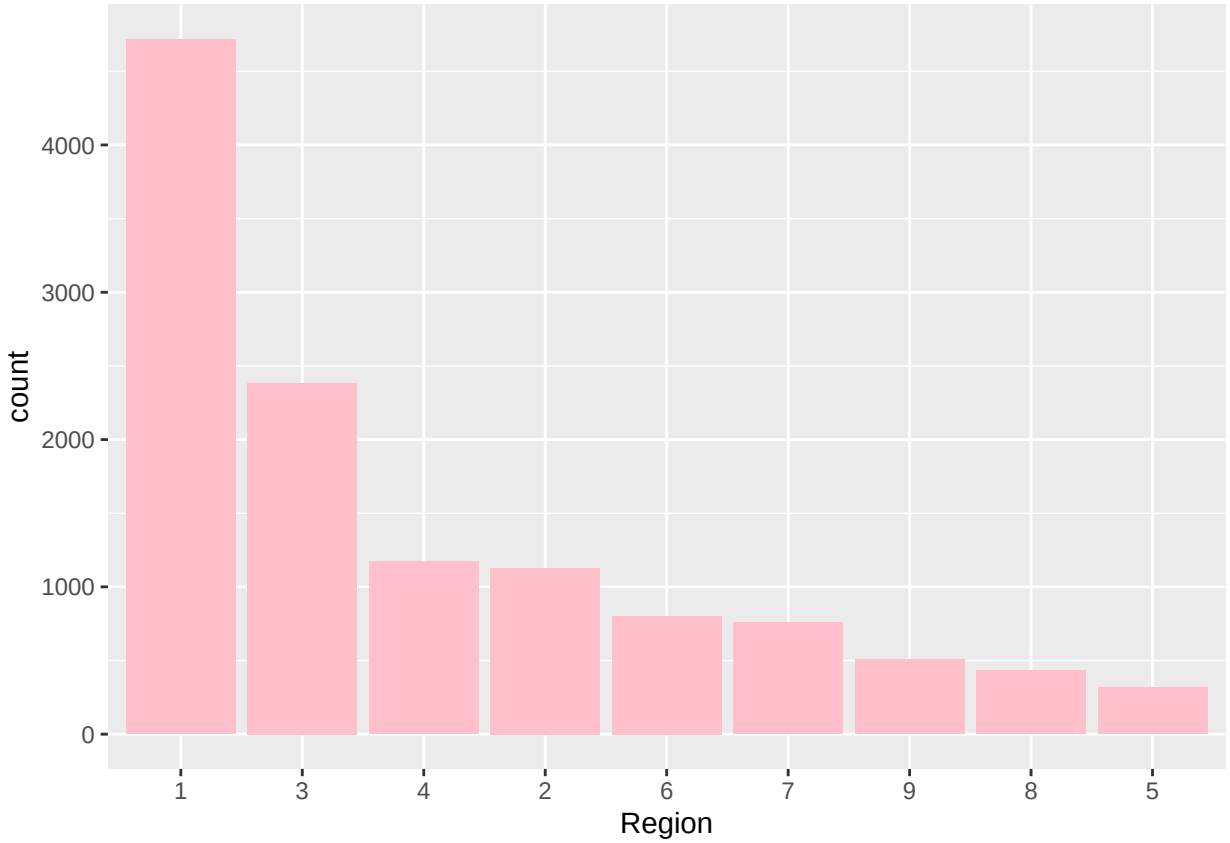
* most visitors used 2 different operating systems to access the brand's pages.

```
## Plotting  Bar plot of Browser
#
ggplot(brand, aes(x=reorder(Browser, Browser, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='browser')
```



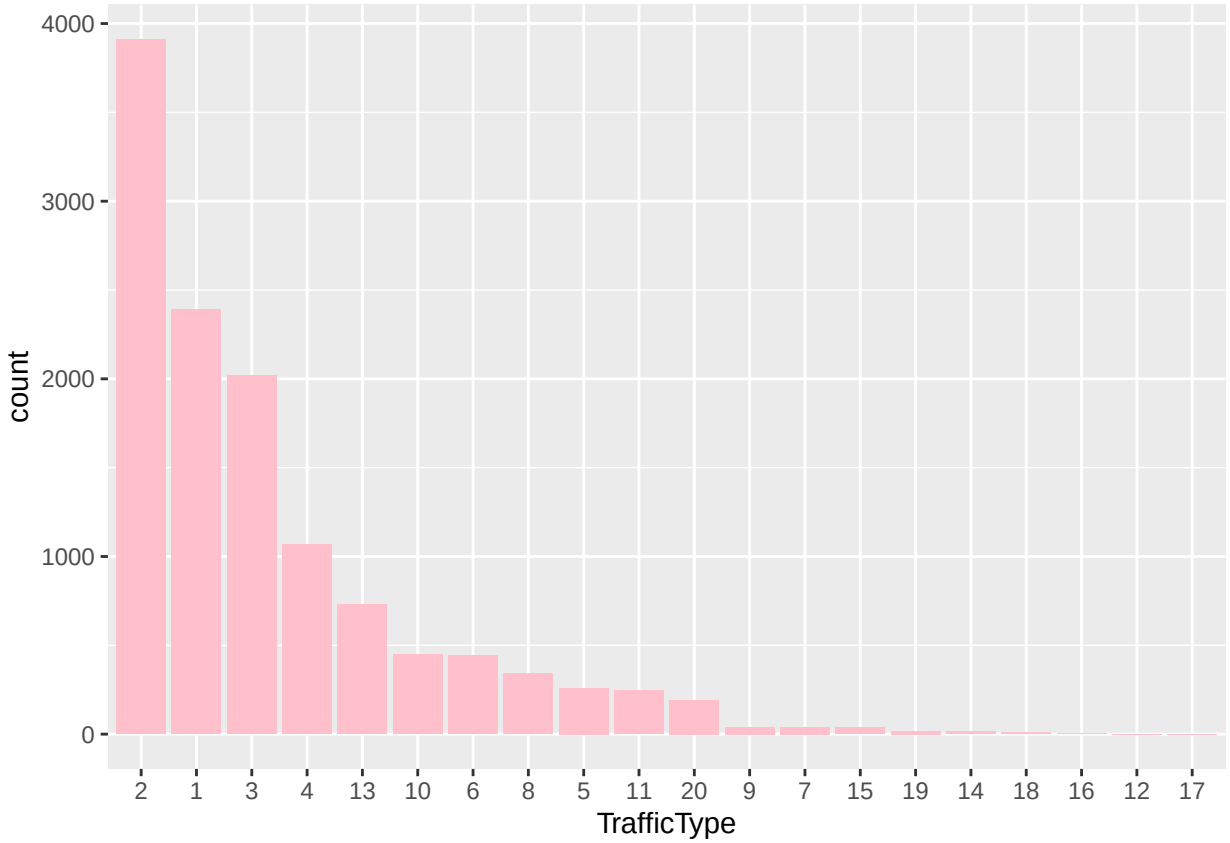
* most visitors used 2 different Browsers to access the brand's pages.

```
## Plotting  Bar plot of Region
#
ggplot(brand, aes(x=reorder(Region, Region, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='Region')
```



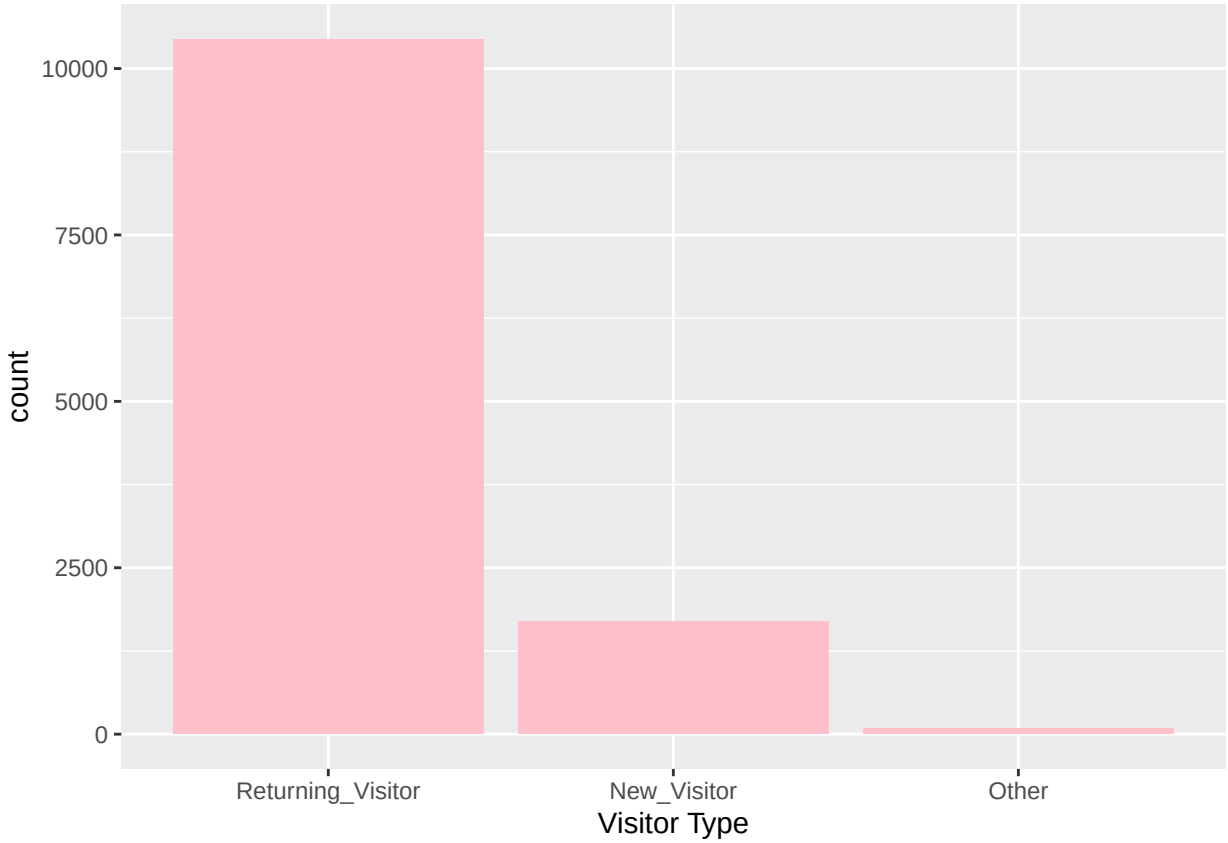
* most customers accessed the brand's pages from the one region.

```
## Plotting  Bar plot of TrafficType
#
ggplot(brand, aes(x=reorder(TrafficType, TrafficType, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='TrafficType')
```



