Unsurprised

kelvin njunge

9/3/2021

Problem definition

** a) Specifying the question**

Perform clustering stating analysis and visualizations.

b) Defining the metrics for success

Bivariate and univariate Exploratory data analysis perform clustering stating insights drawn from your analysis and visualizations.

c) Understanding the context

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

d) Recording the Experimental Design

- Define the question, the metric for success, the context, experimental design taken.
- Read and explore the given dataset.
- Find and deal with outliers, anomalies, and missing data within the dataset.
- Perform univariate and bivariate analysis.
- Perform clustering stating insights drawn from your analysis and visualizations.

e) Relevance of the data

The data used for this project is necessary for understanding their customer's behavior from data that they have collected over the past year. More specifically, to learn the characteristics of customer groups.

** Data analysis**

** Data sourcing**

```
library(data.table)
library(ggplot2)
library(dplyr)
```

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

## The following objects are masked from 'package:data.table':
##
##     between, first, last

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

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

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

** Importing data**

```
df <- read.csv("C:\\Users\\Ricky\\Documents\\online_shoppers_intention.csv")
```

Previewing the top 6 entries

```
head(df)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                      -1             0                     -1
## 4              0                       0             0                      0
## 5              0                       0             0                      0
## 6              0                       0             0                      0
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              1                0.000000  0.20000000 0.2000000          0
## 2              2               64.000000  0.00000000 0.1000000          0
```

```
## 3              1               -1.000000  0.20000000 0.2000000          0
## 4              2                2.666667  0.05000000 0.1400000          0
## 5             10              627.500000  0.02000000 0.0500000          0
## 6             19              154.216667  0.01578947 0.0245614          0
##   SpecialDay Month OperatingSystems Browser Region TrafficType
## 1          0   Feb                1       1      1           1
## 2          0   Feb                2       2      1           2
## 3          0   Feb                4       1      9           3
## 4          0   Feb                3       2      2           4
## 5          0   Feb                3       3      1           4
## 6          0   Feb                2       2      1           3
##         VisitorType Weekend Revenue
## 1 Returning_Visitor   FALSE   FALSE
## 2 Returning_Visitor   FALSE   FALSE
## 3 Returning_Visitor   FALSE   FALSE
## 4 Returning_Visitor   FALSE   FALSE
## 5 Returning_Visitor    TRUE   FALSE
## 6 Returning_Visitor   FALSE   FALSE
```

```
tail(df)
```

Previewing the bottom 6 entries

```
##       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
```

```
names(df)
```

Previewing the columns of our dataset

```
##  [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

Completeness

```
# checking for missing values
colSums(is.na(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
```

```
getmode <- function(v){
  v=v[nchar(as.character(v))>0]
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]}
```

Replacing missing value with mode

```
for (cols in colnames(df)) {
  if (cols %in% names(df[,sapply(df, is.numeric)])) {
    df<-df%>%mutate(!!cols := replace(!!rlang::sym(cols), is.na(!!rlang::sym(cols)), mean(!!rlang::sym(
  }
  else {
    df<-df%>%mutate(!!cols := replace(!!rlang::sym(cols), !!rlang::sym(cols)=="", getmode(!!rlang::sym(
  }
}
```

```
colSums(is.na(df))
```

```
##          Administrative Administrative_Duration           Informational
##                       0                       0                       0
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                       0                       0                       0
##             BounceRates               ExitRates              PageValues
##                       0                       0                       0
##              SpecialDay                   Month        OperatingSystems
##                       0                       0                       0
##                 Browser                  Region             TrafficType
##                       0                       0                       0
##             VisitorType                 Weekend                 Revenue
##                       0                       0                       0
```

Checking for duplicates

```
sum(duplicated(df))
```

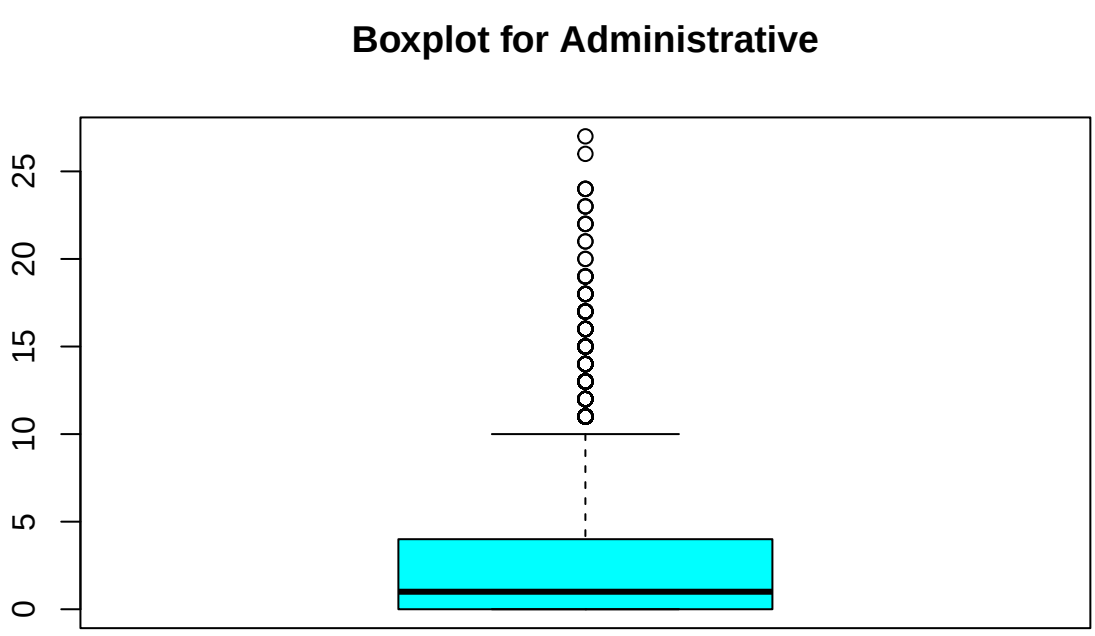
```
## [1] 119
```

We have 119 duplicated rows

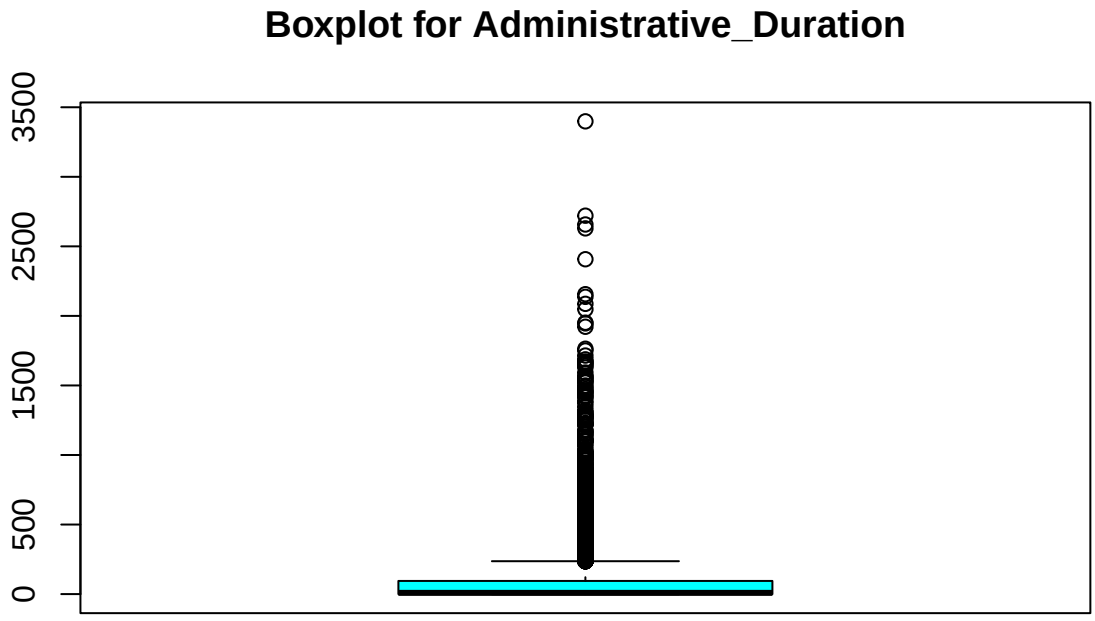
```
# eliminating for duplicates
df <- df[!duplicated(df), ]
```

***checking for outliers**

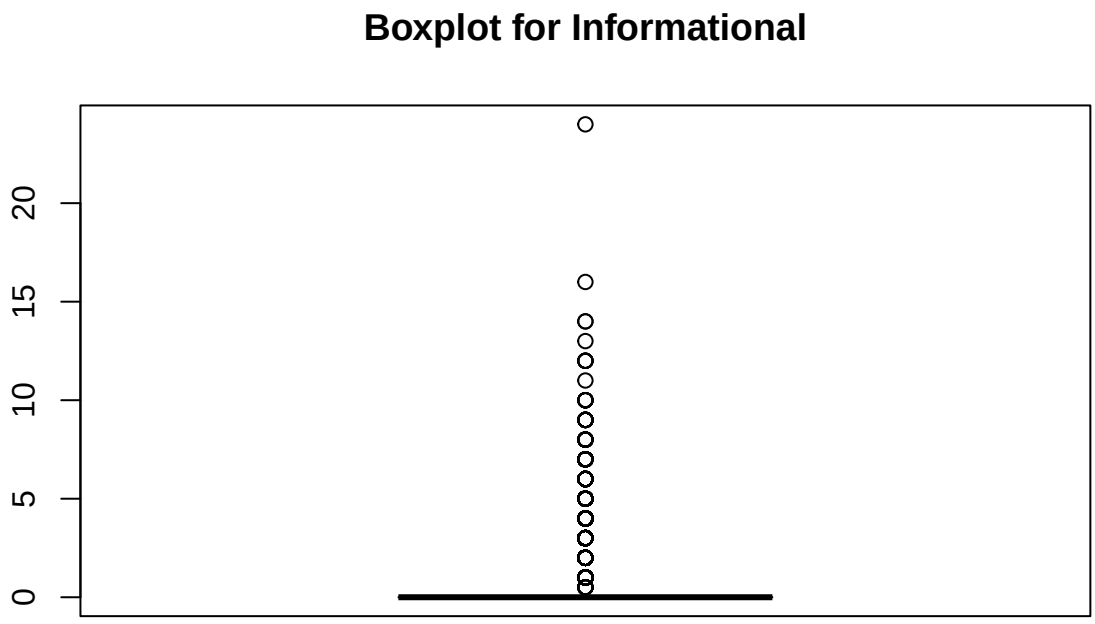
```
boxplot(df$Administrative,main="Boxplot for Administrative",col = "cyan")
```



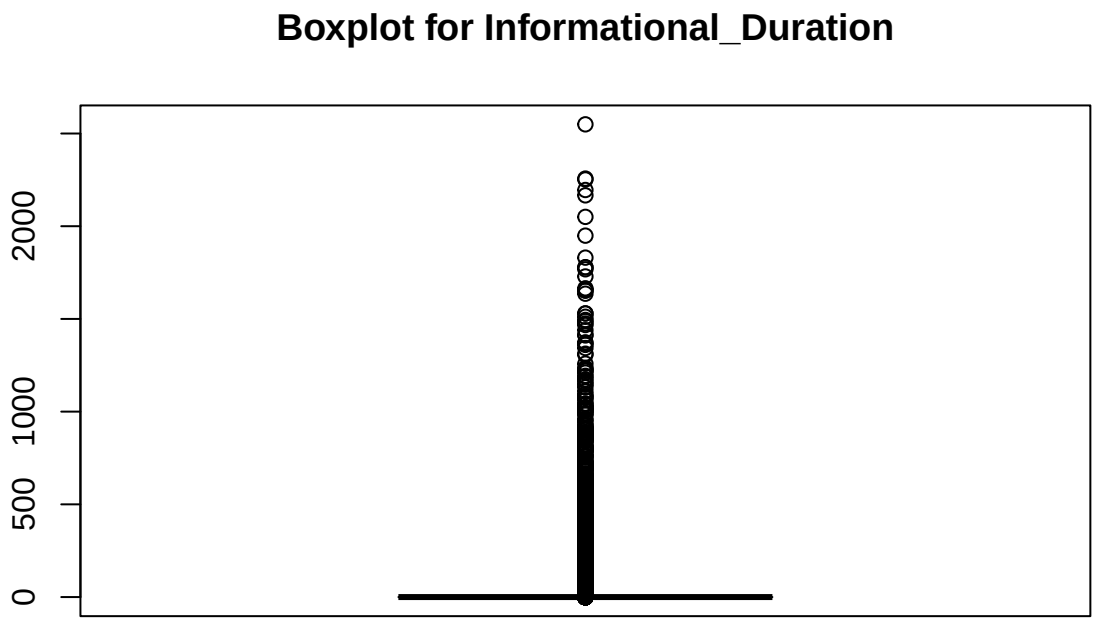

```
boxplot(df$Administrative_Duration,main="Boxplot for Administrative_Duration",col = "cyan")
```



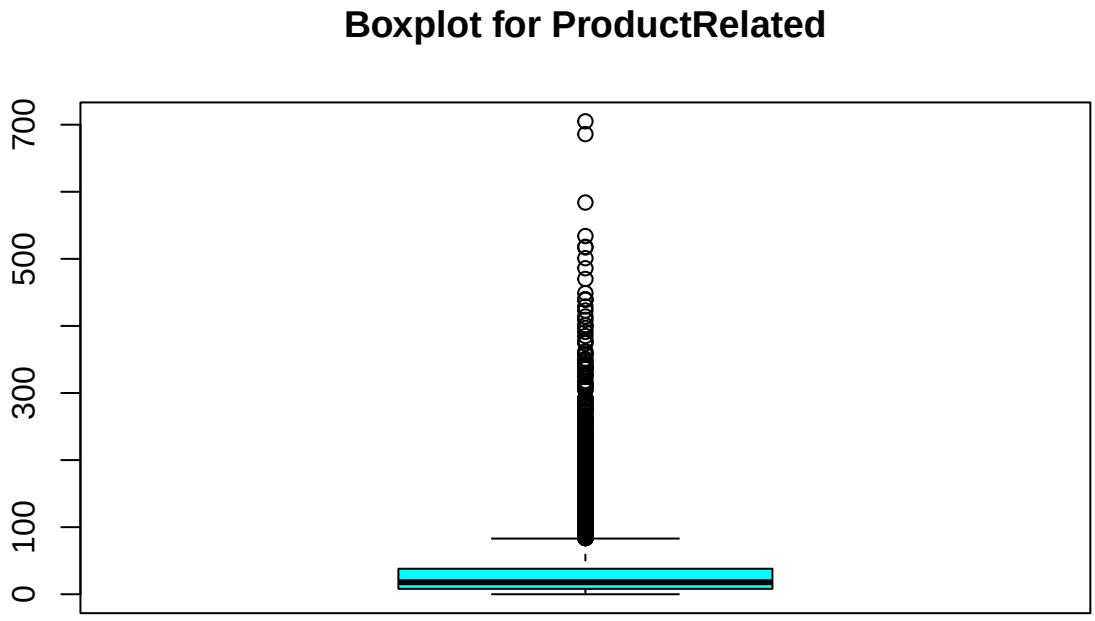

```
boxplot(df$Informational,main="Boxplot for Informational",col = "cyan")
```



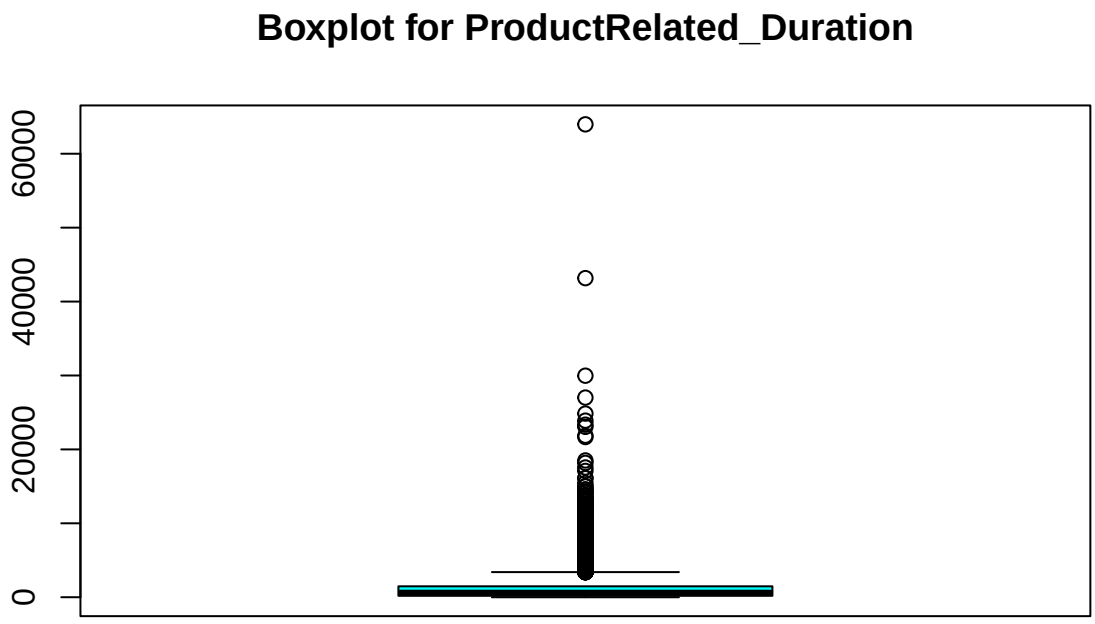

```
boxplot(df$Informational_Duration,main="Boxplot for Informational_Duration",col = "cyan")
```



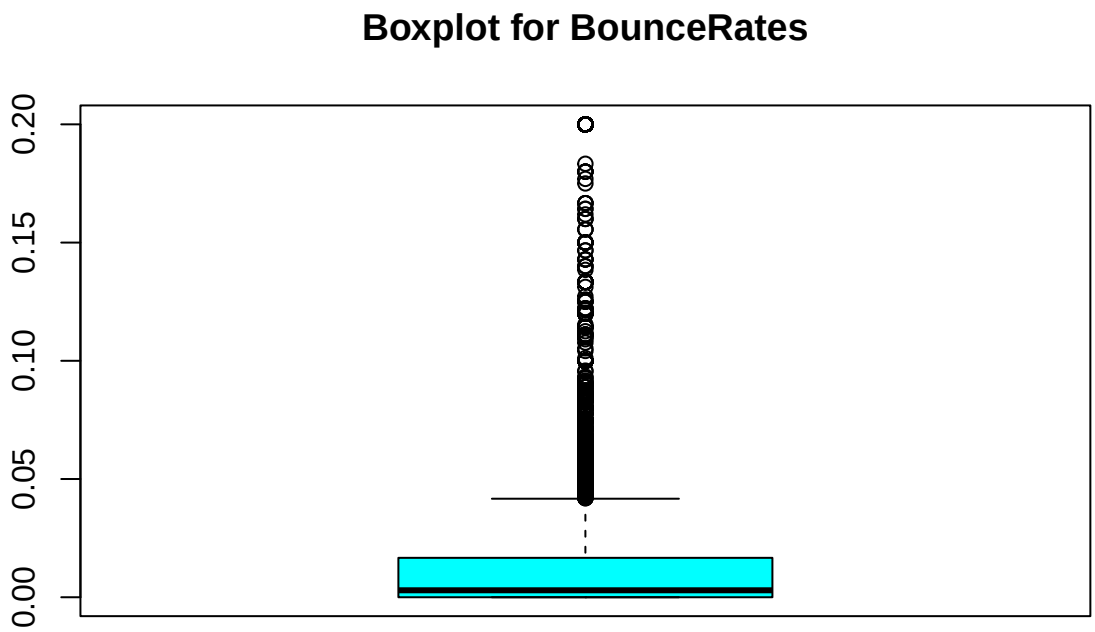

