---
title: “R
Note-
book”
output:
pdf_document:
toc: yes
html_document:
toc: yes
word_document:
toc: yes

#
Import
the
needed
Libraries
```
r
library(reshape2)
library(cluster)
library(fpc)
library(vegan)
library(mclust)
library(scales)
library(Hmisc)
library(corrplot)
library(mice)
library(VIM)
library(sqldf)
library(pROC)
library(ggmosaic)
library(caret)
library(tidyverse)
library(lubridate)
library(tseries)
library(rpart)
library(rpart.plot)
library(randomForest)
library(viridis)
library(hrbrthemes)
library(plyr)
library
(gmodels);
library
(MASS)
library(profvis)
library(factoextra)
r DF1
<-
read.csv("online_shoppers_intention.csv")
```

Data exploration

```
head(DF1)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                       0             0                      0
## 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                0.000000  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
```

```
str(DF1)
```

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ OperatingSystems       : int  1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : int  1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : int  1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : int  1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
```

```
##  $ Weekend                : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ Revenue                : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
```

```
summary(DF1)
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   7.50         Median : 0.0000
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##
##   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 :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##
##      Month      OperatingSystems    Browser           Region
##  May    :3364   Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Nov    :2998   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mar    :1907   Median :2.000    Median : 2.000   Median :3.000
##  Dec    :1727   Mean   :2.124    Mean   : 2.357   Mean   :3.147
##  Oct    : 549   3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##  Sep    : 448   Max.   :8.000    Max.   :13.000   Max.   :9.000
##  (Other):1337
##   TrafficType                VisitorType     Weekend          Revenue
##  Min.   : 1.00   New_Visitor      : 1694   Mode :logical   Mode :logical
##  1st Qu.: 2.00   Other            :   85   FALSE:9462      FALSE:10422
##  Median : 2.00   Returning_Visitor:10551   TRUE :2868      TRUE :1908
##  Mean   : 4.07
##  3rd Qu.: 4.00
##  Max.   :20.00
##
```

#Check Distribution of Variables

```
CrossTable(DF1$Revenue)
```

```
##
##
##    Cell Contents
```

```
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##            |     FALSE |      TRUE |
##            |-----------|-----------|
##            |     10422 |      1908 |
##            |     0.845 |     0.155 |
##            |-----------|-----------|
##
##
##
##
```

```
count(DF1$VisitorType)
```

```
##                   x  freq
## 1       New_Visitor  1694
## 2             Other    85
## 3 Returning_Visitor 10551
```

```
count(DF1$Weekend)
```

```
##       x freq
## 1 FALSE 9462
## 2  TRUE 2868
```

```
count(DF1$Month)
```

```
##       x freq
## 1   Aug  433
## 2   Dec 1727
## 3   Feb  184
## 4   Jul  432
## 5  June  288
## 6   Mar 1907
## 7   May 3364
## 8   Nov 2998
## 9   Oct  549
## 10  Sep  448
```

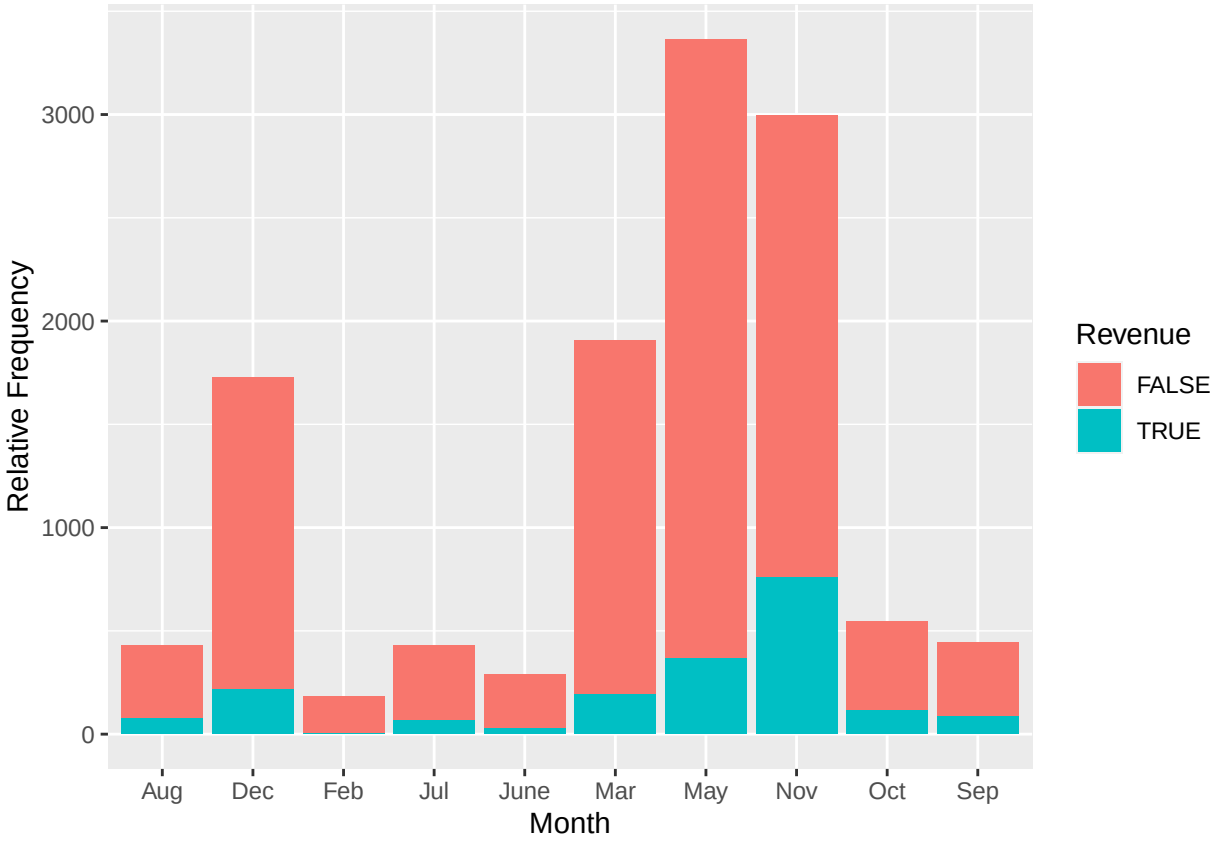
```
CrossTable(DF1$Month, DF1$Revenue)
```

```
##
##
##    Cell Contents
```

```
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##              | DF1$Revenue
##    DF1$Month |     FALSE |      TRUE | Row Total |
## -------------|-----------|-----------|-----------|
##          Aug |       357 |        76 |       433 |
##              |     0.221 |     1.208 |           |
##              |     0.824 |     0.176 |     0.035 |
##              |     0.034 |     0.040 |           |
##              |     0.029 |     0.006 |           |
## -------------|-----------|-----------|-----------|
##          Dec |      1511 |       216 |      1727 |
##              |     1.799 |     9.826 |           |
##              |     0.875 |     0.125 |     0.140 |
##              |     0.145 |     0.113 |           |
##              |     0.123 |     0.018 |           |
## -------------|-----------|-----------|-----------|
##          Feb |       181 |         3 |       184 |
##              |     4.172 |    22.789 |           |
##              |     0.984 |     0.016 |     0.015 |
##              |     0.017 |     0.002 |           |
##              |     0.015 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##          Jul |       366 |        66 |       432 |
##              |     0.002 |     0.011 |           |
##              |     0.847 |     0.153 |     0.035 |
##              |     0.035 |     0.035 |           |
##              |     0.030 |     0.005 |           |
## -------------|-----------|-----------|-----------|
##         June |       259 |        29 |       288 |
##              |     0.995 |     5.437 |           |
##              |     0.899 |     0.101 |     0.023 |
##              |     0.025 |     0.015 |           |
##              |     0.021 |     0.002 |           |
## -------------|-----------|-----------|-----------|
##          Mar |      1715 |       192 |      1907 |
##              |     6.594 |    36.019 |           |
##              |     0.899 |     0.101 |     0.155 |
##              |     0.165 |     0.101 |           |
##              |     0.139 |     0.016 |           |
## -------------|-----------|-----------|-----------|
##          May |      2999 |       365 |      3364 |
##              |     8.511 |    46.487 |           |
##              |     0.891 |     0.109 |     0.273 |
```

```
##              |     0.288 |     0.191 |           |
##              |     0.243 |     0.030 |           |
## -------------|-----------|-----------|-----------|
##          Nov |      2238 |       760 |      2998 |
##              |    34.593 |   188.955 |           |
##              |     0.746 |     0.254 |     0.243 |
##              |     0.215 |     0.398 |           |
##              |     0.182 |     0.062 |           |
## -------------|-----------|-----------|-----------|
##          Oct |       434 |       115 |       549 |
##              |     1.945 |    10.626 |           |
##              |     0.791 |     0.209 |     0.045 |
##              |     0.042 |     0.060 |           |
##              |     0.035 |     0.009 |           |
## -------------|-----------|-----------|-----------|
##          Sep |       362 |        86 |       448 |
##              |     0.734 |     4.011 |           |
##              |     0.808 |     0.192 |     0.036 |
##              |     0.035 |     0.045 |           |
##              |     0.029 |     0.007 |           |
## -------------|-----------|-----------|-----------|
## Column Total |     10422 |      1908 |     12330 |
##              |     0.845 |     0.155 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```r
DF1 %>%
ggplot() +
aes(x = Month, Revenue = ..count../nrow(DF1), fill = Revenue) +
geom_bar() +
ylab("Relative Frequency")
```

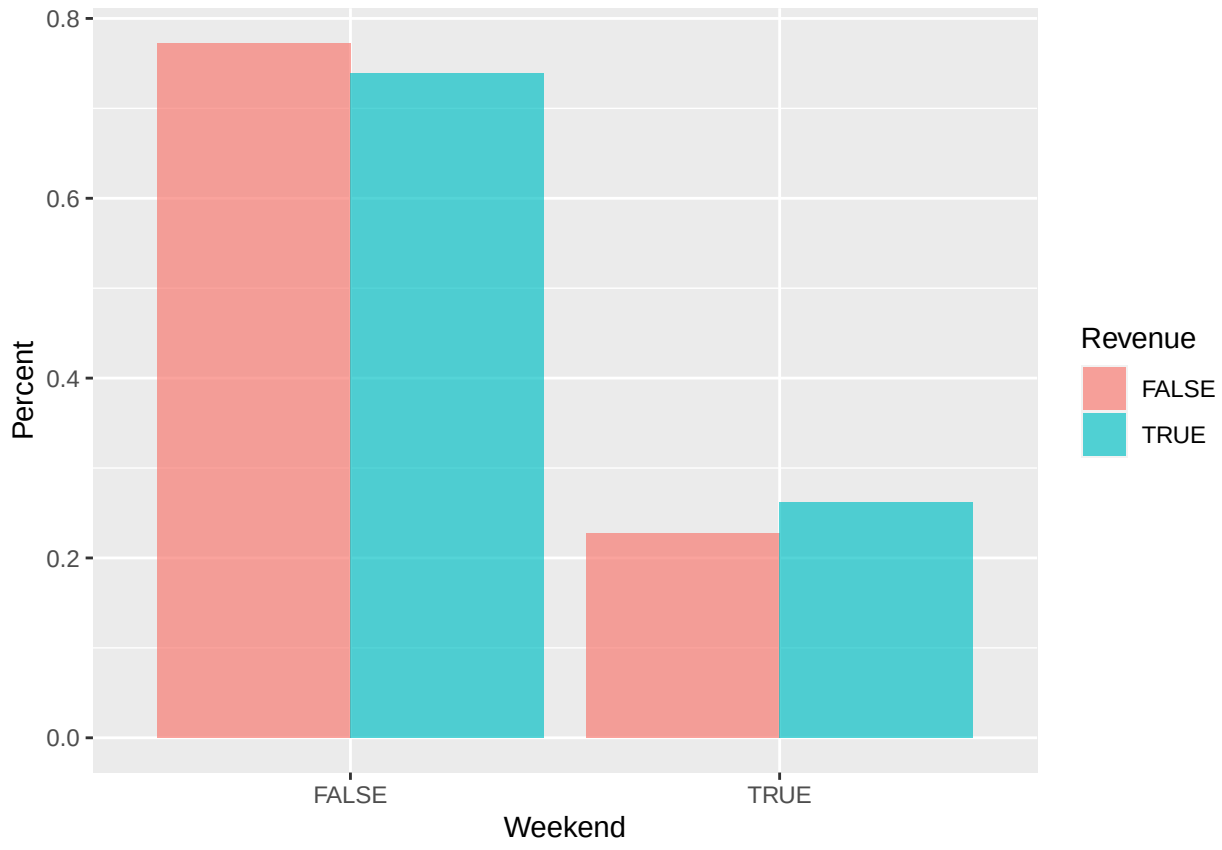



```
CrossTable(DF1$Weekend, DF1$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##              | DF1$Revenue
##  DF1$Weekend |     FALSE |      TRUE | Row Total |
## -------------|-----------|-----------|-----------|
##        FALSE |      8053 |      1409 |      9462 |
##              |     0.381 |     2.080 |           |
##              |     0.851 |     0.149 |     0.767 |
##              |     0.773 |     0.738 |           |
##              |     0.653 |     0.114 |           |
```

```
## -------------|-----------|-----------|-----------|
##         TRUE |      2369 |       499 |      2868 |
##              |     1.257 |     6.864 |           |
##              |     0.826 |     0.174 |     0.233 |
##              |     0.227 |     0.262 |           |
##              |     0.192 |     0.040 |           |
## -------------|-----------|-----------|-----------|
## Column Total |     10422 |      1908 |     12330 |
##              |     0.845 |     0.155 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```r
weekend_table <- table(DF1$Weekend, DF1$Revenue)
weekend_tab <- as.data.frame(prop.table(weekend_table, 2))
colnames(weekend_tab) <- c("Weekend", "Revenue", "Percent")
ggplot(data = weekend_tab, aes(x = Weekend, y = Percent, fill = Revenue)) +
geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3) +
xlab("Weekend")+
ylab("Percent")
```




```r
CrossTable(DF1$TrafficType, DF1$Revenue)
```

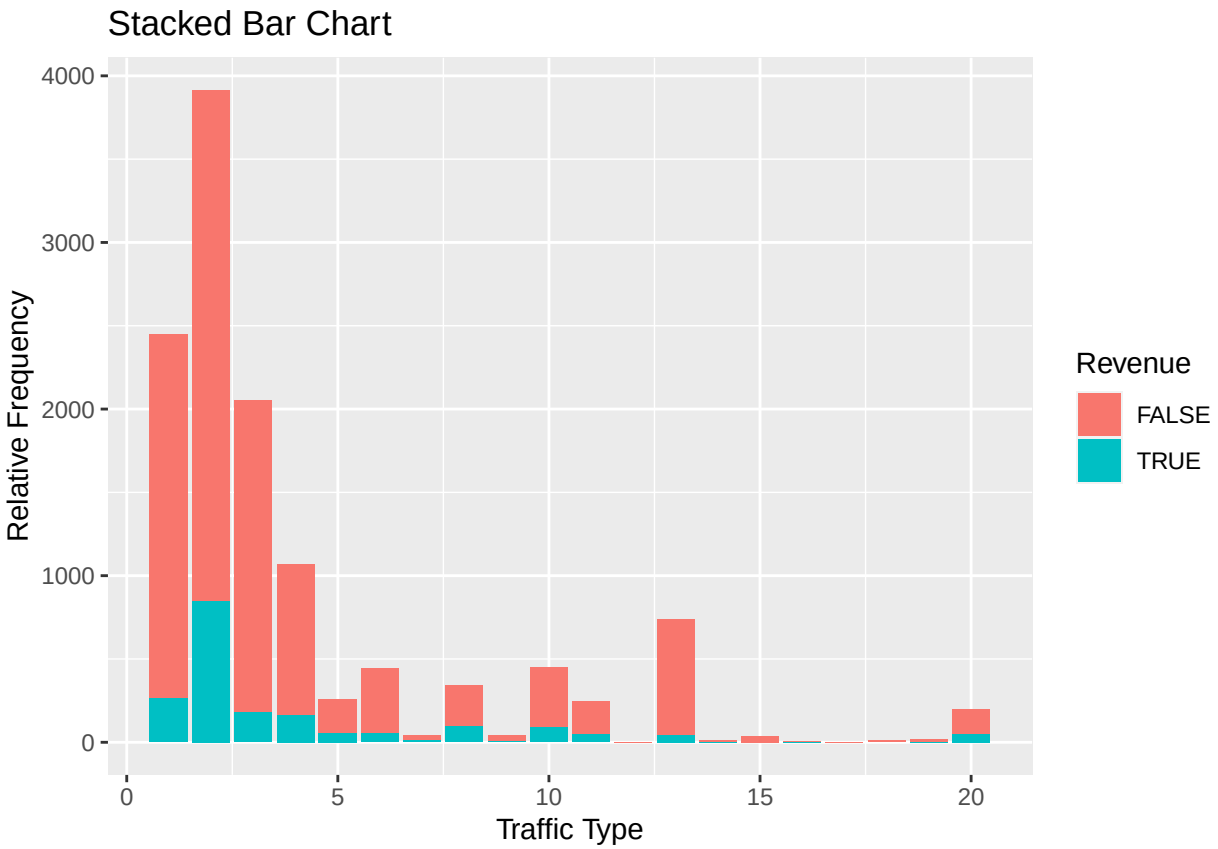
```
##
##
```

```
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##                  | DF1$Revenue
## DF1$TrafficType |     FALSE |      TRUE | Row Total |
## ----------------|-----------|-----------|-----------|
##               1 |      2189 |       262 |      2451 |
##                 |     6.639 |    36.264 |           |
##                 |     0.893 |     0.107 |     0.199 |
##                 |     0.210 |     0.137 |           |
##                 |     0.178 |     0.021 |           |
## ----------------|-----------|-----------|-----------|
##               2 |      3066 |       847 |      3913 |
##                 |    17.631 |    96.306 |           |
##                 |     0.784 |     0.216 |     0.317 |
##                 |     0.294 |     0.444 |           |
##                 |     0.249 |     0.069 |           |
## ----------------|-----------|-----------|-----------|
##               3 |      1872 |       180 |      2052 |
##                 |    10.906 |    59.572 |           |
##                 |     0.912 |     0.088 |     0.166 |
##                 |     0.180 |     0.094 |           |
##                 |     0.152 |     0.015 |           |
## ----------------|-----------|-----------|-----------|
##               4 |       904 |       165 |      1069 |
##                 |     0.000 |     0.001 |           |
##                 |     0.846 |     0.154 |     0.087 |
##                 |     0.087 |     0.086 |           |
##                 |     0.073 |     0.013 |           |
## ----------------|-----------|-----------|-----------|
##               5 |       204 |        56 |       260 |
##                 |     1.131 |     6.178 |           |
##                 |     0.785 |     0.215 |     0.021 |
##                 |     0.020 |     0.029 |           |
##                 |     0.017 |     0.005 |           |
## ----------------|-----------|-----------|-----------|
##               6 |       391 |        53 |       444 |
##                 |     0.657 |     3.591 |           |
##                 |     0.881 |     0.119 |     0.036 |
##                 |     0.038 |     0.028 |           |
##                 |     0.032 |     0.004 |           |
## ----------------|-----------|-----------|-----------|
##               7 |        28 |        12 |        40 |
##                 |     0.998 |     5.454 |           |
```

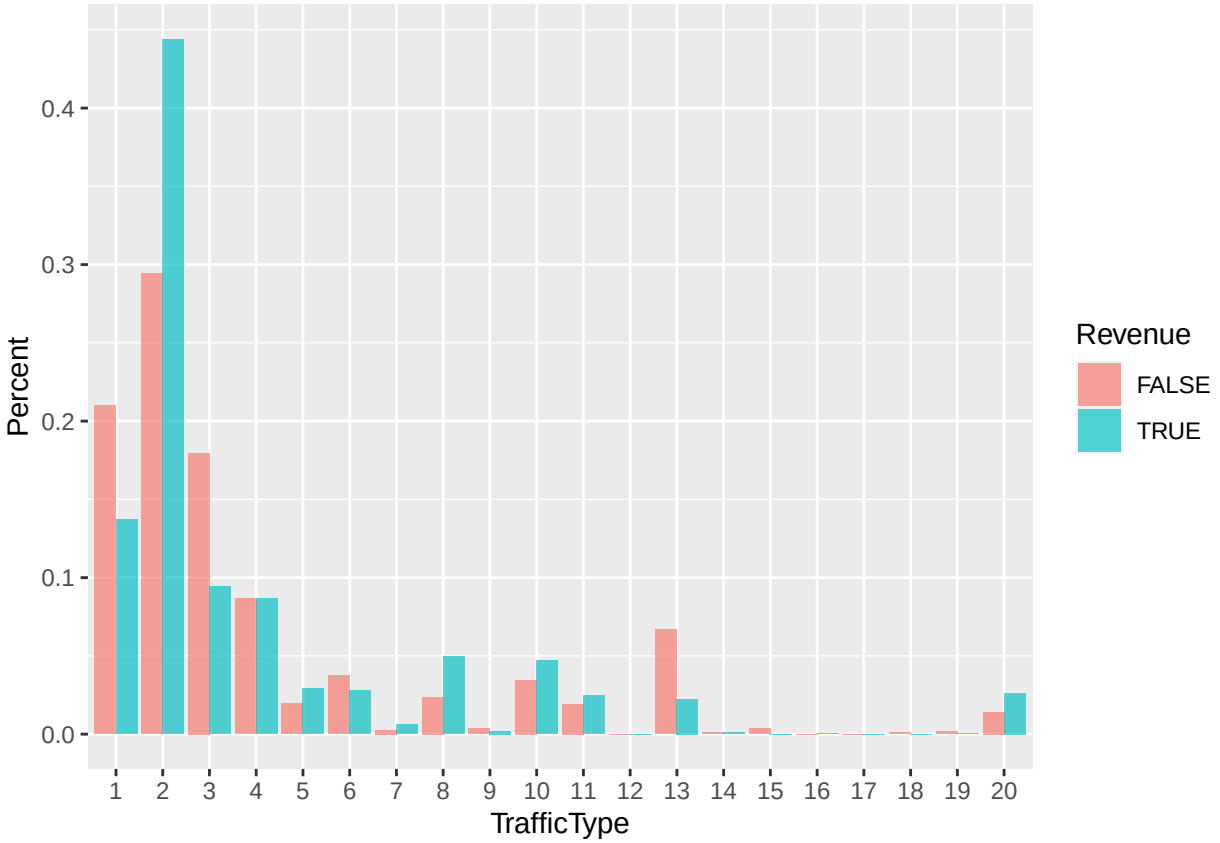
```
##                 |     0.700 |     0.300 |     0.003 |
##                 |     0.003 |     0.006 |           |
##                 |     0.002 |     0.001 |           |
## ----------------|-----------|-----------|-----------|
##               8 |       248 |        95 |       343 |
##                 |     6.062 |    33.112 |           |
##                 |     0.723 |     0.277 |     0.028 |
##                 |     0.024 |     0.050 |           |
##                 |     0.020 |     0.008 |           |
## ----------------|-----------|-----------|-----------|
##               9 |        38 |         4 |        42 |
##                 |     0.176 |     0.961 |           |
##                 |     0.905 |     0.095 |     0.003 |
##                 |     0.004 |     0.002 |           |
##                 |     0.003 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              10 |       360 |        90 |       450 |
##                 |     1.090 |     5.956 |           |
##                 |     0.800 |     0.200 |     0.036 |
##                 |     0.035 |     0.047 |           |
##                 |     0.029 |     0.007 |           |
## ----------------|-----------|-----------|-----------|
##              11 |       200 |        47 |       247 |
##                 |     0.369 |     2.016 |           |
##                 |     0.810 |     0.190 |     0.020 |
##                 |     0.019 |     0.025 |           |
##                 |     0.016 |     0.004 |           |
## ----------------|-----------|-----------|-----------|
##              12 |         1 |         0 |         1 |
##                 |     0.028 |     0.155 |           |
##                 |     1.000 |     0.000 |     0.000 |
##                 |     0.000 |     0.000 |           |
##                 |     0.000 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              13 |       695 |        43 |       738 |
##                 |     8.127 |    44.392 |           |
##                 |     0.942 |     0.058 |     0.060 |
##                 |     0.067 |     0.023 |           |
##                 |     0.056 |     0.003 |           |
## ----------------|-----------|-----------|-----------|
##              14 |        11 |         2 |        13 |
##                 |     0.000 |     0.000 |           |
##                 |     0.846 |     0.154 |     0.001 |
##                 |     0.001 |     0.001 |           |
##                 |     0.001 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              15 |        38 |         0 |        38 |
##                 |     1.077 |     5.880 |           |
##                 |     1.000 |     0.000 |     0.003 |
##                 |     0.004 |     0.000 |           |
##                 |     0.003 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              16 |         2 |         1 |         3 |
##                 |     0.113 |     0.618 |           |
```

```
##                 |     0.667 |     0.333 |     0.000 |
##                 |     0.000 |     0.001 |           |
##                 |     0.000 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              17 |         1 |         0 |         1 |
##                 |     0.028 |     0.155 |           |
##                 |     1.000 |     0.000 |     0.000 |
##                 |     0.000 |     0.000 |           |
##                 |     0.000 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              18 |        10 |         0 |        10 |
##                 |     0.283 |     1.547 |           |
##                 |     1.000 |     0.000 |     0.001 |
##                 |     0.001 |     0.000 |           |
##                 |     0.001 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              19 |        16 |         1 |        17 |
##                 |     0.185 |     1.011 |           |
##                 |     0.941 |     0.059 |     0.001 |
##                 |     0.002 |     0.001 |           |
##                 |     0.001 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              20 |       148 |        50 |       198 |
##                 |     2.240 |    12.234 |           |
##                 |     0.747 |     0.253 |     0.016 |
##                 |     0.014 |     0.026 |           |
##                 |     0.012 |     0.004 |           |
## ----------------|-----------|-----------|-----------|
##    Column Total |     10422 |      1908 |     12330 |
##                 |     0.845 |     0.155 |           |
## ----------------|-----------|-----------|-----------|
##
##
```