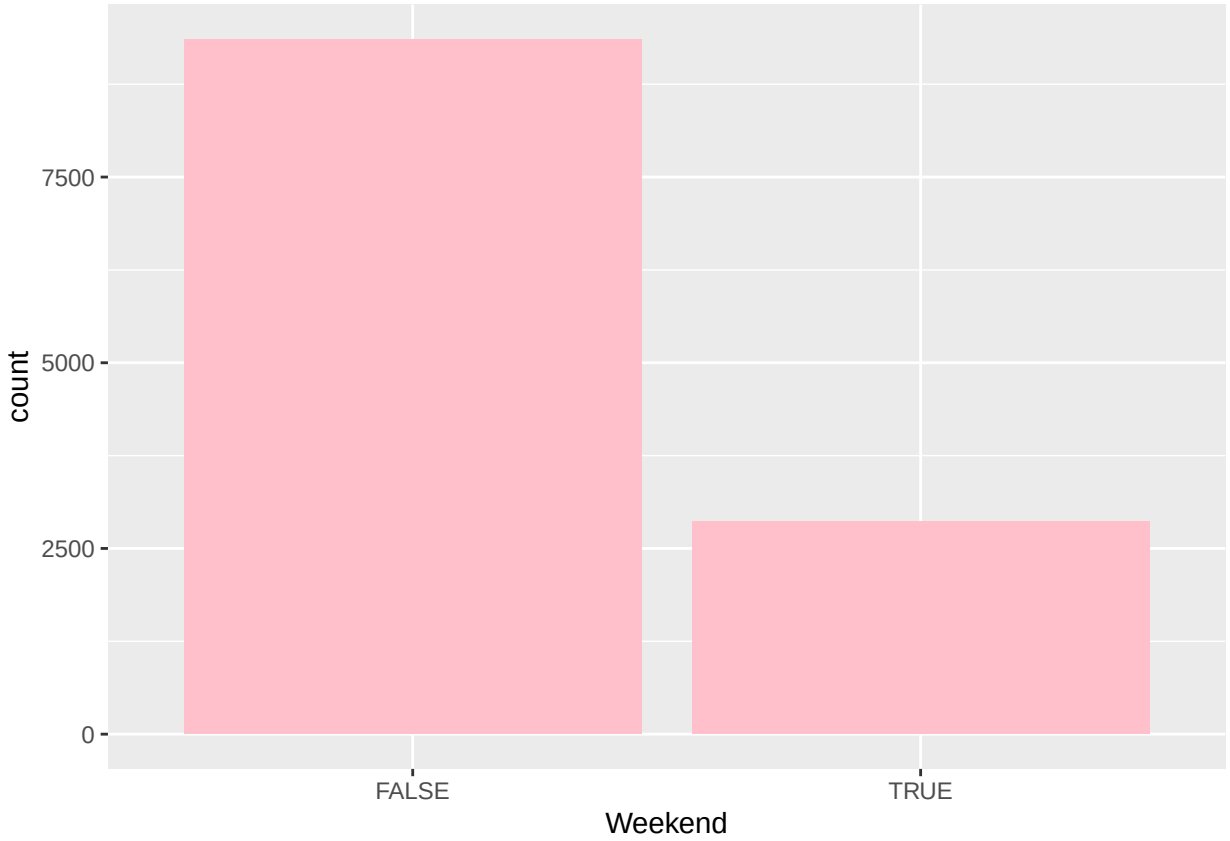
* Most of the visitors had between 1 & 3 traffic types.

```
## Plotting  Bar plot of VisitorType
#
ggplot(brand, aes(x=reorder(VisitorType, VisitorType, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='Visitor Type')
```



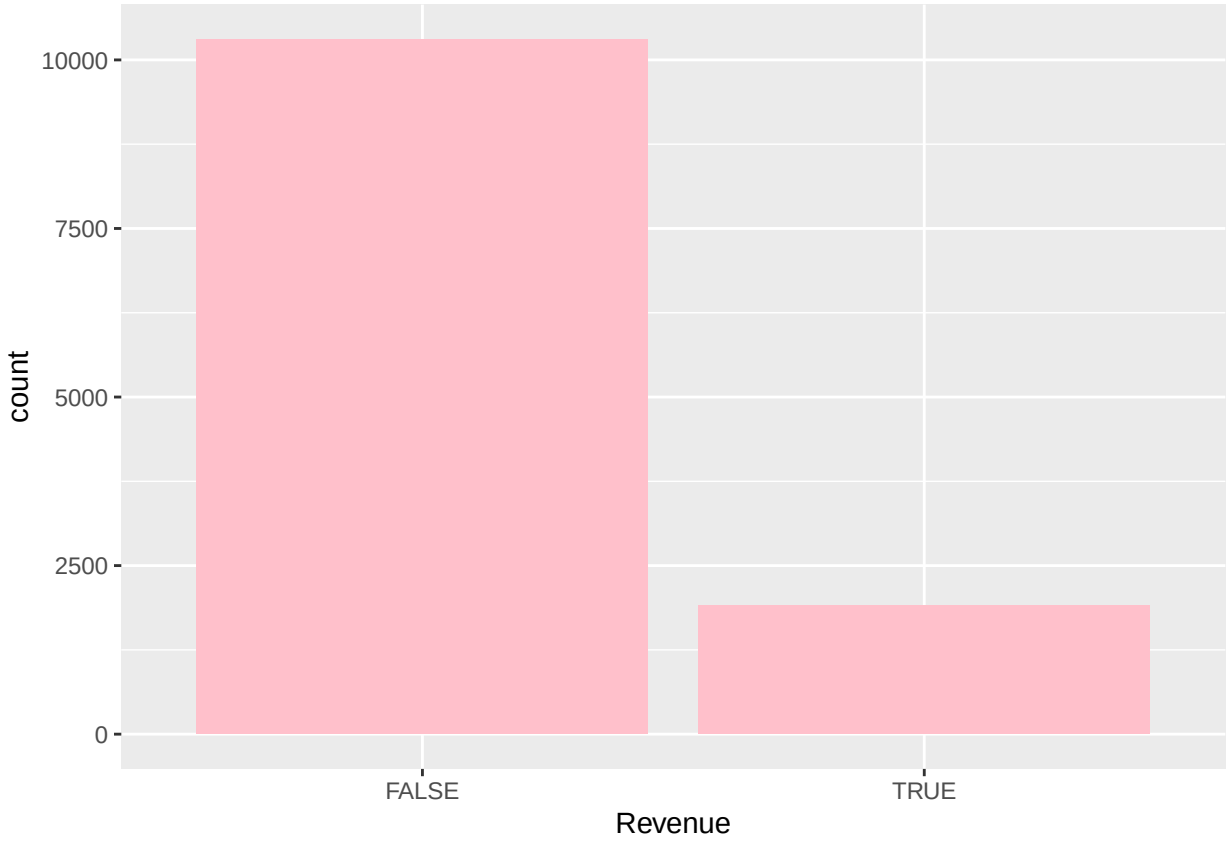
* Majority of page visitors were returning visitors

```
## Plotting  Bar plot of Weekend
#
ggplot(brand, aes(x=reorder(Weekend, Weekend, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='Weekend')
```



* we observe that most page visits were on weekdays.