```
boxplot(df$ProductRelated,main="Boxplot for ProductRelated",col = "cyan")
```



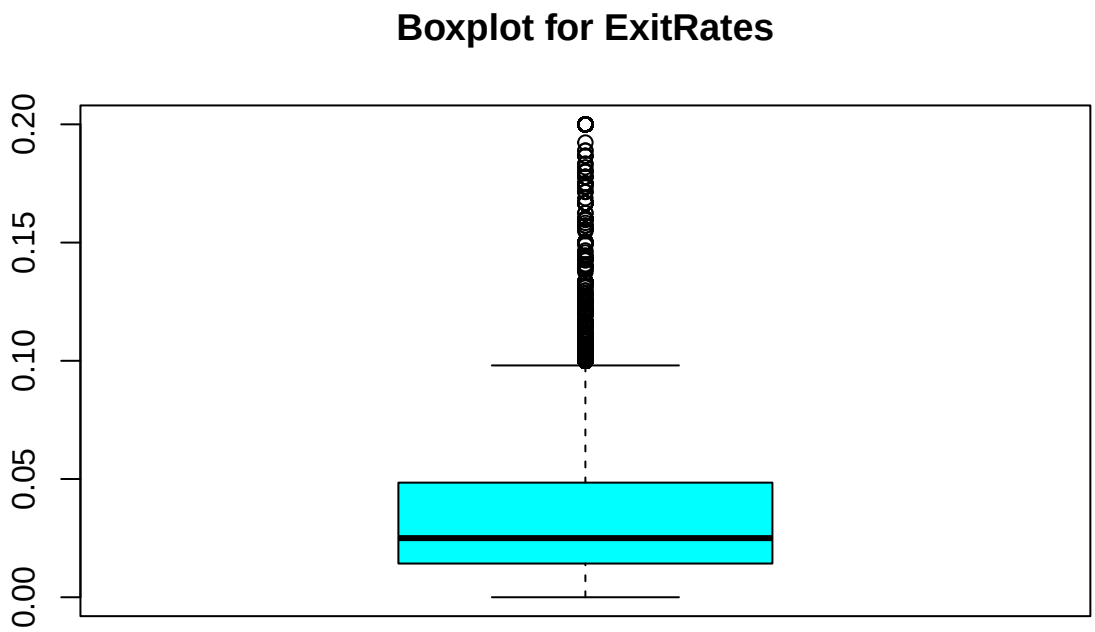

```
boxplot(df$ProductRelated_Duration,main="Boxplot for ProductRelated_Duration",col = "cyan")
```



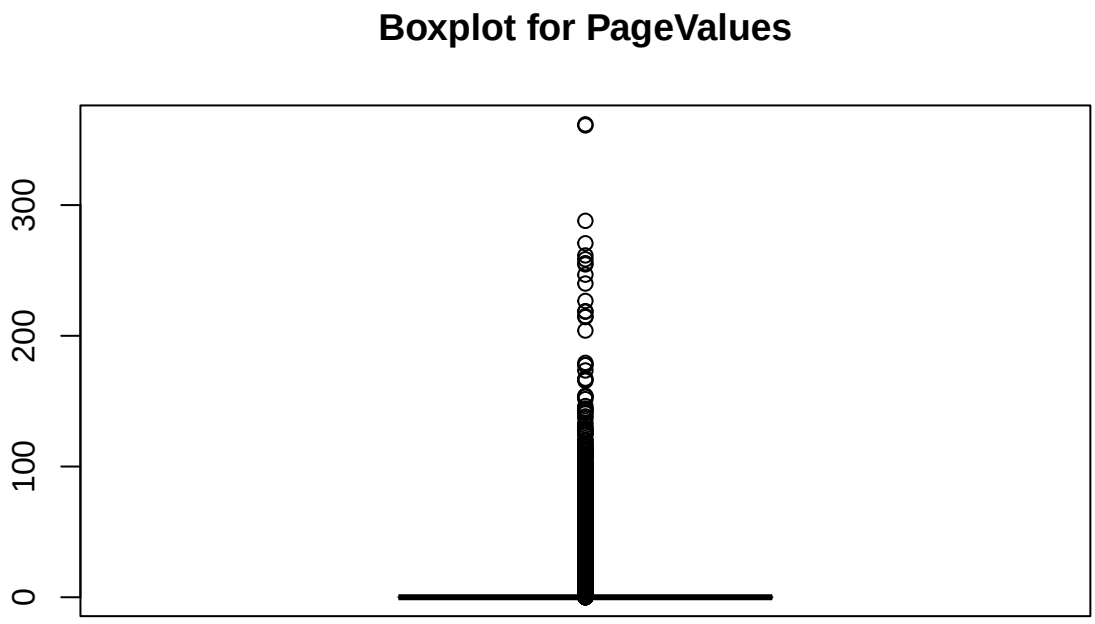

```
boxplot(df$BounceRates,main="Boxplot for BounceRates",col = "cyan")
```



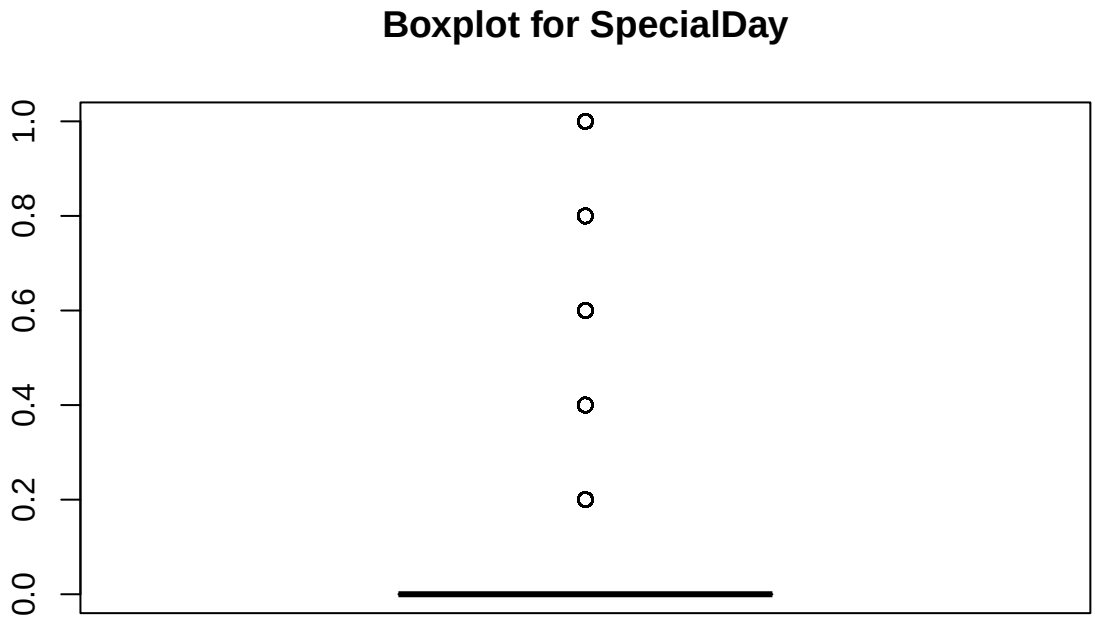

```
boxplot(df$ExitRates,main="Boxplot for ExitRates",col = "cyan")
```



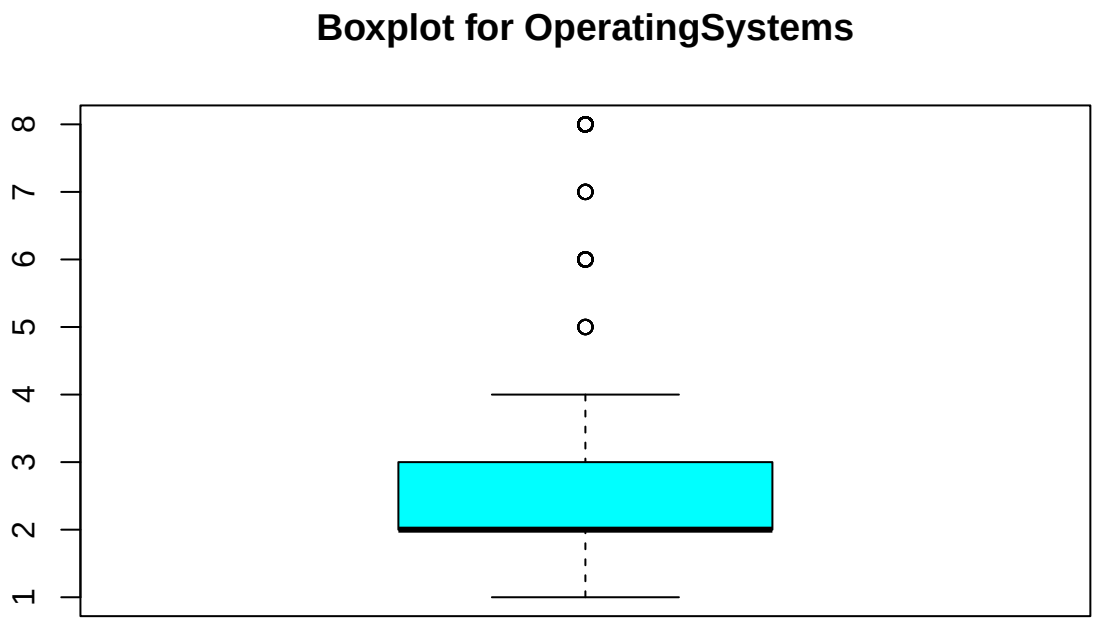

```
boxplot(df$PageValues,main="Boxplot for PageValues",col = "cyan")
```



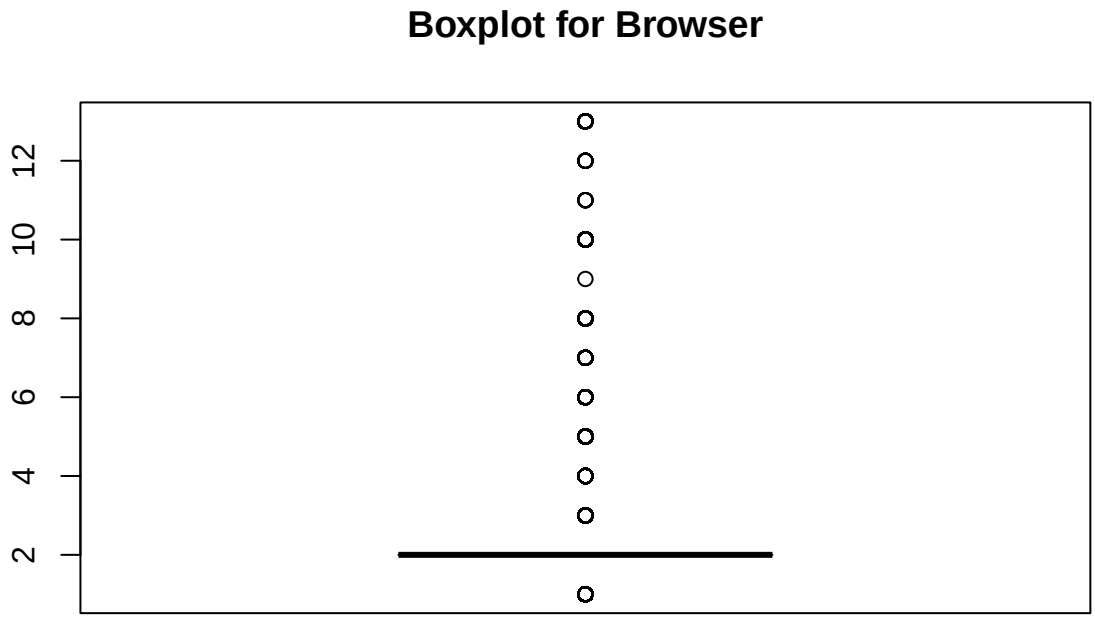

```
boxplot(df$SpecialDay,main="Boxplot for SpecialDay",col = "cyan")
```



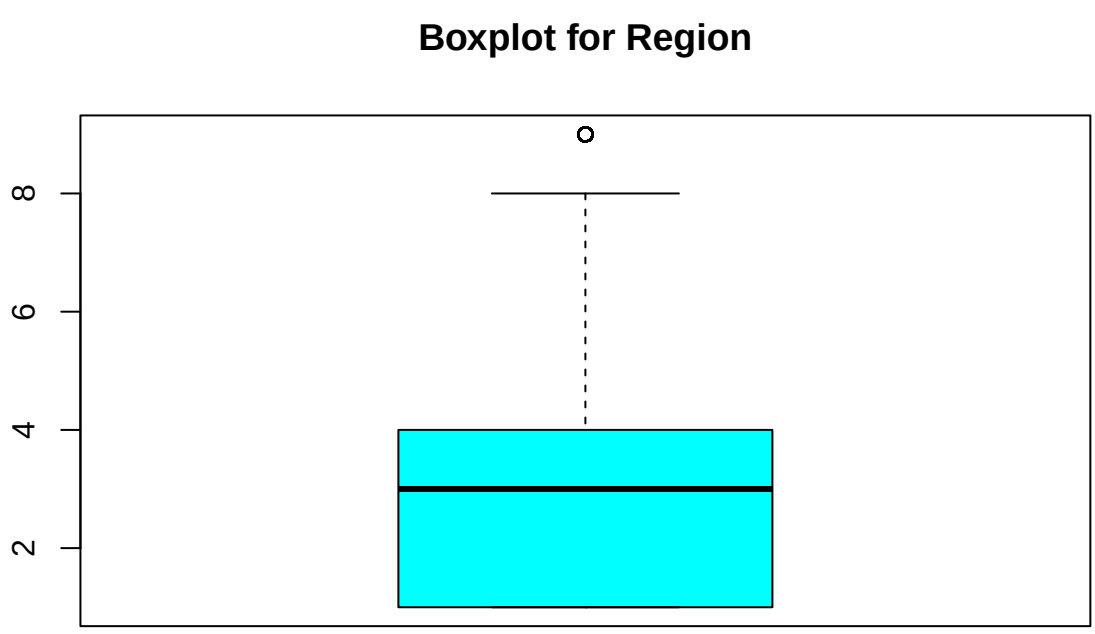

```
boxplot(df$OperatingSystems,main="Boxplot for OperatingSystems",col = "cyan")
boxplot(df$OperatingSystems,main="Boxplot for OperatingSystems",col = "cyan")
```



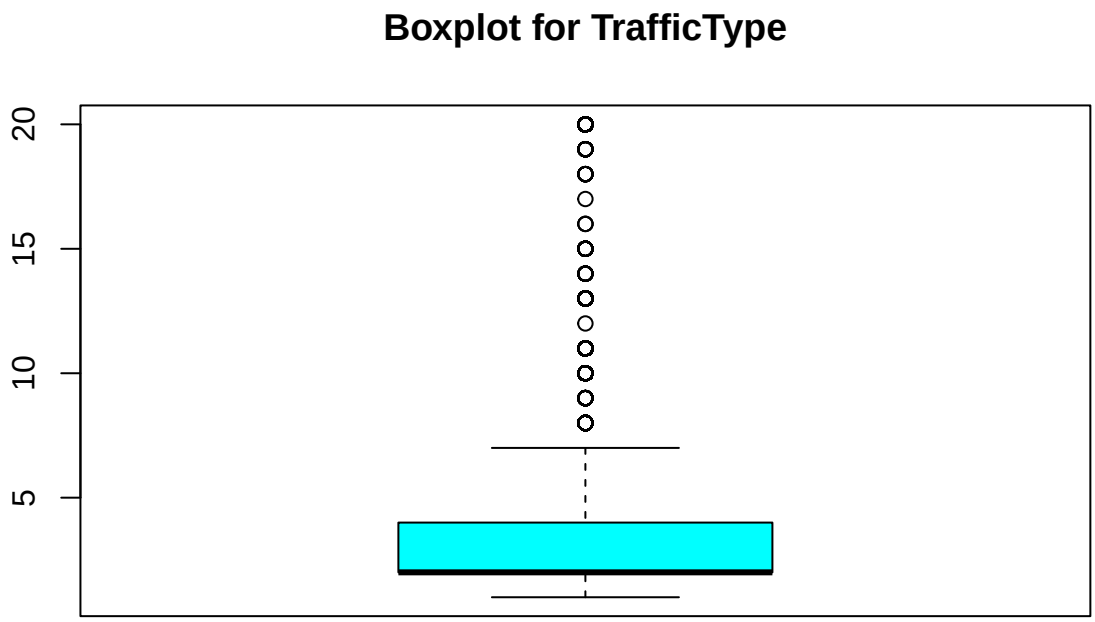

```
boxplot(df$Browser,main="Boxplot for Browser",col = "cyan")
```



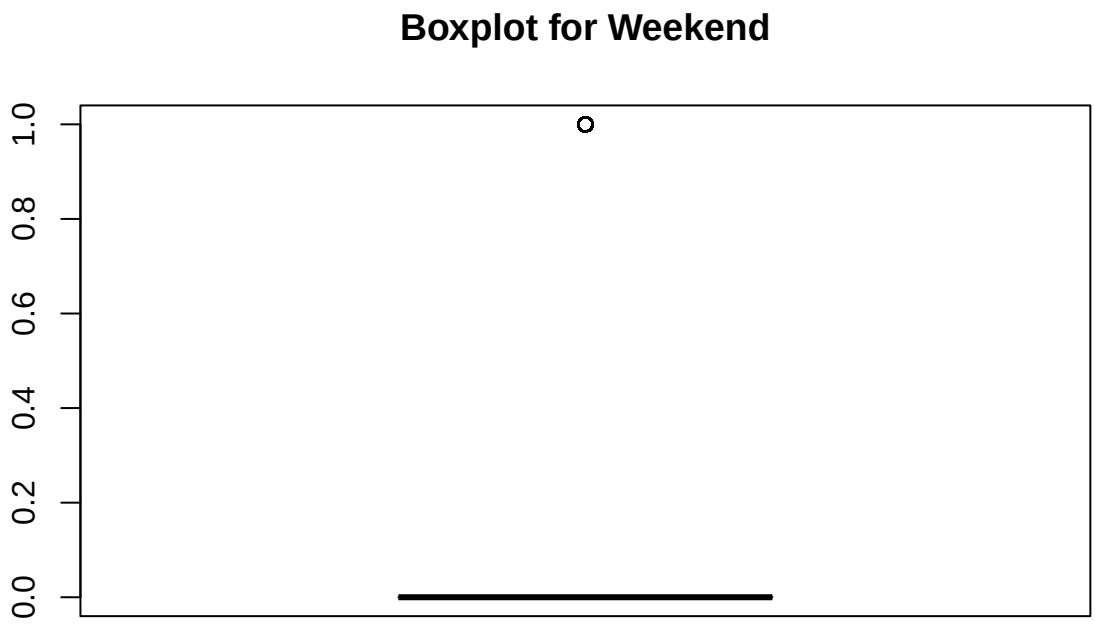