```
ggplot(DF1, aes(TrafficType, Revenue = ..count../nrow(DF1),fill = Revenue)) + geom_bar()+
labs(title = "Stacked Bar Chart", x = "Traffic Type", y = "Relative Frequency")
```




```
TrafficType_table <- table(DF1$TrafficType, DF1$Revenue)
TrafficType_tab <- as.data.frame(prop.table(TrafficType_table, 2))
colnames(TrafficType_tab) <- c("TrafficType", "Revenue", "Percent")
ggplot(data = TrafficType_tab, aes(x = TrafficType, y = Percent, fill = Revenue)) +
geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3) +
xlab("TrafficType")+
ylab("Percent")
```

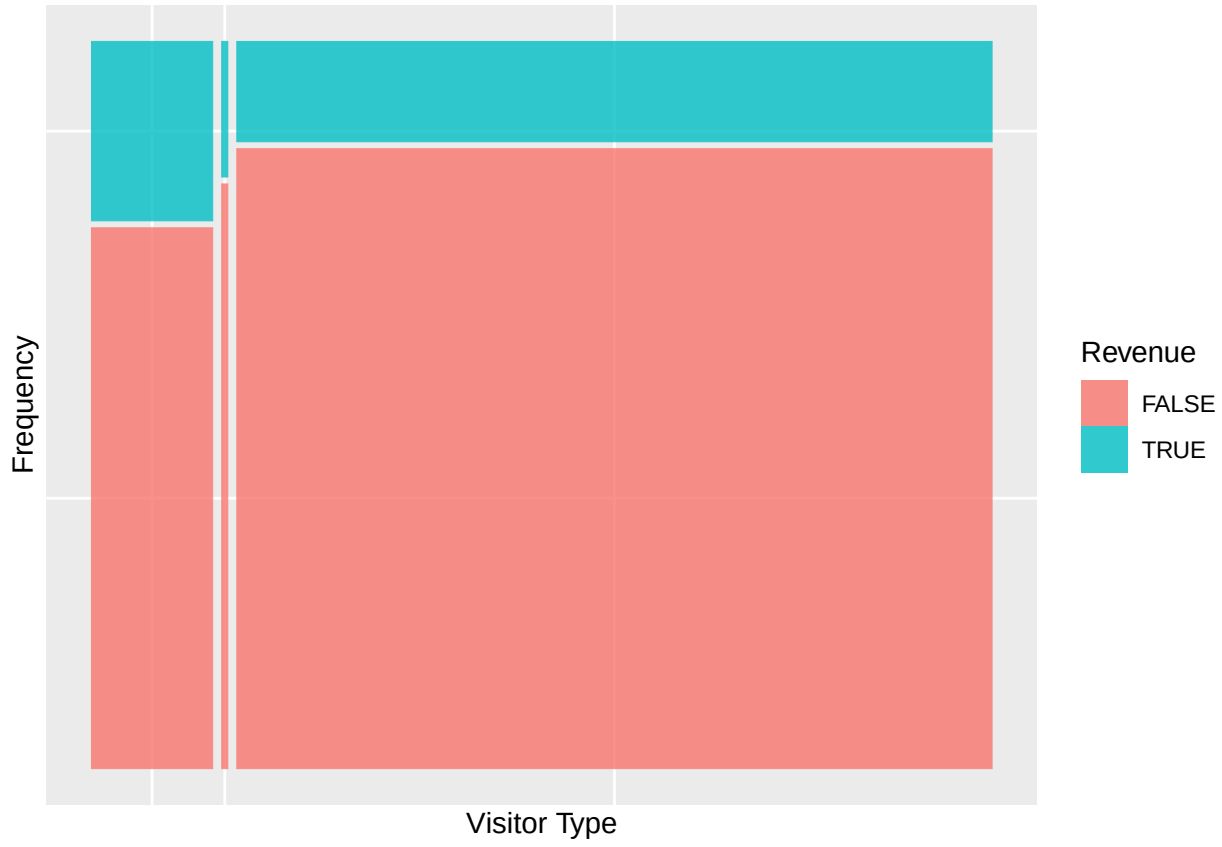



```
CrossTable(DF1$VisitorType, DF1$Revenue)
```

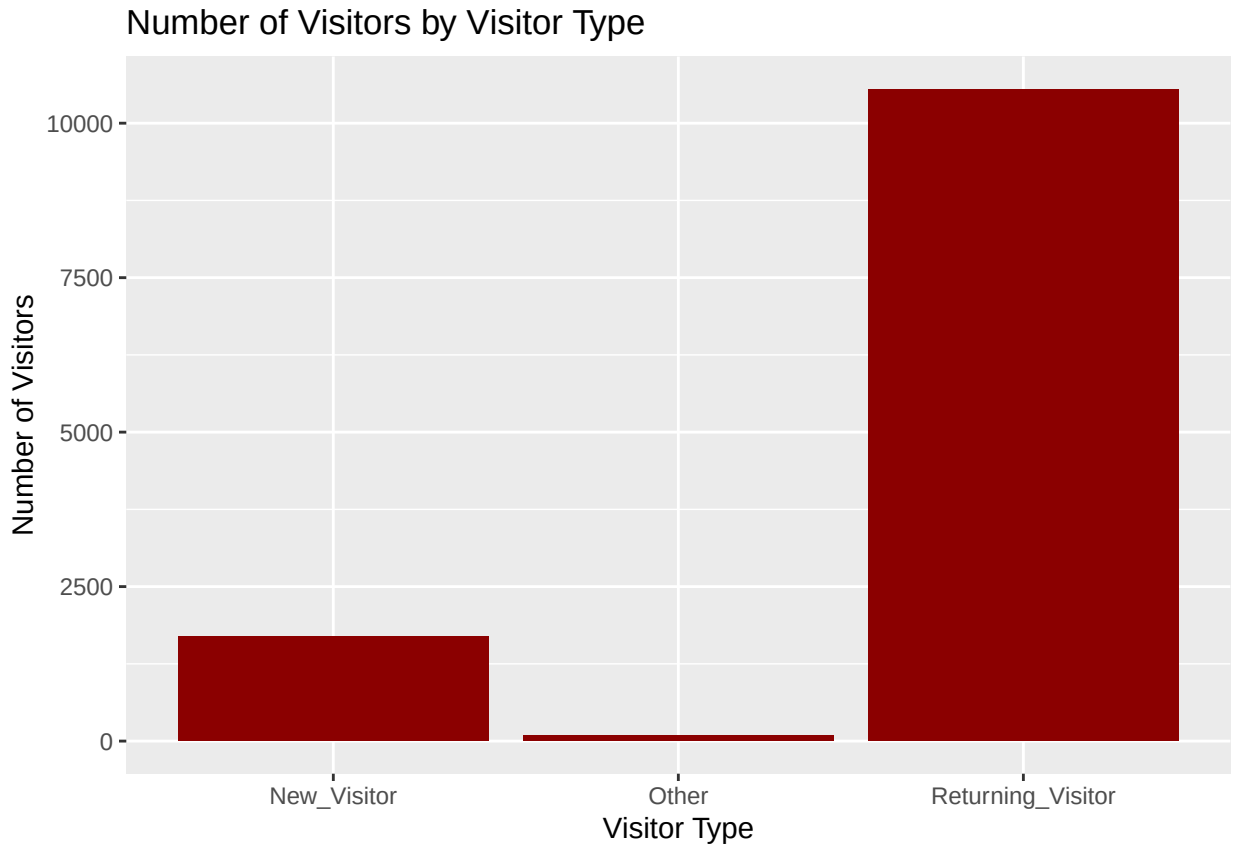
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##                    | DF1$Revenue
##   DF1$VisitorType |     FALSE |      TRUE | Row Total |
## ------------------|-----------|-----------|-----------|
##       New_Visitor |      1272 |       422 |      1694 |
##                   |    17.848 |    97.491 |           |
##                   |     0.751 |     0.249 |     0.137 |
##                   |     0.122 |     0.221 |           |
##                   |     0.103 |     0.034 |           |
```

```
## -------------------|-----------|-----------|-----------|
##              Other |        69 |        16 |        85 |
##                    |     0.113 |     0.616 |           |
##                    |     0.812 |     0.188 |     0.007 |
##                    |     0.007 |     0.008 |           |
##                    |     0.006 |     0.001 |           |
## -------------------|-----------|-----------|-----------|
## Returning_Visitor |      9081 |      1470 |     10551 |
##                    |     2.969 |    16.215 |           |
##                    |     0.861 |     0.139 |     0.856 |
##                    |     0.871 |     0.770 |           |
##                    |     0.736 |     0.119 |           |
## -------------------|-----------|-----------|-----------|
##      Column Total |     10422 |      1908 |     12330 |
##                    |     0.845 |     0.155 |           |
## -------------------|-----------|-----------|-----------|
##
##
```

```
DF1 %>%
ggplot() +
geom_mosaic(aes(x = product(Revenue, VisitorType), fill= Revenue), na.rm=TRUE) +
xlab("Visitor Type") +
ylab("Frequency")
```



```
DF1 %>%
ggplot() +
aes(x = VisitorType) +
geom_bar(fill="darkred") +
ylab("Number of Visitors")+
xlab("Visitor Type")+
ggtitle("Number of Visitors by Visitor Type")
```



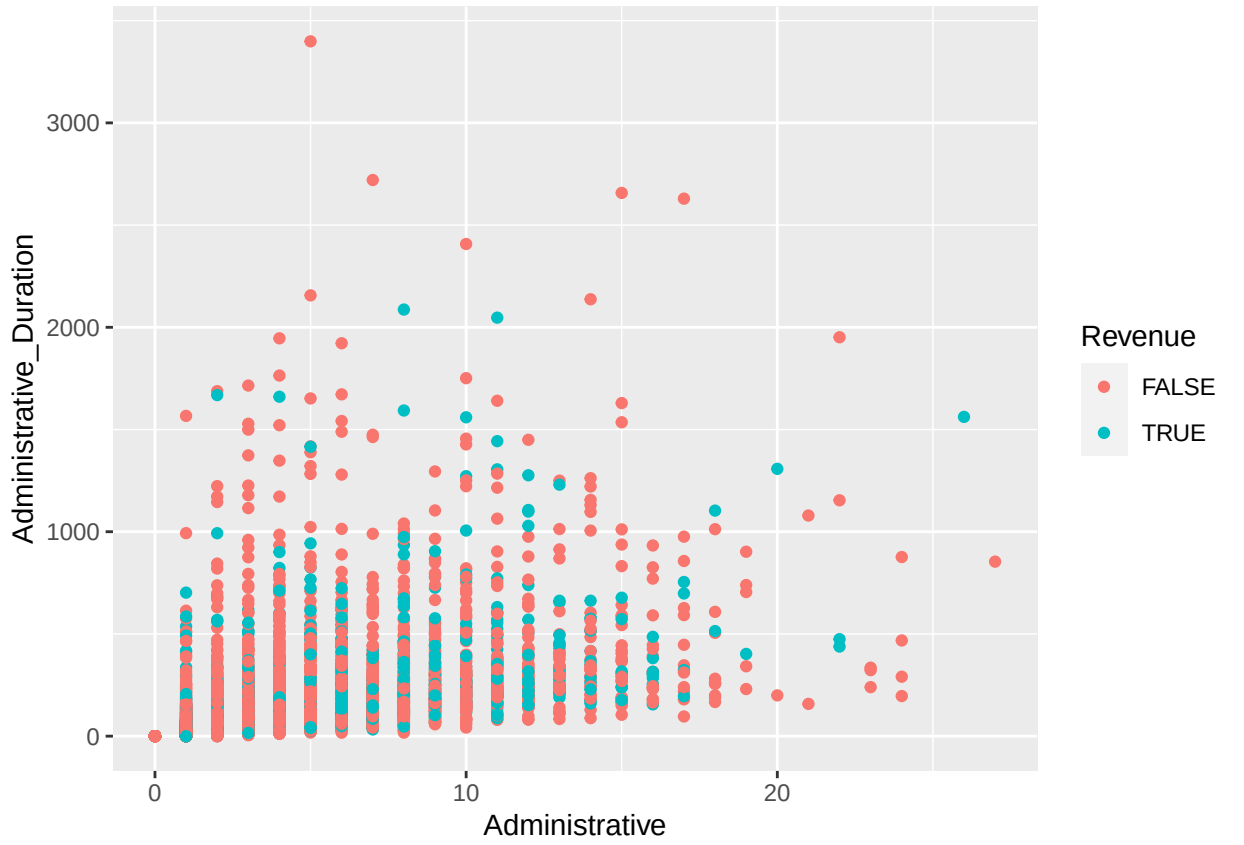

```
summary(DF1$Administrative)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   2.315   4.000  27.000
```

```
summary(DF1$Administrative_Duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    7.50   80.82   93.26 3398.75
```

```
ggplot(data=DF1, aes(x=Administrative, y=Administrative_Duration, group=Revenue, colour=Revenue)
) +
geom_point()
```



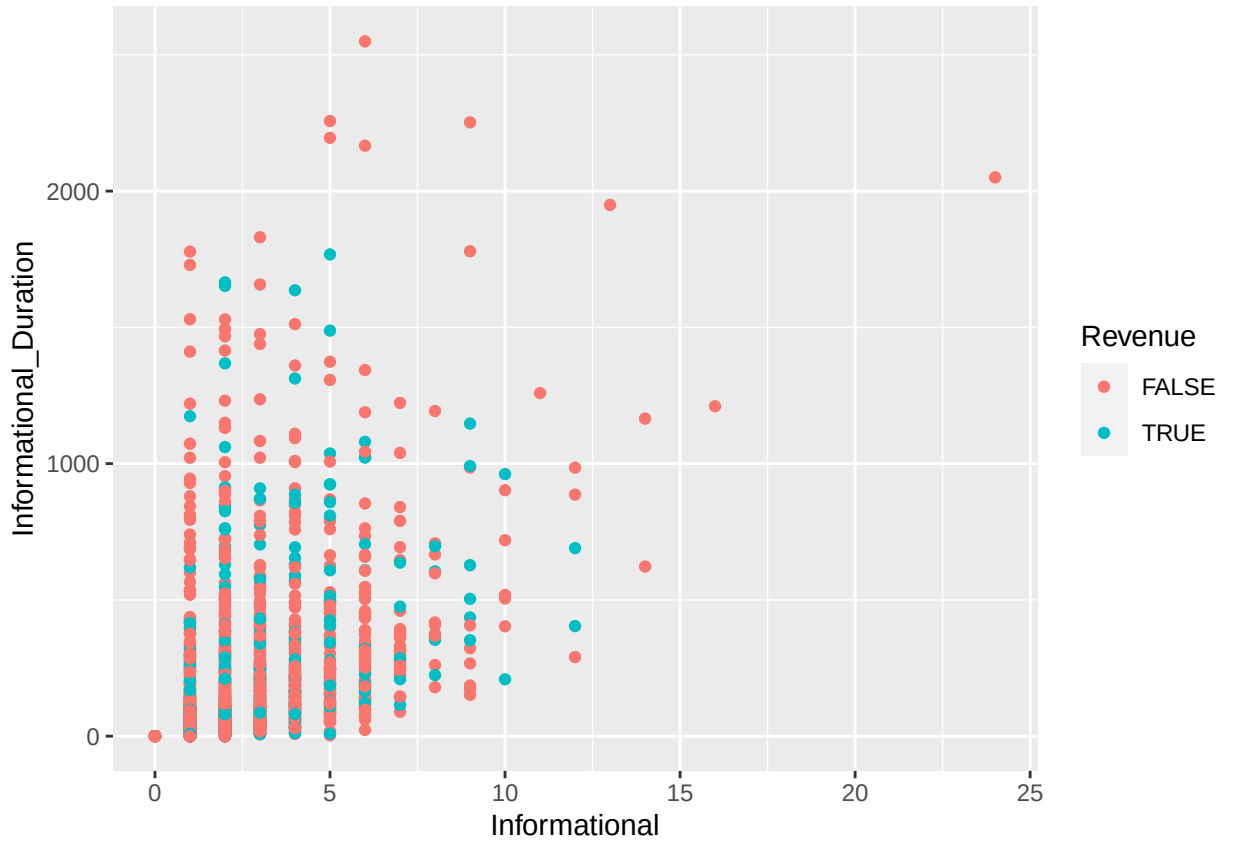

```
summary(DF1$Informational)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.5036  0.0000 24.0000
```

```
summary(DF1$Informational_Duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   34.47    0.00 2549.38
```

```
ggplot(data=DF1, aes(x=Informational, y=Informational_Duration, group=Revenue, colour=Revenue))+
geom_point()
```



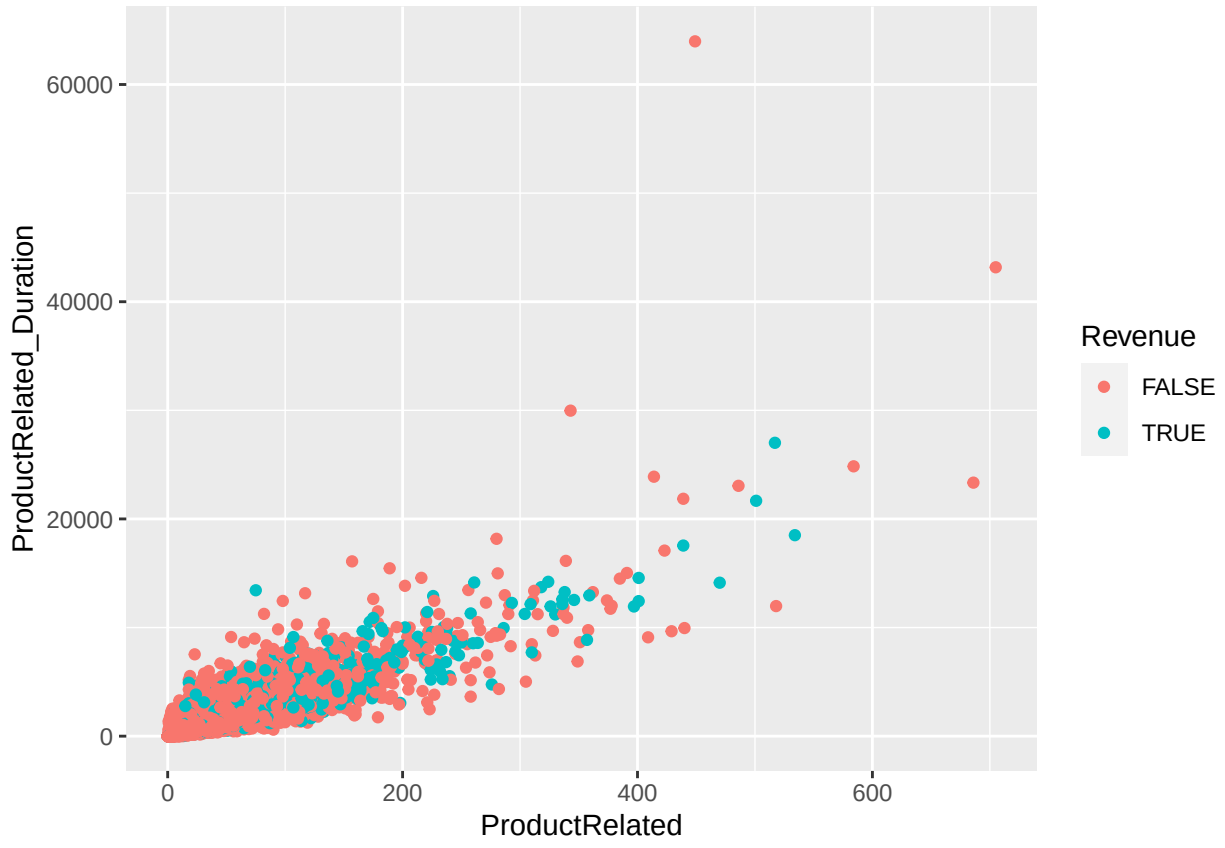

```
summary(DF1$ProductRelated)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    7.00   18.00   31.73   38.00  705.00
```

```
summary(DF1$ProductRelated_Duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   184.1   598.9  1194.8  1464.2 63973.5
```

```
ggplot(data=DF1, aes(x=ProductRelated, y=ProductRelated_Duration, group=Revenue, colour=Revenue)
) +
geom_point()
```



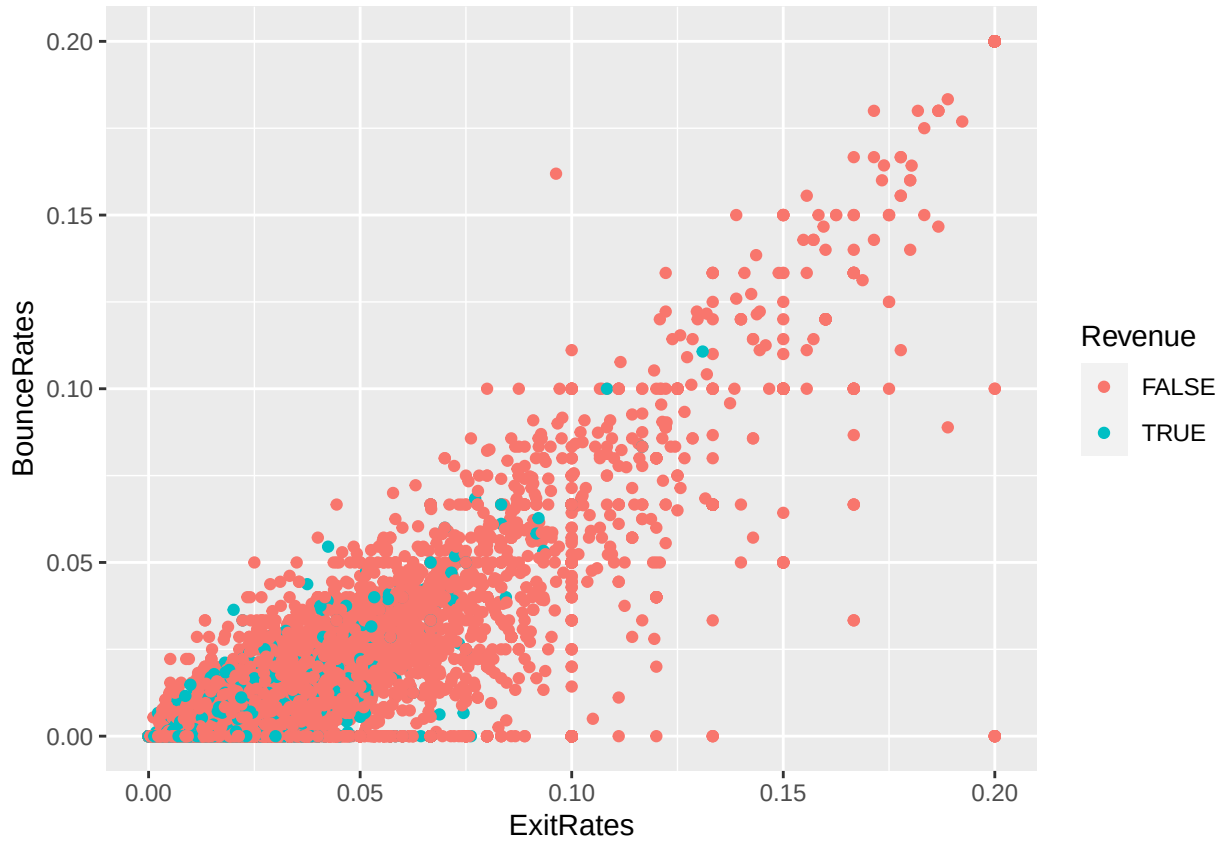

```
summary(DF1$ExitRates)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.01429 0.02516 0.04307 0.05000 0.20000
```

```
summary(DF1$BounceRates)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000000 0.003112 0.022191 0.016813 0.200000
```