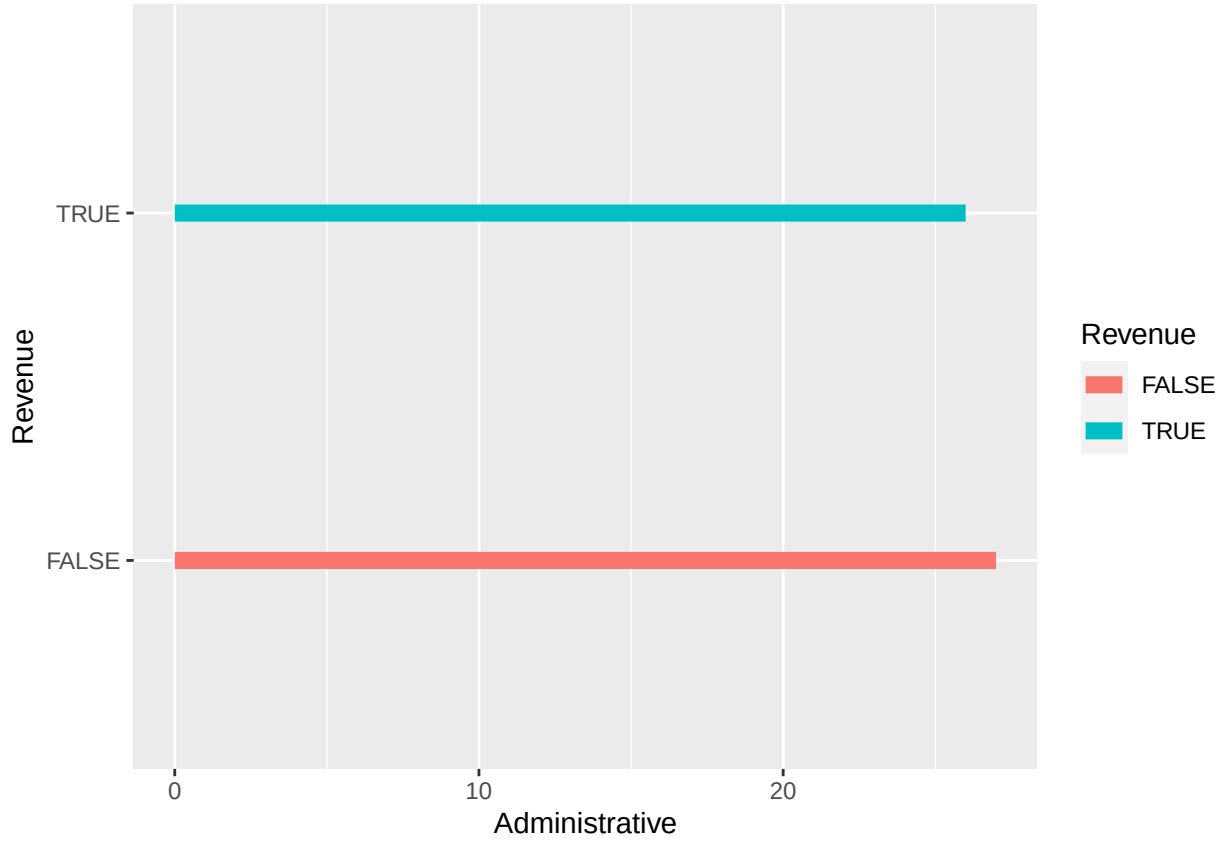
```
## Plotting  Bar plot of Revenue
#
ggplot(brand, aes(x=reorder(Revenue, Revenue, function(x)-length(x)))) +
geom_bar(fill='pink') +  labs(x='Revenue')
```



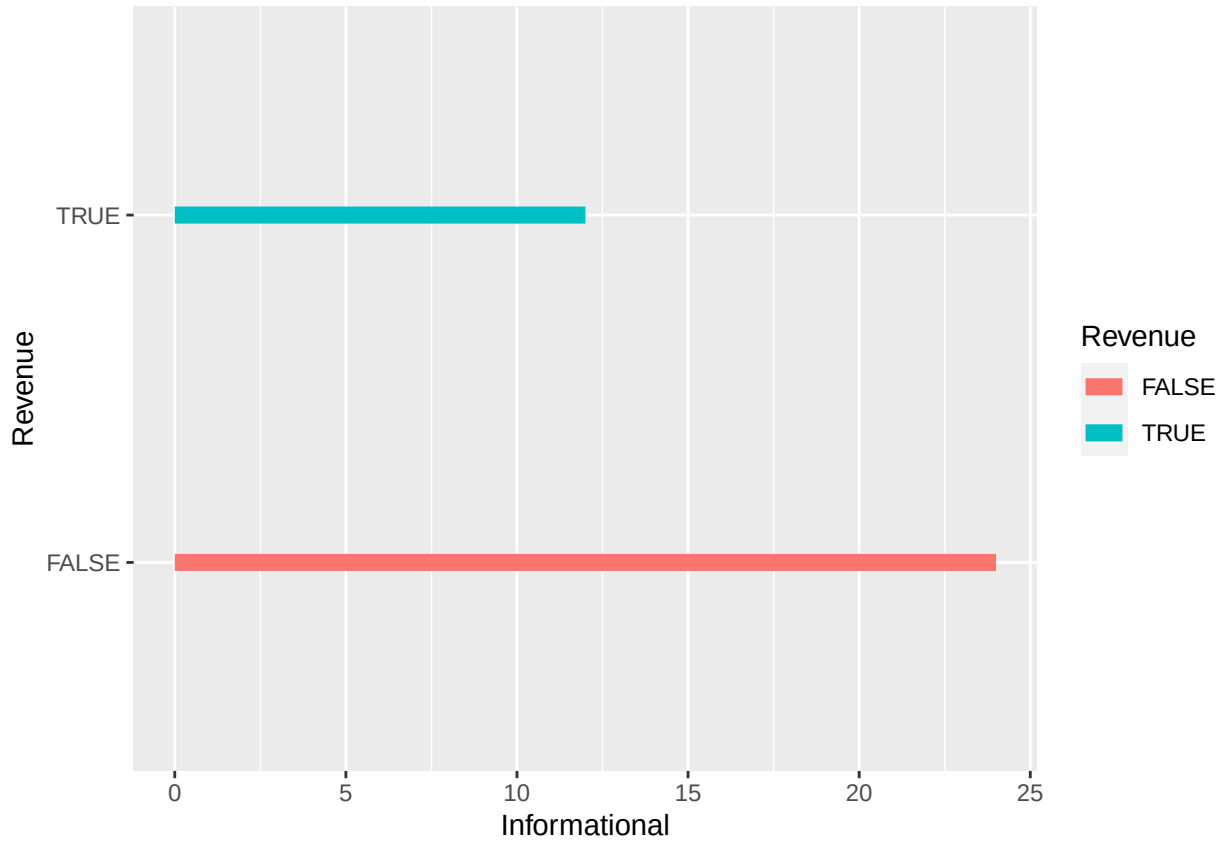
* We observe most of the visits in the brand's pages did not generate any revenue.

Bivariate Analysis

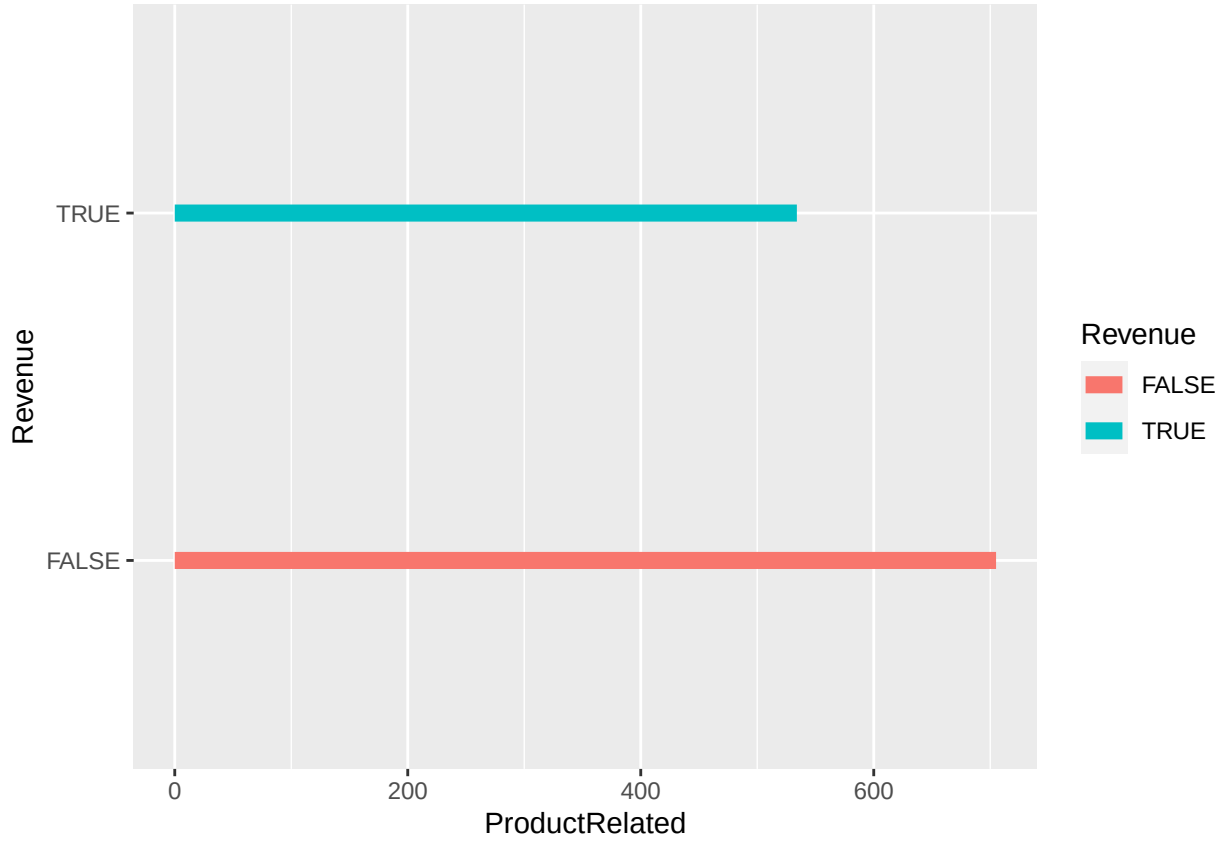
```
# Administrative Vs Revenue
#
ggplot(brand,aes(Administrative, Revenue, colour= Revenue))+
  geom_step(size=3)
```




```
# Informational Vs Revenue
#
ggplot(brand,aes(Informational, Revenue, colour= Revenue))+
  geom_step(size=3)
```