```
boxplot(df$Region,main="Boxplot for Region",col = "cyan")
boxplot(df$Region,main="Boxplot for Region",col = "cyan")
```



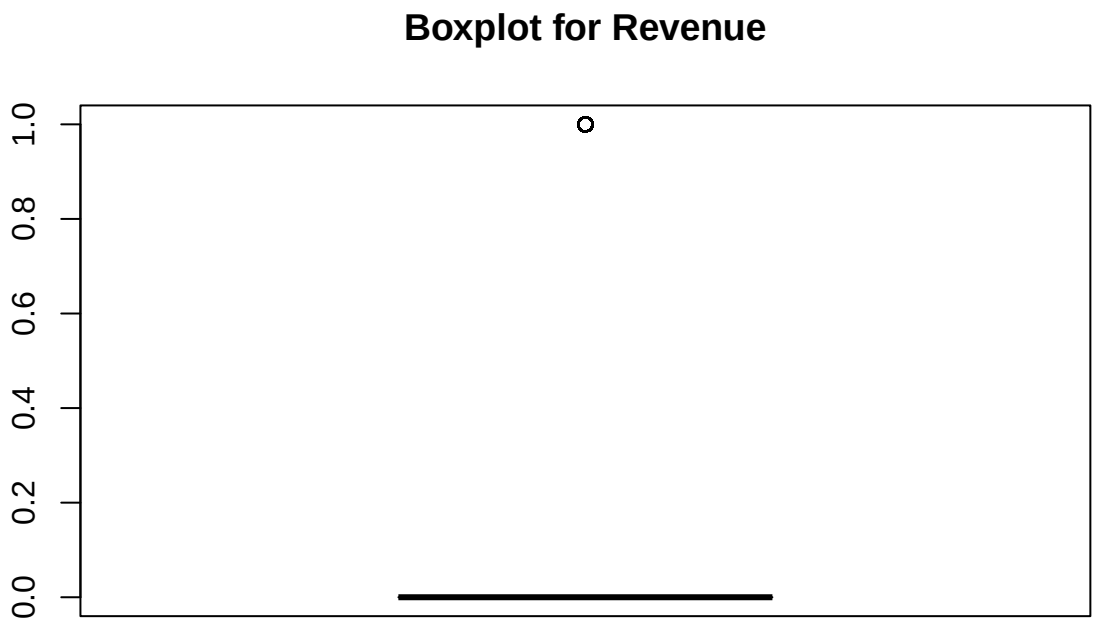

```
boxplot(df$TrafficType,main="Boxplot for TrafficType",col = "cyan")
```




```
boxplot(df$Weekend,main="Boxplot for Weekend",col = "cyan")
```




```
boxplot(df$Revenue,main="Boxplot for Revenue",col = "cyan")
```



4. Univariate

```
cat("The mean for Administrative is",mean(df$Administrative))
```

Measures of central tendency

```
## The mean for Administrative is 2.340006
```

```
cat("\n")
```

```
cat("The mean for Administrative_Duration is",mean(df$Administrative_Duration))
```

```
## The mean for Administrative_Duration is 81.68138
```

```
cat("\n")
```

```
cat("The mean for Informational is",mean(df$Informational))
```

```
## The mean for Informational is 0.5088074
```

```
cat("\n")
```

```
cat("The mean for Informational_Duration is",mean(df$Informational_Duration))
```

```
## The mean for Informational_Duration is 34.83701
```

```
cat("\n")
```

```
cat("The mean for ProductRelated is",mean(df$ProductRelated))
```

```
## The mean for ProductRelated is 32.05816
```

```
cat("\n")
```

```
cat("The mean for ProductRelated_Duration is",mean(df$ProductRelated_Duration))
```

```
## The mean for ProductRelated_Duration is 1207.497
```

```
cat("\n")
```

```
cat("The mean for BounceRates is",mean(df$BounceRates))
```

```
## The mean for BounceRates is 0.02044841
```

```
cat("\n")
```

```
cat("The mean for ExitRates is",mean(df$ExitRates))
```

```
## The mean for ExitRates is 0.04149826
```

```
cat("\n")
```

```
cat("The mean for PageValues is",mean(df$PageValues))
```

```
## The mean for PageValues is 5.946651
```

```
cat("\n")
```

```
cat("The mean for SpecialDay is",mean(df$SpecialDay))
```

```
## The mean for SpecialDay is 0.06191139
```

```
cat("\n")
```

```
cat("The mean for OperatingSystems is",mean(df$OperatingSystems))
```

```
## The mean for OperatingSystems is 2.124232
```

```
cat("\n")
```

```
cat("The mean for Browser is",mean(df$Browser))
```

```
## The mean for Browser is 2.35771
```

```
cat("\n")
```

```
cat("The mean for Region is",mean(df$Region))
```

```
## The mean for Region is 3.152977
```

```
cat("\n")
```

```
cat("The mean for TrafficType is",mean(df$TrafficType))
```

```
## The mean for TrafficType is 4.074032
```

```
cat("\n")
```

```
cat("The mean for Weekend is",mean(df$Weekend))
```

```
## The mean for Weekend is 0.2341332
```

```
cat("\n")
```

```
cat("The mean for Revenue is",mean(df$Revenue))
```

```
## The mean for Revenue is 0.1562526
```

```
cat("The median for Administrative is",median(df$Administrative))
```

Median

```
## The median for Administrative is 1
```

```
cat("\n")
```

```
cat("The median for Administrative_Duration is",mean(df$Administrative_Duration))
```

```
## The median for Administrative_Duration is 81.68138
```

```
cat("\n")
```

```
cat("The median for Informational is",median(df$Informational))
```

```
## The median for Informational is 0
```

```
cat("\n")
```

```
cat("The median for Informational_Duration is",mean(df$Informational_Duration))
```

```
## The median for Informational_Duration is 34.83701
```

```
cat("\n")
```

```
cat("The median for ProductRelated is",median(df$ProductRelated))
```

```
## The median for ProductRelated is 18
```

```
cat("\n")
```

```
cat("The median for ProductRelated_Duration is",median(df$ProductRelated_Duration))
```

```
## The median for ProductRelated_Duration is 611
```

```
cat("\n")
```

```
cat("The median for BounceRates is",median(df$BounceRates))
```

```
## The median for BounceRates is 0.002941176
```

```
cat("\n")
```

```
cat("The median for ExitRates is",median(df$ExitRates))
```

```
## The median for ExitRates is 0.025
```

```
cat("\n")
```

```
cat("The median for PageValues is",median(df$PageValues))
```

```
## The median for PageValues is 0
```

```
cat("\n")
```

```
cat("The median for SpecialDay is",median(df$SpecialDay))
```

```
## The median for SpecialDay is 0
```

```
cat("\n")
```

```
cat("The median for OperatingSystems is",median(df$OperatingSystems))
```

```
## The median for OperatingSystems is 2
```

```
cat("\n")
```

```
cat("The median for Browser is",median(df$Browser))
```

```
## The median for Browser is 2
```

```
cat("\n")
```

```
cat("The median for Region is",median(df$Region))
```

```
## The median for Region is 3
```

```
cat("\n")
```

```
cat("The median for TrafficType is",median(df$TrafficType))
```

```
## The median for TrafficType is 2
```

```
cat("\n")
```

```
cat("The median for Weekend is",median(df$Weekend))
```

```
## The median for Weekend is FALSE
```

```
cat("\n")
```

```
cat("The median for Revenue is",median(df$Revenue))
```

```
## The median for Revenue is FALSE
```

```
cat("\n")
```

Mode

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

```r
cat("The mode for Administrative is",getmode(df$Administrative))
```

```
## The mode for Administrative is 0
```

```r
cat("\n")
```

```r
cat("The mode for Administrative_Duration is",getmode(df$Administrative_Duration))
```

```
## The mode for Administrative_Duration is 0
```

```r
cat("\n")
```

```r
cat("The mode for Informational is",getmode(df$Informational))
```

```
## The mode for Informational is 0
```

```r
cat("\n")
```

```r
cat("The mode for Informational_Duration is",getmode(df$Informational_Duration))
```

```
## The mode for Informational_Duration is 0
```

```r
cat("\n")
```

```r
cat("The mode for ProductRelated is",getmode(df$ProductRelated))
```

```
## The mode for ProductRelated is 1
```

```r
cat("\n")
```

```r
cat("The mode for ProductRelated_Duration is",getmode(df$ProductRelated_Duration))
```

```
## The mode for ProductRelated_Duration is 0
```

```r
cat("\n")
```

```r
cat("The mode for BounceRates is",getmode(df$BounceRates))
```

```
## The mode for BounceRates is 0
```

```
cat("\n")
```

```
cat("The mode for ExitRates is",getmode(df$ExitRates))
```

```
## The mode for ExitRates is 0.2
```

```
cat("\n")
```

```
cat("The mode for PageValues is",getmode(df$PageValues))
```

```
## The mode for PageValues is 0
```

```
cat("\n")
```

```
cat("The mode for SpecialDay is",getmode(df$SpecialDay))
```

```
## The mode for SpecialDay is 0
```

```
cat("\n")
```

```
cat("The mode for OperatingSystems is",getmode(df$OperatingSystems))
```

```
## The mode for OperatingSystems is 2
```

```
cat("\n")
```

```
cat("The mode for Browser is",getmode(df$Browser))
```

```
## The mode for Browser is 2
```

```
cat("\n")
```

```
cat("The mode for Region is",getmode(df$Region))
```

```
## The mode for Region is 1
```

```
cat("\n")
```

```
cat("The mode for TrafficType is",getmode(df$TrafficType))
```

```
## The mode for TrafficType is 2
```

```
cat("\n")
```

```
cat("The standard deviation for Administrative is",sd(df$Administrative))
```

Standard deviation

```
## The standard deviation for Administrative is 3.329214
```

```
cat("\n")
```

```
cat("The standard deviation for Administrative_Duration is",sd(df$Administrative_Duration))
```

```
## The standard deviation for Administrative_Duration is 177.4409
```

```
cat("\n")
```

```
cat("The standard deviation for Informational is",sd(df$Informational))
```

```
## The standard deviation for Informational is 1.27519
```

```
cat("\n")
```

```
cat("The standard deviation for Informational_Duration is",sd(df$Informational_Duration))
```

```
## The standard deviation for Informational_Duration is 141.389
```

```
cat("\n")
```

```
cat("The standard deviation for ProductRelated is",sd(df$ProductRelated))
```

```
## The standard deviation for ProductRelated is 44.57899
```

```
cat("\n")
```

```
cat("The standard deviation for ProductRelated_Duration is",sd(df$ProductRelated_Duration))
```

```
## The standard deviation for ProductRelated_Duration is 1918.984
```

```
cat("\n")
```

```
cat("The standard deviation for BounceRates is",sd(df$BounceRates))
```

```
## The standard deviation for BounceRates is 0.04538022
```

```
cat("\n")
```

```
cat("The standard deviation for ExitRates is",sd(df$ExitRates))
```

```
## The standard deviation for ExitRates is 0.04622445
```

```
cat("\n")
```

```
cat("The standard deviation for PageValues is",sd(df$PageValues))
```

```
## The standard deviation for PageValues is 18.64955
```

```
cat("\n")
```

```
cat("The standard deviation for SpecialDay is",sd(df$SpecialDay))
```

```
## The standard deviation for SpecialDay is 0.1996219
```

```
cat("\n")
```

```
cat("The standard deviation for OperatingSystems is",sd(df$OperatingSystems))
```

```
## The standard deviation for OperatingSystems is 0.9068192
```

```
cat("\n")
```

```
cat("The standard deviation for Browser is",sd(df$Browser))
```

```
## The standard deviation for Browser is 1.709958
```

```
cat("\n")
```

```
cat("The standard deviation for Region is",sd(df$Region))
```

```
## The standard deviation for Region is 2.401853
```

```
cat("\n")
```

```
cat("The standard deviation for TrafficType is",sd(df$TrafficType))
```

```
## The standard deviation for TrafficType is 4.016643
```

```
cat("\n")
```

```
cat("The Variance for Administrative is",var(df$Administrative))
```

***Variance**

```
## The Variance for Administrative is 11.08367
```

```
cat("\n")
```

```
cat("The Variance for Administrative_Duration is",var(df$Administrative_Duration))
```

```
## The Variance for Administrative_Duration is 31485.28
```

```
cat("\n")
```

```
cat("The Variance for Informational is",var(df$Informational))
```

```
## The Variance for Informational is 1.62611
```

```
cat("\n")
```

```
cat("The Variance for Informational_Duration is",var(df$Informational_Duration))
```

```
## The Variance for Informational_Duration is 19990.84
```

```
cat("\n")
```

```
cat("The Variance for ProductRelated is",var(df$ProductRelated))
```

```
## The Variance for ProductRelated is 1987.286
```

```
cat("\n")
```

```
cat("The Variance for ProductRelated_Duration is",var(df$ProductRelated_Duration))
```

```
## The Variance for ProductRelated_Duration is 3682499
```

```
cat("\n")
```

```
cat("The Variance for BounceRates is",var(df$BounceRates))
```

```
## The Variance for BounceRates is 0.002059364
```

```
cat("\n")
```

```
cat("The Variance for ExitRates is",var(df$ExitRates))
```

```
## The Variance for ExitRates is 0.0021367
```

```
cat("\n")
```

```
cat("The Variance for PageValues is",var(df$PageValues))
```

```
## The Variance for PageValues is 347.8058
```

```
cat("\n")
```

```
cat("The Variance for SpecialDay is",var(df$SpecialDay))
```

```
## The Variance for SpecialDay is 0.03984889
```

```
cat("\n")
```

```
cat("The Variance for OperatingSystems is",var(df$OperatingSystems))
```

```
## The Variance for OperatingSystems is 0.822321
```

```
cat("\n")
```

```
cat("The Variance for Browser is",var(df$Browser))
```

```
## The Variance for Browser is 2.923958
```

```
cat("\n")
```

```
cat("The Variance for Region is",var(df$Region))
```

```
## The Variance for Region is 5.768898
```

```
cat("\n")
```

```
cat("The Variance for TrafficType is",var(df$TrafficType))
```

```
## The Variance for TrafficType is 16.13342
```

```
cat("\n")
```

Measures of Dispersion

```
library(dplyr)
df %>% summarise_if(is.numeric,min)
```

Minimum

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       -1             0                      -1
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1              0                       -1           0         0          0
##   SpecialDay OperatingSystems Browser Region TrafficType
## 1          0                1       1      1           1
```

```
#Maximum of the columns
df %>% summarise_if(is.numeric,max)
```

Maximum

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1             27                 3398.75            24                2549.375
##   ProductRelated ProductRelated_Duration BounceRates ExitRates PageValues
## 1            705                63973.52         0.2       0.2   361.7637
##   SpecialDay OperatingSystems Browser Region TrafficType
## 1          1                8      13      9          20
```

Quantile

```
cat("The quantile for Administrative is",quantile(df$Administrative))
```

```
## The quantile for Administrative is 0 0 1 4 27
```

```
cat("\n")
```

```
cat("The quantile for Administrative_Duration is",quantile(df$Administrative_Duration))
```

```
## The quantile for Administrative_Duration is -1 0 9 94.6 3398.75
```

```
cat("\n")
```

```
cat("The quantile for Informational is",quantile(df$Informational))
```

```
## The quantile for Informational is 0 0 0 0 24
```

```
cat("\n")
```

```
cat("The quantile for Informational_Duration is",range(df$Informational_Duration))
```

```
## The quantile for Informational_Duration is -1 2549.375
```

```
cat("\n")
```

```
cat("The quantile for ProductRelated is",quantile(df$ProductRelated))
```

```
## The quantile for ProductRelated is 0 8 18 38 705
```

```
cat("\n")
```

```
cat("The quantile for ProductRelated_Duration is",quantile(df$ProductRelated_Duration))
```

```
## The quantile for ProductRelated_Duration is -1 194 611 1476.4 63973.52
```

```
cat("\n")
```

```
cat("The quantile for BounceRates is",quantile(df$BounceRates))
```

```
## The quantile for BounceRates is 0 0 0.002941176 0.01666667 0.2
```

```
cat("\n")
```

```
cat("The quantile for ExitRates is",quantile(df$ExitRates))
```

```
## The quantile for ExitRates is 0 0.01425523 0.025 0.04846603 0.2
```

```
cat("\n")
```

```
cat("The quantile for PageValues is",quantile(df$PageValues))
```

```
## The quantile for PageValues is 0 0 0 0 361.7637
```

```
cat("\n")
```

```
cat("The quantile for SpecialDay is",quantile(df$SpecialDay))
```

```
## The quantile for SpecialDay is 0 0 0 0 1
```

```
cat("\n")
```

```
cat("The quantile for OperatingSystems is",quantile(df$OperatingSystems))
```

```
## The quantile for OperatingSystems is 1 2 2 3 8
```

```
cat("\n")
```

```
cat("The quantile for Browser is",quantile(df$Browser))
```

```
## The quantile for Browser is 1 2 2 2 13
```

```
cat("\n")
```

```
cat("The quantile for Region is",quantile(df$Region))
```

```
## The quantile for Region is 1 1 3 4 9
```

```
cat("\n")
```

```
cat("The quantile for TrafficType is",quantile(df$TrafficType))
```

```
## The quantile for TrafficType is 1 2 2 4 20
```

```
cat("\n")
```

```
cat("The range for Administrative is",range(df$Administrative))
```

Range

```
## The range for Administrative is 0 27
```

```
cat("\n")
```

```
cat("The range for Administrative_Duration is",range(df$Administrative_Duration))
```

```
## The range for Administrative_Duration is -1 3398.75
```

```
cat("\n")
```

```
cat("The range for Informational is",range(df$Informational))
```

```
## The range for Informational is 0 24
```

```
cat("\n")
```

```
cat("The range for Informational_Duration is",range(df$Informational_Duration))
```

```
## The range for Informational_Duration is -1 2549.375
```

```
cat("\n")
```

```
cat("The range for ProductRelated is",range(df$ProductRelated))
```

```
## The range for ProductRelated is 0 705
```

```
cat("\n")
```

```
cat("The range for ProductRelated_Duration is",range(df$ProductRelated_Duration))
```