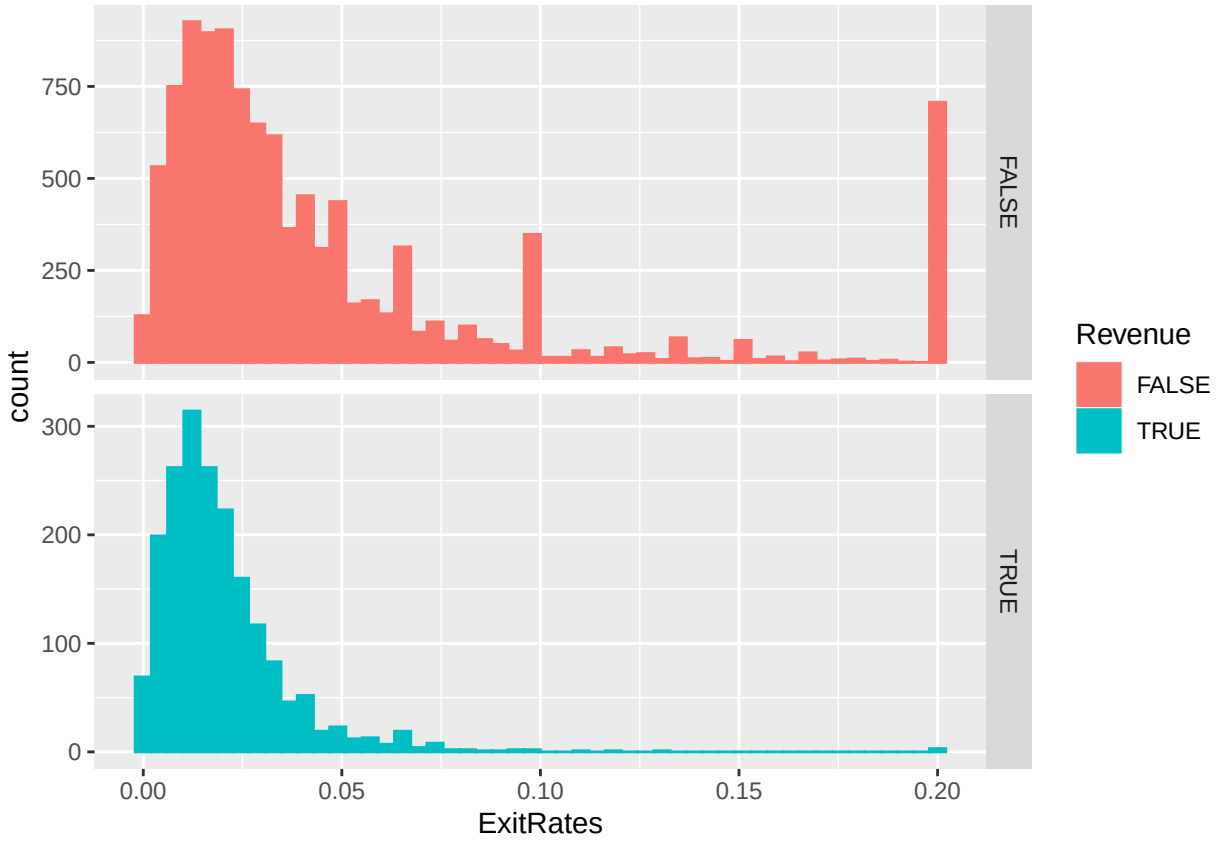
```
ggplot(data=DF1, aes(x=ExitRates, y=BounceRates, group=Revenue, colour=Revenue)
) +
geom_point()
```



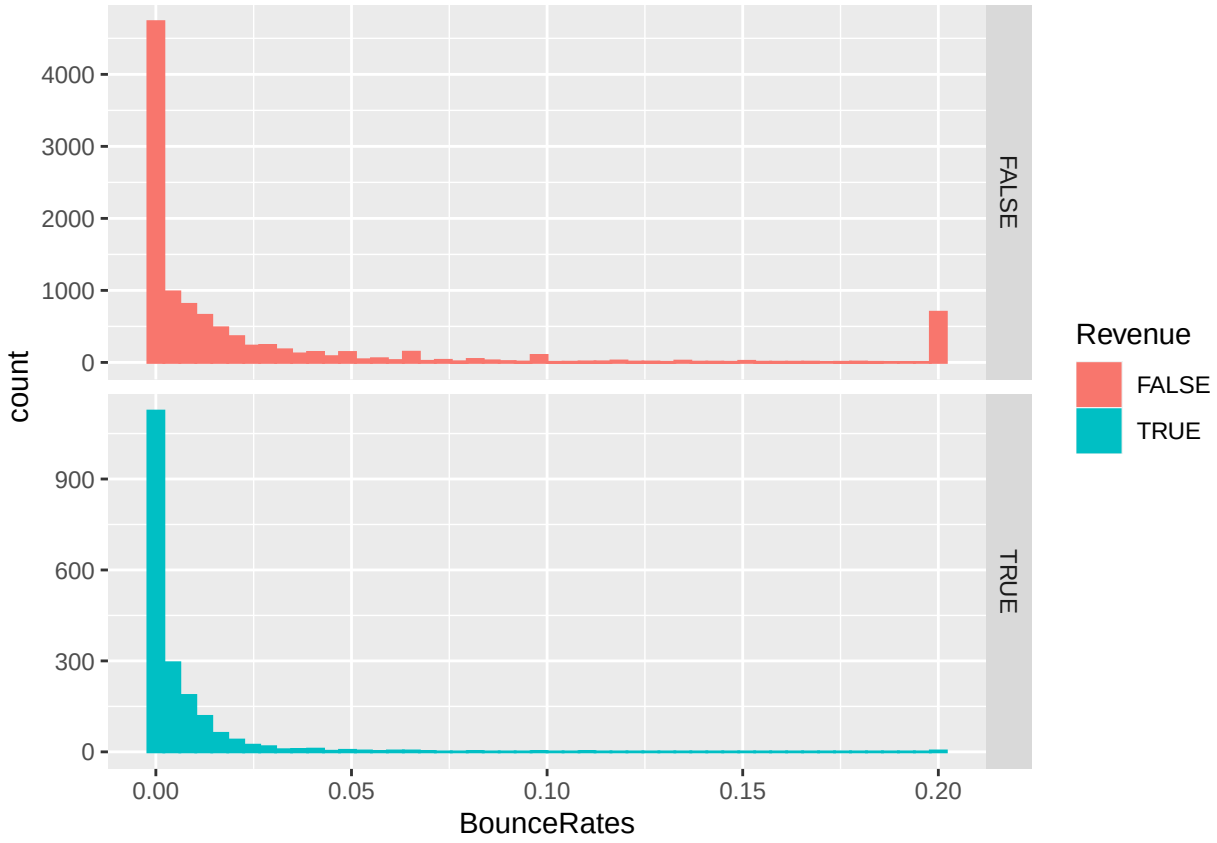

```
summary(DF1$PageValues)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   5.889   0.000 361.764
```

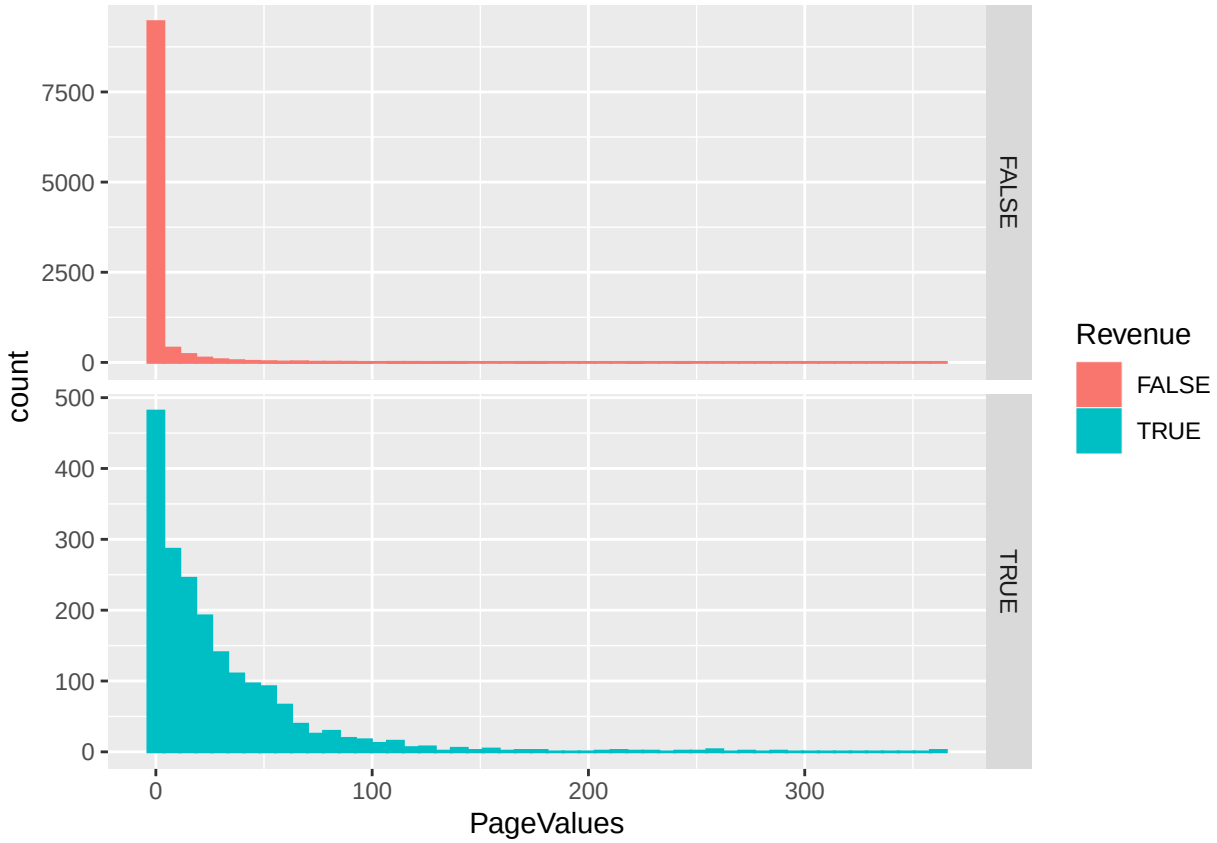
```
DF1 %>%
ggplot() +
aes(x = ExitRates, fill = Revenue, color=Revenue) +
geom_histogram(bins = 50) +
facet_grid(Revenue ~ .,
scales = "free_y")
```




```
DF1 %>%
ggplot() +
aes(x = BounceRates, fill = Revenue, color=Revenue) +
geom_histogram(bins = 50) +
facet_grid(Revenue ~ .,
scales = "free_y")
```




```
DF1 %>%
ggplot() +
aes(x = PageValues, fill = Revenue, color=Revenue) +
geom_histogram(bins = 50) +
facet_grid(Revenue ~ .,
scales = "free_y")
```



Feature Engineering

```
#Dropping Month, the column is missing values for Jan and April.
DF2 <- DF1[c(1:10,12:18)]
head(DF2)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1              0                       0             0                      0
## 2              0                       0             0                      0
## 3              0                       0             0                      0
## 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                0.000000  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 OperatingSystems Browser Region TrafficType       VisitorType
## 1          0                1       1      1           1 Returning_Visitor
## 2          0                2       2      1           2 Returning_Visitor
## 3          0                4       1      9           3 Returning_Visitor
```

```
## 4          0               3       2      2           4 Returning_Visitor
## 5          0               3       3      1           4 Returning_Visitor
## 6          0               2       2      1           3 Returning_Visitor
##   Weekend Revenue
## 1   FALSE   FALSE
## 2   FALSE   FALSE
## 3   FALSE   FALSE
## 4   FALSE   FALSE
## 5    TRUE   FALSE
## 6   FALSE   FALSE
```

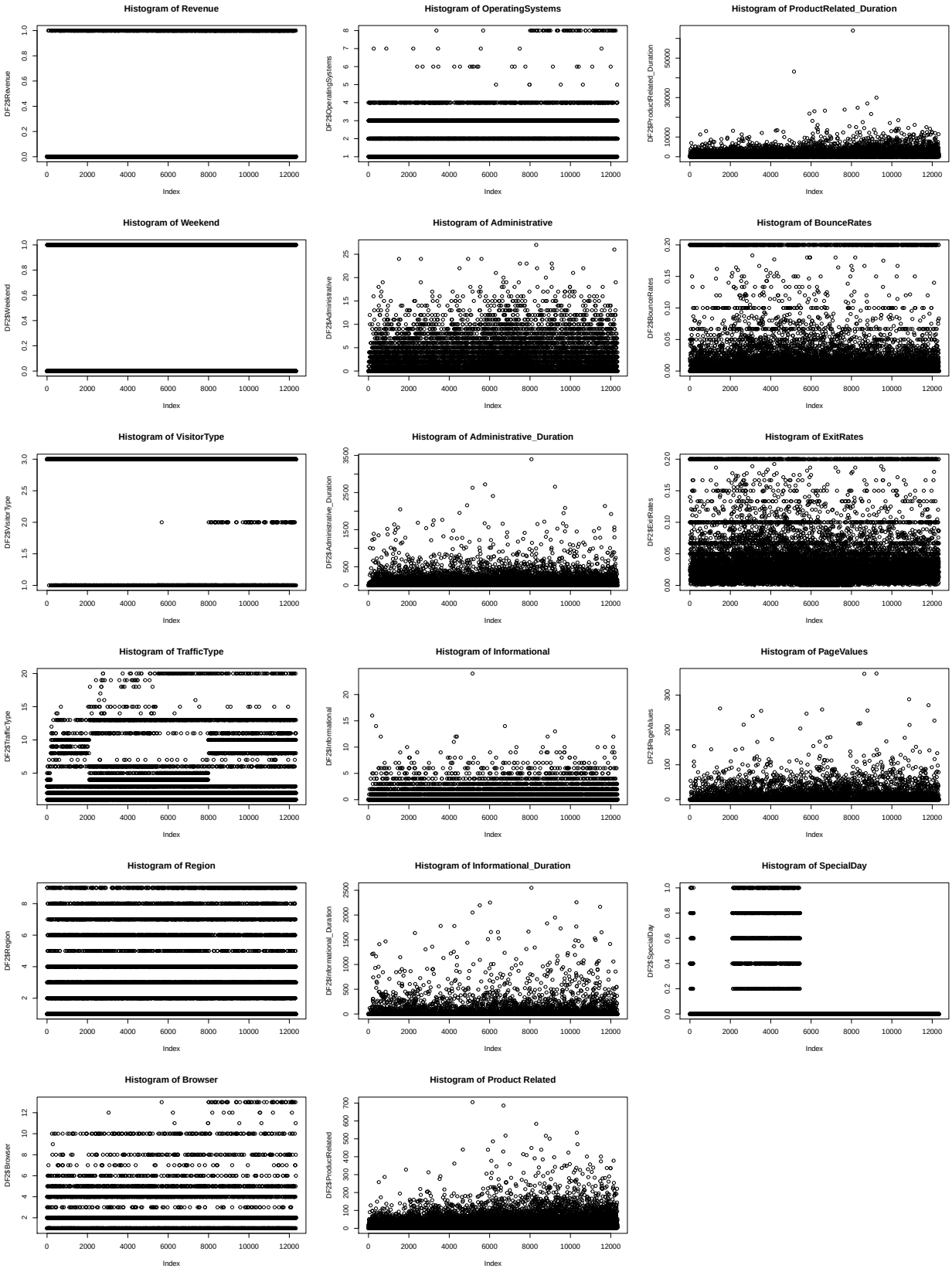
```
#Convert VisitorType, Weekend and Revenue to numeric
VisitorType <- factor(DF2$VisitorType)
DF2$VisitorType <- as.numeric(VisitorType)
DF2$Weekend=ifelse(DF2$Weekend=="TRUE",1,0)
DF2$Revenue=ifelse(DF2$Revenue=="TRUE",1,0)
```

```
CrossTable(DF2$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##           |         0 |         1 |
##           |-----------|-----------|
##           |     10422 |      1908 |
##           |     0.845 |     0.155 |
##           |-----------|-----------|
##
##
##
##
```

```
par(mfcol=c(6,3))
plot(DF2$Revenue,main="Histogram of Revenue")
plot(DF2$Weekend,main="Histogram of Weekend")
plot (DF2$VisitorType,main="Histogram of VisitorType")
plot(DF2$TrafficType,main="Histogram of TrafficType")
plot(DF2$Region,main="Histogram of Region")
plot(DF2$Browser,main="Histogram of Browser")
plot(DF2$OperatingSystems,main="Histogram of OperatingSystems")

plot(DF2$Administrative, main = "Histogram of Administrative")
plot(DF2$Administrative_Duration, main = "Histogram of Administrative_Duration")
plot(DF2$Informational, main = "Histogram of Informational")
```

```r
plot(DF2$Informational_Duration, main = "Histogram of Informational_Duration")
plot(DF2$ProductRelated, main = "Histogram of Product Related")
plot(DF2$ProductRelated_Duration, main = "Histogram of ProductRelated_Duration")
plot(DF2$BounceRates, main = "Histogram of BounceRates")
plot(DF2$ExitRates, main = "Histogram of ExitRates")
plot(DF2$PageValues, main = "Histogram of PageValues")
plot(DF2$SpecialDay, main = "Histogram of SpecialDay")
```


Histogram of Revenue
DF2$Revenue
Index
Histogram of OperatingSystems
DF2$OperatingSystems
Index
Histogram of ProductRelated_Duration
DF2$ProductRelated_Duration
Index
Histogram of Weekend
DF2$Weekend
Index
Histogram of Administrative
DF2$Administrative
Index
Histogram of BounceRates
DF2$BounceRates
Index
Histogram of VisitorType
DF2$VisitorType
Index
Histogram of Administrative_Duration
DF2$Administrative_Duration
Index
Histogram of ExitRates
DF2$ExitRates
Index
Histogram of TrafficType
DF2$TrafficType
Index
Histogram of Informational
DF2$Informational
Index
Histogram of PageValues
DF2$PageValues
Index
Histogram of Region
DF2$Region
Index
Histogram of Informational_Duration
DF2$Informational_Duration
Index
Histogram of SpecialDay
DF2$SpecialDay
Index
Histogram of Browser
DF2$Browser
Index
Histogram of Product Related
DF2$ProductRelated
Index

```
CrossTable(DF2$Region,DF2$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##              | DF2$Revenue
##   DF2$Region |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |      4009 |       771 |      4780 |
##              |     0.243 |     1.326 |           |
##              |     0.839 |     0.161 |     0.388 |
##              |     0.385 |     0.404 |           |
##              |     0.325 |     0.063 |           |
## -------------|-----------|-----------|-----------|
##            2 |       948 |       188 |      1136 |
##              |     0.155 |     0.848 |           |
##              |     0.835 |     0.165 |     0.092 |
##              |     0.091 |     0.099 |           |
##              |     0.077 |     0.015 |           |
## -------------|-----------|-----------|-----------|
##            3 |      2054 |       349 |      2403 |
##              |     0.257 |     1.404 |           |
##              |     0.855 |     0.145 |     0.195 |
##              |     0.197 |     0.183 |           |
##              |     0.167 |     0.028 |           |
## -------------|-----------|-----------|-----------|
##            4 |      1007 |       175 |      1182 |
##              |     0.063 |     0.342 |           |
##              |     0.852 |     0.148 |     0.096 |
##              |     0.097 |     0.092 |           |
##              |     0.082 |     0.014 |           |
## -------------|-----------|-----------|-----------|
##            5 |       266 |        52 |       318 |
##              |     0.029 |     0.158 |           |
##              |     0.836 |     0.164 |     0.026 |
##              |     0.026 |     0.027 |           |
##              |     0.022 |     0.004 |           |
## -------------|-----------|-----------|-----------|
##            6 |       693 |       112 |       805 |
##              |     0.232 |     1.268 |           |
##              |     0.861 |     0.139 |     0.065 |
```

```
##              |     0.066 |     0.059 |           |
##              |     0.056 |     0.009 |           |
## -------------|-----------|-----------|-----------|
##            7 |       642 |       119 |       761 |
##              |     0.002 |     0.013 |           |
##              |     0.844 |     0.156 |     0.062 |
##              |     0.062 |     0.062 |           |
##              |     0.052 |     0.010 |           |
## -------------|-----------|-----------|-----------|
##            8 |       378 |        56 |       434 |
##              |     0.339 |     1.854 |           |
##              |     0.871 |     0.129 |     0.035 |
##              |     0.036 |     0.029 |           |
##              |     0.031 |     0.005 |           |
## -------------|-----------|-----------|-----------|
##            9 |       425 |        86 |       511 |
##              |     0.111 |     0.607 |           |
##              |     0.832 |     0.168 |     0.041 |
##              |     0.041 |     0.045 |           |
##              |     0.034 |     0.007 |           |
## -------------|-----------|-----------|-----------|
## Column Total |     10422 |      1908 |     12330 |
##              |     0.845 |     0.155 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
CrossTable(DF2$Browser,DF2$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##              | DF2$Revenue
##  DF2$Browser |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |      2097 |       365 |      2462 |
##              |     0.123 |     0.670 |           |
##              |     0.852 |     0.148 |     0.200 |
##              |     0.201 |     0.191 |           |
##              |     0.170 |     0.030 |           |
## -------------|-----------|-----------|-----------|
##            2 |      6738 |      1223 |      7961 |
```

```
##              |     0.012 |     0.065 |           |
##              |     0.846 |     0.154 |     0.646 |
##              |     0.647 |     0.641 |           |
##              |     0.546 |     0.099 |           |
## -------------|-----------|-----------|-----------|
##            3 |       100 |         5 |       105 |
##              |     1.426 |     7.787 |           |
##              |     0.952 |     0.048 |     0.009 |
##              |     0.010 |     0.003 |           |
##              |     0.008 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            4 |       606 |       130 |       736 |
##              |     0.417 |     2.278 |           |
##              |     0.823 |     0.177 |     0.060 |
##              |     0.058 |     0.068 |           |
##              |     0.049 |     0.011 |           |
## -------------|-----------|-----------|-----------|
##            5 |       381 |        86 |       467 |
##              |     0.478 |     2.610 |           |
##              |     0.816 |     0.184 |     0.038 |
##              |     0.037 |     0.045 |           |
##              |     0.031 |     0.007 |           |
## -------------|-----------|-----------|-----------|
##            6 |       154 |        20 |       174 |
##              |     0.326 |     1.781 |           |
##              |     0.885 |     0.115 |     0.014 |
##              |     0.015 |     0.010 |           |
##              |     0.012 |     0.002 |           |
## -------------|-----------|-----------|-----------|
##            7 |        43 |         6 |        49 |
##              |     0.060 |     0.330 |           |
##              |     0.878 |     0.122 |     0.004 |
##              |     0.004 |     0.003 |           |
##              |     0.003 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            8 |       114 |        21 |       135 |
##              |     0.000 |     0.001 |           |
##              |     0.844 |     0.156 |     0.011 |
##              |     0.011 |     0.011 |           |
##              |     0.009 |     0.002 |           |
## -------------|-----------|-----------|-----------|
##            9 |         1 |         0 |         1 |
##              |     0.028 |     0.155 |           |
##              |     1.000 |     0.000 |     0.000 |
##              |     0.000 |     0.000 |           |
##              |     0.000 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##           10 |       131 |        32 |       163 |
##              |     0.333 |     1.821 |           |
##              |     0.804 |     0.196 |     0.013 |
##              |     0.013 |     0.017 |           |
##              |     0.011 |     0.003 |           |
## -------------|-----------|-----------|-----------|
##           11 |         5 |         1 |         6 |
```

```
##               |     0.001 |     0.006 |           |
##               |     0.833 |     0.167 |     0.000 |
##               |     0.000 |     0.001 |           |
##               |     0.000 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            12 |         7 |         3 |        10 |
##               |     0.250 |     1.363 |           |
##               |     0.700 |     0.300 |     0.001 |
##               |     0.001 |     0.002 |           |
##               |     0.001 |     0.000 |           |
## -------------|-----------|-----------|-----------|
##            13 |        45 |        16 |        61 |
##               |     0.835 |     4.560 |           |
##               |     0.738 |     0.262 |     0.005 |
##               |     0.004 |     0.008 |           |
##               |     0.004 |     0.001 |           |
## -------------|-----------|-----------|-----------|
## Column Total |     10422 |      1908 |     12330 |
##               |     0.845 |     0.155 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
CrossTable(DF2$TrafficType,DF2$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##                  | DF2$Revenue
## DF2$TrafficType |         0 |         1 | Row Total |
## ----------------|-----------|-----------|-----------|
##               1 |      2189 |       262 |      2451 |
##                 |     6.639 |    36.264 |           |
##                 |     0.893 |     0.107 |     0.199 |
##                 |     0.210 |     0.137 |           |
##                 |     0.178 |     0.021 |           |
## ----------------|-----------|-----------|-----------|
##               2 |      3066 |       847 |      3913 |
##                 |    17.631 |    96.306 |           |
##                 |     0.784 |     0.216 |     0.317 |
##                 |     0.294 |     0.444 |           |
##                 |     0.249 |     0.069 |           |
```

```
## ----------------|-----------|-----------|-----------|
##               3 |      1872 |       180 |      2052 |
##                 |    10.906 |    59.572 |           |
##                 |     0.912 |     0.088 |     0.166 |
##                 |     0.180 |     0.094 |           |
##                 |     0.152 |     0.015 |           |
## ----------------|-----------|-----------|-----------|
##               4 |       904 |       165 |      1069 |
##                 |     0.000 |     0.001 |           |
##                 |     0.846 |     0.154 |     0.087 |
##                 |     0.087 |     0.086 |           |
##                 |     0.073 |     0.013 |           |
## ----------------|-----------|-----------|-----------|
##               5 |       204 |        56 |       260 |
##                 |     1.131 |     6.178 |           |
##                 |     0.785 |     0.215 |     0.021 |
##                 |     0.020 |     0.029 |           |
##                 |     0.017 |     0.005 |           |
## ----------------|-----------|-----------|-----------|
##               6 |       391 |        53 |       444 |
##                 |     0.657 |     3.591 |           |
##                 |     0.881 |     0.119 |     0.036 |
##                 |     0.038 |     0.028 |           |
##                 |     0.032 |     0.004 |           |
## ----------------|-----------|-----------|-----------|
##               7 |        28 |        12 |        40 |
##                 |     0.998 |     5.454 |           |
##                 |     0.700 |     0.300 |     0.003 |
##                 |     0.003 |     0.006 |           |
##                 |     0.002 |     0.001 |           |
## ----------------|-----------|-----------|-----------|
##               8 |       248 |        95 |       343 |
##                 |     6.062 |    33.112 |           |
##                 |     0.723 |     0.277 |     0.028 |
##                 |     0.024 |     0.050 |           |
##                 |     0.020 |     0.008 |           |
## ----------------|-----------|-----------|-----------|
##               9 |        38 |         4 |        42 |
##                 |     0.176 |     0.961 |           |
##                 |     0.905 |     0.095 |     0.003 |
##                 |     0.004 |     0.002 |           |
##                 |     0.003 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              10 |       360 |        90 |       450 |
##                 |     1.090 |     5.956 |           |
##                 |     0.800 |     0.200 |     0.036 |
##                 |     0.035 |     0.047 |           |
##                 |     0.029 |     0.007 |           |
## ----------------|-----------|-----------|-----------|
##              11 |       200 |        47 |       247 |
##                 |     0.369 |     2.016 |           |
##                 |     0.810 |     0.190 |     0.020 |
##                 |     0.019 |     0.025 |           |
##                 |     0.016 |     0.004 |           |
```

```
## ----------------|-----------|-----------|-----------|
##              12 |         1 |         0 |         1 |
##                 |     0.028 |     0.155 |           |
##                 |     1.000 |     0.000 |     0.000 |
##                 |     0.000 |     0.000 |           |
##                 |     0.000 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              13 |       695 |        43 |       738 |
##                 |     8.127 |    44.392 |           |
##                 |     0.942 |     0.058 |     0.060 |
##                 |     0.067 |     0.023 |           |
##                 |     0.056 |     0.003 |           |
## ----------------|-----------|-----------|-----------|
##              14 |        11 |         2 |        13 |
##                 |     0.000 |     0.000 |           |
##                 |     0.846 |     0.154 |     0.001 |
##                 |     0.001 |     0.001 |           |
##                 |     0.001 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              15 |        38 |         0 |        38 |
##                 |     1.077 |     5.880 |           |
##                 |     1.000 |     0.000 |     0.003 |
##                 |     0.004 |     0.000 |           |
##                 |     0.003 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              16 |         2 |         1 |         3 |
##                 |     0.113 |     0.618 |           |
##                 |     0.667 |     0.333 |     0.000 |
##                 |     0.000 |     0.001 |           |
##                 |     0.000 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              17 |         1 |         0 |         1 |
##                 |     0.028 |     0.155 |           |
##                 |     1.000 |     0.000 |     0.000 |
##                 |     0.000 |     0.000 |           |
##                 |     0.000 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              18 |        10 |         0 |        10 |
##                 |     0.283 |     1.547 |           |
##                 |     1.000 |     0.000 |     0.001 |
##                 |     0.001 |     0.000 |           |
##                 |     0.001 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              19 |        16 |         1 |        17 |
##                 |     0.185 |     1.011 |           |
##                 |     0.941 |     0.059 |     0.001 |
##                 |     0.002 |     0.001 |           |
##                 |     0.001 |     0.000 |           |
## ----------------|-----------|-----------|-----------|
##              20 |       148 |        50 |       198 |
##                 |     2.240 |    12.234 |           |
##                 |     0.747 |     0.253 |     0.016 |
##                 |     0.014 |     0.026 |           |
##                 |     0.012 |     0.004 |           |
```

```
## -----------------|-----------|-----------|-----------|
##    Column Total |     10422 |      1908 |     12330 |
##                 |     0.845 |     0.155 |           |
## ----------------|-----------|-----------|-----------|
##
##
```

```
CrossTable(DF2$VisitorType,DF2$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##                  | DF2$Revenue
## DF2$VisitorType |         0 |         1 | Row Total |
## -----------------|-----------|-----------|-----------|
##                1 |      1272 |       422 |      1694 |
##                  |    17.848 |    97.491 |           |
##                  |     0.751 |     0.249 |     0.137 |
##                  |     0.122 |     0.221 |           |
##                  |     0.103 |     0.034 |           |
## -----------------|-----------|-----------|-----------|
##                2 |        69 |        16 |        85 |
##                  |     0.113 |     0.616 |           |
##                  |     0.812 |     0.188 |     0.007 |
##                  |     0.007 |     0.008 |           |
##                  |     0.006 |     0.001 |           |
## -----------------|-----------|-----------|-----------|
##                3 |      9081 |      1470 |     10551 |
##                  |     2.969 |    16.215 |           |
##                  |     0.861 |     0.139 |     0.856 |
##                  |     0.871 |     0.770 |           |
##                  |     0.736 |     0.119 |           |
## -----------------|-----------|-----------|-----------|
##    Column Total |     10422 |      1908 |     12330 |
##                 |     0.845 |     0.155 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