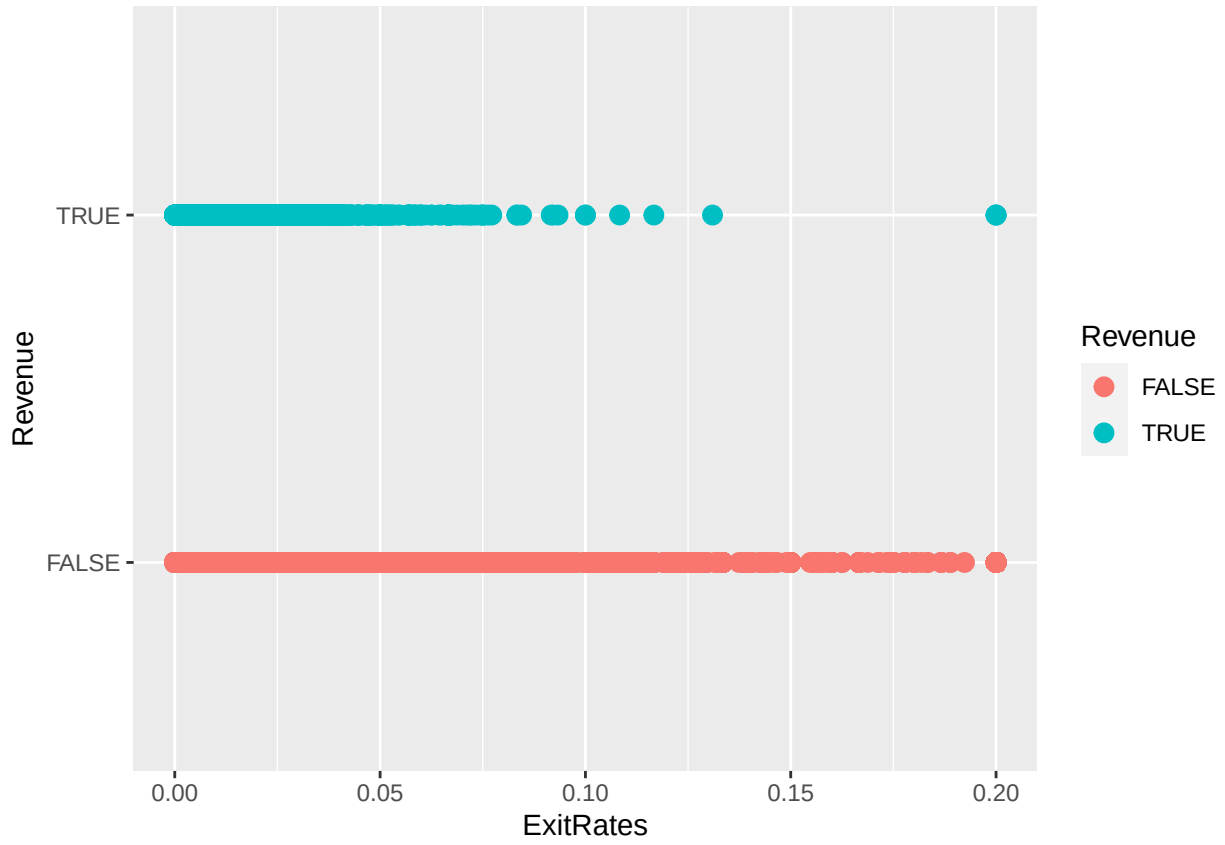



```
# ProductRelated Vs Revenue
#
ggplot(brand,aes(ProductRelated, Revenue, colour= Revenue))+
  geom_step(size=3)
```

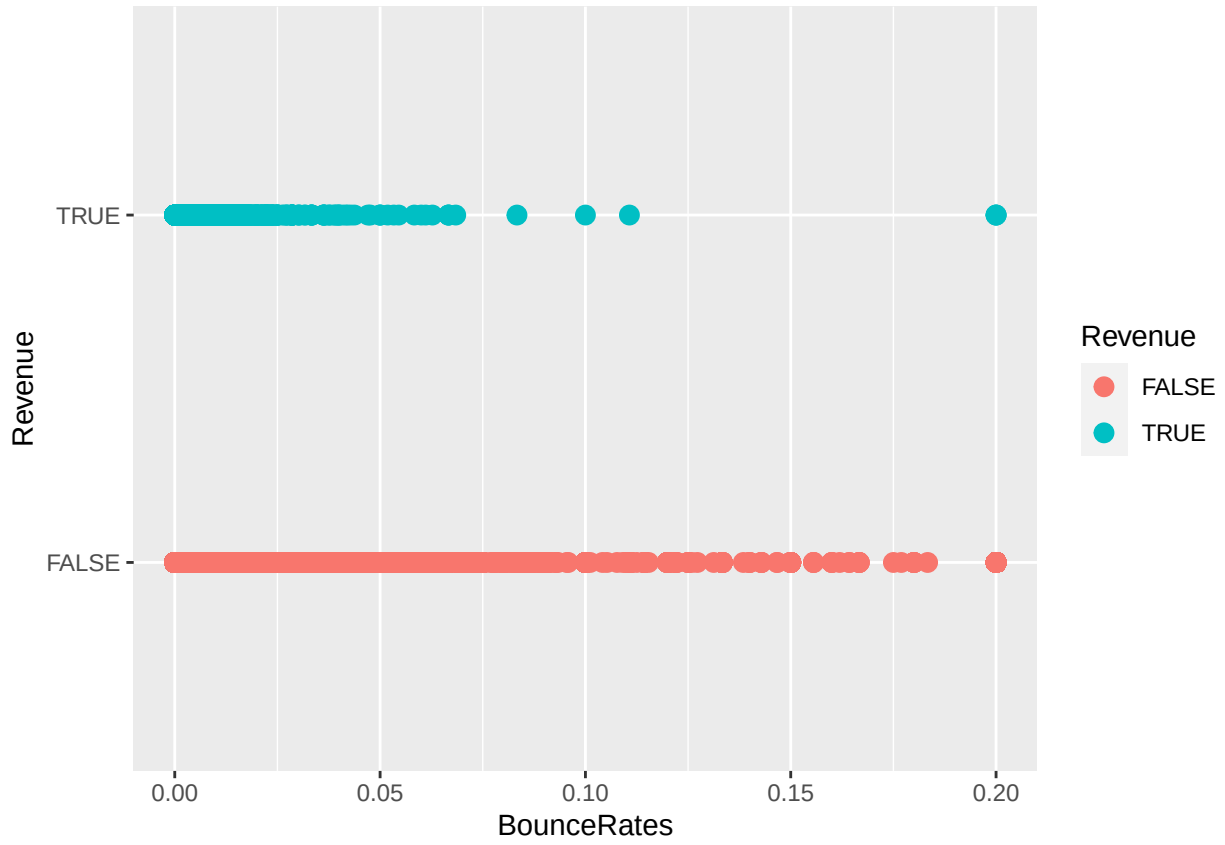


* We observe that most of Administrative,Informational & ProductRelated page visits did not generate much revenue.