```
## The range for ProductRelated_Duration is -1 63973.52
```

```
cat("\n")
```

```
cat("The range for BounceRates is",range(df$BounceRates))
```

```
## The range for BounceRates is 0 0.2
```

```
cat("\n")
```

```
cat("The range for ExitRates is",range(df$ExitRates))
```

```
## The range for ExitRates is 0 0.2
```

```
cat("\n")
```

```
cat("The range for PageValues is",range(df$PageValues))
```

```
## The range for PageValues is 0 361.7637
```

```
cat("\n")
```

```
cat("The range for SpecialDay is",range(df$SpecialDay))
```

```
## The range for SpecialDay is 0 1
```

```
cat("\n")
```

```
cat("The range for OperatingSystems is",range(df$OperatingSystems))
```

```
## The range for OperatingSystems is 1 8
```

```
cat("\n")
```

```
cat("The range for Browser is",range(df$Browser))
```

```
## The range for Browser is 1 13
```

```
cat("\n")
```

```
cat("The range for Region is",range(df$Region))
```

```
## The range for Region is 1 9
```

```
cat("\n")
```

```
cat("The range for TrafficType is",range(df$TrafficType))
```

```
## The range for TrafficType is 1 20
```

```
cat("\n")
```

Summary

```
summary(df)
```

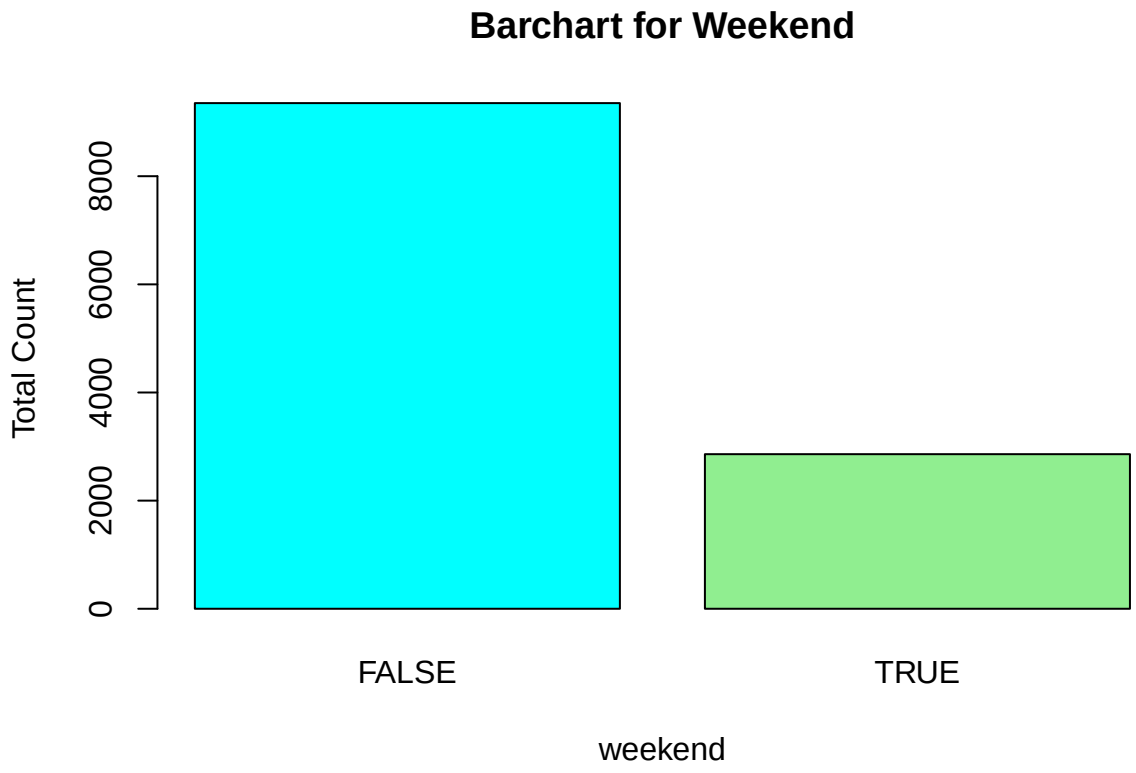
```
##  Administrative  Administrative_Duration Informational
##  Min.   : 0.00   Min.   :  -1.00         Min.   : 0.0000
##  1st Qu.: 0.00   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.00   Median :   9.00         Median : 0.0000
##  Mean   : 2.34   Mean   :  81.68         Mean   : 0.5088
##  3rd Qu.: 4.00   3rd Qu.:  94.60         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
##  1st Qu.:   0.00        1st Qu.:  8.00   1st Qu.:  194
##  Median :   0.00        Median : 18.00   Median :  611
##  Mean   :  34.84        Mean   : 32.06   Mean   : 1208
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1476
##  Max.   :2549.38        Max.   :705.00   Max.   :63974
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01426   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.002941   Median :0.02500   Median :  0.000   Median :0.00000
##  Mean   :0.020448   Mean   :0.04150   Mean   :  5.947   Mean   :0.06191
##  3rd Qu.:0.016667   3rd Qu.:0.04847   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##     Month           OperatingSystems    Browser           Region
##  Length:12211       Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mode  :character   Median :2.000    Median : 2.000   Median :3.000
##                     Mean   :2.124    Mean   : 2.358   Mean   :3.153
##                     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.   : 1.000   Length:12211       Mode :logical   Mode :logical
##  1st Qu.: 2.000   Class :character   FALSE:9352      FALSE:10303
##  Median : 2.000   Mode  :character   TRUE :2859      TRUE :1908
##  Mean   : 4.074
##  3rd Qu.: 4.000
##  Max.   :20.000
```

```
frequency <- table(df$Weekend)
frequency
```

Barcharts

```
##
## FALSE  TRUE
##  9352  2859
```

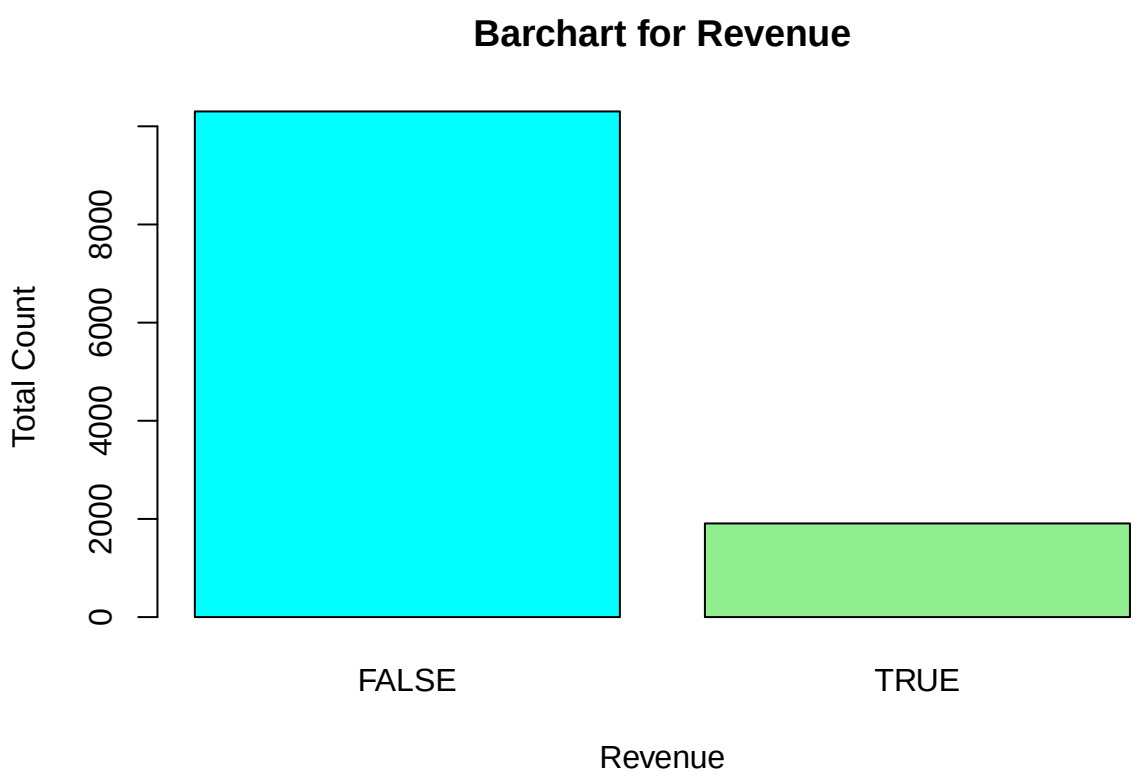
```
barplot(frequency,col=c("Cyan","lightgreen"),main="Barchart for Weekend",xlab = "weekend",ylab = "Total
```




```
frequency <- table(df$Revenue)
frequency
```

```
##
## FALSE  TRUE
## 10303  1908
```

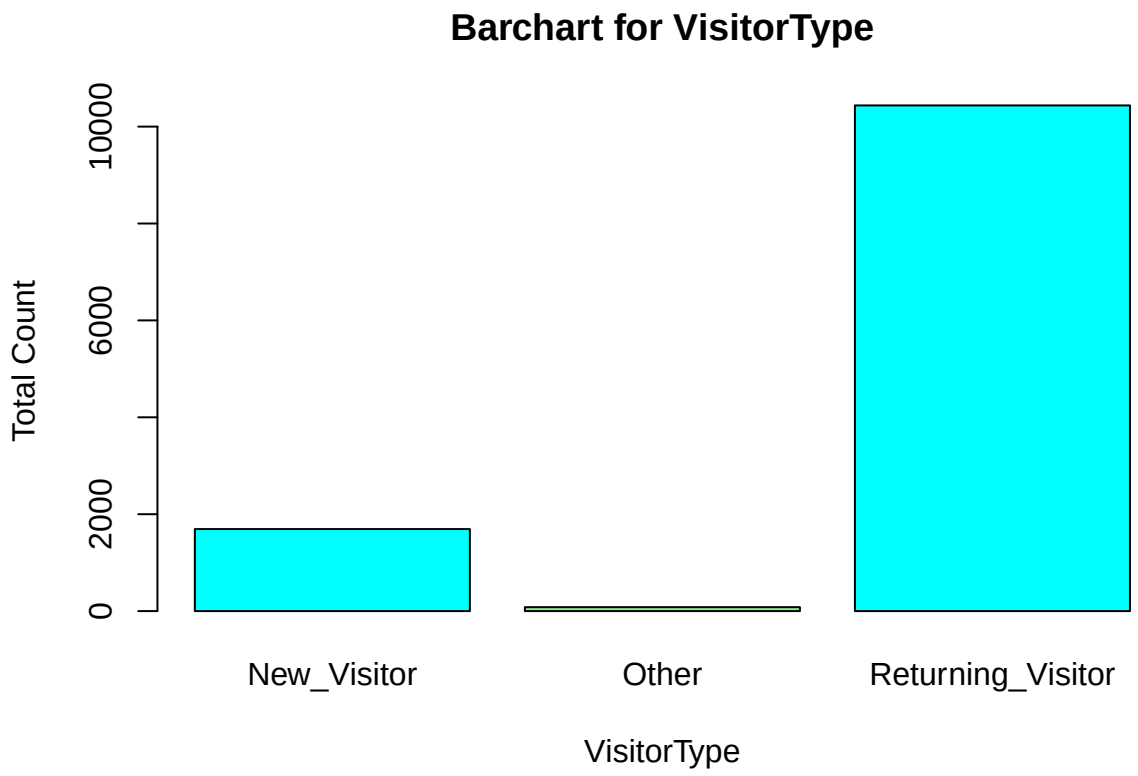
```
barplot(frequency,col=c("Cyan","lightgreen"),main="Barchart for Revenue",xlab = "Revenue",ylab = "Total
```




```
frequency <- table(df$VisitorType)
frequency
```

```
##
##       New_Visitor             Other Returning_Visitor
##              1693                81             10437
```

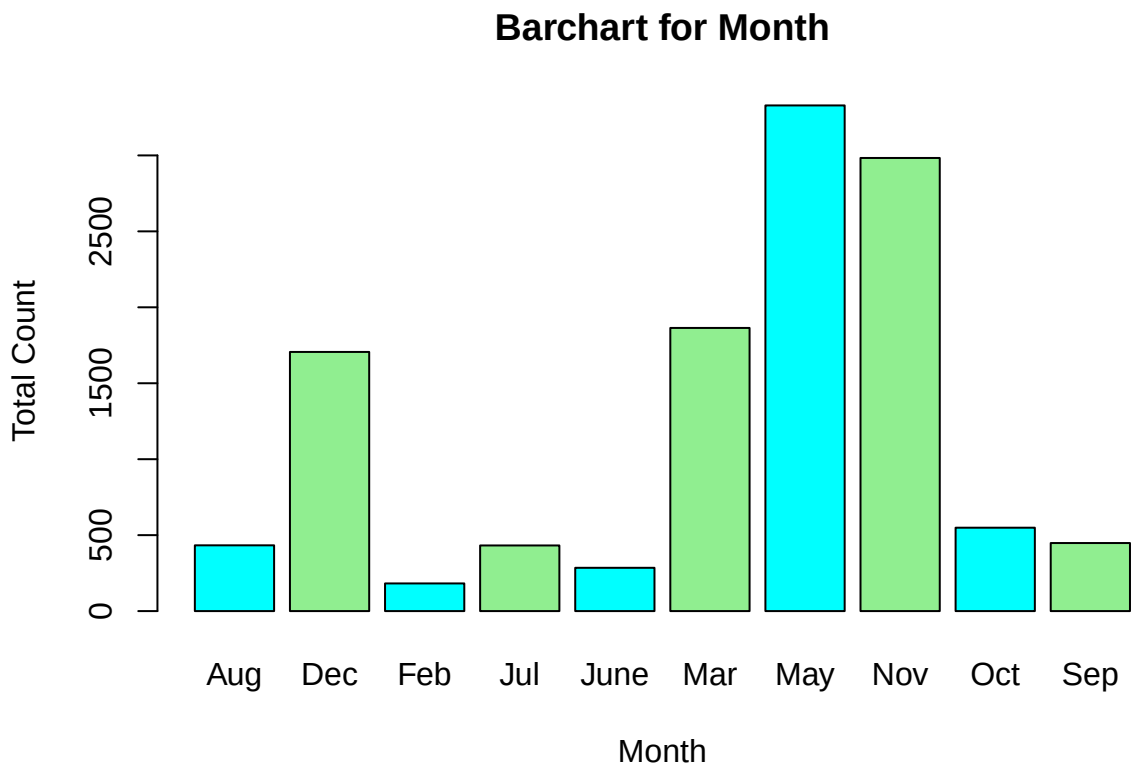
```
barplot(frequency,col=c("Cyan","lightgreen"),main="Barchart for VisitorType",xlab = "VisitorType",ylab =
```




```
frequency <- table(df$Month)
frequency
```

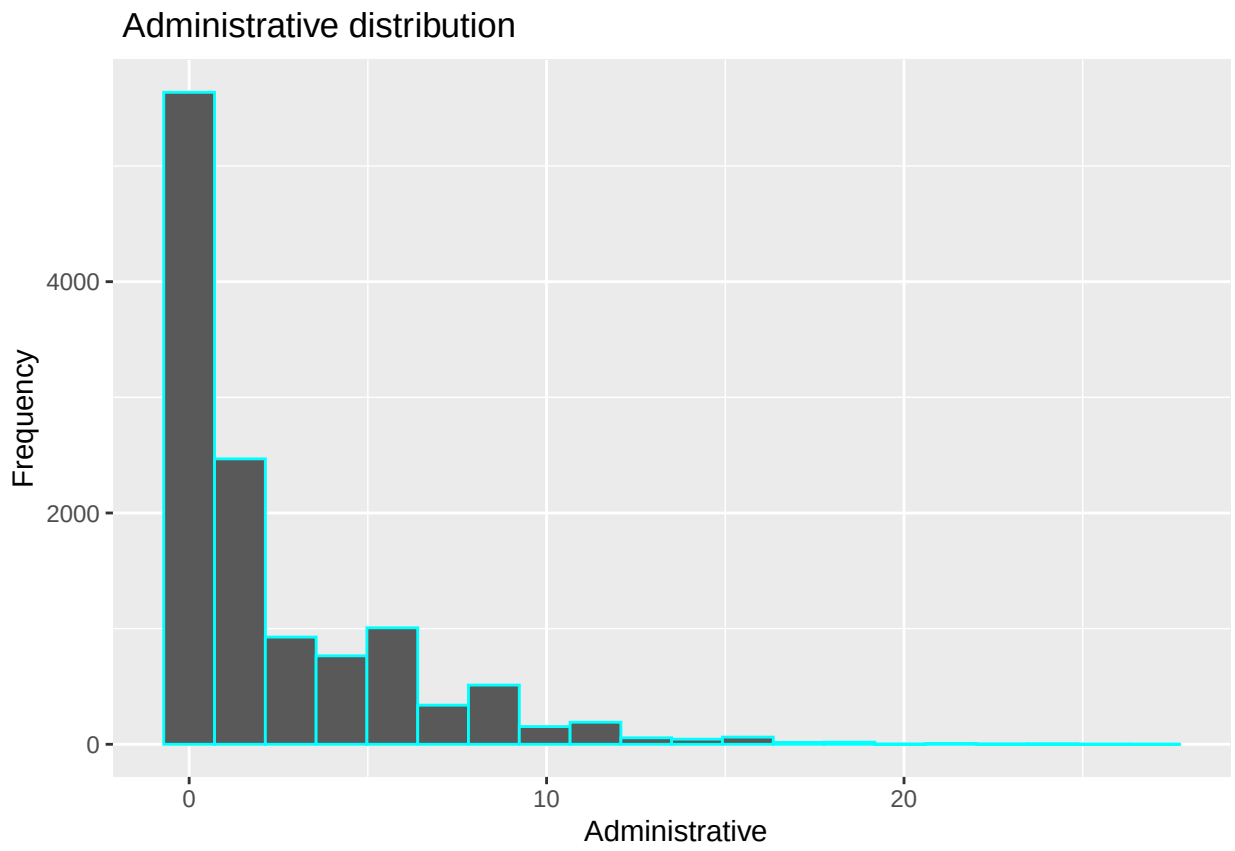
```
##
##  Aug  Dec  Feb  Jul June  Mar  May  Nov  Oct  Sep
##  433 1706  182  432  285 1864 3329 2983  549  448
```