```
CrossTable(DF1$Month,DF2$VisitorType)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##              | DF2$VisitorType
##    DF1$Month |         1 |         2 |         3 | Row Total |
## -------------|-----------|-----------|-----------|-----------|
##          Aug |        72 |         0 |       361 |       433 |
##              |     2.631 |     2.985 |     0.245 |           |
##              |     0.166 |     0.000 |     0.834 |     0.035 |
##              |     0.043 |     0.000 |     0.034 |           |
##              |     0.006 |     0.000 |     0.029 |           |
## -------------|-----------|-----------|-----------|-----------|
##          Dec |       335 |        62 |      1330 |      1727 |
##              |    40.254 |   210.781 |    14.787 |           |
##              |     0.194 |     0.036 |     0.770 |     0.140 |
##              |     0.198 |     0.729 |     0.126 |           |
##              |     0.027 |     0.005 |     0.108 |           |
## -------------|-----------|-----------|-----------|-----------|
##          Feb |         1 |         0 |       183 |       184 |
##              |    23.319 |     1.268 |     4.145 |           |
##              |     0.005 |     0.000 |     0.995 |     0.015 |
##              |     0.001 |     0.000 |     0.017 |           |
##              |     0.000 |     0.000 |     0.015 |           |
## -------------|-----------|-----------|-----------|-----------|
##          Jul |        54 |         0 |       378 |       432 |
##              |     0.483 |     2.978 |     0.188 |           |
##              |     0.125 |     0.000 |     0.875 |     0.035 |
##              |     0.032 |     0.000 |     0.036 |           |
##              |     0.004 |     0.000 |     0.031 |           |
## -------------|-----------|-----------|-----------|-----------|
##         June |        30 |         1 |       257 |       288 |
##              |     2.314 |     0.489 |     0.452 |           |
##              |     0.104 |     0.003 |     0.892 |     0.023 |
##              |     0.018 |     0.012 |     0.024 |           |
##              |     0.002 |     0.000 |     0.021 |           |
## -------------|-----------|-----------|-----------|-----------|
##          Mar |       232 |         0 |      1675 |      1907 |
##              |     3.435 |    13.146 |     1.141 |           |
##              |     0.122 |     0.000 |     0.878 |     0.155 |
##              |     0.137 |     0.000 |     0.159 |           |
##              |     0.019 |     0.000 |     0.136 |           |
## -------------|-----------|-----------|-----------|-----------|
```

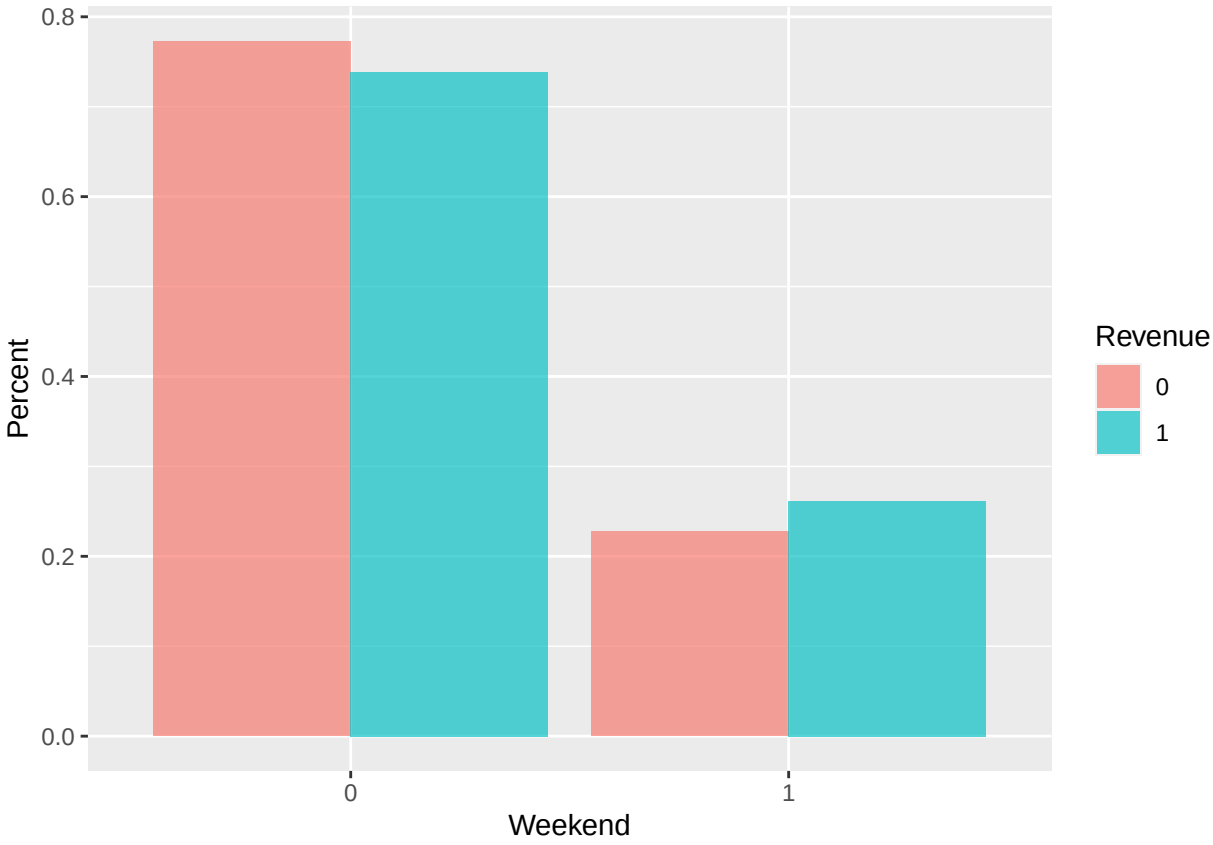
```
##           May |       319 |         0 |      3045 |      3364 |
##               |    44.353 |    23.191 |     9.615 |           |
##               |     0.095 |     0.000 |     0.905 |     0.273 |
##               |     0.188 |     0.000 |     0.289 |           |
##               |     0.026 |     0.000 |     0.247 |           |
## -------------|-----------|-----------|-----------|-----------|
##           Nov |       419 |        22 |      2557 |      2998 |
##               |     0.123 |     0.086 |     0.028 |           |
##               |     0.140 |     0.007 |     0.853 |     0.243 |
##               |     0.247 |     0.259 |     0.242 |           |
##               |     0.034 |     0.002 |     0.207 |           |
## -------------|-----------|-----------|-----------|-----------|
##           Oct |       124 |         0 |       425 |       549 |
##               |    31.281 |     3.785 |     4.270 |           |
##               |     0.226 |     0.000 |     0.774 |     0.045 |
##               |     0.073 |     0.000 |     0.040 |           |
##               |     0.010 |     0.000 |     0.034 |           |
## -------------|-----------|-----------|-----------|-----------|
##           Sep |       108 |         0 |       340 |       448 |
##               |    35.054 |     3.088 |     4.905 |           |
##               |     0.241 |     0.000 |     0.759 |     0.036 |
##               |     0.064 |     0.000 |     0.032 |           |
##               |     0.009 |     0.000 |     0.028 |           |
## -------------|-----------|-----------|-----------|-----------|
## Column Total |      1694 |        85 |     10551 |     12330 |
##               |     0.137 |     0.007 |     0.856 |           |
## -------------|-----------|-----------|-----------|-----------|
##
##
```

```
CrossTable(DF2$SpecialDay,DF2$Revenue)
```

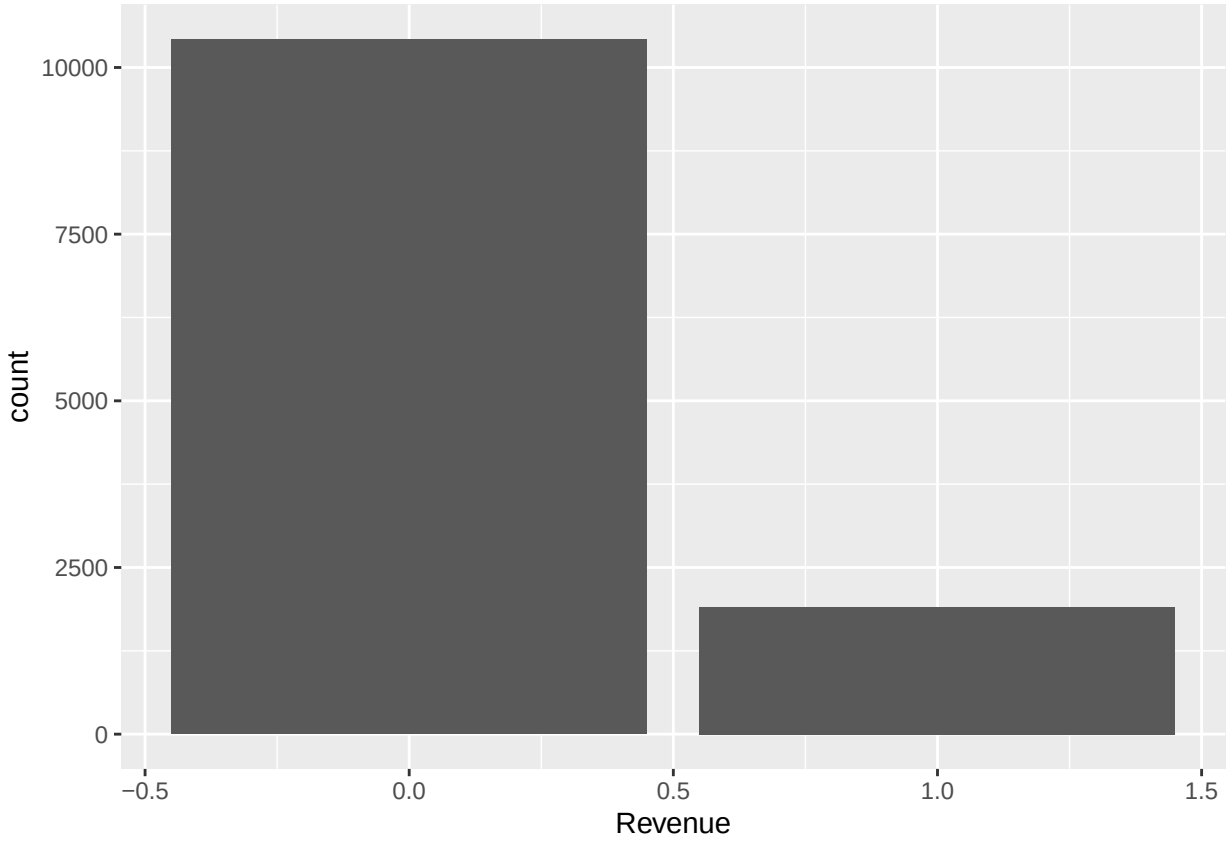
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##                | DF2$Revenue
## DF2$SpecialDay |         0 |         1 | Row Total |
## ---------------|-----------|-----------|-----------|
##              0 |      9248 |      1831 |     11079 |
##                |     1.451 |     7.928 |           |
##                |     0.835 |     0.165 |     0.899 |
##                |     0.887 |     0.960 |           |
```

```
##                |     0.750 |     0.148 |           |
## ---------------|-----------|-----------|-----------|
##            0.2 |       164 |        14 |       178 |
##                |     1.219 |     6.660 |           |
##                |     0.921 |     0.079 |     0.014 |
##                |     0.016 |     0.007 |           |
##                |     0.013 |     0.001 |           |
## ---------------|-----------|-----------|-----------|
##            0.4 |       230 |        13 |       243 |
##                |     2.947 |    16.097 |           |
##                |     0.947 |     0.053 |     0.020 |
##                |     0.022 |     0.007 |           |
##                |     0.019 |     0.001 |           |
## ---------------|-----------|-----------|-----------|
##            0.6 |       322 |        29 |       351 |
##                |     2.160 |    11.799 |           |
##                |     0.917 |     0.083 |     0.028 |
##                |     0.031 |     0.015 |           |
##                |     0.026 |     0.002 |           |
## ---------------|-----------|-----------|-----------|
##            0.8 |       314 |        11 |       325 |
##                |     5.620 |    30.698 |           |
##                |     0.966 |     0.034 |     0.026 |
##                |     0.030 |     0.006 |           |
##                |     0.025 |     0.001 |           |
## ---------------|-----------|-----------|-----------|
##              1 |       144 |        10 |       154 |
##                |     1.470 |     8.027 |           |
##                |     0.935 |     0.065 |     0.012 |
##                |     0.014 |     0.005 |           |
##                |     0.012 |     0.001 |           |
## ---------------|-----------|-----------|-----------|
##   Column Total |     10422 |      1908 |     12330 |
##                |     0.845 |     0.155 |           |
## ---------------|-----------|-----------|-----------|
##
##
```

```r
weekend_table <- table(DF2$Weekend, DF2$Revenue)
weekend_tab <- as.data.frame(prop.table(weekend_table, 2))
colnames(weekend_tab) <- c("Weekend", "Revenue", "Percent")
ggplot(data = weekend_tab, aes(x = Weekend, y = Percent, fill = Revenue)) +
geom_bar(stat = 'identity', position = 'dodge', alpha = 2/3) +
xlab("Weekend")+
ylab("Percent")
```




```
ggplot(data = DF2, mapping = aes(x = Revenue)) + geom_bar()
```




```
prop.table(table(DF2$Revenue))
```

```
##
##         0         1
## 0.8452555 0.1547445
```

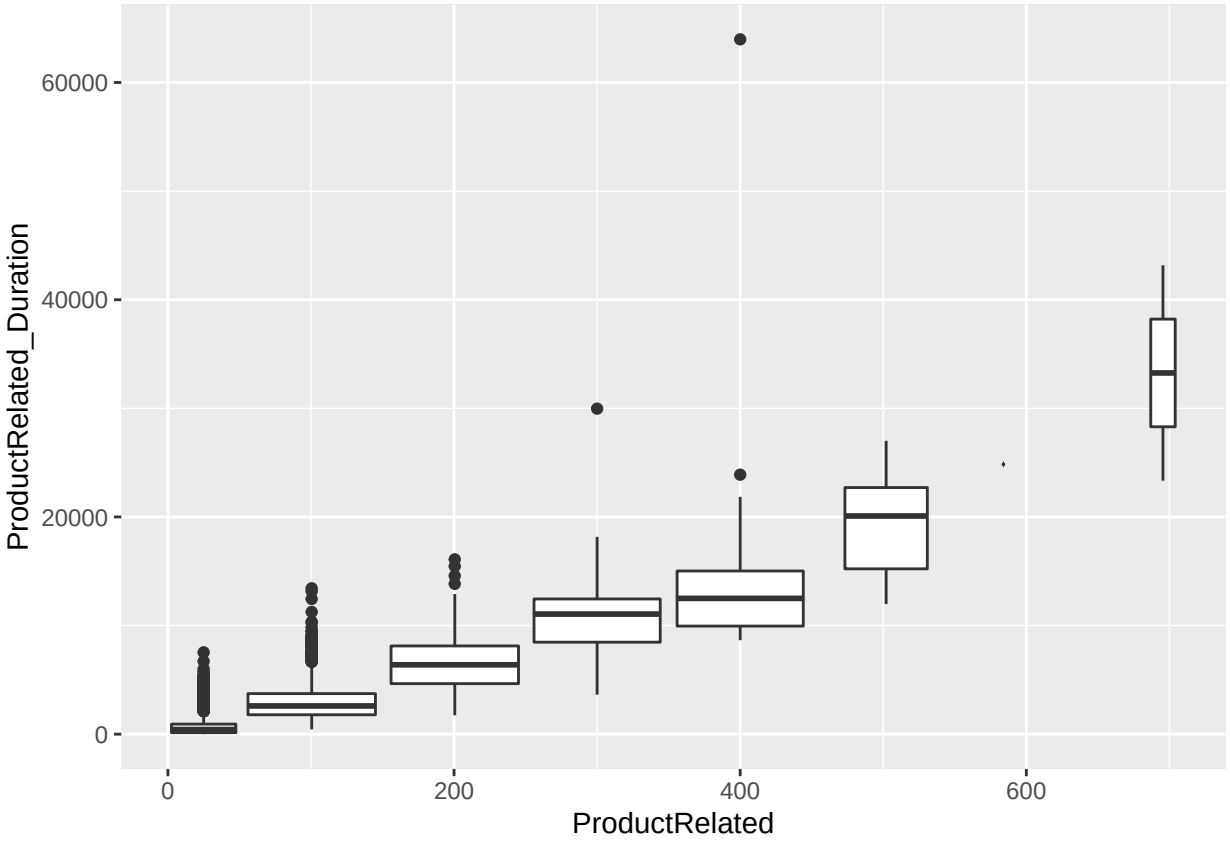
Check for missing data

```
r pMiss = function(x){sum(is.na(x))/length(x)*100}
apply(DF2,2,pMiss)
```

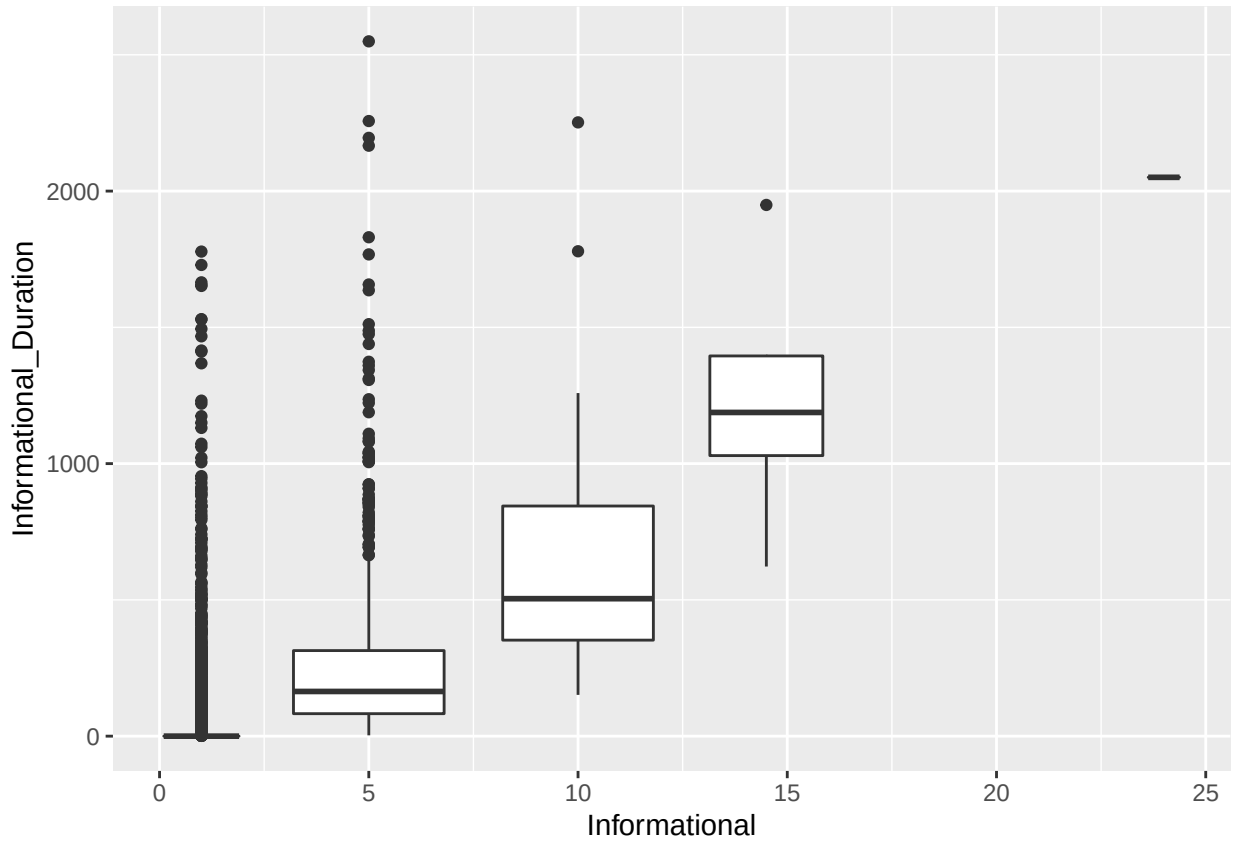
```
_______
##
Administrative
Administrative_Duration
Informational
##
0
0
0 ##
Informational_Duration
ProductRelated
ProductRelated_Duration
##
0
0
0 ##
BounceRates
ExitRates
PageValues
##
0
0
0 ##
SpecialDay
OperatingSystems
Browser
##
0
0
0 ##
Region
TrafficType
VisitorType
##
0
0
0 ##
Weekend
Revenue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

Check for outliers

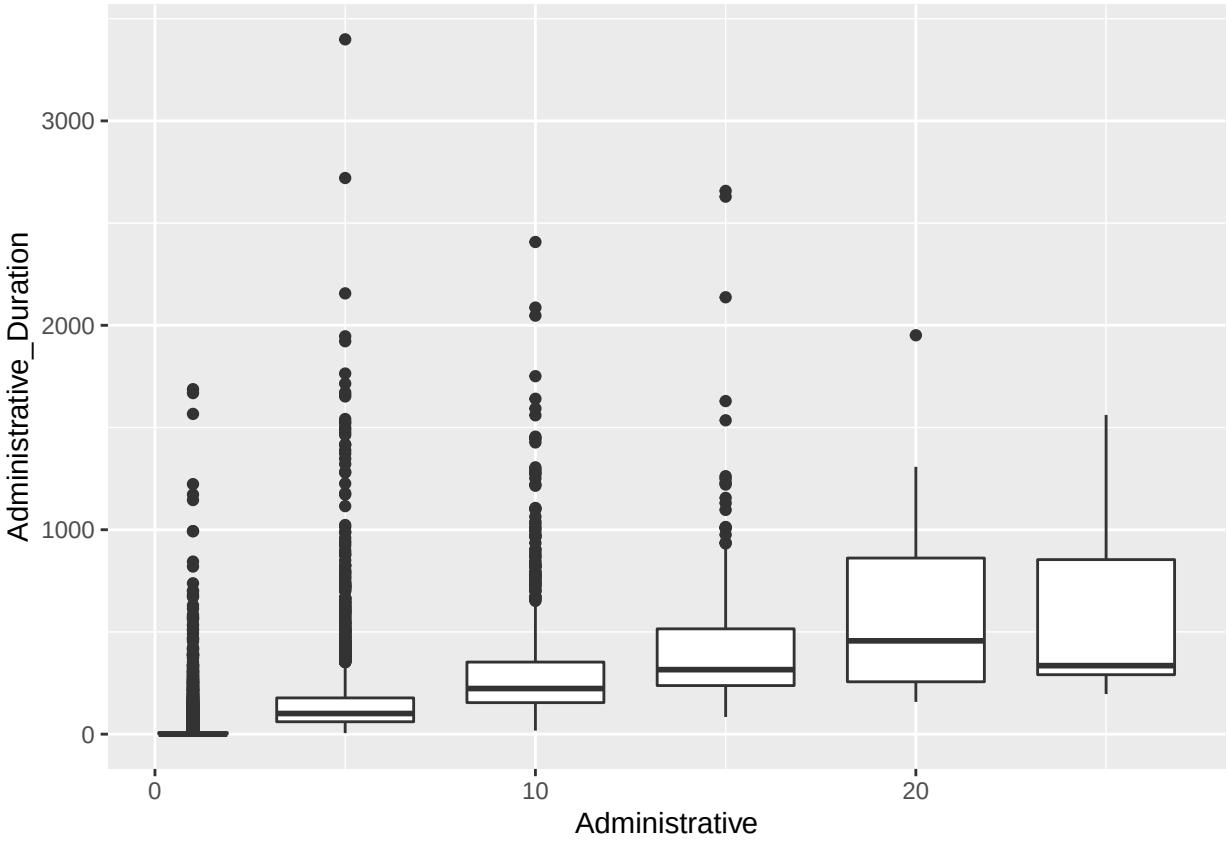
```
ggplot(DF2, aes(ProductRelated, ProductRelated_Duration))+
  geom_boxplot(aes(group=cut_width(ProductRelated,100)))
```