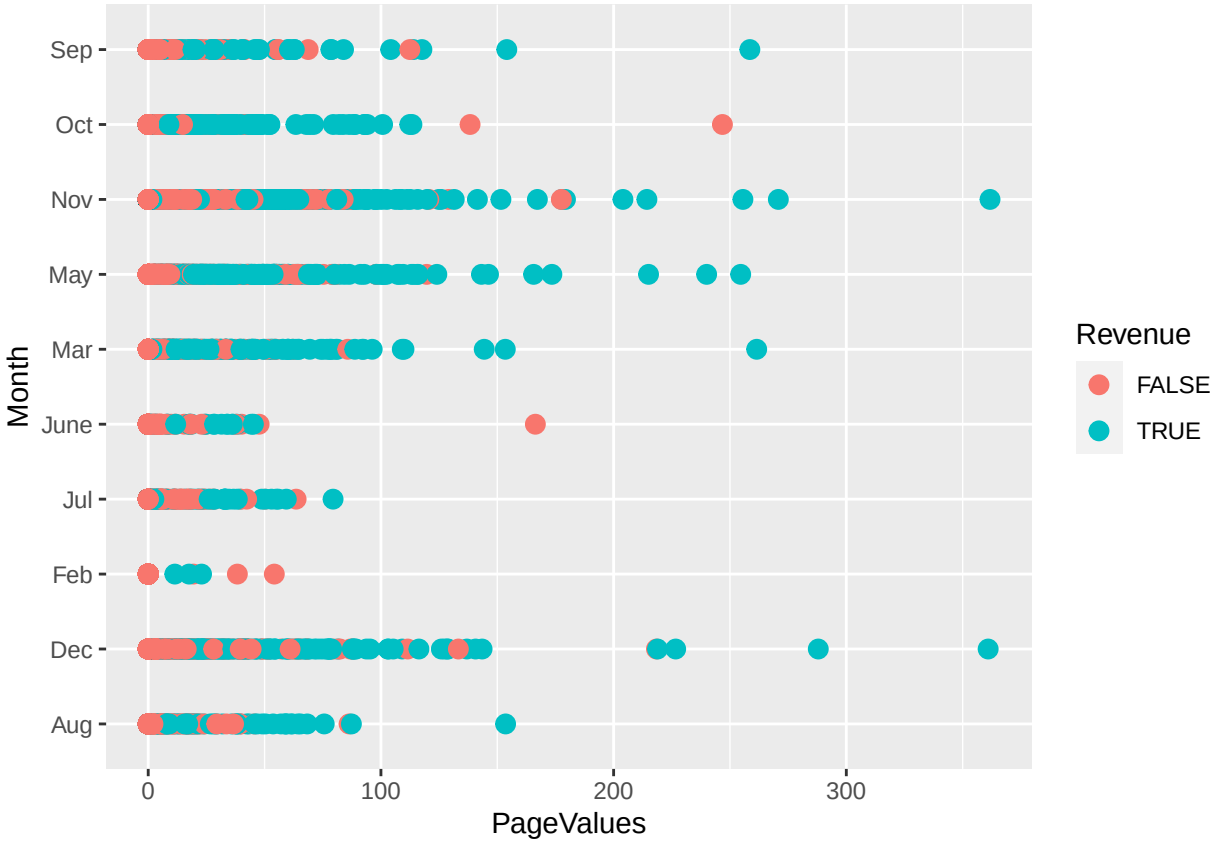
```
# ExitRate Vs Revenue
#
ggplot(brand,aes(ExitRates, Revenue, colour= Revenue))+
  geom_point(size=3)
```




```
# Bouncerate Vs Revenue
#
ggplot(brand,aes(BounceRates, Revenue, colour= Revenue))+
  geom_point(size=3)
```

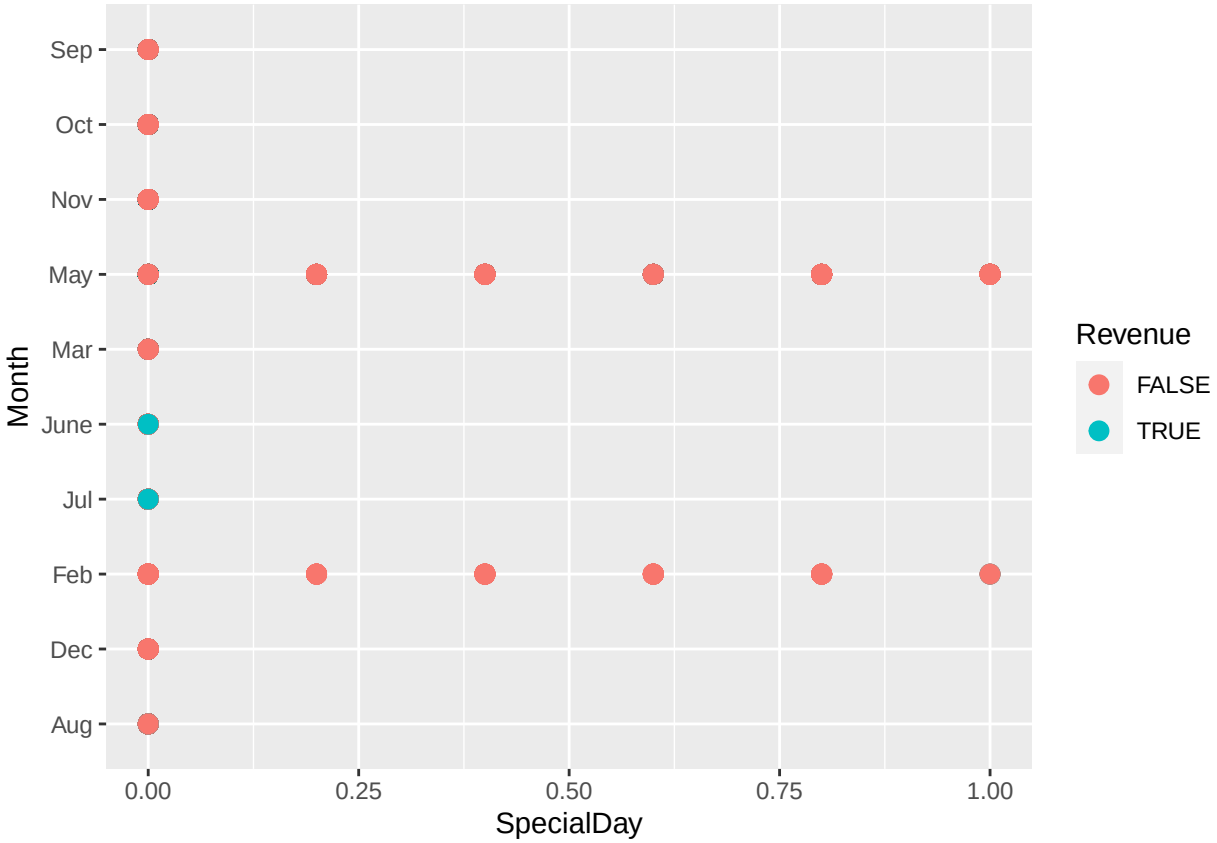



```
# Pagevalues Vs Month
#
ggplot(brand,aes(PageValues, Month, colour= Revenue))+
  geom_point(size=3)
```



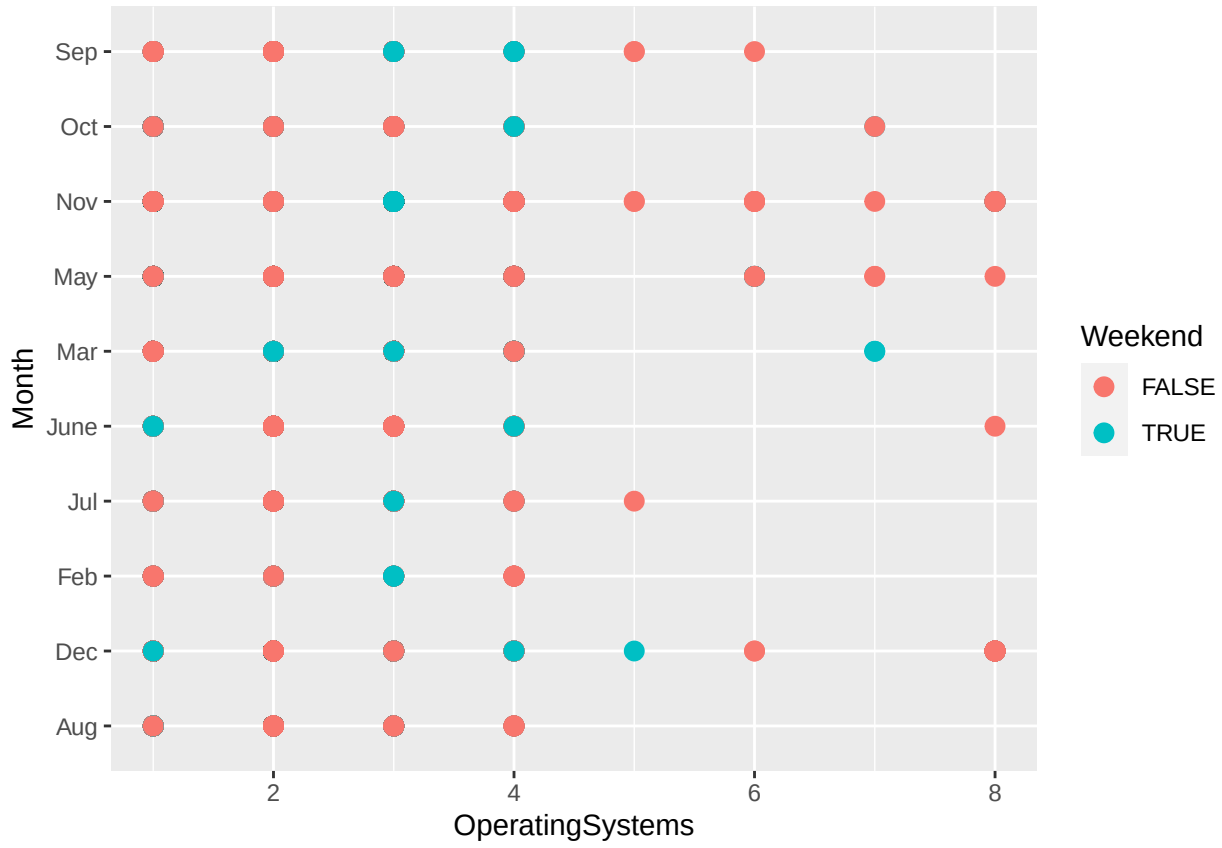
* Page value mostly attracted revenue through out the year.

```
# Month Vs Specialday,  Revenue
#
ggplot(brand,aes(SpecialDay, Month, colour= Revenue))+
  geom_point(size=3)
```