```
barplot(frequency,col=c("Cyan","lightgreen"),main="Barchart for Month",xlab = "Month",ylab = "Total Cou
```

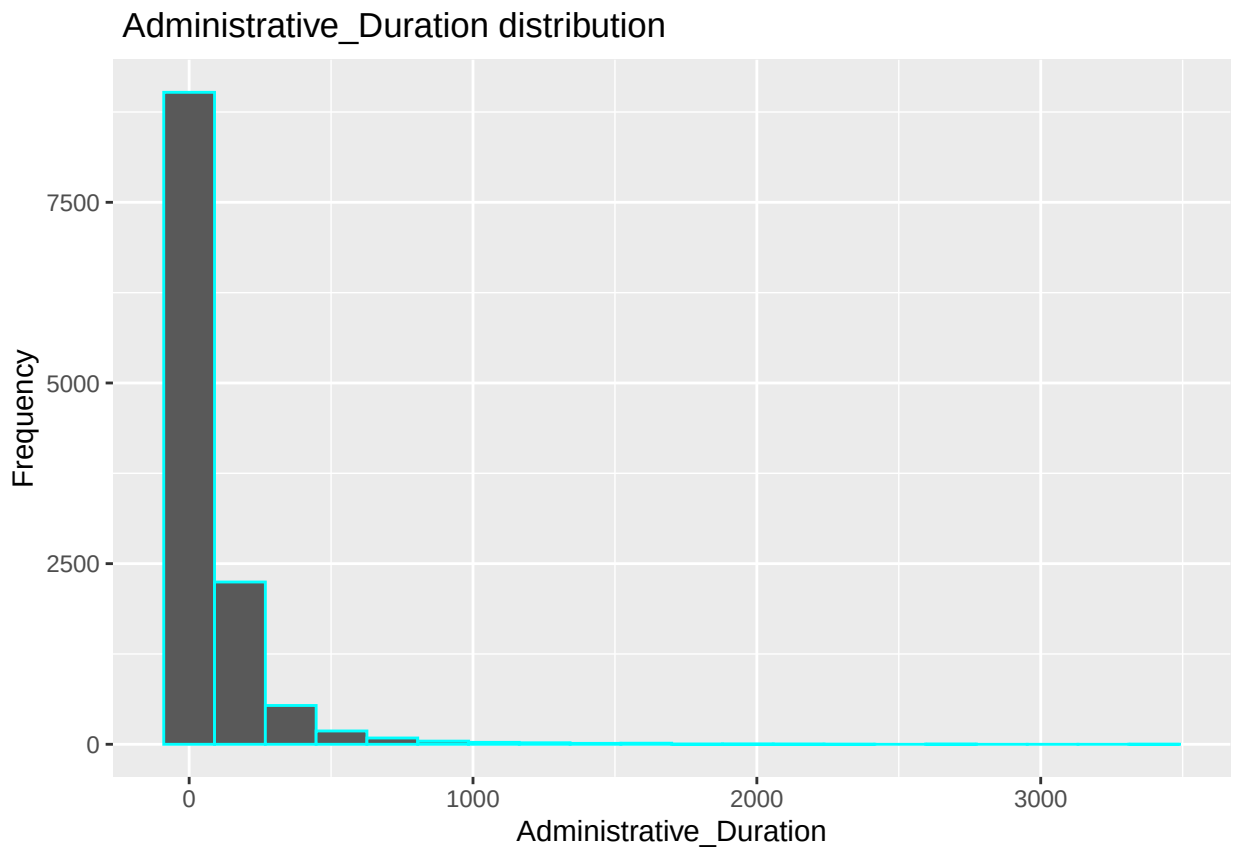


Histograms

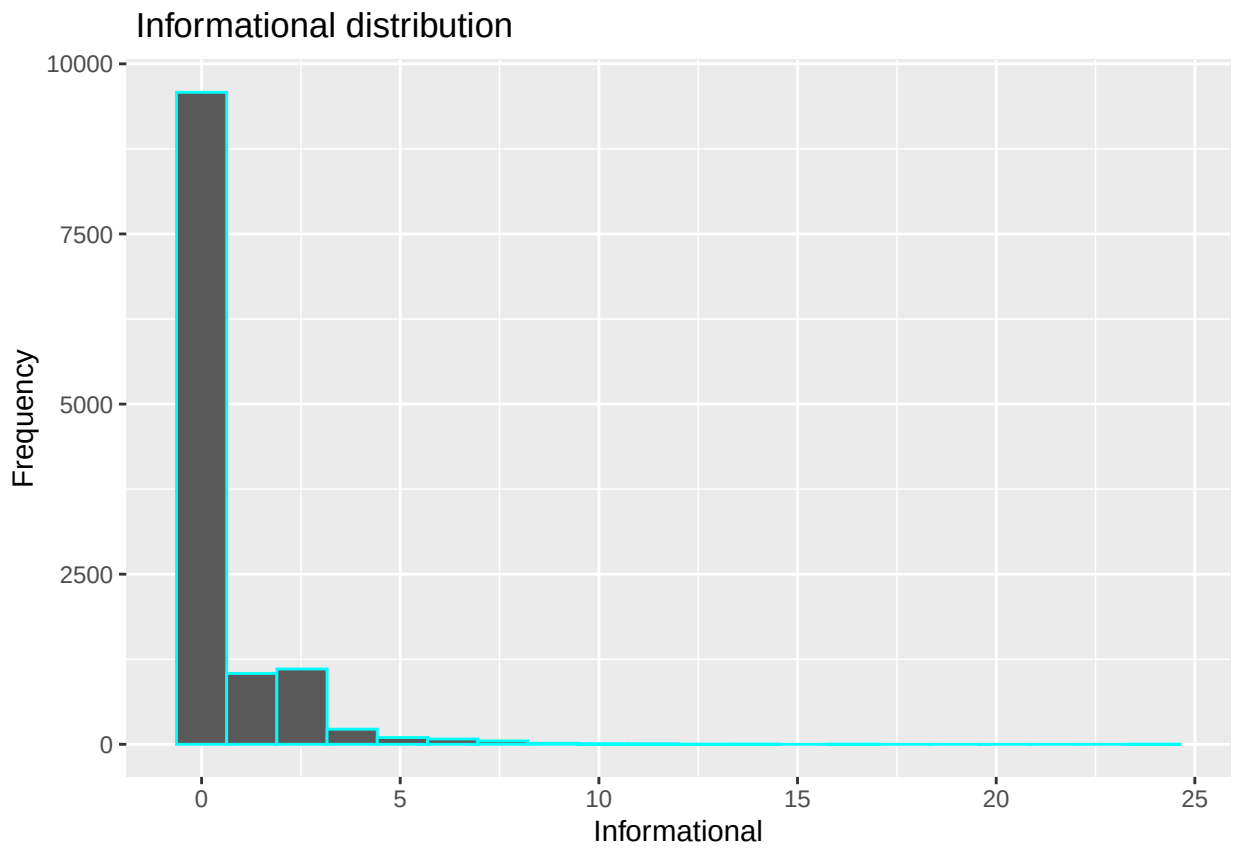
```
library(ggplot2)
ggplot(df, aes( Administrative)) + geom_histogram(bins = 20, color = 'cyan') +
    labs(title = ' Administrative distribution', x = ' Administrative', y = 'Frequency')
```



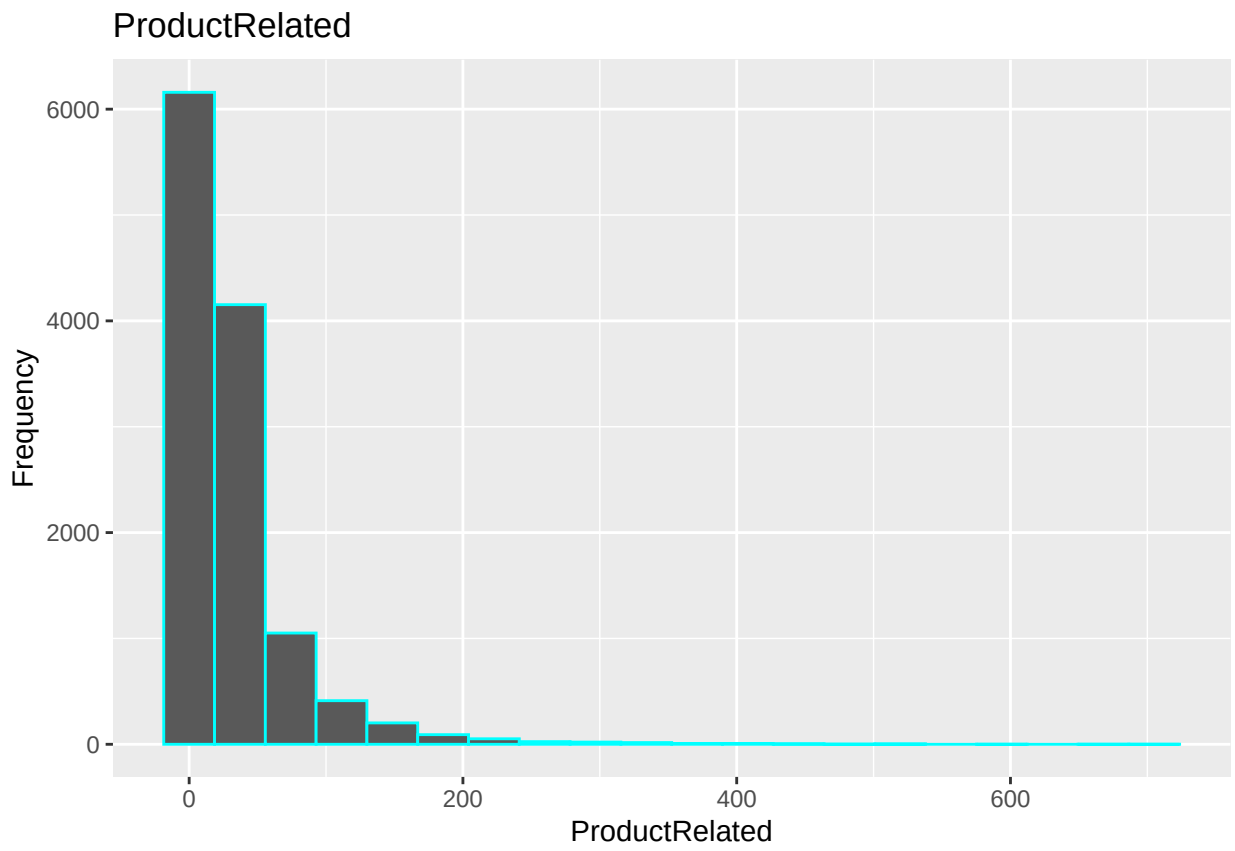

```
library(ggplot2)
ggplot(df, aes( Administrative_Duration)) + geom_histogram(bins = 20, color = 'cyan') +
    labs(title = ' Administrative_Duration distribution', x = 'Administrative_Duration', y = 'Frequency
```



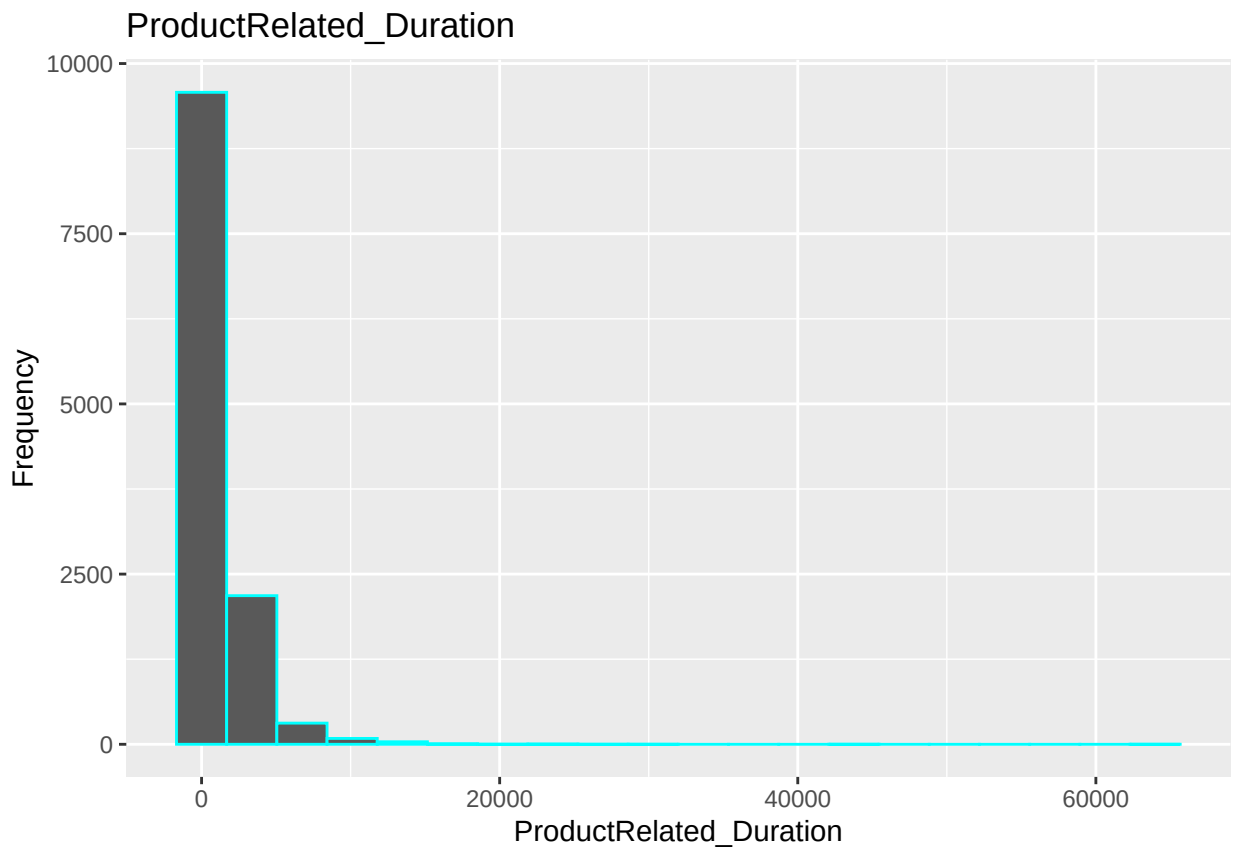

```
library(ggplot2)
ggplot(df, aes(Informational)) + geom_histogram(bins = 20, color = 'cyan') +
    labs(title = ' Informational distribution', x = 'Informational', y = 'Frequency')
```



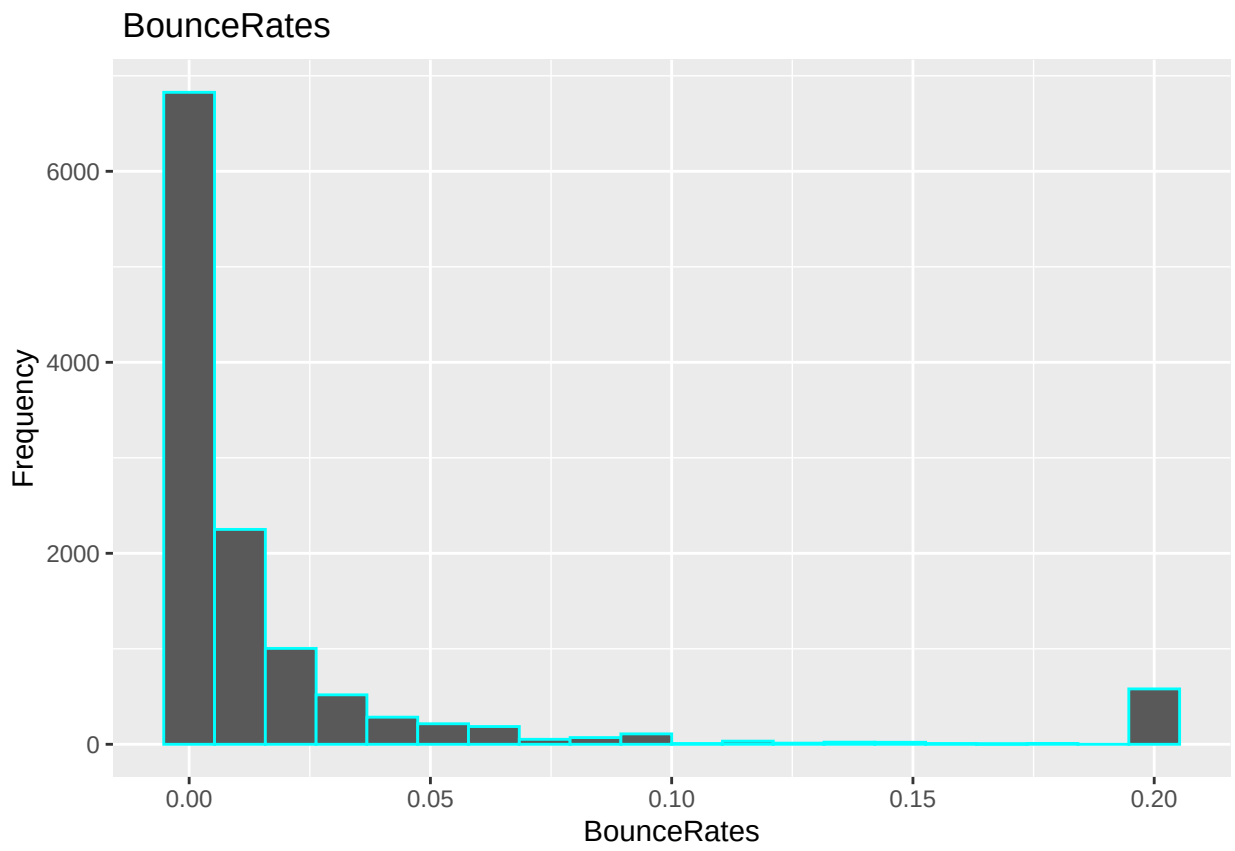

```
library(ggplot2)
ggplot(df, aes(ProductRelated)) + geom_histogram(bins = 20, color = 'cyan') +
    labs(title = 'ProductRelated', x = 'ProductRelated', y = 'Frequency')
```



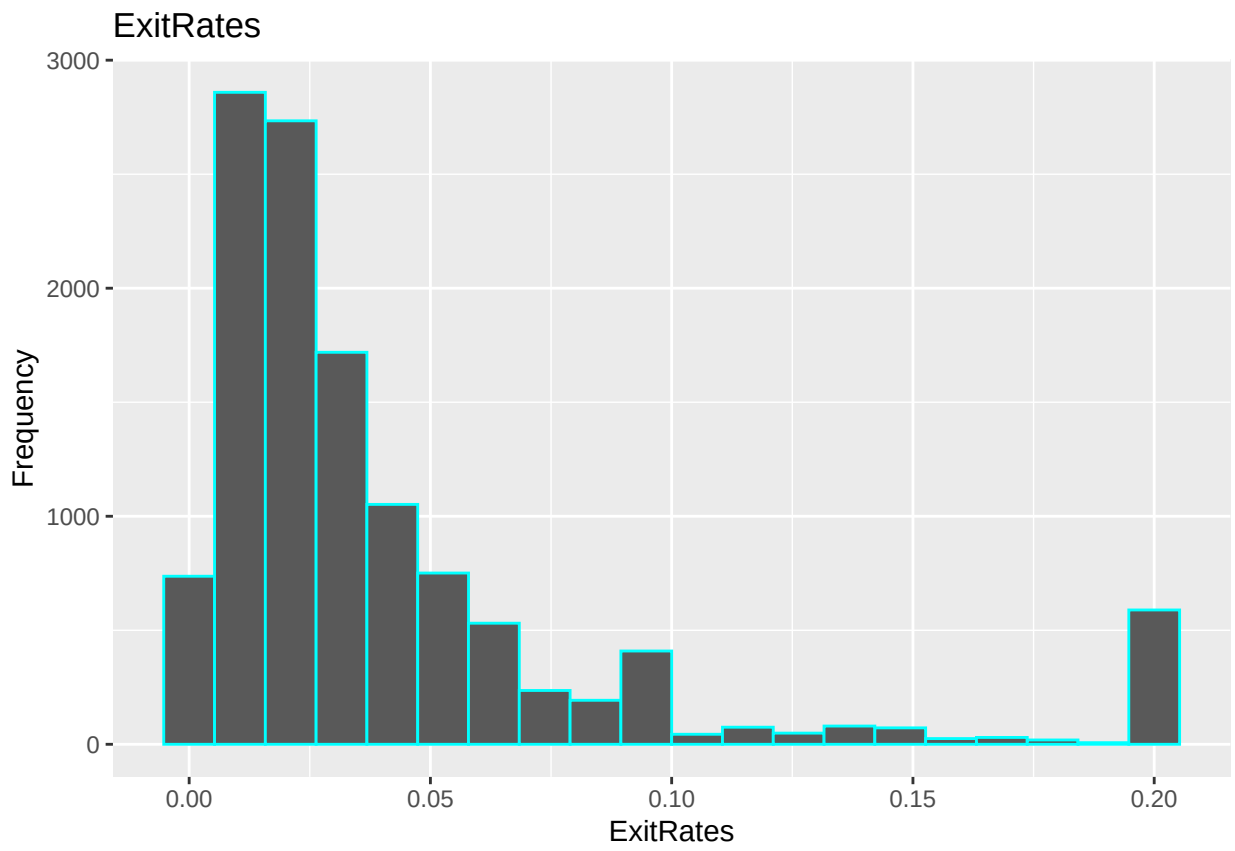