```
ggplot(DF2, aes(Informational, Informational_Duration))+
  geom_boxplot(aes(group=cut_width(Informational,5)))
```

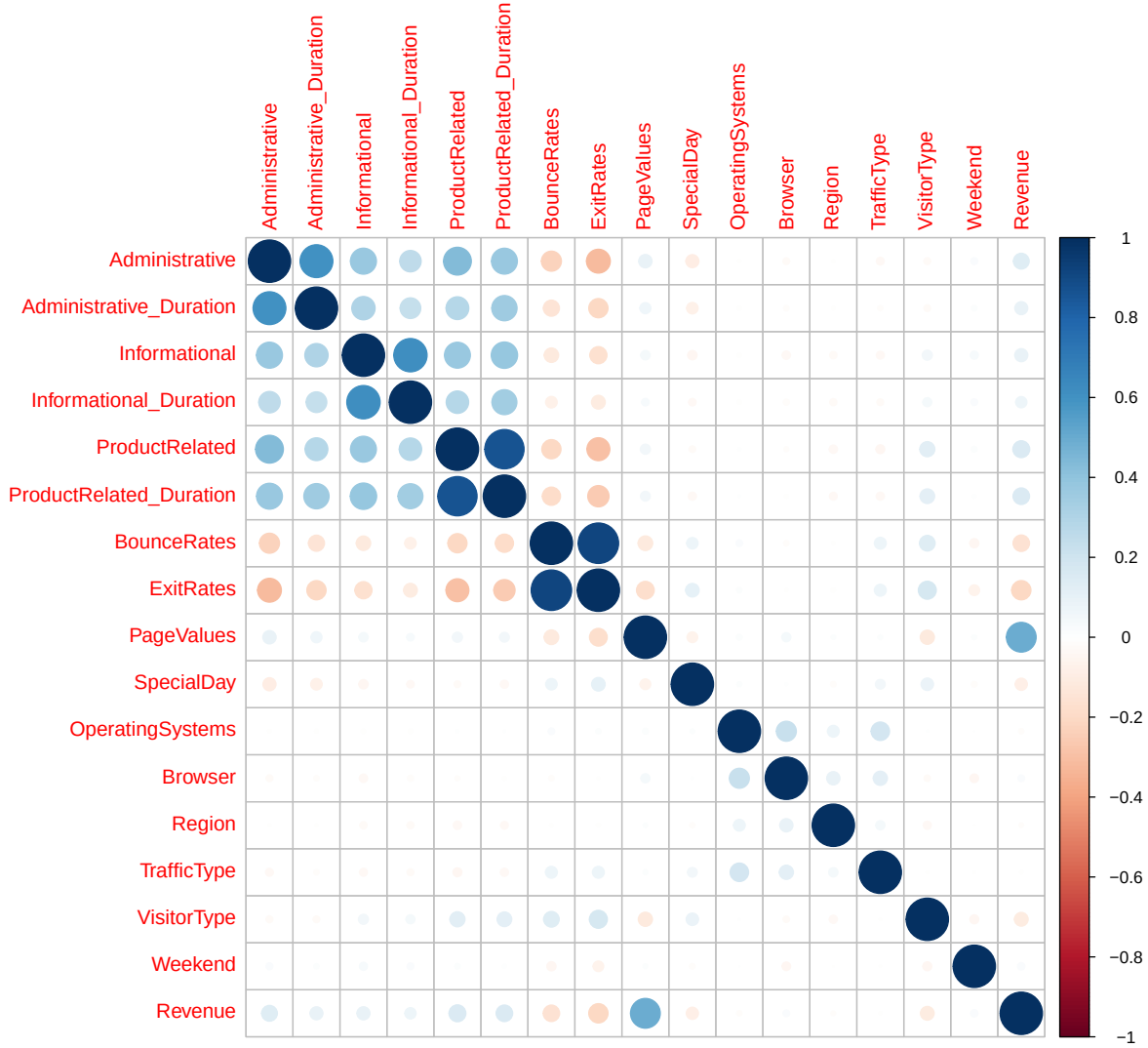



```
ggplot(DF2, aes(Administrative, Administrative_Duration))+
  geom_boxplot(aes(group=cut_width(Administrative,5)))
```



Correlation Analysis

```
cor.matrix <- cor(DF2, method = "pearson", use = "complete.obs")
corrplot(cor.matrix)
```




```
#Numeric features only
cor(DF2)
```

```
##                          Administrative Administrative_Duration Informational
## Administrative             1.000000000             0.601583342   0.376850429
## Administrative_Duration    0.601583342             1.000000000   0.302709709
## Informational              0.376850429             0.302709709   1.000000000
## Informational_Duration     0.255848140             0.238030789   0.618954862
## ProductRelated             0.431119340             0.289086621   0.374164291
## ProductRelated_Duration    0.373939013             0.355421954   0.387505306
## BounceRates               -0.223562630            -0.144170410  -0.116113616
## ExitRates                 -0.316482998            -0.205797757  -0.163666061
## PageValues                 0.098989585             0.067608481   0.048631692
## SpecialDay                -0.094777598            -0.073303725  -0.048219254
```

```
## OperatingSystems          -0.006347063            -0.007343418  -0.009526668
## Browser                   -0.025034572            -0.015391527  -0.038234678
## Region                    -0.005486805            -0.005560563  -0.029168638
## TrafficType               -0.033560713            -0.014376431  -0.034490754
## VisitorType               -0.025819710            -0.023939717   0.055827573
## Weekend                    0.026416750             0.014990142   0.035784725
## Revenue                    0.138917094             0.093586719   0.095200343
##                         Informational_Duration ProductRelated
## Administrative                     0.255848140    0.431119340
## Administrative_Duration            0.238030789    0.289086621
## Informational                      0.618954862    0.374164291
## Informational_Duration             1.000000000    0.280046268
## ProductRelated                     0.280046268    1.000000000
## ProductRelated_Duration            0.347363577    0.860926836
## BounceRates                       -0.074066610   -0.204577633
## ExitRates                         -0.105275683   -0.292526283
## PageValues                         0.030860874    0.056281794
## SpecialDay                        -0.030576549   -0.023958175
## OperatingSystems                  -0.009578676    0.004289621
## Browser                           -0.019284981   -0.013145721
## Region                            -0.027144112   -0.038121842
## TrafficType                       -0.024674908   -0.043064304
## VisitorType                        0.044676760    0.126655811
## Weekend                            0.024078486    0.016091964
## Revenue                            0.070344502    0.158537984
##                         ProductRelated_Duration  BounceRates    ExitRates
## Administrative                       0.373939013 -0.223562630 -0.316482998
## Administrative_Duration              0.355421954 -0.144170410 -0.205797757
## Informational                        0.387505306 -0.116113616 -0.163666061
## Informational_Duration               0.347363577 -0.074066610 -0.105275683
## ProductRelated                       0.860926836 -0.204577633 -0.292526283
## ProductRelated_Duration              1.000000000 -0.184541115 -0.251984097
## BounceRates                         -0.184541115  1.000000000  0.913004396
## ExitRates                           -0.251984097  0.913004396  1.000000000
## PageValues                           0.052823063 -0.119386026 -0.174498310
## SpecialDay                          -0.036379845  0.072702253  0.102241802
## OperatingSystems                     0.002975790  0.023823182  0.014566735
## Browser                             -0.007380440 -0.015772209 -0.004442355
## Region                              -0.033090520 -0.006485347 -0.008907006
## TrafficType                         -0.036377170  0.078285541  0.078616331
## VisitorType                          0.119329172  0.135536393  0.179143931
## Weekend                              0.007310614 -0.046513997 -0.062587048
## Revenue                              0.152372611 -0.150672912 -0.207071082
##                          PageValues   SpecialDay OperatingSystems      Browser
## Administrative           0.09898959 -0.094777598    -0.0063470633 -0.025034572
## Administrative_Duration  0.06760848 -0.073303725    -0.0073434175 -0.015391527
## Informational            0.04863169 -0.048219254    -0.0095266679 -0.038234678
## Informational_Duration   0.03086087 -0.030576549    -0.0095786764 -0.019284981
## ProductRelated           0.05628179 -0.023958175     0.0042896206 -0.013145721
## ProductRelated_Duration  0.05282306 -0.036379845     0.0029757898 -0.007380440
## BounceRates             -0.11938603  0.072702253     0.0238231825 -0.015772209
## ExitRates               -0.17449831  0.102241802     0.0145667353 -0.004442355
## PageValues               1.00000000 -0.063541272     0.0185079466  0.045591919
## SpecialDay              -0.06354127  1.000000000     0.0126522347  0.003498747
```

```
## OperatingSystems         0.01850795  0.012652235     1.0000000000  0.223012888
## Browser                  0.04559192  0.003498747     0.2230128882  1.000000000
## Region                   0.01131530 -0.016097975     0.0767754856  0.097392849
## TrafficType              0.01253169  0.052301443     0.1891536121  0.111938224
## VisitorType             -0.11122783  0.085556612     0.0015042220 -0.021866988
## Weekend                  0.01200164 -0.016767155     0.0002842506 -0.040260864
## Revenue                  0.49256930 -0.082304598    -0.0146675596  0.023984289
##                                Region  TrafficType  VisitorType       Weekend
## Administrative          -0.0054868053 -0.033560713 -0.025819710  0.0264167503
## Administrative_Duration -0.0055605628 -0.014376431 -0.023939717  0.0149901419
## Informational           -0.0291686379 -0.034490754  0.055827573  0.0357847251
## Informational_Duration  -0.0271441124 -0.024674908  0.044676760  0.0240784862
## ProductRelated          -0.0381218417 -0.043064304  0.126655811  0.0160919642
## ProductRelated_Duration -0.0330905198 -0.036377170  0.119329172  0.0073106138
## BounceRates             -0.0064853474  0.078285541  0.135536393 -0.0465139965
## ExitRates               -0.0089070060  0.078616331  0.179143931 -0.0625870480
## PageValues               0.0113152995  0.012531693 -0.111227826  0.0120016392
## SpecialDay              -0.0160979746  0.052301443  0.085556612 -0.0167671553
## OperatingSystems         0.0767754856  0.189153612  0.001504222  0.0002842506
## Browser                  0.0973928492  0.111938224 -0.021866988 -0.0402608638
## Region                   1.0000000000  0.047520231 -0.036190794 -0.0006906703
## TrafficType              0.0475202313  1.000000000 -0.002839178 -0.0022212292
## VisitorType             -0.0361907939 -0.002839178  1.000000000 -0.0436792493
## Weekend                 -0.0006906703 -0.002221229 -0.043679249  1.0000000000
## Revenue                 -0.0115950678 -0.005112971 -0.104725722  0.0292953680
##                              Revenue
## Administrative           0.138917094
## Administrative_Duration  0.093586719
## Informational            0.095200343
## Informational_Duration   0.070344502
## ProductRelated           0.158537984
## ProductRelated_Duration  0.152372611
## BounceRates             -0.150672912
## ExitRates               -0.207071082
## PageValues               0.492569295
## SpecialDay              -0.082304598
## OperatingSystems        -0.014667560
## Browser                  0.023984289
## Region                  -0.011595068
## TrafficType             -0.005112971
## VisitorType             -0.104725722
## Weekend                  0.029295368
## Revenue                  1.000000000
```

Creating Training and Test Data

```
# Create Training Data
input_ones <- DF2[which(DF2$Revenue == 1), ]  # all 1's
input_zeros <- DF2[which(DF2$Revenue == 0), ]  # all 0's
set.seed(100)  # for repeatability of samples
input_ones_training_rows <- sample(1:nrow(input_ones), 0.7*nrow(input_ones))  # 1's for training
```

```
input_zeros_training_rows <- sample(1:nrow(input_zeros), 0.7*nrow(input_ones))  # 0's for training. Pic
training_ones <- input_ones[input_ones_training_rows, ]
training_zeros <- input_zeros[input_zeros_training_rows, ]
trainingData <- rbind(training_ones, training_zeros)  # row bind the 1's and 0's

# Create Test Data
test_ones <- input_ones[-input_ones_training_rows, ]
test_zeros <- input_zeros[-input_zeros_training_rows, ]
testData <- rbind(test_ones, test_zeros)  # row bind the 1's and 0's

#REF: http://r-statistics.co/Logistic-Regression-With-R.html
```

```
head(trainingData)
```

```
##       Administrative Administrative_Duration Informational
## 11444              0                   0.00             0
## 4954               7                  67.00             2
## 9693               1                 113.75             0
## 11008              0                   0.00             0
## 7818               0                   0.00             0
## 11456              0                   0.00             0
##       Informational_Duration ProductRelated ProductRelated_Duration BounceRates
## 11444                    0.0             19                577.5417 0.000000000
## 4954                   725.5            101               4713.8000 0.009444444
## 9693                     0.0             26               1905.0786 0.022666667
## 11008                    0.0             39               1262.6515 0.005128205
## 7818                     0.0              1                 15.6000 0.000000000
## 11456                    0.0             23                662.5000 0.000000000
##         ExitRates PageValues SpecialDay OperatingSystems Browser Region
## 11444 0.011764706   90.92852          0                2       2      1
## 4954  0.010481956   20.62474          0                1       1      6
## 9693  0.034600000   11.70615          0                3       2      2
## 11008 0.007264957    0.00000          0                3       2      2
## 7818  0.000000000   47.79177          0                1       1      6
## 11456 0.013043478    0.00000          0                2       2      6
##       TrafficType VisitorType Weekend Revenue
## 11444           2           3       0       1
## 4954            3           3       1       1
## 9693            1           3       0       1
## 11008          10           3       0       1
## 7818            3           1       0       1
## 11456           2           1       1       1
```

Dimension Reduction

PCA

```
head(DF2)
```

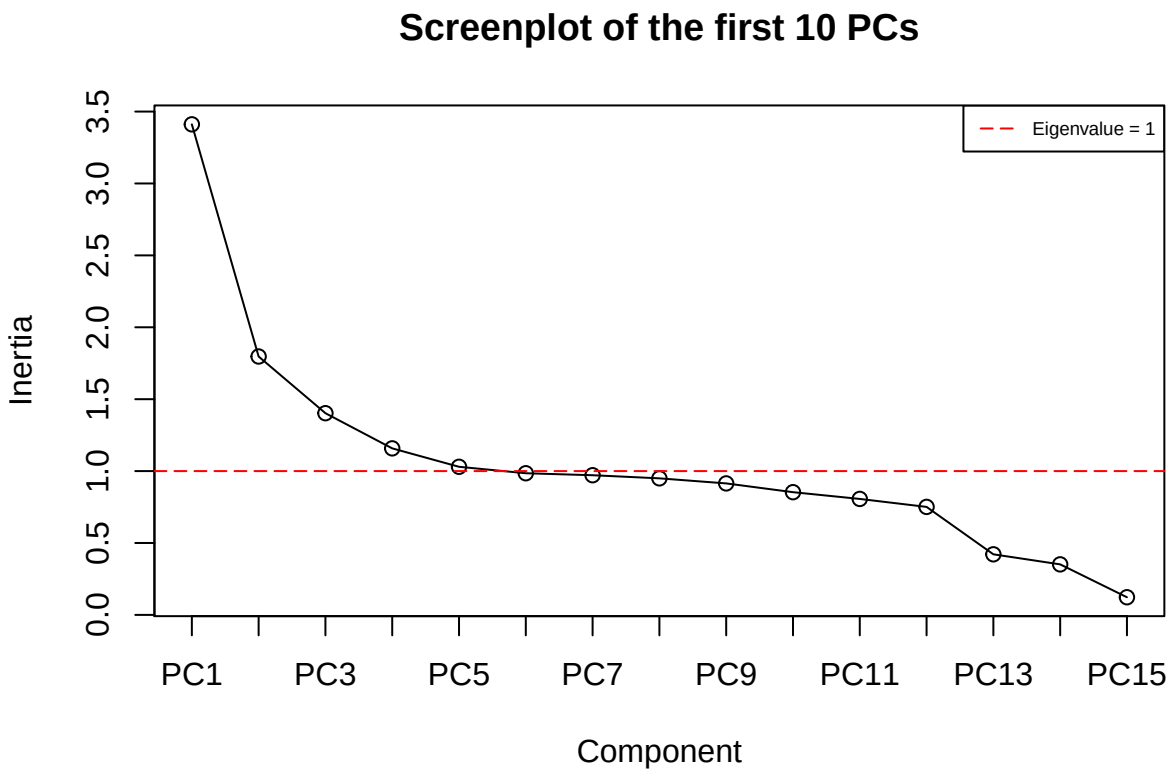
```
##   Administrative Administrative_Duration Informational Informational_Duration
```

```
## 1              0                      0            0                     0
## 2              0                      0            0                     0
## 3              0                      0            0                     0
## 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                0.000000  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 OperatingSystems Browser Region TrafficType VisitorType Weekend
## 1          0                1       1      1           1           3       0
## 2          0                2       2      1           2           3       0
## 3          0                4       1      9           3           3       0
## 4          0                3       2      2           4           3       0
## 5          0                3       3      1           4           3       1
## 6          0                2       2      1           3           3       0
##   Revenue
## 1       0
## 2       0
## 3       0
## 4       0
## 5       0
## 6       0
```

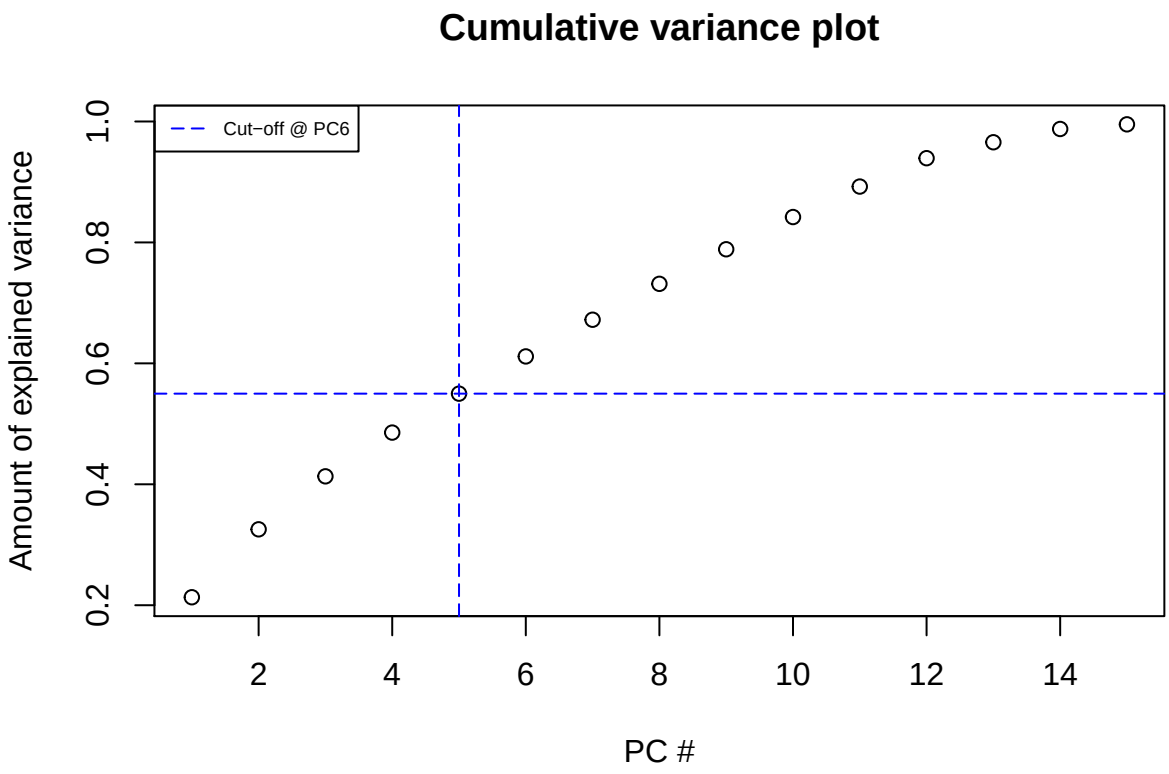
```
DF2.pca<-prcomp(DF2[c(1:16)], center=TRUE, scale=TRUE)
summary(DF2.pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6    PC7
## Standard deviation     1.8469 1.3405 1.18427 1.07608 1.01483 0.99243 0.9855
## Proportion of Variance 0.2132 0.1123 0.08766 0.07237 0.06437 0.06156 0.0607
## Cumulative Proportion  0.2132 0.3255 0.41315 0.48552 0.54989 0.61145 0.6722
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.97448 0.95616 0.92371 0.89795 0.86655 0.64898 0.59280
## Proportion of Variance 0.05935 0.05714 0.05333 0.05039 0.04693 0.02632 0.02196
## Cumulative Proportion  0.73150 0.78864 0.84197 0.89236 0.93929 0.96562 0.98758
##                           PC15    PC16
## Standard deviation     0.35041 0.27557
## Proportion of Variance 0.00767 0.00475
## Cumulative Proportion  0.99525 1.00000
```

```
screeplot(DF2.pca, type="l", npcs =15, main="Screenplot of the first 10 PCs")
abline(h=1, col="red", lty=5)
legend("topright", legend=c("Eigenvalue = 1"),
       col=c("red"), lty=5, cex=0.6)
```

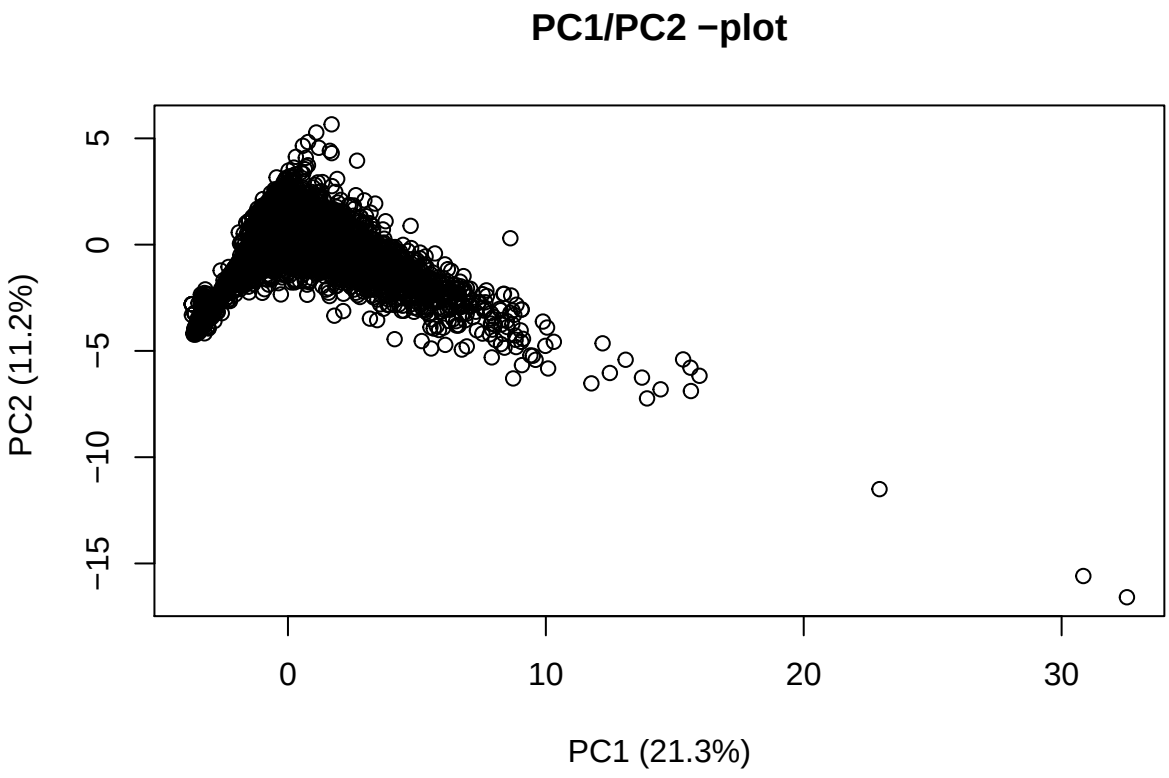