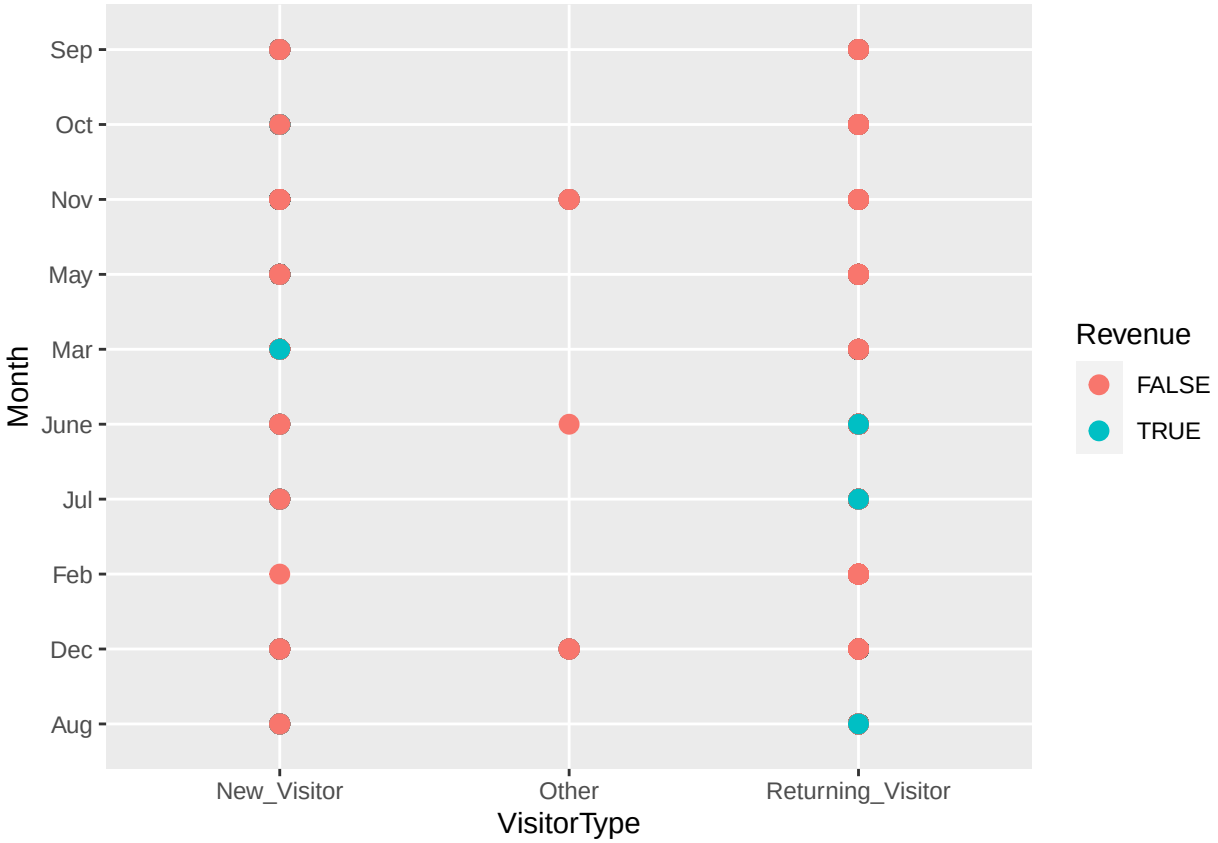
* Revenue was generated on special day in the months of June and July

```
# OperatingSytems Vs Month
#
ggplot(brand,aes(OperatingSystems, Month, colour= Weekend))+
  geom_point(size=3)
```



*

```
# Pagevalues Vs Month
#
ggplot(brand,aes(VisitorType, Month, colour= Revenue))+
  geom_point(size=3)
```



Modelling

```
# converting months to numeric
#
brand$Month <- match(Month,month.abb)
tail(brand)
```

```
##       Administrative Administrative_Duration Informational
## 12208              0                      0             1
## 12209              3                    145             0
## 12210              0                      0             0
## 12211              0                      0             0
## 12212              4                     75             0
## 12213              0                      0             0
##       Informational_Duration ProductRelated ProductRelated_Duration BounceRates
## 12208                      0             16                503.000 0.000000000
## 12209                      0             53               1783.792 0.007142857
## 12210                      0              5                465.750 0.000000000
## 12211                      0              6                184.250 0.083333333
## 12212                      0             15                346.000 0.000000000
## 12213                      0              3                 21.250 0.000000000
##        ExitRates PageValues SpecialDay Month OperatingSystems Browser Region
## 12208 0.03764706    0.00000          0    11                2       2      1
## 12209 0.02903061   12.24172          0    12                4       6      1
## 12210 0.02133333    0.00000          0    11                3       2      1
## 12211 0.08666667    0.00000          0    11                3       2      1
## 12212 0.02105263    0.00000          0    11                2       2      3
```

```
## 12213 0.06666667    0.00000          0    11               3       2      1
##       TrafficType       VisitorType Weekend Revenue
## 12208           1 Returning_Visitor   FALSE   FALSE
## 12209           1 Returning_Visitor    TRUE   FALSE
## 12210           8 Returning_Visitor    TRUE   FALSE
## 12211          13 Returning_Visitor    TRUE   FALSE
## 12212          11 Returning_Visitor   FALSE   FALSE
## 12213           2       New_Visitor    TRUE   FALSE
```

```
# Removing class label before label encoding
#
df <- brand[,c(1:17)]
head(df)
```

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

```
# Standardising the data
#
data <- df %>% select(-VisitorType, -Weekend, -Month) %>%  scale()
head(data)
```

```
##      Administrative Administrative_Duration Informational
## [1,]     -0.7023197              -0.4600000    -0.3986122
## [2,]     -0.7023197              -0.4600000    -0.3986122
## [3,]     -0.7023197              -0.4656356    -0.3986122
## [4,]     -0.7023197              -0.4600000    -0.3986122
```