```
library(ggplot2)
ggplot(df, aes(ProductRelated_Duration)) + geom_histogram(bins = 20, color = 'cyan') +
    labs(title = 'ProductRelated_Duration', x = 'ProductRelated_Duration', y = 'Frequency')
```



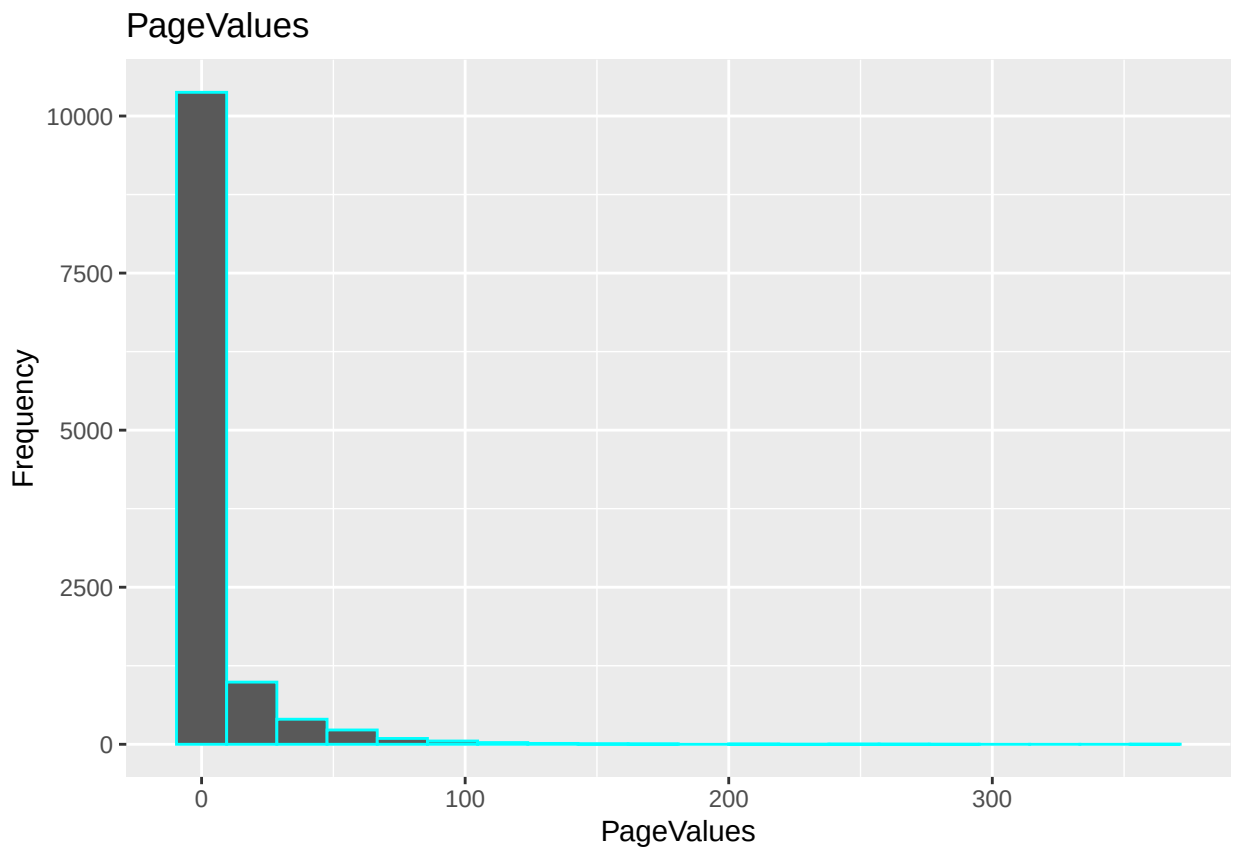

```
library(ggplot2)
ggplot(df, aes(BounceRates)) + geom_histogram(bins = 20, color = 'cyan') +
    labs(title = ' BounceRates', x = 'BounceRates', y = 'Frequency')
```




```
library(ggplot2)
ggplot(df, aes(ExitRates)) + geom_histogram(bins = 20, color = 'cyan') +
    labs(title = 'ExitRates', x = 'ExitRates', y = 'Frequency')
```

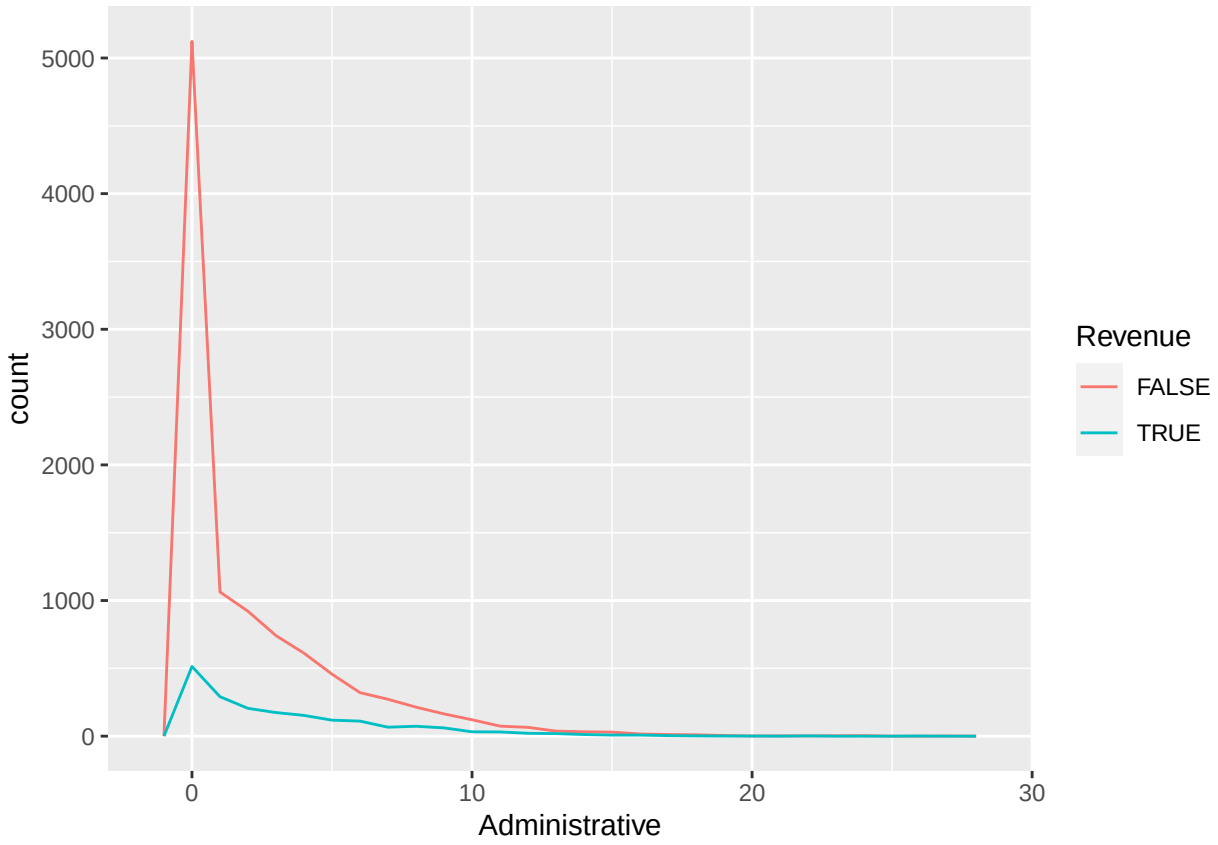



```
library(ggplot2)
ggplot(df, aes(PageValues)) + geom_histogram(bins = 20, color = 'cyan') +
    labs(title = 'PageValues', x = 'PageValues', y = 'Frequency')
```



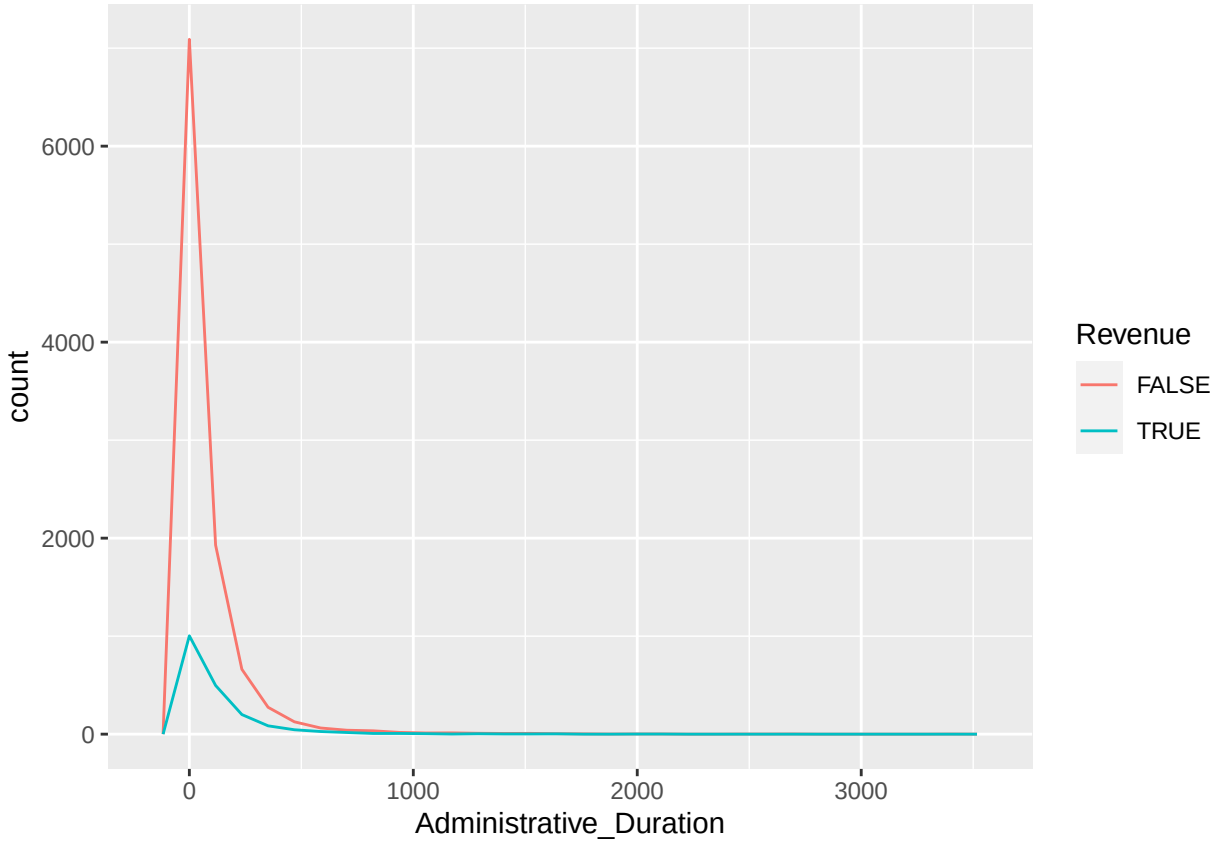
Bivariate Analysis Examining how different variables affect the labels

```
# Administrative sites and Revenue
ggplot(df, aes(Administrative, color=Revenue)) +
  geom_freqpoly(binwidth=1)
```



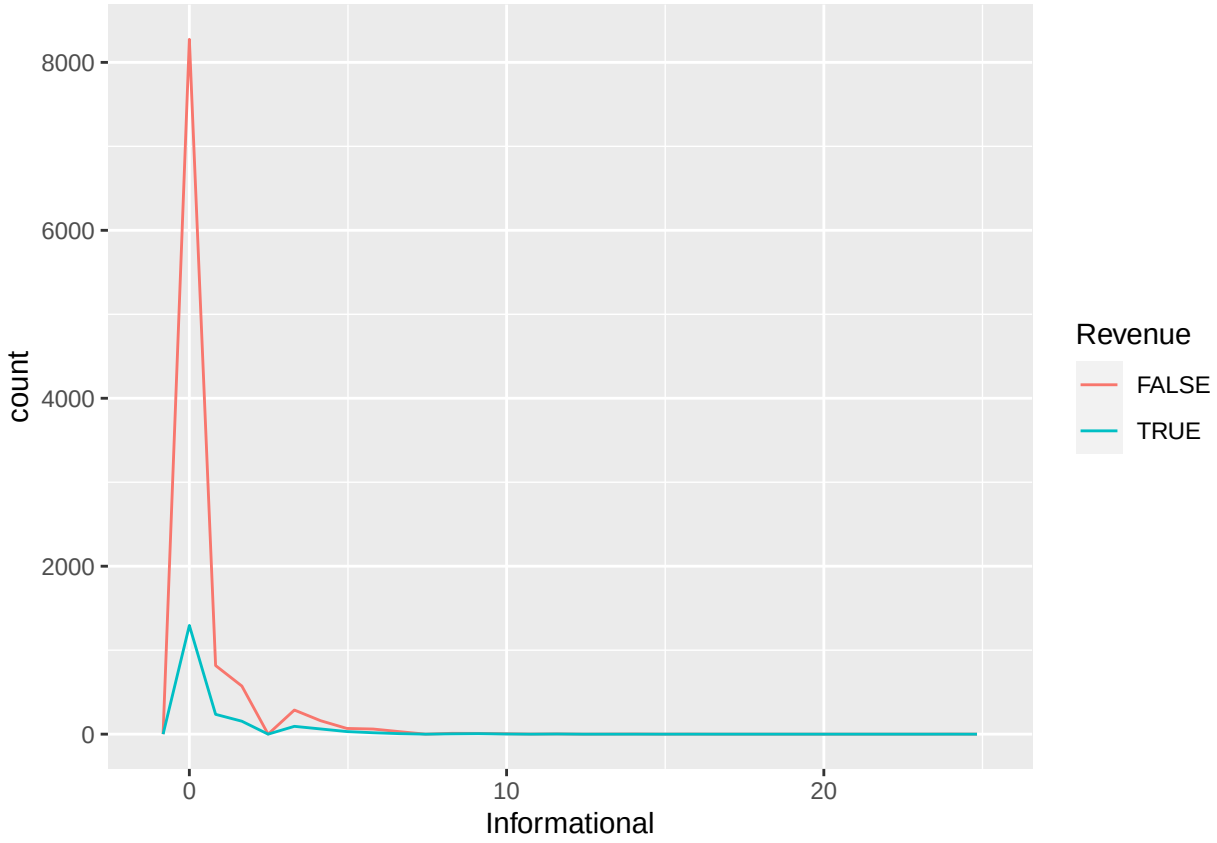

```
ggplot(df, aes(Administrative_Duration, color=Revenue)) +
  geom_freqpoly()
```

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



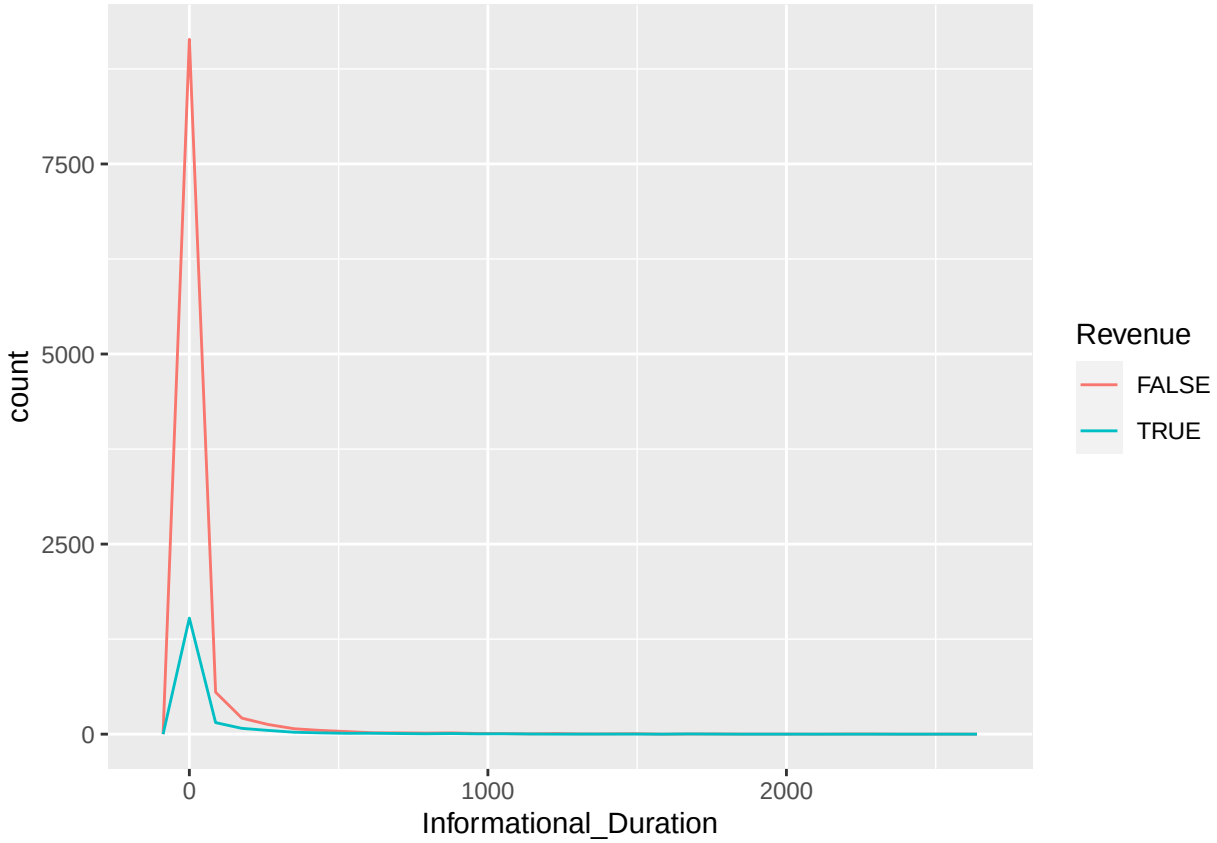

```
ggplot(df, aes(Informational, color=Revenue)) +
  geom_freqpoly()
```

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



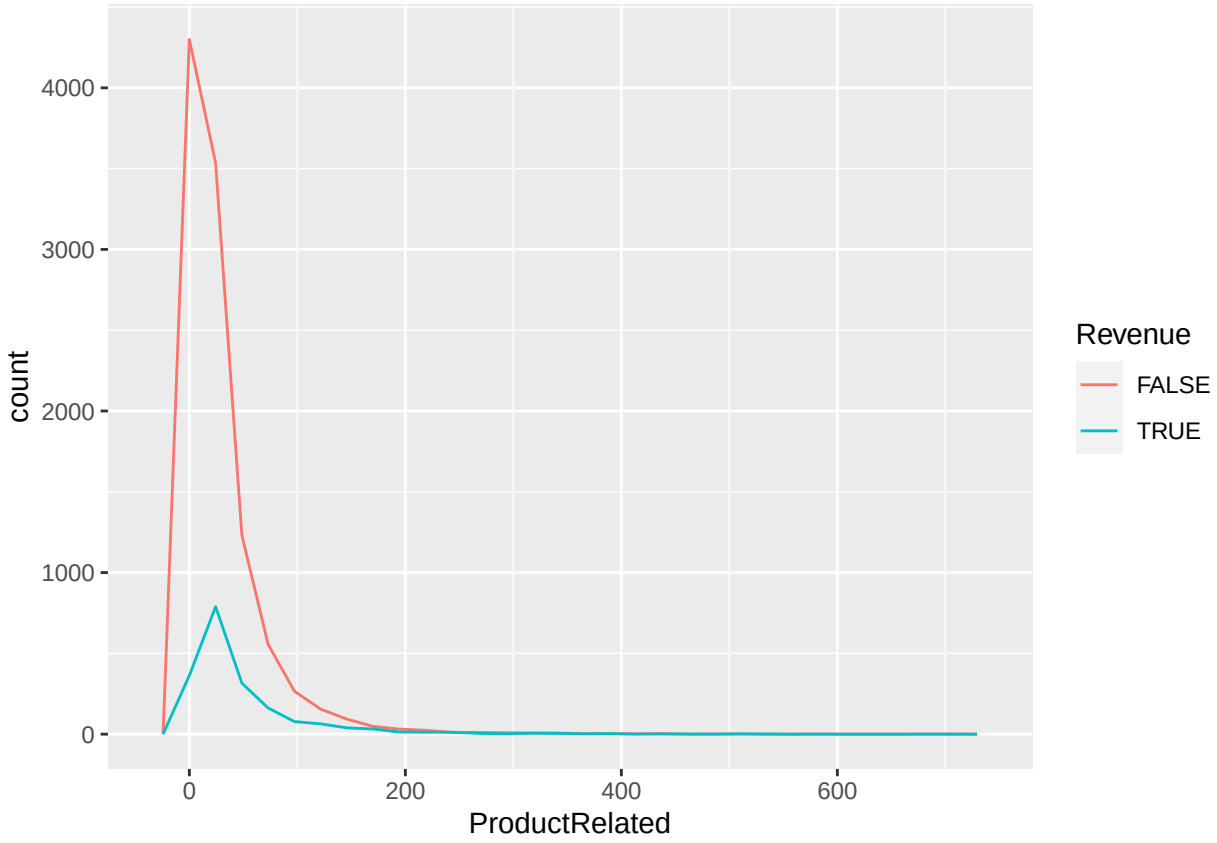