```
cumpro<-cumsum(DF2.pca$sdev^2/sum(DF2.pca$sdev^2))
plot(cumpro[0:15], xlab = "PC #", ylab = "Amount of explained variance", main = "Cumulative variance pl
abline(v=5, col="blue", lty=5)
abline(h=0.54989, col="blue", lty=5)
legend("topleft", legend=c("Cut-off @ PC6"),
       col=c("blue"), lty=5, cex=0.6)
```

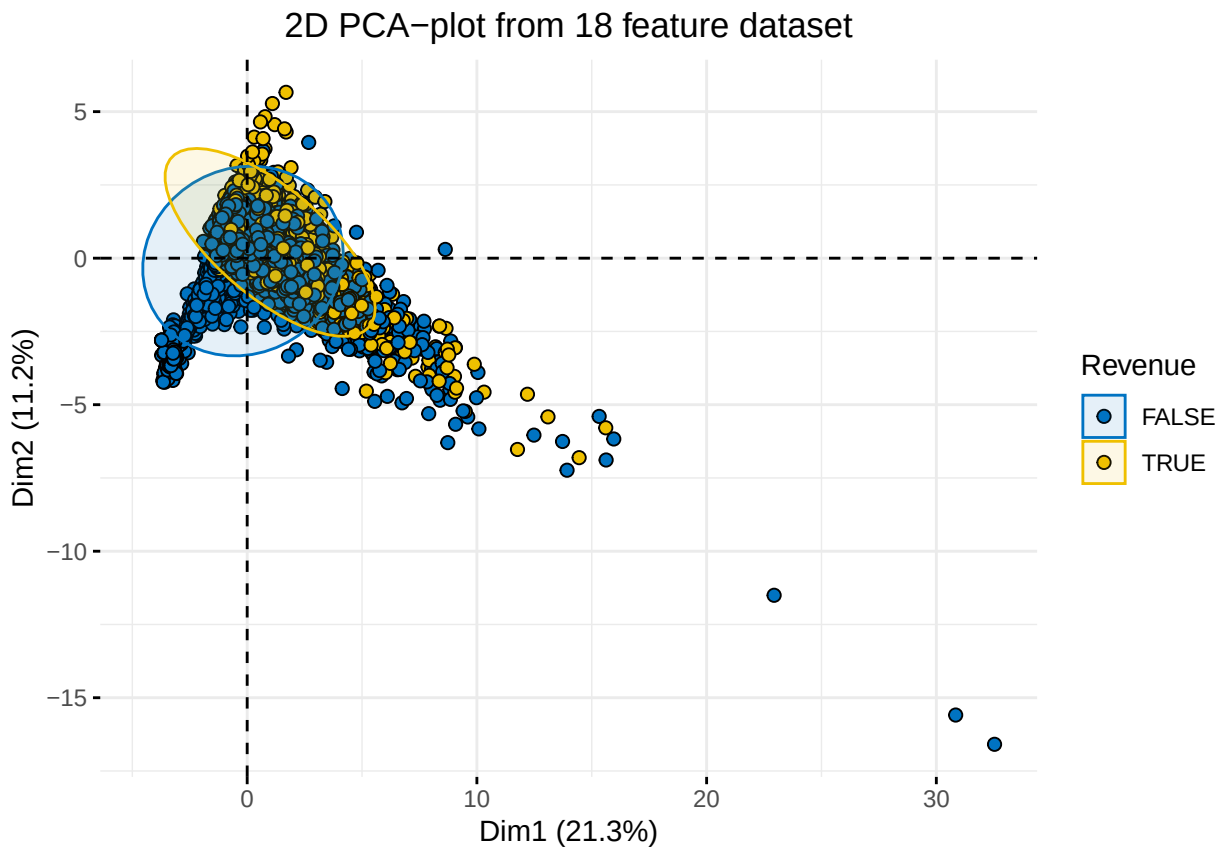


#When the PCs that have a eigenvalue < 1 are discarded we are left with a cumulative variance 55%. Not a #very reassuring number to explain the data set.

```
plot(DF2.pca$x[,1],DF2.pca$x[,2], xlab="PC1 (21.3%)", ylab="PC2 (11.2%) ", main = "PC1/PC2 -plot")
```




```
fviz_pca_ind(DF2.pca, geom.ind = "point", pointshape = 21,
             pointsize=2,
             fill.ind=DF1$Revenue,
             col.ind ="black",
             palette="jco",
             addEllipses = TRUE,
             label = "var",
             col.var="black",
             repel = TRUE,
             legend.title = "Revenue")+
  ggtitle("2D PCA-plot from 18 feature dataset")+
  theme(plot.title= element_text(hjust=0.5))
```



Classification

Logistic Regression

```
# Logistics Regression
glm.fit <- glm(Revenue ~ Administrative + Administrative_Duration + Informational + Informational_Durati
```

```
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Revenue ~ Administrative + Administrative_Duration +
##     Informational + Informational_Duration + ProductRelated +
##     ProductRelated_Duration + BounceRates + ExitRates + PageValues +
##     SpecialDay + OperatingSystems + Browser + Region + TrafficType +
##     VisitorType + Weekend, family = binomial(link = "logit"),
##     data = trainingData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7990  -0.7790  -0.0685   0.6928   2.7585
##
```

```
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.723e-01  2.557e-01  -2.629  0.00857 **
## Administrative            3.390e-02  1.897e-02   1.787  0.07399 .
## Administrative_Duration -4.102e-04  3.468e-04  -1.183  0.23688
## Informational             5.601e-02  4.727e-02   1.185  0.23599
## Informational_Duration  -1.023e-06  4.312e-04  -0.002  0.99811
## ProductRelated            1.350e-03  2.225e-03   0.606  0.54419
## ProductRelated_Duration  1.832e-04  5.624e-05   3.258  0.00112 **
## BounceRates              -3.530e+00  4.120e+00  -0.857  0.39156
## ExitRates                -1.049e+01  3.507e+00  -2.992  0.00277 **
## PageValues                1.103e-01  6.189e-03  17.824  < 2e-16 ***
## SpecialDay               -1.030e+00  3.157e-01  -3.262  0.00111 **
## OperatingSystems         -1.381e-02  5.875e-02  -0.235  0.81422
## Browser                   1.002e-02  3.060e-02   0.327  0.74338
## Region                   -5.756e-03  2.101e-02  -0.274  0.78406
## TrafficType               1.285e-03  1.329e-02   0.097  0.92300
## VisitorType              -9.871e-02  7.163e-02  -1.378  0.16819
## Weekend                   1.451e-01  1.163e-01   1.248  0.21215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3701.4  on 2669  degrees of freedom
## Residual deviance: 2402.1  on 2653  degrees of freedom
## AIC: 2436.1
##
## Number of Fisher Scoring iterations: 7
```

```
glm.probs<- predict(glm.fit, type="response")
glm.probs[1:5]
```

```
##     11444      4954      9693     11008      7818
## 0.9998838 0.9305577 0.5546925 0.3102787 0.9886937
```

```
glm.pred<-ifelse(glm.probs>0.5,1,0)
attach(trainingData)
table(glm.pred,Revenue)
```

```
##         Revenue
## glm.pred    0    1
##        0 1203  327
##        1  132 1008
```

```
predicted <- plogis(predict(glm.fit, testData))

library(InformationValue)
optCutOff <- optimalCutoff(testData$Revenue, predicted)[1]
optCutOff
```

```
## [1] 0.98
```

```
sensitivity(testData$Revenue, predicted, threshold = optCutOff)
```

```
## [1] 0.2303665
```

```
specificity(testData$Revenue, predicted, threshold = optCutOff)
```

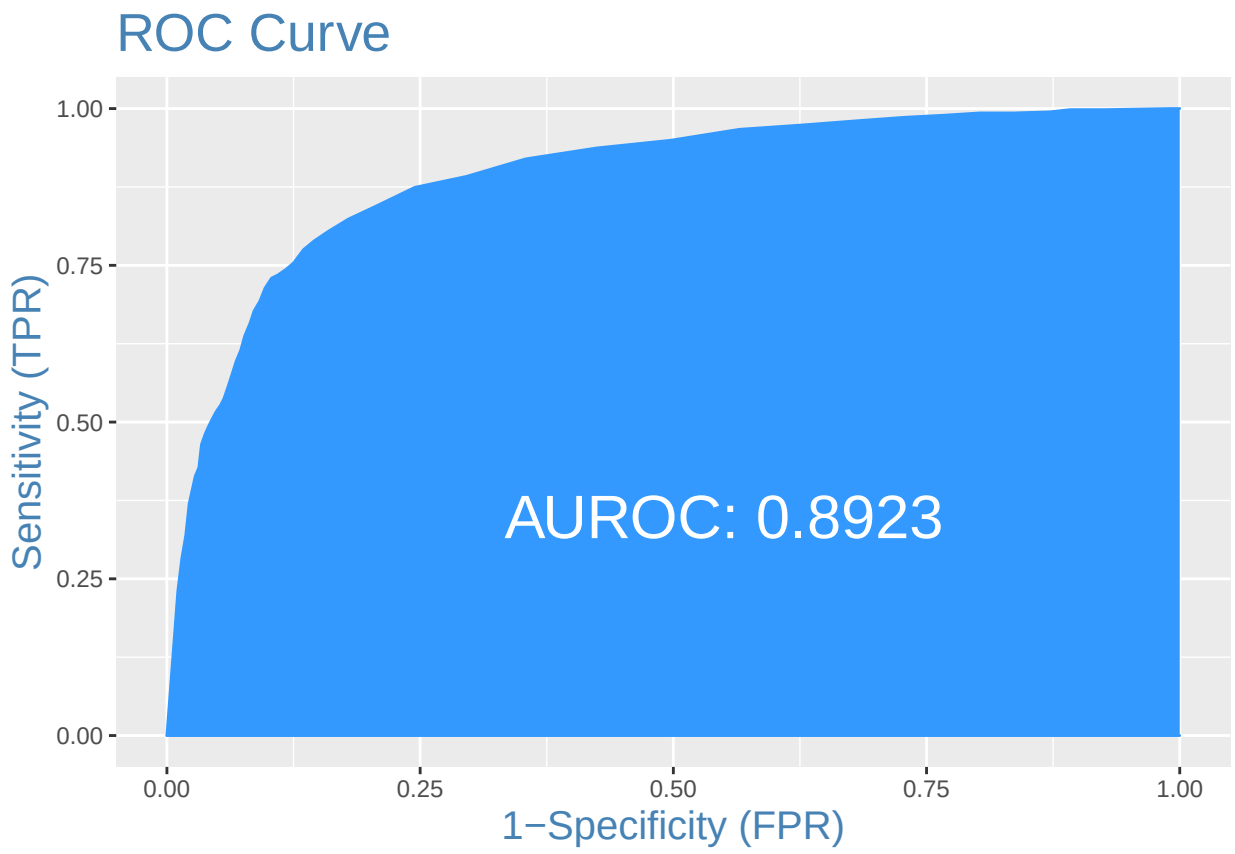
```
## [1] 0.9892154
```

```
CM = confusionMatrix(testData$Revenue, predicted, threshold = optCutOff)
CM
```

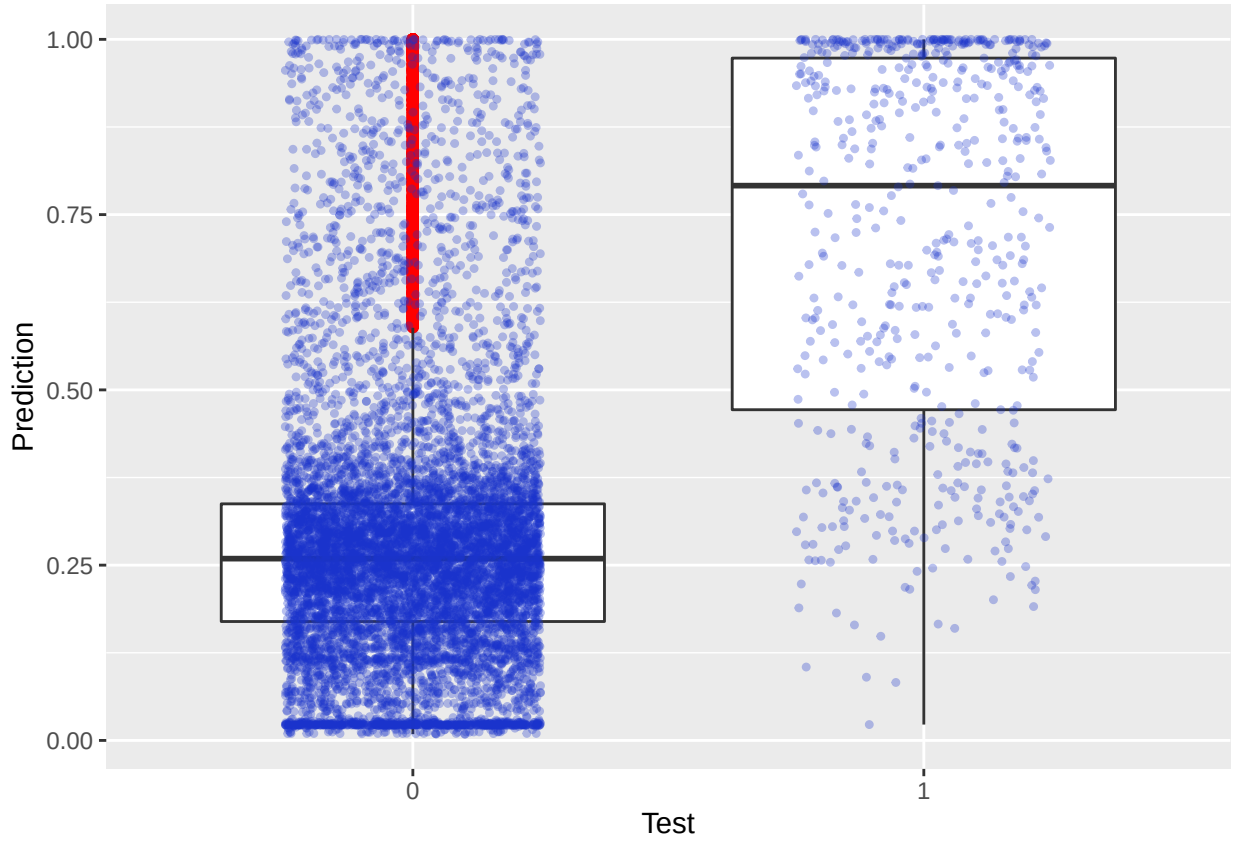
```
##      0   1
## 0 8989 441
## 1   98 132
```

```
##          Administrative Administrative_Duration           Informational
##                1.822938                1.621724                1.776756
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                1.641124                4.428455                4.304519
##             BounceRates               ExitRates             PageValues
##                2.520927                2.813807                1.045826
##              SpecialDay        OperatingSystems                 Browser
##                1.023848                1.040524                1.041691
##                  Region             TrafficType             VisitorType
##                1.017883                1.034015                1.132343
##                 Weekend
##                1.008243
```

```
plotROC(testData$Revenue, predicted)
```




```
df4 = data.frame(as.factor(testData$Revenue), predicted)
colnames(df4) <- c("Test","Prediction")
ggplot(df4, aes(x = Test, y = Prediction)) +
geom_boxplot(outlier.colour = "red") +
geom_jitter(width = 0.25, pch=20, col=rgb(0.1, 0.2, 0.8, 0.3))
```




```
confint(glm.fit)
```

```
## Waiting for profiling to be done...
```

```
##                                 2.5 %        97.5 %
## (Intercept)             -1.175058e+00 -0.1719033763
## Administrative          -3.179051e-03  0.0712656114
## Administrative_Duration -1.104218e-03  0.0002584491
## Informational           -3.647418e-02  0.1491767172
## Informational_Duration  -8.318361e-04  0.0008689682
## ProductRelated          -3.003438e-03  0.0057306448
## ProductRelated_Duration  7.380127e-05  0.0002945067
## BounceRates             -1.176832e+01  4.4497993839
## ExitRates               -1.754647e+01 -3.7967152765
## PageValues               9.856520e-02  0.1228360104
## SpecialDay              -1.664972e+00 -0.4247490027
## OperatingSystems        -1.299755e-01  0.1006274518
## Browser                 -5.028288e-02  0.0698140792
## Region                  -4.704865e-02  0.0353358985
## TrafficType             -2.490159e-02  0.0272393124
## VisitorType             -2.388193e-01  0.0421203003
## Weekend                 -8.334827e-02  0.3728058388
```

```
exp(coef(glm.fit))
```

```
##             (Intercept)          Administrative Administrative_Duration
##            5.105473e-01            1.034483e+00            9.995899e-01
##           Informational  Informational_Duration          ProductRelated
##            1.057613e+00            9.999990e-01            1.001350e+00
## ProductRelated_Duration             BounceRates               ExitRates
##            1.000183e+00            2.931501e-02            2.768541e-05
##              PageValues             SpecialDay        OperatingSystems
##            1.116633e+00            3.571662e-01            9.862893e-01
##                 Browser                  Region             TrafficType
##            1.010068e+00            9.942601e-01            1.001285e+00
##             VisitorType                 Weekend
##            9.060039e-01            1.156187e+00
```

```
## odds ratios and 95% CI
exp(cbind(OR = coef(glm.fit), confint(glm.fit)))
```

```
## Waiting for profiling to be done...
```

```
##                                   OR        2.5 %      97.5 %
## (Intercept)             5.105473e-01 3.088010e-01  0.84206053
## Administrative          1.034483e+00 9.968260e-01  1.07386642
## Administrative_Duration 9.995899e-01 9.988964e-01  1.00025848
## Informational           1.057613e+00 9.641830e-01  1.16087812
## Informational_Duration  9.999990e-01 9.991685e-01  1.00086935
## ProductRelated          1.001350e+00 9.970011e-01  1.00574710
## ProductRelated_Duration 1.000183e+00 1.000074e+00  1.00029455
## BounceRates             2.931501e-02 7.746106e-06 85.60976758
## ExitRates               2.768541e-05 2.396986e-08  0.02244437
## PageValues              1.116633e+00 1.103586e+00  1.13069898
## SpecialDay              3.571662e-01 1.891959e-01  0.65393390
## OperatingSystems        9.862893e-01 8.781170e-01  1.10586458
## Browser                 1.010068e+00 9.509604e-01  1.07230880
## Region                  9.942601e-01 9.540410e-01  1.03596763
## TrafficType             1.001285e+00 9.754059e-01  1.02761369
## VisitorType             9.060039e-01 7.875572e-01  1.04301995
## Weekend                 1.156187e+00 9.200307e-01  1.45180243
```

```
preproc1 <- preProcess(DF2, method=c("center", "scale"))

norm1 <- predict(preproc1, DF2)

summary(norm1)
```

```
##  Administrative    Administrative_Duration Informational
##  Min.   :-0.6970   Min.   :-0.45717        Min.   :-0.3965
##  1st Qu.:-0.6970   1st Qu.:-0.45717        1st Qu.:-0.3965
##  Median :-0.3959   Median :-0.41475        Median :-0.3965
##  Mean   : 0.0000   Mean   : 0.00000        Mean   : 0.0000
##  3rd Qu.: 0.5072   3rd Qu.: 0.07036        3rd Qu.:-0.3965
```

```
##  Max.   : 7.4312   Max.   :18.76880        Max.   :18.4988
##  Informational_Duration ProductRelated    ProductRelated_Duration
##  Min.   :-0.2449        Min.   :-0.7135   Min.   :-0.6243
##  1st Qu.:-0.2449        1st Qu.:-0.5561   1st Qu.:-0.5281
##  Median :-0.2449        Median :-0.3087   Median :-0.3113
##  Mean   : 0.0000        Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.2449        3rd Qu.: 0.1409   3rd Qu.: 0.1408
##  Max.   :17.8680        Max.   :15.1380   Max.   :32.8055
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :-0.4577   Min.   :-0.8863   Min.   :-0.3172   Min.   :-0.3088
##  1st Qu.:-0.4577   1st Qu.:-0.5924   1st Qu.:-0.3172   1st Qu.:-0.3088
##  Median :-0.3935   Median :-0.3687   Median :-0.3172   Median :-0.3088
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.1109   3rd Qu.: 0.1425   3rd Qu.:-0.3172   3rd Qu.:-0.3088
##  Max.   : 3.6670   Max.   : 3.2292   Max.   :19.1656   Max.   : 4.7184
##  OperatingSystems     Browser            Region          TrafficType
##  Min.   :-1.2334   Min.   :-0.7903   Min.   :-0.89414   Min.   :-0.76260
##  1st Qu.:-0.1361   1st Qu.:-0.2079   1st Qu.:-0.89414   1st Qu.:-0.51416
##  Median :-0.1361   Median :-0.2079   Median :-0.06136   Median :-0.51416
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.9612   3rd Qu.:-0.2079   3rd Qu.: 0.35503   3rd Qu.:-0.01729
##  Max.   : 6.4477   Max.   : 6.1975   Max.   : 2.43698   Max.   : 3.95770
##   VisitorType         Weekend           Revenue
##  Min.   :-2.4876   Min.   :-0.5505   Min.   :-0.4279
##  1st Qu.: 0.4078   1st Qu.:-0.5505   1st Qu.:-0.4279
##  Median : 0.4078   Median :-0.5505   Median :-0.4279
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4078   3rd Qu.:-0.5505   3rd Qu.:-0.4279
##  Max.   : 0.4078   Max.   : 1.8163   Max.   : 2.3371
```

bias the model

now let's try to balance the data by minmizing the zeros of the training set

```
# Create Training Data
input_ones <- DF2[which(DF2$Revenue == 1), ]  # all 1's

input_zeros <- DF2[which(DF2$Revenue == 0), ]  # all 0's

set.seed(100)  # for repeatability of samples
input_ones_training_rows <- sample(1:nrow(input_ones), 0.7*nrow(input_ones))  # 1's for training
input_zeros_training_rows_half <- sample(1:nrow(input_zeros), 0.35*nrow(input_ones))  # 0's for trainin
training_ones <- input_ones[input_ones_training_rows, ]
training_zeros_Half <- input_zeros[input_zeros_training_rows_half, ]
trainingData_Half <- rbind(training_ones, training_zeros_Half)  # row bind the 1's and 0's

# Create Test Data
test_ones <- input_ones[-input_ones_training_rows, ]
test_zeros_Half <- input_zeros[-input_zeros_training_rows_half, ]
testData_Half <- rbind(test_ones, test_zeros_Half)  # row bind the 1's and 0's

#REF: http://r-statistics.co/Logistic-Regression-With-R.html
```

Biased Logistics Regression (Balanced Training Data)

```
# Logistics Regression
glm.fit_Half <- glm(Revenue ~ Administrative + Administrative_Duration + Informational + Informational_
```

```
predicted_Half <- plogis(predict(glm.fit, testData_Half))

library(InformationValue)
optCutOff_Half <- optimalCutoff(testData_Half$Revenue, predicted)[1]
optCutOff
```

```
## [1] 0.98
```

```
sensitivity(testData_Half$Revenue, predicted_Half, threshold = optCutOff_Half)
```

```
## [1] 0.001745201
```

```
specificity(testData_Half$Revenue, predicted_Half, threshold = optCutOff_Half)
```

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

```
CM_Half =confusionMatrix(testData_Half$Revenue, predicted_Half, threshold = optCutOff_Half)
CM_Half
```