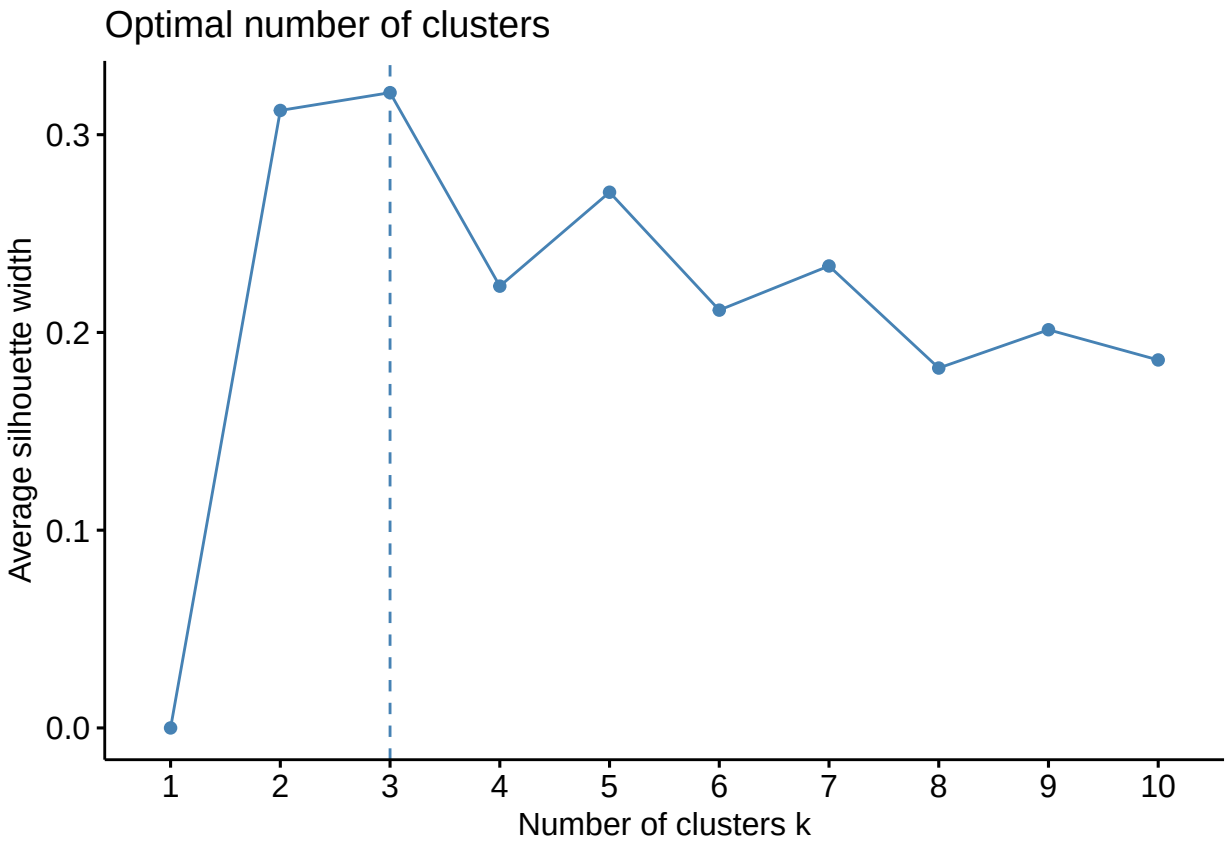
```
## [5,]     -0.7023197              -0.4600000    -0.3986122
## [6,]     -0.7023197              -0.4600000    -0.3986122
##      Informational_Duration ProductRelated ProductRelated_Duration  BounceRates
## [1,]             -0.2461227     -0.6962701              -0.6289255  3.954168629
## [2,]             -0.2461227     -0.6738388              -0.5955748 -0.450250350
## [3,]             -0.2531958     -0.6962701              -0.6294466  3.954168629
## [4,]             -0.2461227     -0.6738388              -0.6275359  0.650854395
## [5,]             -0.2461227     -0.4943884              -0.3019325 -0.009808452
## [6,]             -0.2461227     -0.2925068              -0.5485625 -0.102533055
##       ExitRates PageValues SpecialDay OperatingSystems    Browser     Region
## [1,]  3.4273560 -0.3188341 -0.3101155       -1.2398307 -0.7940299 -0.8962493
## [2,]  1.2650593 -0.3188341 -0.3101155       -0.1369864 -0.2091745 -0.8962493
## [3,]  3.4273560 -0.3188341 -0.3101155        2.0687022 -0.7940299  2.4345662
## [4,]  2.1299780 -0.3188341 -0.3101155        0.9658579 -0.2091745 -0.4798973
## [5,]  0.1839110 -0.3188341 -0.3101155        0.9658579  0.3756809 -0.8962493
## [6,] -0.3661469 -0.3188341 -0.3101155       -0.1369864 -0.2091745 -0.8962493
##      TrafficType
## [1,] -0.76528498
## [2,] -0.51630590
## [3,] -0.26732683
## [4,] -0.01834776
## [5,] -0.01834776
## [6,] -0.26732683
```

K-Means Clustering

```
# Determining the optimal value of k
#
library(factoextra)
```

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

```
fviz_nbclust(data, kmeans, method= "silhouette")
```

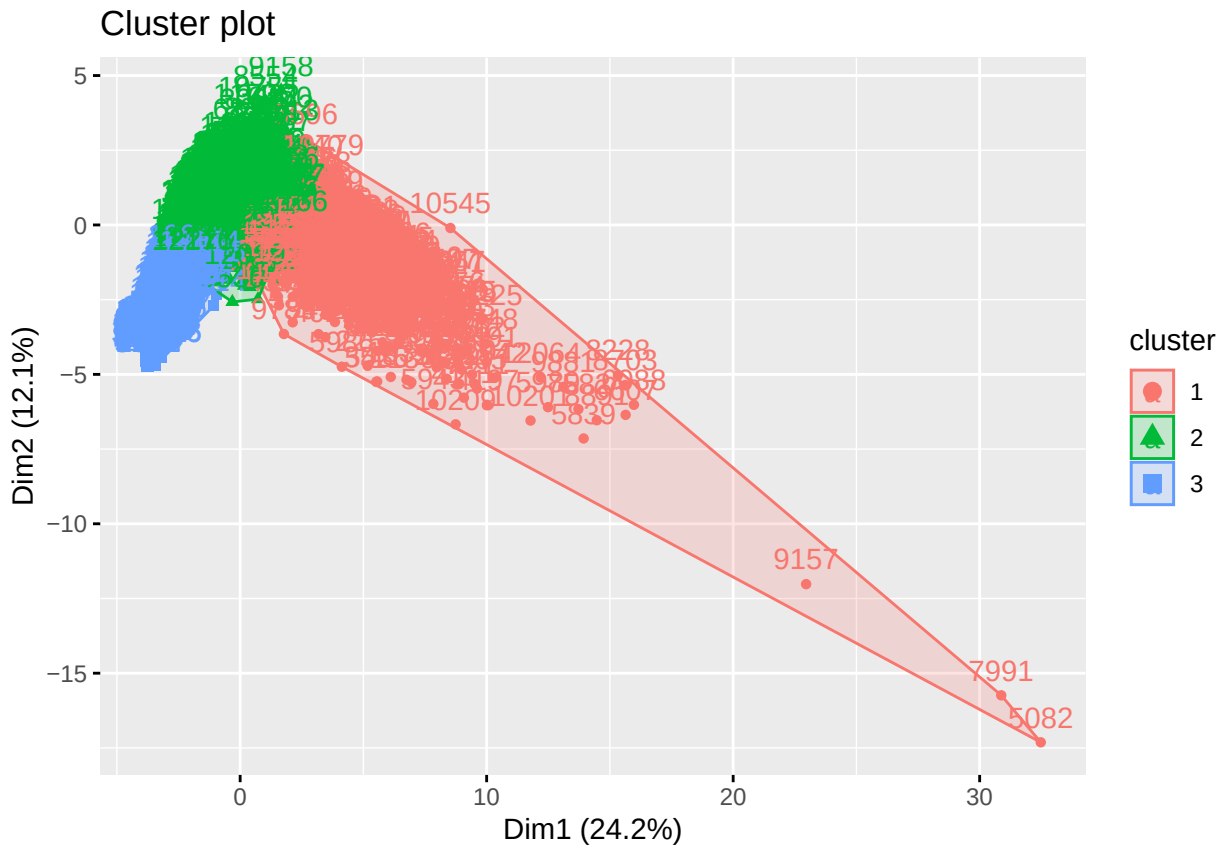


* The optimal k value is 3.

```
# Applying the K-means at k=3
# ---
#
k_clust<- kmeans(data,3)

# Previewing the no. of records in each cluster
#
k_clust$size
```

```
## [1]   952   879 10382
```

```
# Visualizing the  clustering results
#
fviz_cluster(kmeans(data, centers = 3), data = data)
```




```
# displaying the results in a table
#
brand.class<- brand[, "VisitorType"]
table(k_clust$cluster, brand.class)
```

```
##    brand.class
##     New_Visitor Other Returning_Visitor
##   1          48     0               904
##   2          30    12               837
##   3        1615    69              8698
```