```
ggplot(df, aes(Informational_Duration, color=Revenue)) +
  geom_freqpoly()
```

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



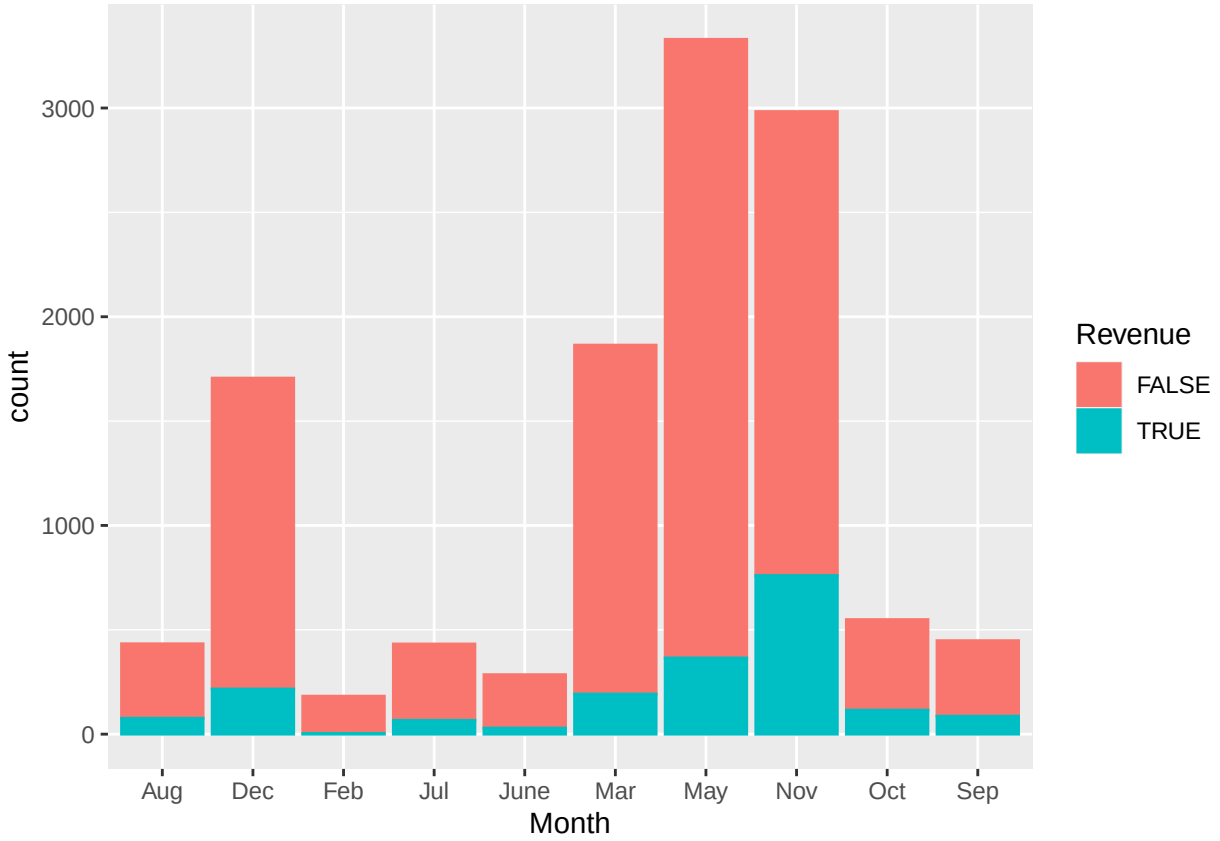

```
ggplot(df, aes(ProductRelated, color=Revenue)) +
  geom_freqpoly()
```

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




```
# Months vs GeneratingRevenue
ggplot(df, aes(Month, color=Revenue, fill=Revenue)) +
  geom_bar(binwidth=1)
```

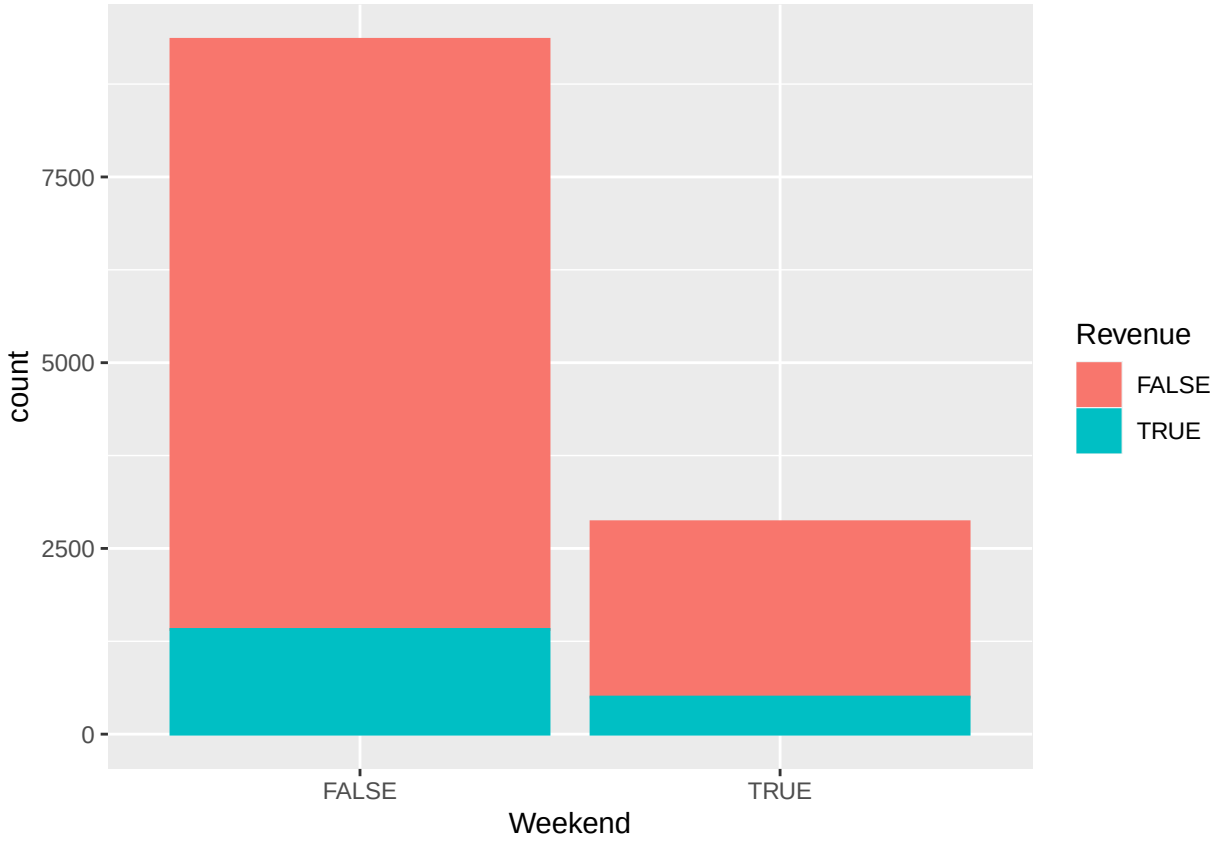
```
## Warning: Ignoring unknown parameters: binwidth
```



March, May, November and December are the months which generate significantly more revenue for the business.

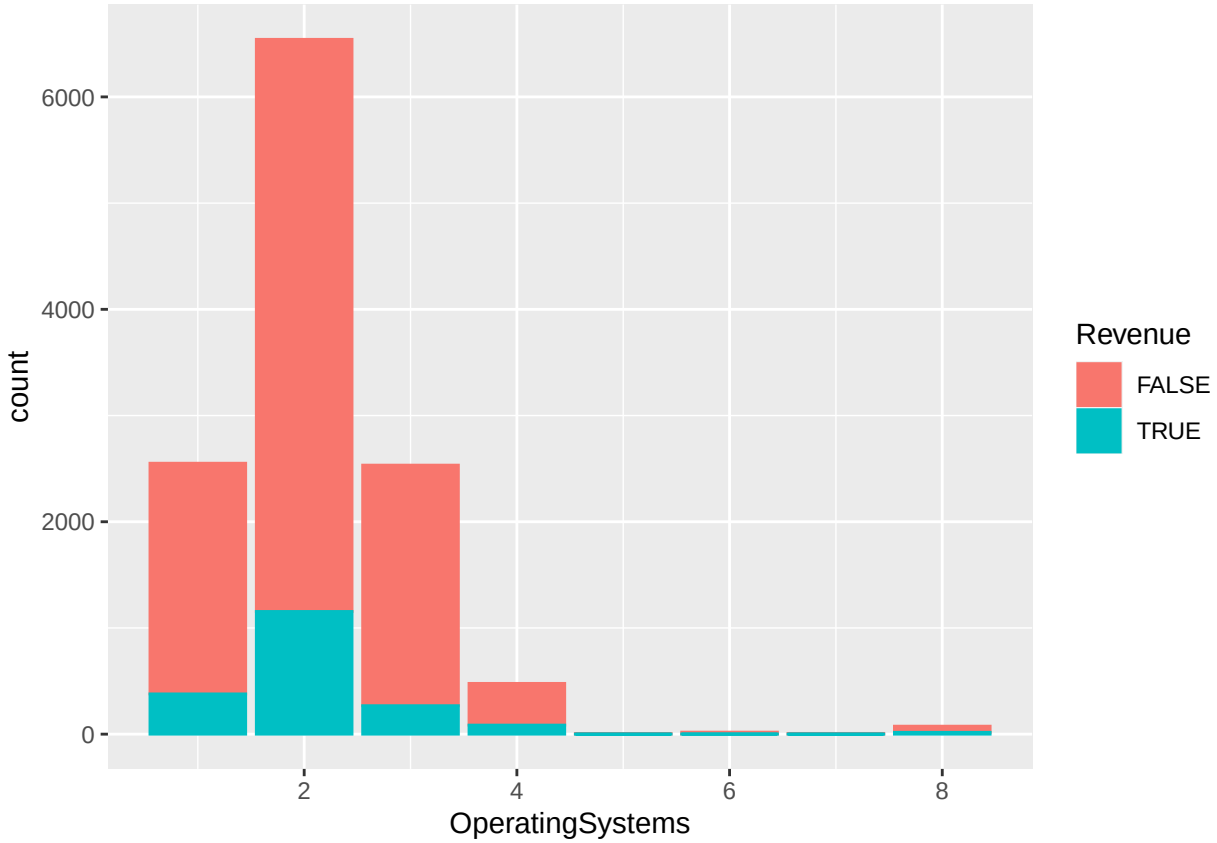
```
# Day type vs Generating Revenue
ggplot(df, aes(Weekend, color=Revenue, fill=Revenue)) +
  geom_bar(binwidth=1)
```

```
## Warning: Ignoring unknown parameters: binwidth
```



Weekdays generate slightly more Revenue than weekends.

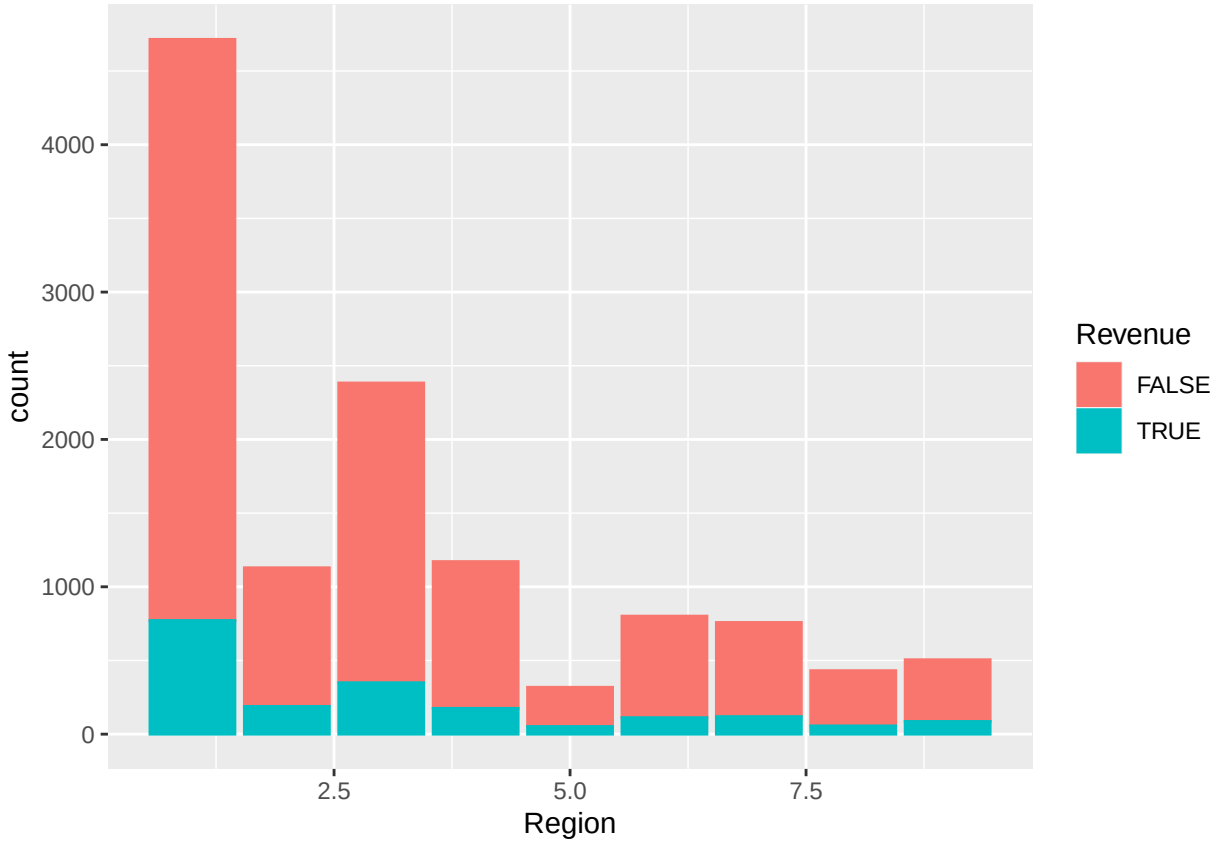
```
# Operating systems vs Generating Revenue
ggplot(df, aes(OperatingSystems, color=Revenue, fill=Revenue)) +
  geom_bar()
```



Users of type 2 OS generated the most revenue for the site, while 1, and 3 followed.

```
ggplot(df, aes(Region, fill=Revenue, color=Revenue)) +
  geom_bar(binwidth=1)
```

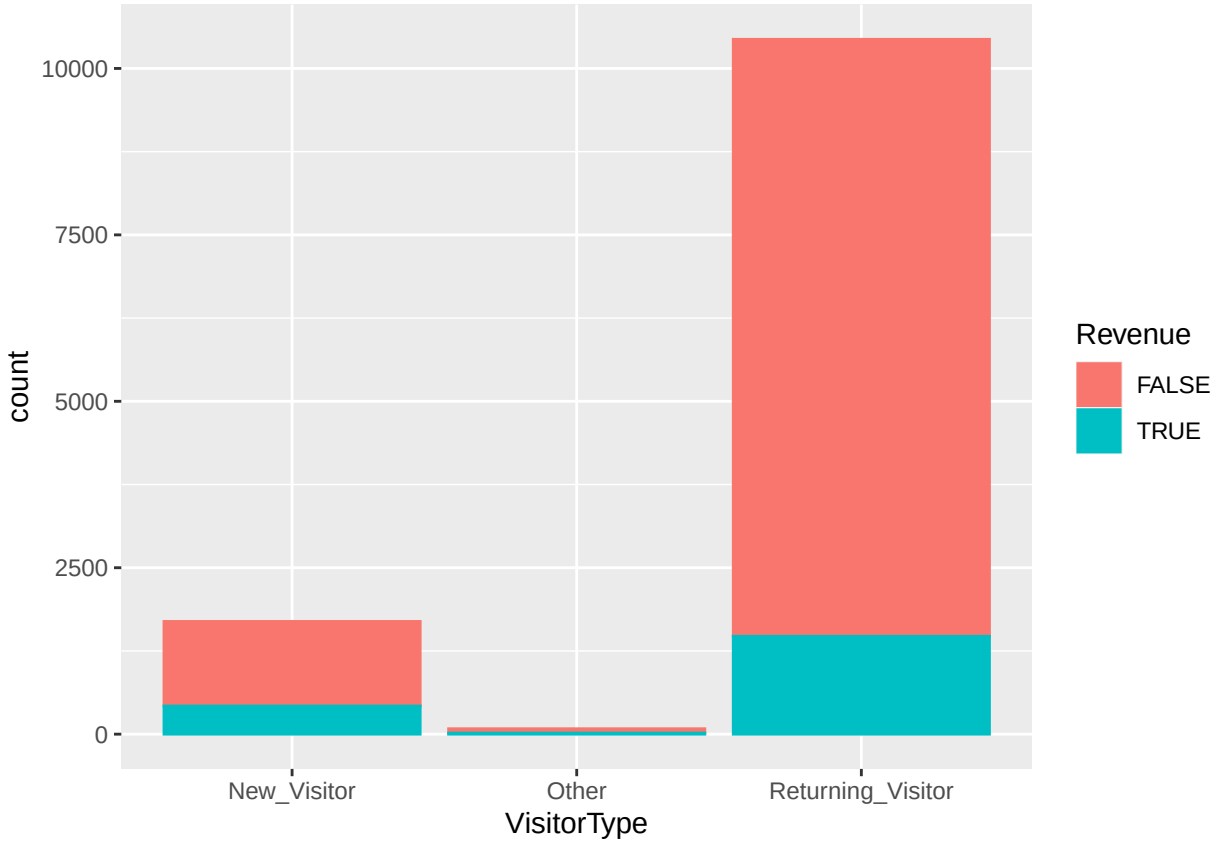
```
## Warning: Ignoring unknown parameters: binwidth
```



Region 1 produced the most revenue out of all the others with region 5 producing the least.