```
##      0   1
## 0 9755 572
## 1    0   1
```

```
#FIrst CM
print(CM )
```

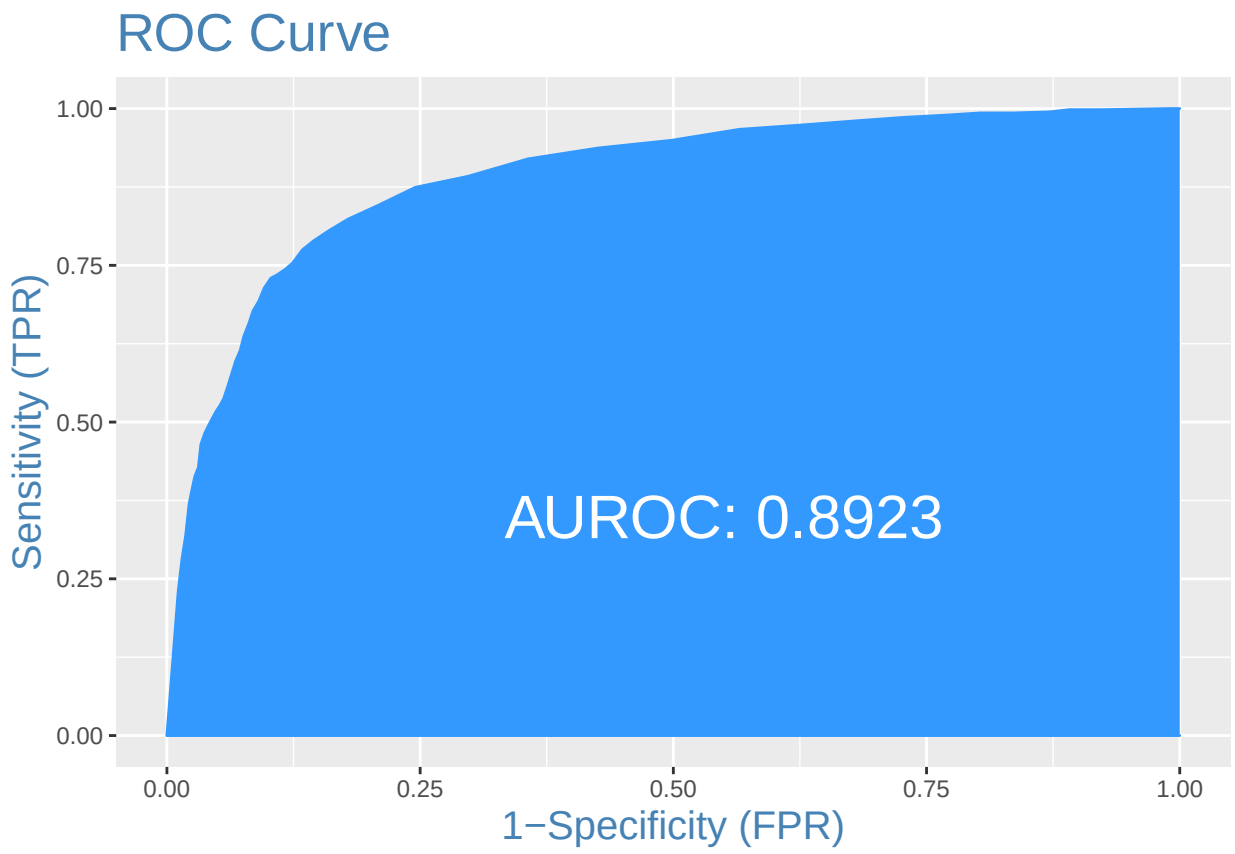
```
##      0   1
## 0 8989 441
## 1   98 132
```

```
summary(glm.fit_Half)
```

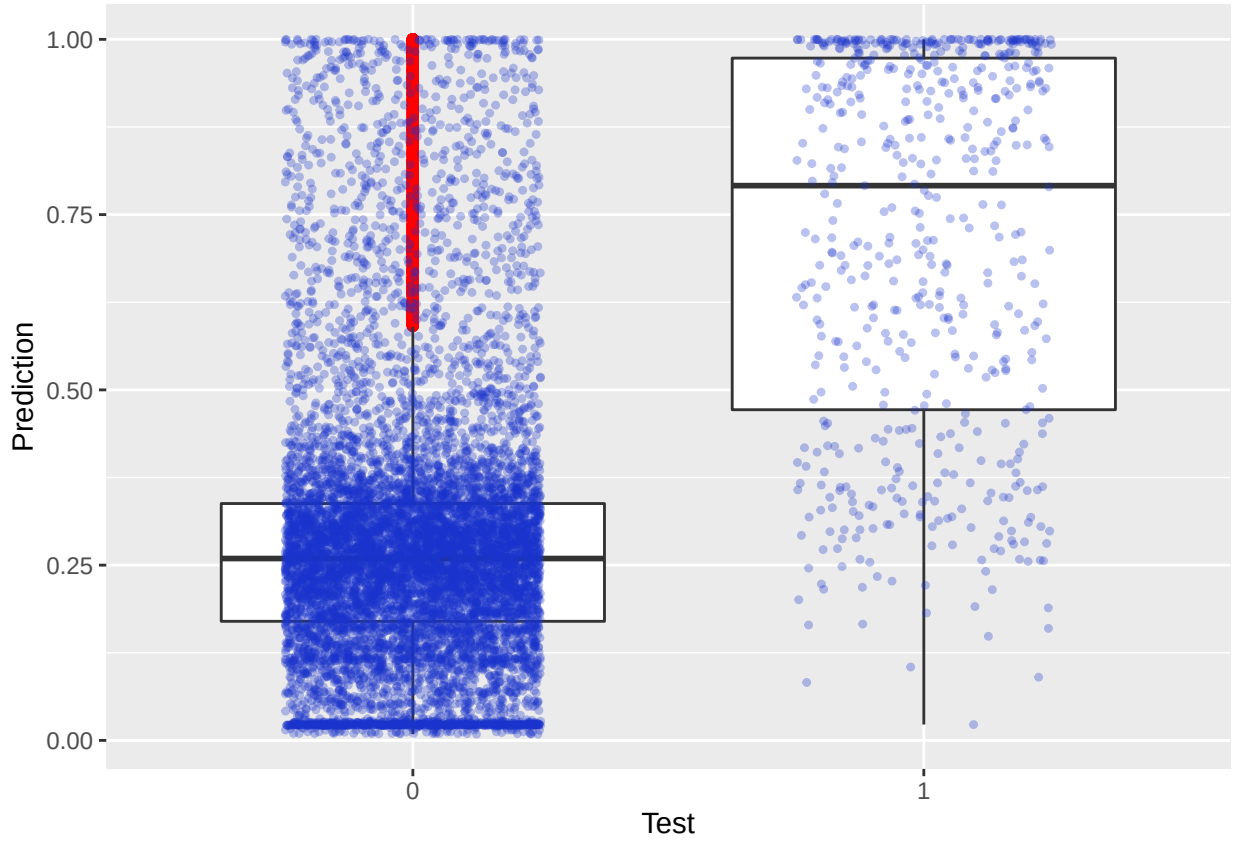
```
##
## Call:
## glm(formula = Revenue ~ Administrative + Administrative_Duration +
##     Informational + Informational_Duration + ProductRelated +
##     ProductRelated_Duration + BounceRates + ExitRates + PageValues +
##     SpecialDay + OperatingSystems + Browser + Region + TrafficType +
##     VisitorType + Weekend, family = "binomial", data = trainingData_Half)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.0534  -0.8000   0.1475   0.6733   2.5272
```

```
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.1147605  0.3068514  -0.374   0.7084
## Administrative            0.0336951  0.0245609   1.372   0.1701
## Administrative_Duration  -0.0003363  0.0004471  -0.752   0.4520
## Informational             0.1264286  0.0640812   1.973   0.0485 *
## Informational_Duration   -0.0003108  0.0005525  -0.562   0.5738
## ProductRelated            0.0061021  0.0030889   1.976   0.0482 *
## ProductRelated_Duration   0.0001459  0.0000738   1.977   0.0480 *
## BounceRates              -7.6806237  4.6157768  -1.664   0.0961 .
## ExitRates                -5.4810975  3.9868023  -1.375   0.1692
## PageValues                0.1162501  0.0082113  14.157   <2e-16 ***
## SpecialDay               -0.8604806  0.3733133  -2.305   0.0212 *
## OperatingSystems         -0.0061384  0.0684751  -0.090   0.9286
## Browser                   0.0018834  0.0370022   0.051   0.9594
## Region                   -0.0331115  0.0249131  -1.329   0.1838
## TrafficType               0.0047324  0.0157099   0.301   0.7632
## VisitorType              -0.1423616  0.0855985  -1.663   0.0963 .
## Weekend                   0.1402929  0.1401397   1.001   0.3168
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2548.1  on 2001  degrees of freedom
## Residual deviance: 1657.6  on 1985  degrees of freedom
## AIC: 1691.6
##
## Number of Fisher Scoring iterations: 7

##          Administrative Administrative_Duration           Informational
##                1.900602                1.705712                1.750472
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                1.614359                3.816092                3.625127
##             BounceRates               ExitRates              PageValues
##                2.610819                2.940988                1.054255
##              SpecialDay        OperatingSystems                 Browser
##                1.031540                1.044147                1.051313
##                  Region             TrafficType             VisitorType
##                1.020869                1.033556                1.138404
##                 Weekend
##                1.009583
```

```
plotROC(testData_Half$Revenue, predicted_Half)
```




```
df5 = data.frame(as.factor(testData_Half$Revenue), predicted_Half)
colnames(df5) <- c("Test","Prediction")
ggplot(df5, aes(x = Test, y = Prediction)) +
geom_boxplot(outlier.colour = "red") +
geom_jitter(width = 0.25, pch=20, col=rgb(0.1, 0.2, 0.8, 0.3))
```




```
confint(glm.fit_Half)
```

```
## Waiting for profiling to be done...
```

```
##                                 2.5 %        97.5 %
## (Intercept)             -7.159647e-01  0.4881370328
## Administrative          -1.407080e-02  0.0823239864
## Administrative_Duration -1.216077e-03  0.0005449216
## Informational            3.130551e-03  0.2549334015
## Informational_Duration  -1.353997e-03  0.0008364809
## ProductRelated           1.191976e-04  0.0122479042
## ProductRelated_Duration  2.825008e-06  0.0002925571
## BounceRates             -1.691379e+01  1.2710834804
## ExitRates               -1.350103e+01  2.1506502722
## PageValues               1.008064e-01  0.1332032082
## SpecialDay              -1.604991e+00 -0.1381909620
## OperatingSystems        -1.411308e-01  0.1278831556
## Browser                 -7.023998e-02  0.0750908761
## Region                  -8.198166e-02  0.0157412333
## TrafficType             -2.600299e-02  0.0356488075
## VisitorType             -3.107803e-01  0.0250083096
## Weekend                 -1.338052e-01  0.4158693455
```

```
exp(coef(glm.fit_Half))
```

```
##             (Intercept)          Administrative Administrative_Duration
##            0.8915796653            1.0342691675            0.9996638097
##           Informational  Informational_Duration          ProductRelated
##            1.1347683827            0.9996892738            1.0061208043
## ProductRelated_Duration             BounceRates               ExitRates
##            1.0001459455            0.0004616869            0.0041647566
##              PageValues             SpecialDay        OperatingSystems
##            1.1232768030            0.4229587504            0.9938803727
##                 Browser                  Region             TrafficType
##            1.0018851902            0.9674306509            1.0047436227
##             VisitorType                 Weekend
##            0.8673076180            1.1506107405
```

```
## odds ratios and 95% CI
exp(cbind(OR = coef(glm.fit_Half), confint(glm.fit_Half)))
```

```
## Waiting for profiling to be done...
```

```
##                                   OR        2.5 %    97.5 %
## (Intercept)             0.8915796653 4.887204e-01 1.6292781
## Administrative          1.0342691675 9.860277e-01 1.0858075
## Administrative_Duration 0.9996638097 9.987847e-01 1.0005451
## Informational           1.1347683827 1.003135e+00 1.2903757
## Informational_Duration  0.9996892738 9.986469e-01 1.0008368
## ProductRelated          1.0061208043 1.000119e+00 1.0123232
## ProductRelated_Duration 1.0001459455 1.000003e+00 1.0002926
## BounceRates             0.0004616869 4.512662e-08 3.5647128
## ExitRates               0.0041647566 1.369549e-06 8.5904427
## PageValues              1.1232768030 1.106063e+00 1.1424821
## SpecialDay              0.4229587504 2.008914e-01 0.8709324
## OperatingSystems        0.9938803727 8.683757e-01 1.1364202
## Browser                 1.0018851902 9.321701e-01 1.0779821
## Region                  0.9674306509 9.212889e-01 1.0158658
## TrafficType             1.0047436227 9.743322e-01 1.0362918
## VisitorType             0.8673076180 7.328749e-01 1.0253236
## Weekend                 1.1506107405 8.747605e-01 1.5156878
```

```
glm.probs_Half <- predict(glm.fit_Half,type = "response")
```

```
glm.pred_Half <- ifelse(glm.probs_Half > 0.5, "True", "False")
```

```
preproc1_Half <- preProcess(DF2, method=c("center", "scale"))

norm1_Half <- predict(preproc1_Half, DF2)

summary(norm1_Half)
```

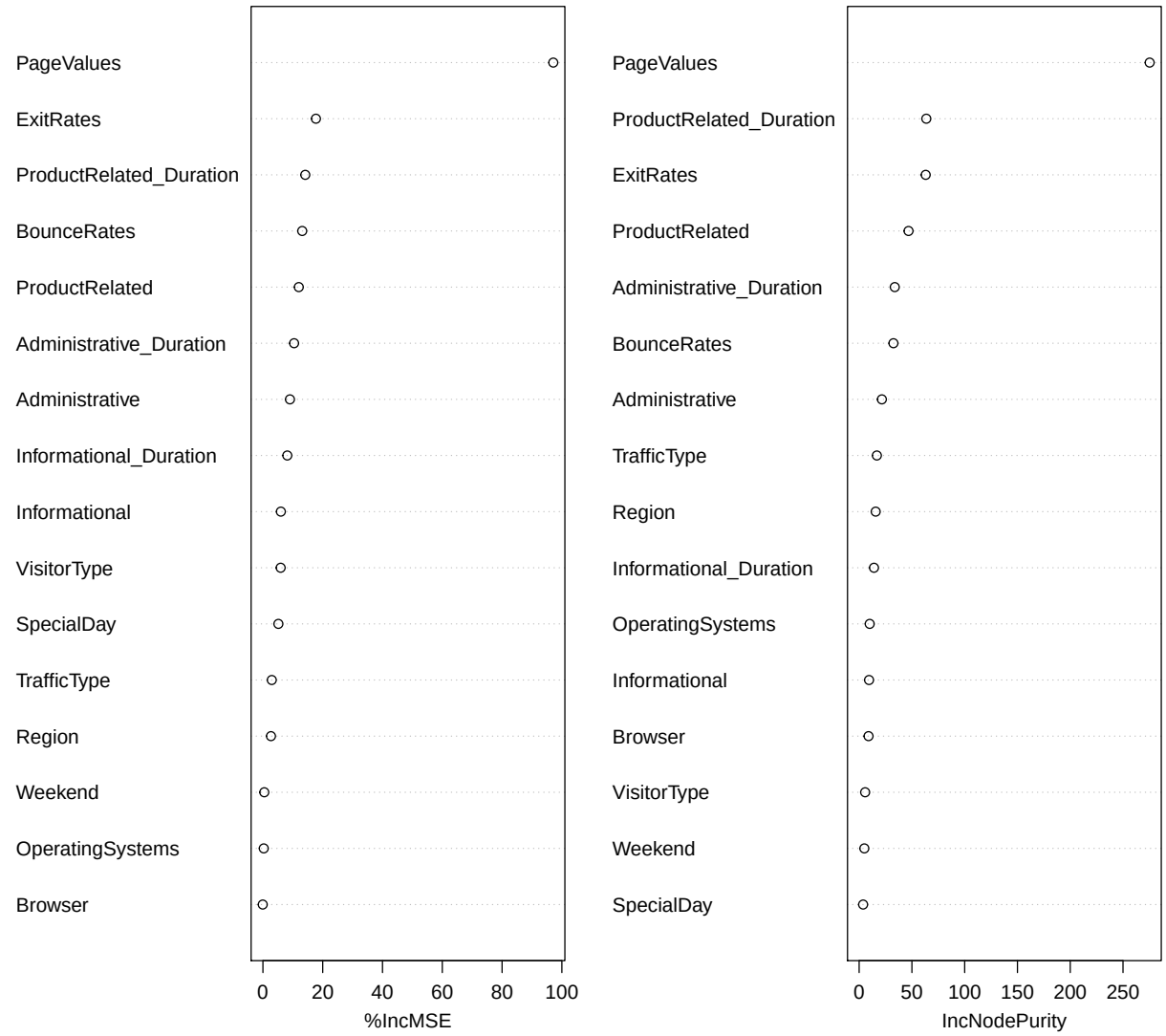
```
##  Administrative    Administrative_Duration Informational
```

```
##  Min.   :-0.6970   Min.   :-0.45717        Min.   :-0.3965
##  1st Qu.:-0.6970   1st Qu.:-0.45717        1st Qu.:-0.3965
##  Median :-0.3959   Median :-0.41475        Median :-0.3965
##  Mean   : 0.0000   Mean   : 0.00000        Mean   : 0.0000
##  3rd Qu.: 0.5072   3rd Qu.: 0.07036        3rd Qu.:-0.3965
##  Max.   : 7.4312   Max.   :18.76880        Max.   :18.4988
##  Informational_Duration ProductRelated    ProductRelated_Duration
##  Min.   :-0.2449        Min.   :-0.7135   Min.   :-0.6243
##  1st Qu.:-0.2449        1st Qu.:-0.5561   1st Qu.:-0.5281
##  Median :-0.2449        Median :-0.3087   Median :-0.3113
##  Mean   : 0.0000        Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.2449        3rd Qu.: 0.1409   3rd Qu.: 0.1408
##  Max.   :17.8680        Max.   :15.1380   Max.   :32.8055
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :-0.4577   Min.   :-0.8863   Min.   :-0.3172   Min.   :-0.3088
##  1st Qu.:-0.4577   1st Qu.:-0.5924   1st Qu.:-0.3172   1st Qu.:-0.3088
##  Median :-0.3935   Median :-0.3687   Median :-0.3172   Median :-0.3088
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.1109   3rd Qu.: 0.1425   3rd Qu.:-0.3172   3rd Qu.:-0.3088
##  Max.   : 3.6670   Max.   : 3.2292   Max.   :19.1656   Max.   : 4.7184
##  OperatingSystems     Browser            Region          TrafficType
##  Min.   :-1.2334   Min.   :-0.7903   Min.   :-0.89414   Min.   :-0.76260
##  1st Qu.:-0.1361   1st Qu.:-0.2079   1st Qu.:-0.89414   1st Qu.:-0.51416
##  Median :-0.1361   Median :-0.2079   Median :-0.06136   Median :-0.51416
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.9612   3rd Qu.:-0.2079   3rd Qu.: 0.35503   3rd Qu.:-0.01729
##  Max.   : 6.4477   Max.   : 6.1975   Max.   : 2.43698   Max.   : 3.95770
##   VisitorType         Weekend           Revenue
##  Min.   :-2.4876   Min.   :-0.5505   Min.   :-0.4279
##  1st Qu.: 0.4078   1st Qu.:-0.5505   1st Qu.:-0.4279
##  Median : 0.4078   Median :-0.5505   Median :-0.4279
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4078   3rd Qu.:-0.5505   3rd Qu.:-0.4279
##  Max.   : 0.4078   Max.   : 1.8163   Max.   : 2.3371
```

Random Forest

```
fitRF1 <- randomForest(
Revenue ~ ., method="anova",
data=trainingData[1:17], importance=TRUE, ntree=100)
```

```
varImpPlot(fitRF1, main="")
```




```
PredictionRF1 <- predict(fitRF1, testData)
```

```
summary(fitRF1)
```

```
##                 Length Class  Mode
## call               6   -none- call
## type               1   -none- character
## predicted       2670   -none- numeric
## mse              100   -none- numeric
## rsq              100   -none- numeric
## oob.times       2670   -none- numeric
## importance        32   -none- numeric
## importanceSD      16   -none- numeric
```

```
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            11   -none- list
## coefs              0   -none- NULL
## y               2670   -none- numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

```
sensitivity(testData$Revenue, PredictionRF1)
```

```
## [1] 0.8307155
```

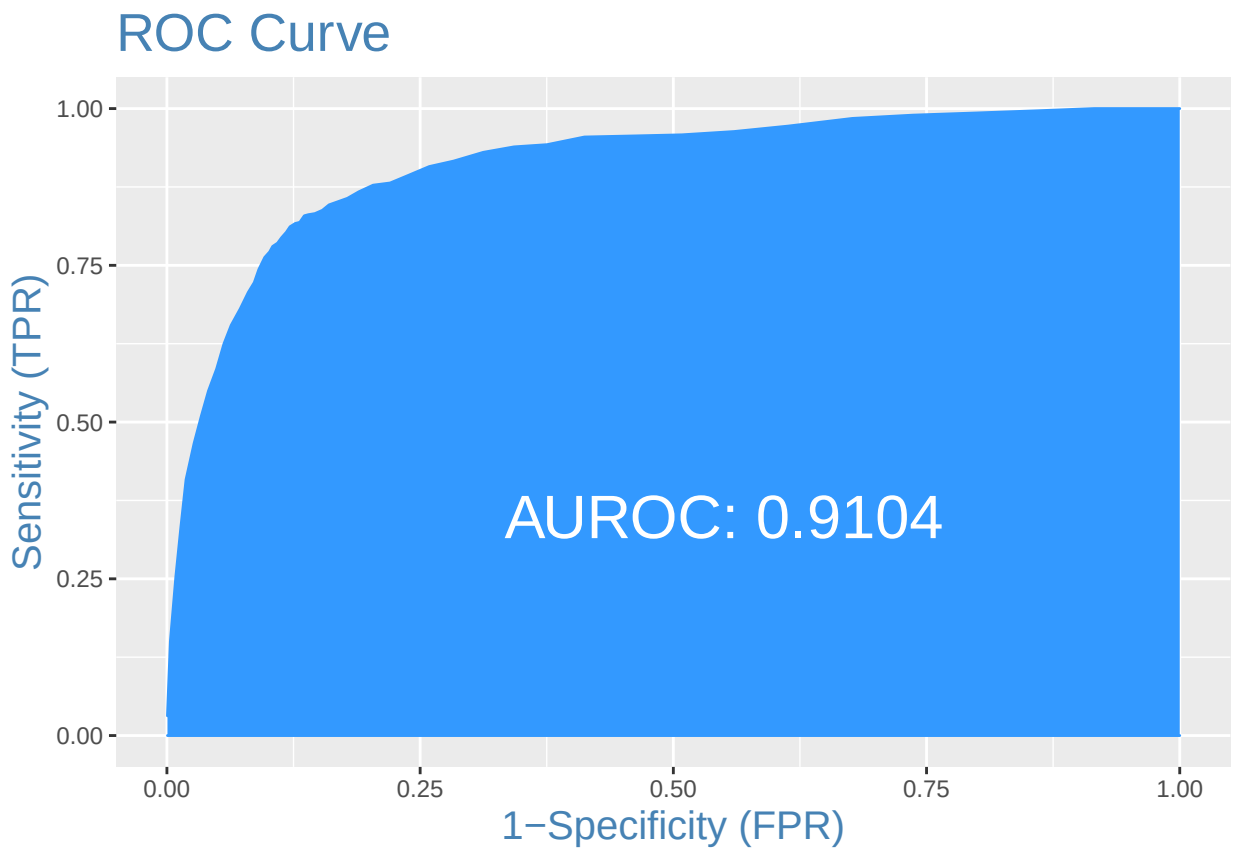
```
specificity(testData$Revenue, PredictionRF1)
```

```
## [1] 0.86035
```

```
CM_RF = confusionMatrix(testData$Revenue, PredictionRF1)
CM_RF
```

```
##      0   1
## 0 7818  97
## 1 1269 476
```

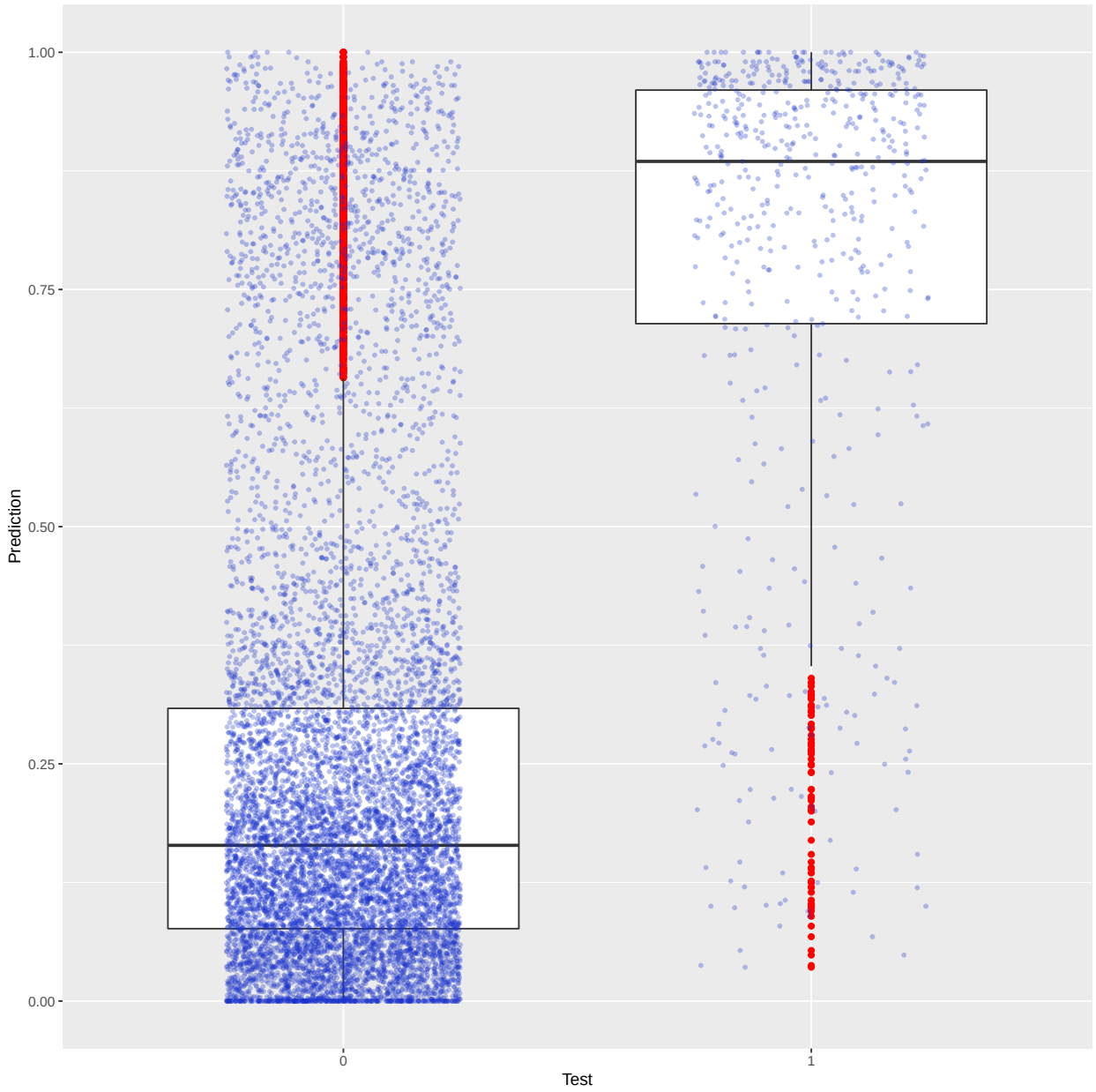
```
plotROC(testData$Revenue, PredictionRF1)
```




```
cor(PredictionRF1,testData$Revenue)
```

```
## [1] 0.4582964
```