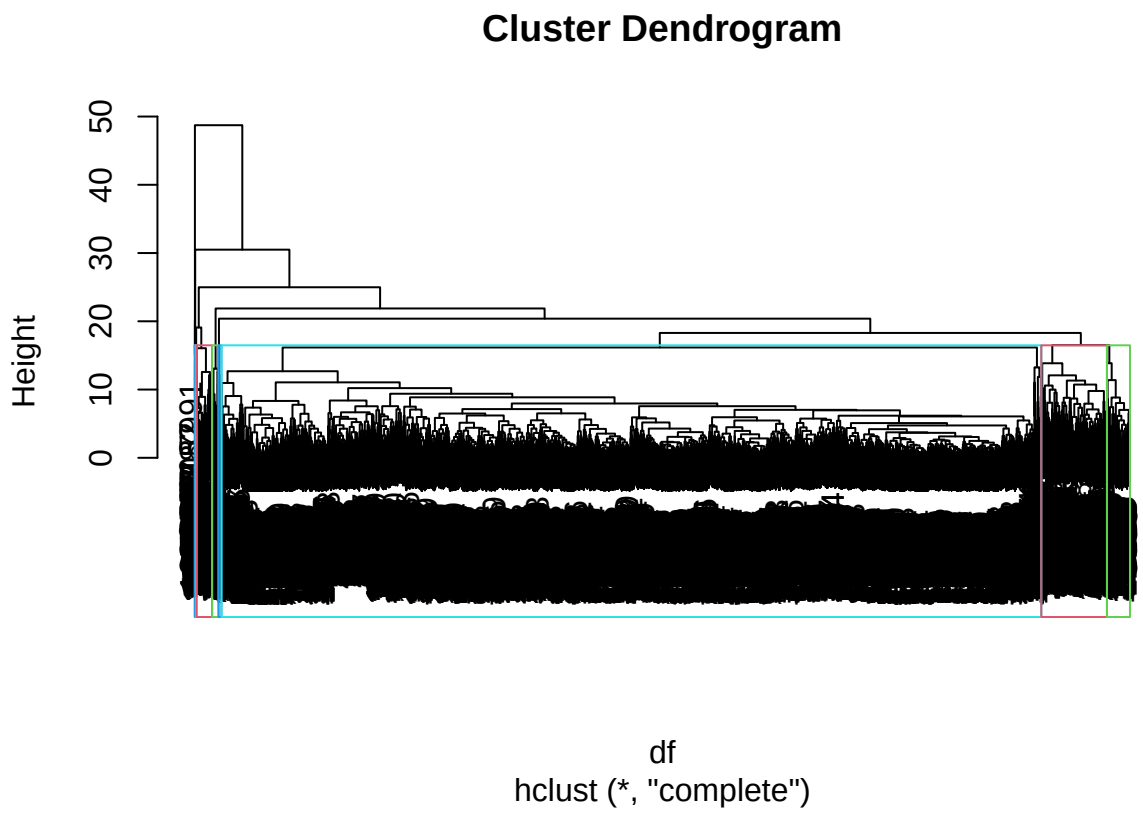
Hierarchical

```
# computing the Euclidean distance between observations,
#
df <- dist(data, method = "euclidean")

# hierarchical clustering using the complete method
#
hc <- hclust(df, method = "complete" )

hc
```

```
##
```

```
## Call:
## hclust(d = df, method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 12213
```

```
# plotting adendrogram
#
plot(hc)
rect.hclust(hc, k = 10, border = 2:5)
```



DBSCAN Clustering

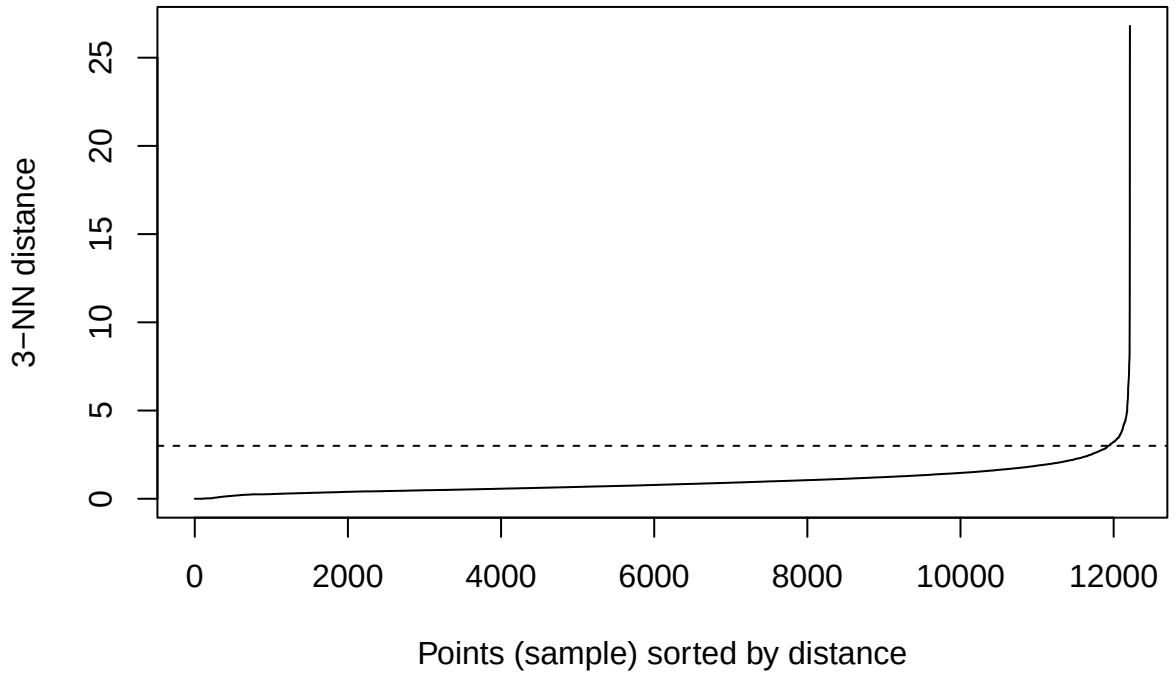
```
# getting optimum eps
#
library(fpc)
library(dbscan)
```

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:fpc':
```

```
##
##     dbscan

## The following object is masked from 'package:VIM':
##
##     kNN
```

```
kNNdistplot(data, k=3)
abline(h = 3, lty=2)
```




```
# Applying dbscan algorithm with the optimal eps = 3
#
db <- dbscan(data,eps=3,MinPts = 4)
```

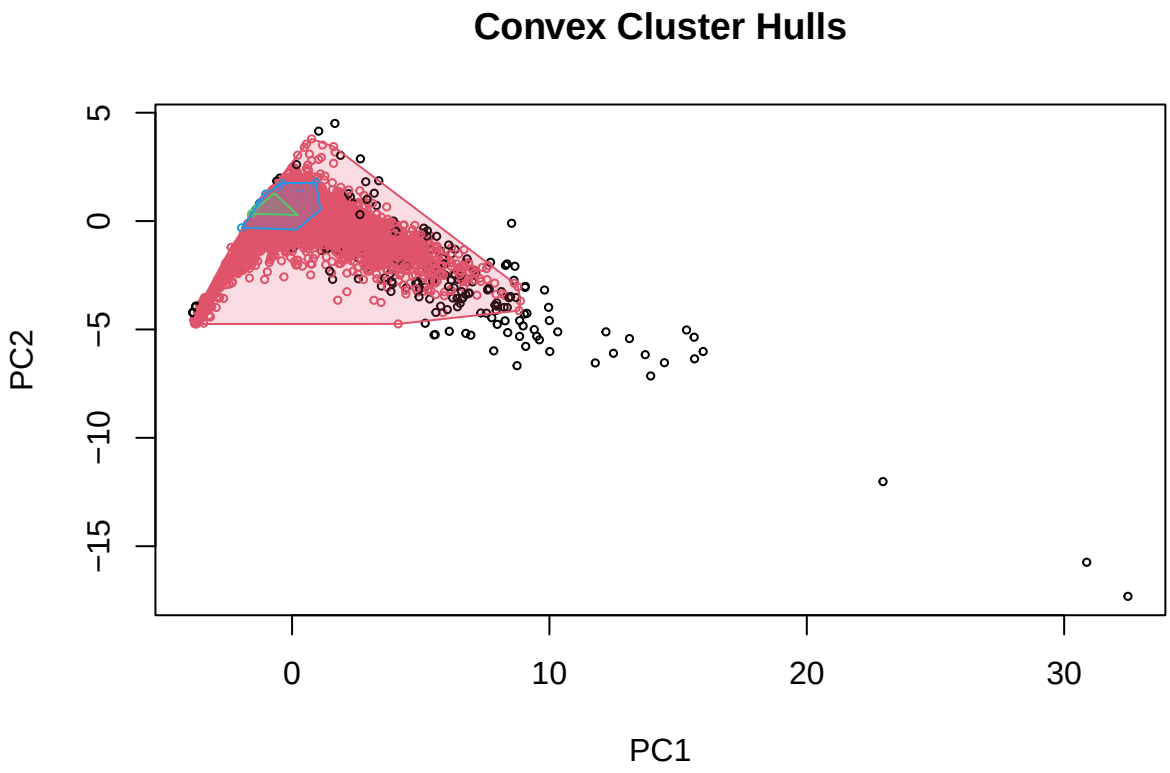
```
## Warning in dbscan(data, eps = 3, MinPts = 4): converting argument MinPts (fpc)
## to minPts (dbscan)!
```

```
db
```

```
## DBSCAN clustering for 12213 objects.
## Parameters: eps = 3, minPts = 4
## The clustering contains 3 cluster(s) and 181 noise points.
##
##     0     1     2     3
```

```
##   181 11985     5    42
##
## Available fields: cluster, eps, minPts
```

```
# Plotting DBSCAN
#

hullplot(data,db$cluster)
```




```
#dispalying the cluster results in a table

table(db$cluster, brand.class)
```

```
##    brand.class
##     New_Visitor Other Returning_Visitor
##   0           9     6               166
##   1        1683    33             10269
##   2           1     0                 4
##   3           0    42                 0
```

conclusion

We conclude that returning visitors are most likely to generate revenue Pagevalue through out the year showed revenue was being genenaated.