```
# Visitor type and revenue
ggplot(df, aes(VisitorType, color=Revenue, fill=Revenue)) +
  geom_bar(binwidth=2)
```

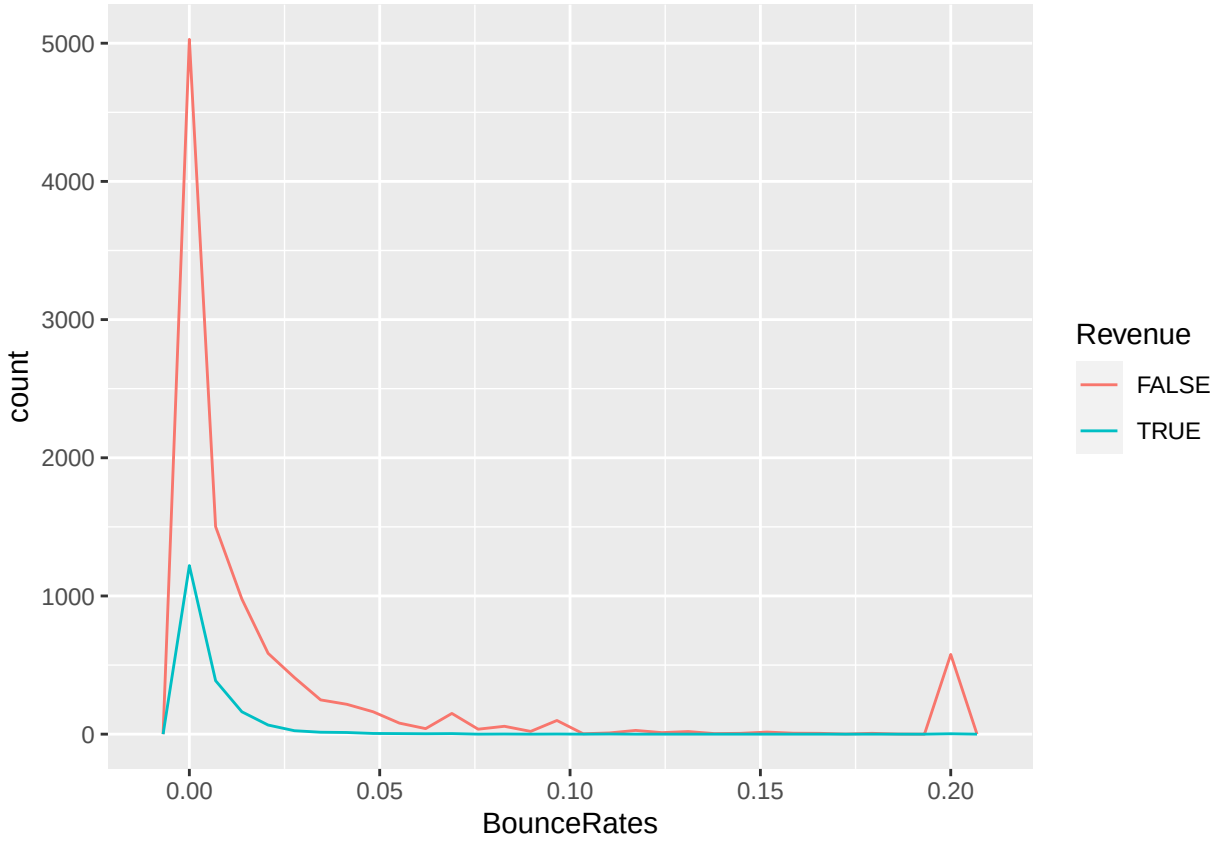
```
## Warning: Ignoring unknown parameters: binwidth
```



Returning visitors generated a lot more revenue than new ones

```
ggplot(df, aes(BounceRates, color=Revenue)) +
  geom_freqpoly()
```

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



A lot of sites had a high percentage of visitors just leaving without triggering any requests from our target website. ### **Correlations**

```
cor(df[,unlist(lapply(df, is.numeric))])
```

```
##                          Administrative Administrative_Duration Informational
## Administrative             1.000000000              0.600409663   0.375287625
## Administrative_Duration    0.600409663              1.000000000   0.301436307
## Informational              0.375287625              0.301436307   1.000000000
## Informational_Duration     0.254786030              0.237189867   0.618677950
## ProductRelated             0.428191539              0.286783934   0.372604735
## ProductRelated_Duration    0.371027248              0.353513809   0.386083730
## BounceRates               -0.213666729             -0.137333462  -0.109505362
## ExitRates                 -0.311274177             -0.202024485  -0.159566852
## PageValues                 0.096918211              0.066166426   0.047388971
## SpecialDay                -0.097065467             -0.074732016  -0.049373282
## OperatingSystems          -0.006694960             -0.007607782  -0.009622395
## Browser                   -0.025758626             -0.015830515  -0.038760823
## Region                    -0.007259381             -0.006721474  -0.030469277
## TrafficType               -0.034758903             -0.015063876  -0.035161595
##                          Informational_Duration ProductRelated
## Administrative                      0.254786030     0.42819154
## Administrative_Duration             0.237189867     0.28678393
## Informational                       0.618677950     0.37260473
## Informational_Duration              1.000000000     0.27906196
## ProductRelated                      0.279061956     1.00000000
```

```
## ProductRelated_Duration            0.346580698     0.86030819
## BounceRates                       -0.070159509    -0.19351588
## ExitRates                         -0.102932699    -0.28616327
## PageValues                         0.030063390     0.05411486
## SpecialDay                        -0.031290846    -0.02592738
## OperatingSystems                  -0.009746665     0.00408998
## Browser                           -0.019606214    -0.01370276
## Region                            -0.027912876    -0.04009570
## TrafficType                       -0.025145674    -0.04431244
##                          ProductRelated_Duration  BounceRates    ExitRates
## Administrative                        0.371027248 -0.213666729 -0.311274177
## Administrative_Duration               0.353513809 -0.137333462 -0.202024485
## Informational                         0.386083730 -0.109505362 -0.159566852
## Informational_Duration                0.346580698 -0.070159509 -0.102932699
## ProductRelated                        0.860308191 -0.193515878 -0.286163267
## ProductRelated_Duration               1.000000000 -0.174375596 -0.245334071
## BounceRates                          -0.174375596  1.000000000  0.903358297
## ExitRates                            -0.245334071  0.903358297  1.000000000
## PageValues                            0.050839954 -0.115997874 -0.173572979
## SpecialDay                           -0.038207021  0.087824321  0.116768246
## OperatingSystems                      0.002775738  0.026826492  0.016472540
## Browser                              -0.007835801 -0.016025821 -0.003573335
## Region                               -0.034853079  0.001426743 -0.001841360
## TrafficType                          -0.037479918  0.089131424  0.087320603
##                           PageValues   SpecialDay OperatingSystems      Browser
## Administrative            0.09691821 -0.097065467     -0.006694960 -0.025758626
## Administrative_Duration   0.06616643 -0.074732016     -0.007607782 -0.015830515
## Informational             0.04738897 -0.049373282     -0.009622395 -0.038760823
## Informational_Duration    0.03006339 -0.031290846     -0.009746665 -0.019606214
## ProductRelated            0.05411486 -0.025927380      0.004089980 -0.013702762
## ProductRelated_Duration   0.05083995 -0.038207021      0.002775738 -0.007835801
## BounceRates              -0.11599787  0.087824321      0.026826492 -0.016025821
## ExitRates                -0.17357298  0.116768246      0.016472540 -0.003573335
## PageValues                1.00000000 -0.064429141      0.018620103  0.045917739
## SpecialDay               -0.06442914  1.000000000      0.012795031  0.003544066
## OperatingSystems          0.01862010  0.012795031      1.000000000  0.212345154
## Browser                   0.04591774  0.003544066      0.212345154  1.000000000
## Region                    0.01062946 -0.016407006      0.071931190  0.091965004
## TrafficType               0.01227262  0.052832168      0.182956729  0.102831383
##                                Region TrafficType
## Administrative           -0.007259381 -0.03475890
## Administrative_Duration  -0.006721474 -0.01506388
## Informational            -0.030469277 -0.03516159
## Informational_Duration   -0.027912876 -0.02514567
## ProductRelated           -0.040095701 -0.04431244
## ProductRelated_Duration  -0.034853079 -0.03747992
## BounceRates               0.001426743  0.08913142
## ExitRates                -0.001841360  0.08732060
## PageValues                0.010629461  0.01227262
## SpecialDay               -0.016407006  0.05283217
## OperatingSystems          0.071931190  0.18295673
## Browser                   0.091965004  0.10283138
## Region                    1.000000000  0.04271596
## TrafficType               0.042715962  1.00000000
```

The rates were significantly correlated while types of number of sites were strongly correlated with how much time was spent in them.

```
df$Weekend<- as.numeric(df$Weekend)
df$Revenue<- as.numeric(df$Revenue)
```

```
# casting categorical columns as factors
df$Month <- factor(df$Month)
df$OperatingSystems <- factor(df$OperatingSystems)
df$Browser <- factor(df$Browser)
df$Region <- factor(df$Region)
df$TrafficType <- factor(df$TrafficType)
df$VisitorType <- factor(df$VisitorType)
df$Weekend <- factor(df$Weekend)
df$Revenue <- factor(df$Revenue)
```

```
library(superml)
```

Data Preparation

```
## Warning: package 'superml' was built under R version 4.1.1
```

```
## Loading required package: R6
```

```
## Warning: package 'R6' was built under R version 4.1.1
```

```
label <- LabelEncoder$new()
df$Month <- label$fit_transform(df$Month)
df$VisitorType <- label$fit_transform(df$VisitorType)
df$Weekend <- label$fit_transform(df$Weekend)
df$Revenue <- label$fit_transform(df$Revenue)
```

KNN

```
# separating features from Revenue labels
x <- df[, -18]
# normalizing
normalize <- function(x){
  return ((x-min(x)) / (max(x)-min(x)))
}
x$Administrative <- normalize(x$Administrative)
x$Administrative_Duration <- normalize(x$Administrative_Duration)
x$Informational <- normalize(x$Informational)
x$Informational_Duration <- normalize(x$Informational_Duration)
```

```
x$ProductRelated <- normalize(x$ProductRelated)
x$ProductRelated_Duration <- normalize(x$ProductRelated_Duration)
x$BounceRates <- normalize(x$BounceRates)
x$ExitRates <- normalize(x$ExitRates)
x$PageValues <- normalize(x$PageValues)
x$SpecialDay <- normalize(x$SpecialDay)
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.1
```

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

```
## Warning: package 'lattice' was built under R version 4.1.1
```

```
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 4.1.1
```

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

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

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.1.1
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.1.1
```

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

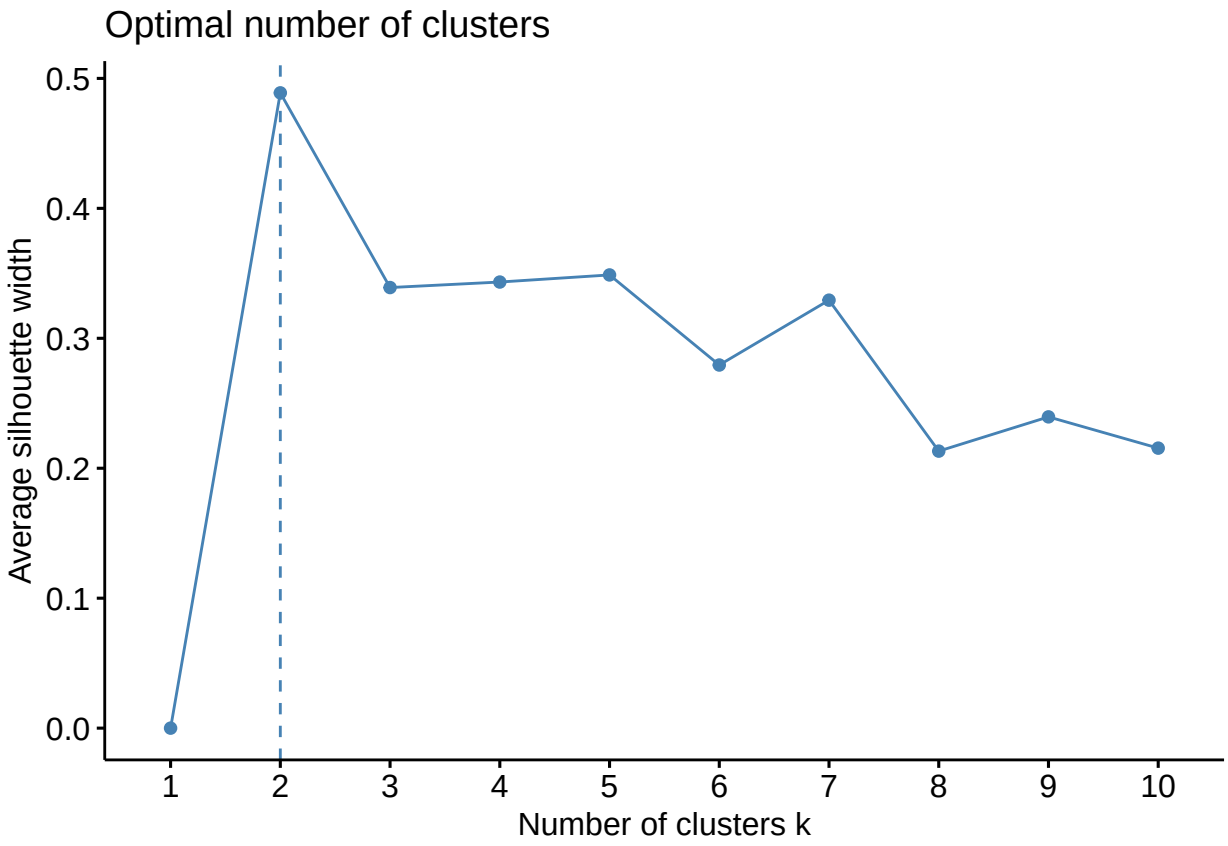
```
library(cluster)
library(purrr)
```

```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:caret':
##
##     lift
```

```
## The following object is masked from 'package:data.table':
##
##     transpose
```

```
# finding optimum k
fviz_nbclust(x, kmeans, method="silhouette")
```



According to the silhouette method above, only 2 clusters are sufficient.

2 clusters shall be used.

```
# using 2 clusters
k <- kmeans(x,centers=3,nstart=25)
```

```
# Number of records in each cluster
k$size
```

```
## [1] 1995 7794 2422
```

```
df$cluster <- as.factor(k$cluster)
```

```
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 VisitorType
## 1          0     2                1       1      1           1           2
## 2          0     2                2       2      1           2           2
## 3          0     2                4       1      9           3           2
## 4          0     2                3       2      2           4           2
## 5          0     2                3       3      1           4           2
## 6          0     2                2       2      1           3           2
##   Weekend Revenue cluster
## 1       0       0       3
## 2       0       0       3
## 3       0       0       3
## 4       0       0       3
## 5       1       0       3
## 6       0       0       3
```

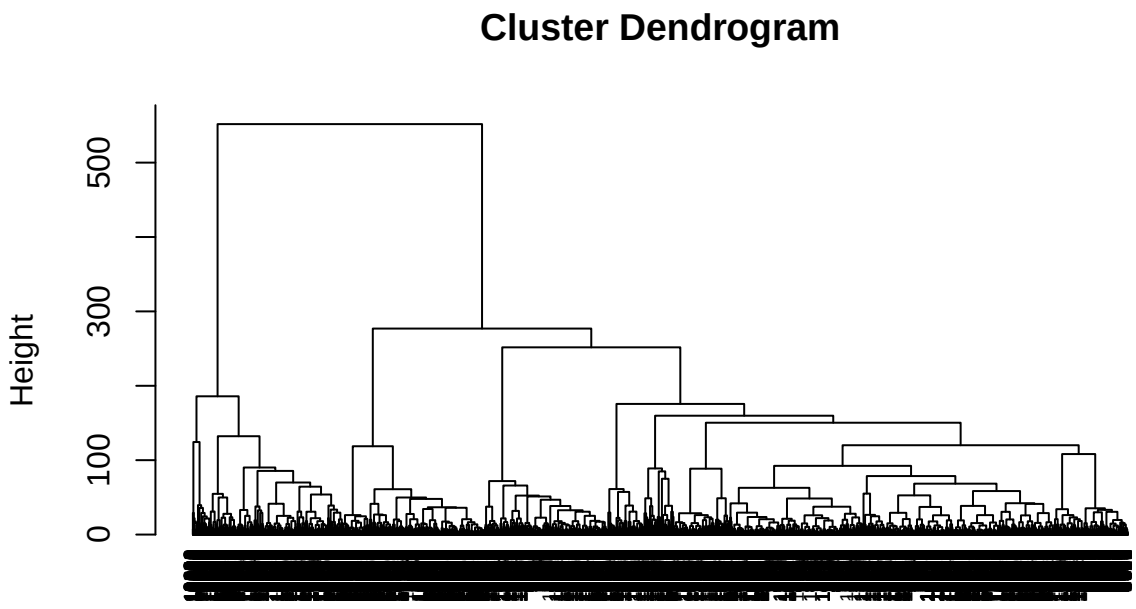
Hierarchical clustering

```
r <- df[,1:17]
```

```
# scaling the data
r$Administrative <- scale(r$Administrative)
r$Administrative_Duration <- scale(r$Administrative_Duration)
r$Informational <- scale(r$Informational)
r$Informational_Duration <- scale(r$Informational_Duration)
r$ProductRelated <- scale(r$ProductRelated)
r$ProductRelated_Duration <- scale(r$ProductRelated_Duration)
r$BounceRates <- scale(r$BounceRates)
r$ExitRates <- scale(r$ExitRates)
r$PageValues <- scale(r$PageValues)
r$SpecialDay <- scale(r$SpecialDay)
```

```
# computing the distance
d <- dist(r, method="euclidean")
# Clustering  algorithm deployment
model <- hclust(d, method="ward.D2")
```

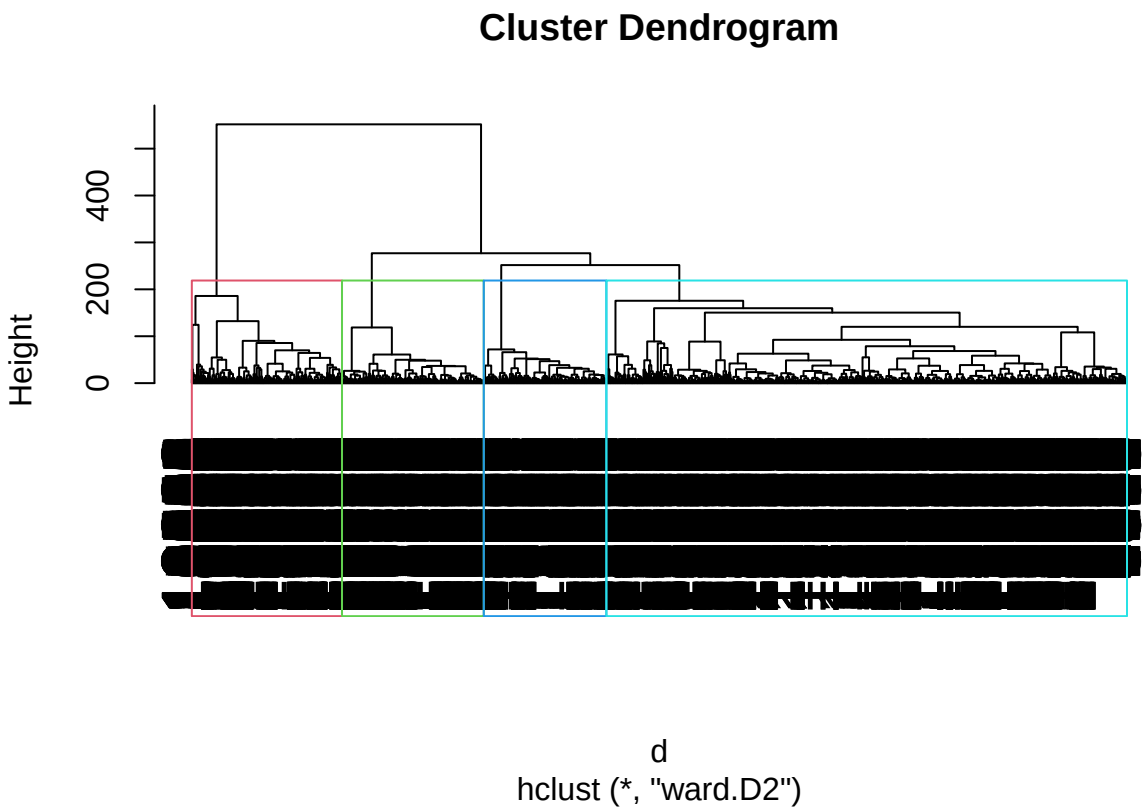
```
plot(model, cex=0.6, hang=-1)
```




```
# Ward's method
hc <- hclust(d, method="ward.D2")
# cut the tree into 5 parts
sub_grp <- cutree(hc, k=4)
table(sub_grp)
```

```
## sub_grp
##    1    2    3    4
## 6798 1851 1961 1601
```

```
plot(hc, cex=2, hang=-1 )
rect.hclust(hc, k=4, border=2:5)
```



Conclusions

1.) Weekdays have the highest flux of customers on site.

2.) Returning visitors will always generate revenue for the site

3.) Most customers prefers to use the second operating system

Recommendations

1.) Improve customer engagement on operating systems

2.) Come up with promotional offers