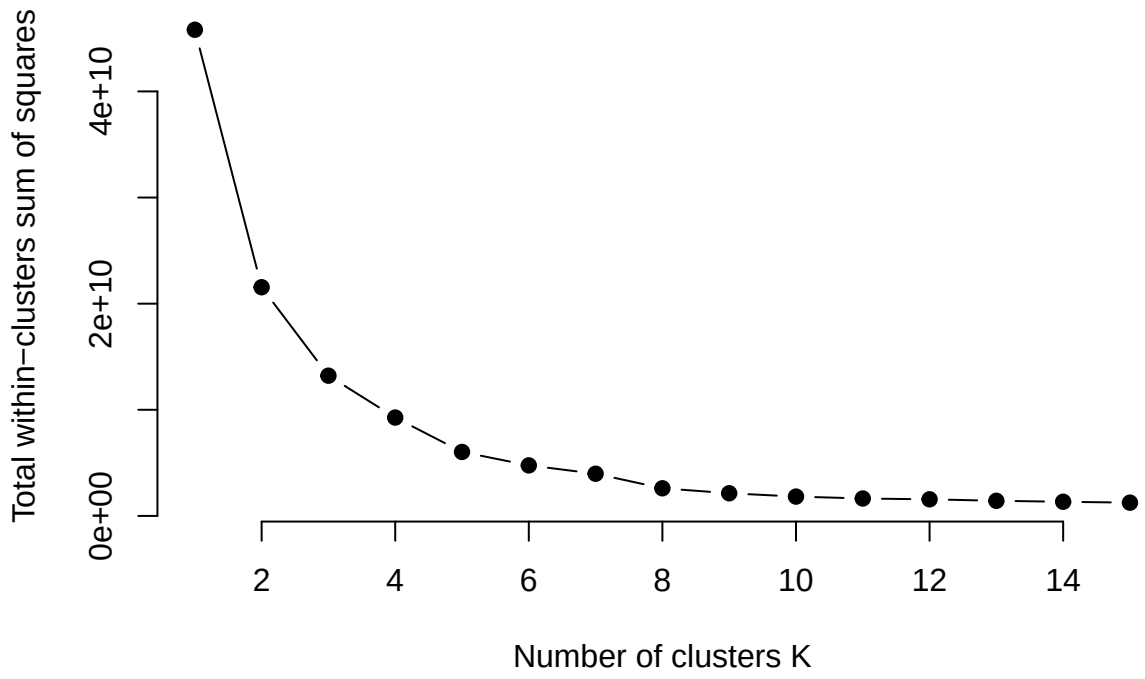
```
df3 = data.frame(as.factor(testData$Revenue), PredictionRF1)
colnames(df3) <- c("Test","Prediction")
ggplot(df3, aes(x = Test, y = Prediction)) +
geom_boxplot(outlier.colour = "red") +
geom_jitter(width = 0.25, pch=20, col=rgb(0.1, 0.2, 0.8, 0.3))
```



Clustering ## K Mean Clustering

```
set.seed(123)

# function to compute total within-cluster sum of square
wss <- function(k) {
  kmeans(DF2, k, nstart = 10 )$tot.withinss
}

# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15

# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)
```

```
plot(k.values, wss_values,
       type="b", pch = 19, frame = FALSE,
       xlab="Number of clusters K",
       ylab="Total within-clusters sum of squares")
```




```
# As we saw above 4 was the optimal number of clusters.

set.seed(240)
clusters <- kmeans(DF2[, 1:17],4 , nstart = 10)

# Save the cluster number in the dataset as column 'Borough'
DF2$Borough <- as.factor(clusters$cluster)
```

```
str(clusters)
```

```
## List of 9
##  $ cluster     : int [1:12330] 2 2 2 2 2 2 2 2 2 2 ...
##  $ centers     : num [1:4, 1:17] 3.75 1.64 7.65 5.89 138.27 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "1" "2" "3" "4"
##   .. ..$ : chr [1:17] "Administrative" "Administrative_Duration" "Informational" "Informational_Dura
##  $ totss       : num 4.58e+10
##  $ withinss    : num [1:4] 1.70e+09 1.60e+09 4.35e+09 1.63e+09
##  $ tot.withinss: num 9.27e+09
```

```
##  $ betweenss    : num 3.65e+10
##  $ size         : int [1:4] 2609 9084 66 571
##  $ iter         : int 3
##  $ ifault       : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
DF2$KMEAN = clusters$cluster
```

```
clusters$withinss
```

```
## [1] 1698803331 1596734986 4345284451 1628132463
```

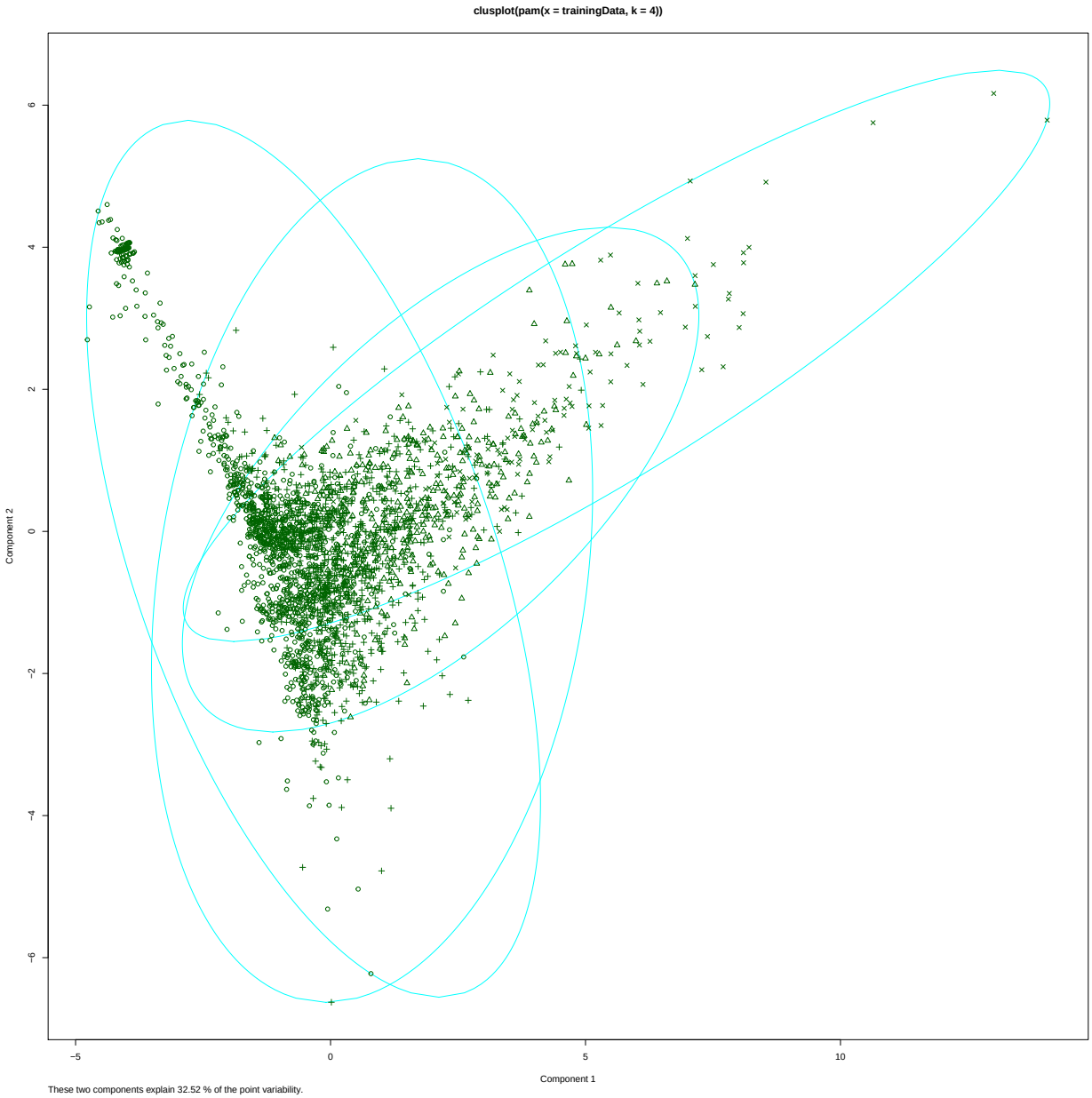
```
clusters$tot.withinss/clusters$betweenss
```

```
## [1] 0.2536653
```

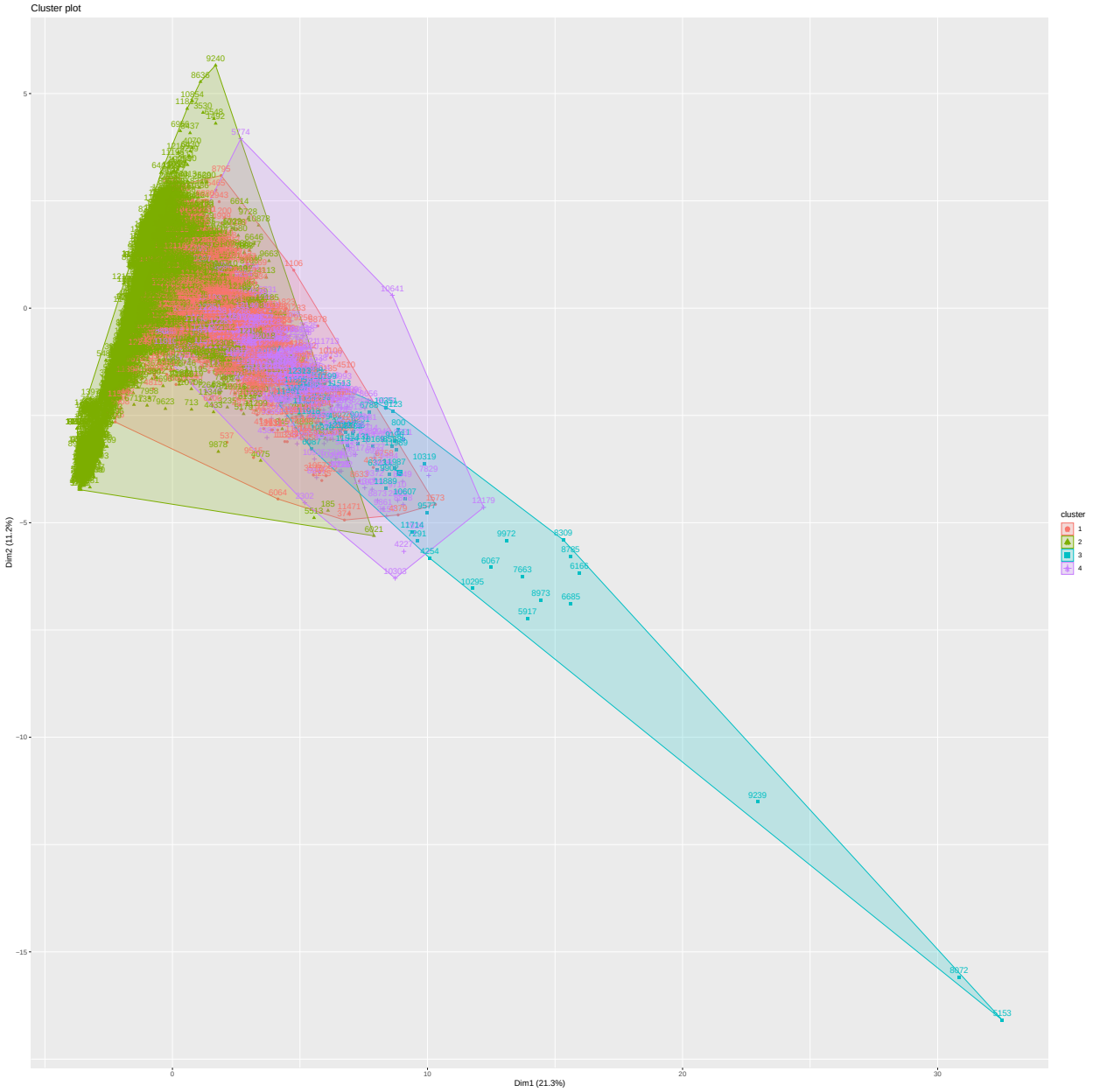
```
data.frame(clusters$centers)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1       3.753545               138.27140     0.9398237               67.17332
## 2       1.638595                53.65965     0.2761999               15.71935
## 3       7.651515               416.47911     3.3787879              364.76162
## 4       5.889667               211.57876     1.7950963              145.22002
##   ProductRelated ProductRelated_Duration BounceRates  ExitRates PageValues
## 1       56.93292                2331.938 0.008004385 0.02308015   8.560955
## 2       15.78622                 450.526 0.027370608 0.05038192   5.070271
## 3      332.33333               15772.493 0.005424551 0.01864796   4.108254
## 4      135.50788                6153.472 0.006556316 0.02096564   6.916867
##   SpecialDay OperatingSystems  Browser   Region TrafficType VisitorType
## 1 0.05473361         2.131468 2.315830 3.204676    3.811422    2.864316
## 2 0.06442096         2.120321 2.370211 3.157530    4.178446    2.658851
## 3 0.01212121         2.045455 2.393939 2.484848    3.893939    2.969697
## 4 0.05008757         2.157618 2.332750 2.800350    3.537653    2.968476
##     Weekend   Revenue
## 1 0.2364891 0.2150249
## 2 0.2317261 0.1247248
## 3 0.2727273 0.3787879
## 4 0.2241681 0.3309982
```

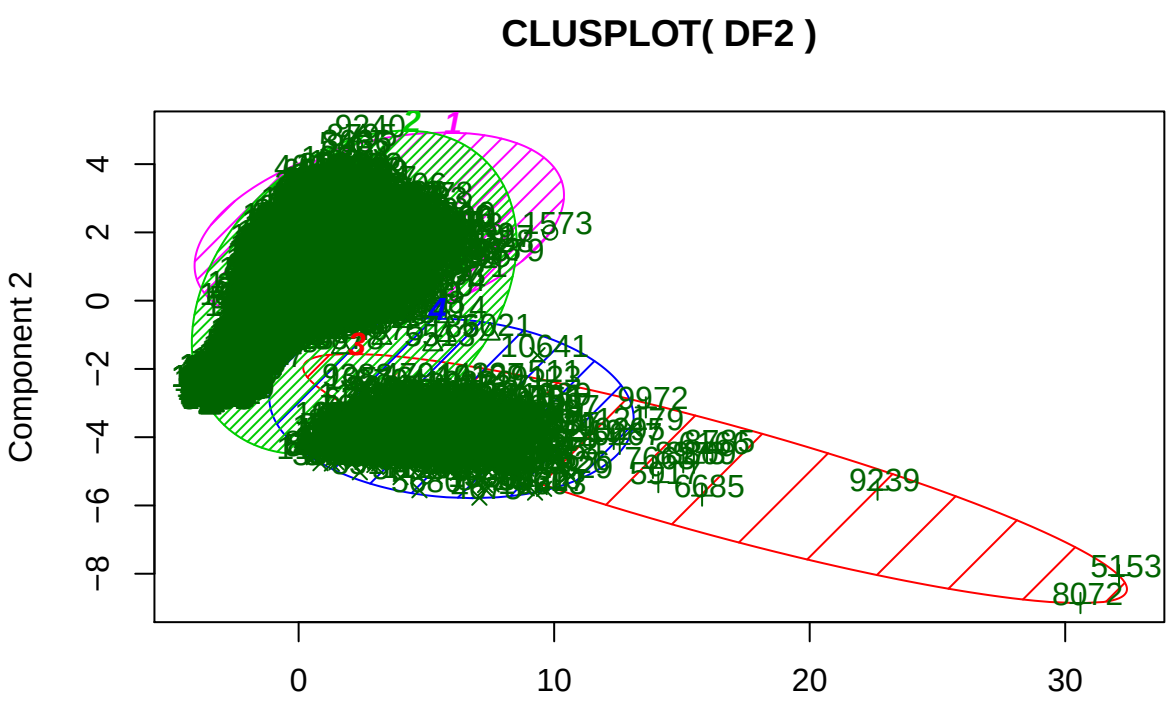
```
library("cluster")
clusplot(pam(trainingData,4))
```



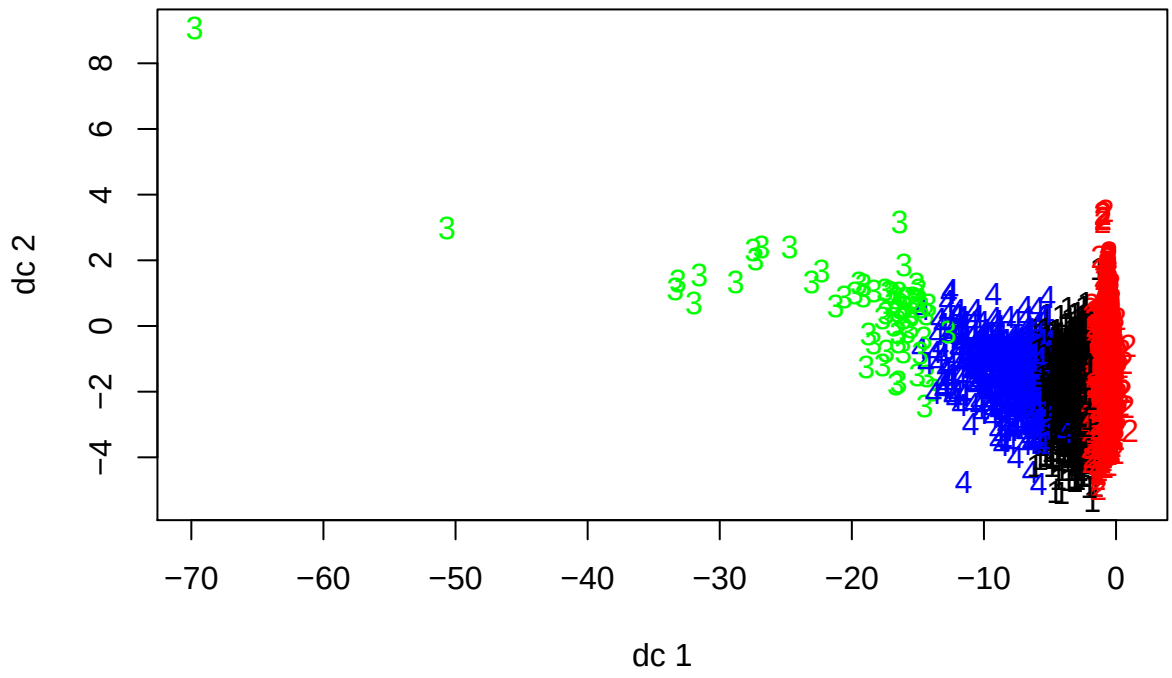

```
fviz_cluster(clusters, data = DF2[1:16],ellipse=TRUE)
```




```
# vary parameters for most readable graph
library(cluster)
clusplot(DF2, clusters$cluster, color=TRUE, shade=TRUE,
   labels=2, lines=0)
```




```
# Centroid Plot against 1st 2 discriminant functions
library(fpc)
plotcluster(DF2[1:17], clusters$cluster)
```

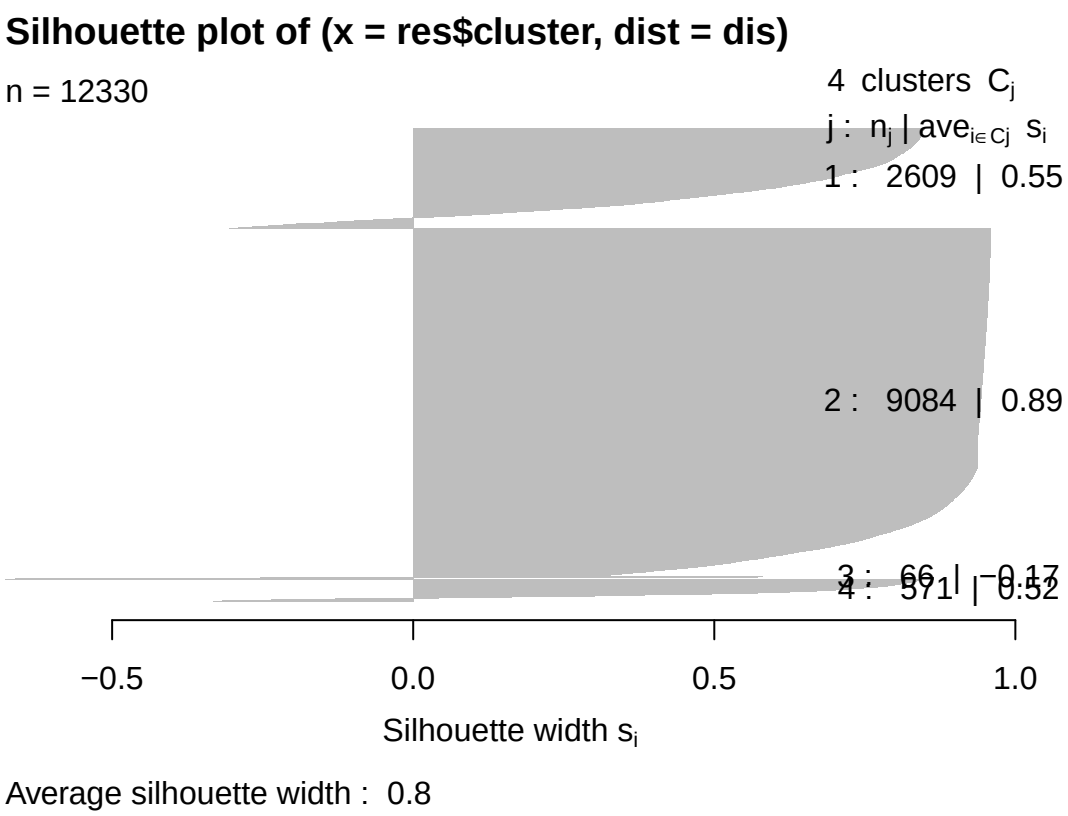



```
dis = dist(DF2)^2
res = clusters
sil = silhouette (res$cluster, dis)
```

```
summary(sil)
```

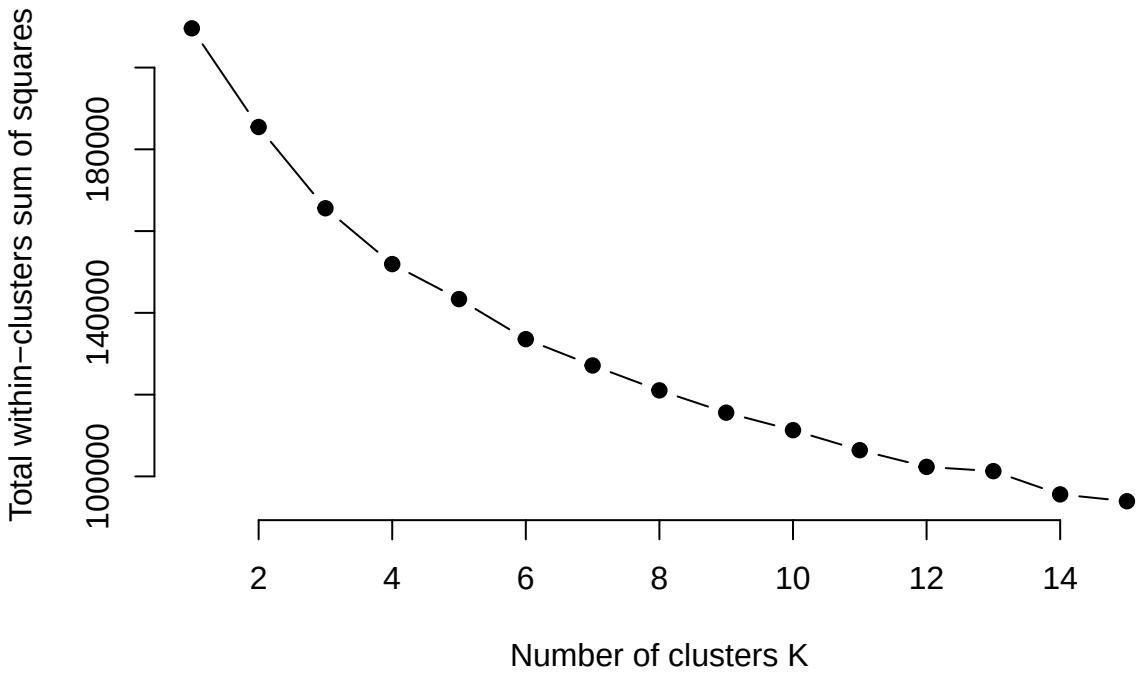
```
## Silhouette of 12330 units in 4 clusters from silhouette.default(x = res$cluster, dist = dis) :
##  Cluster sizes and average silhouette widths:
##       2609       9084         66        571
##  0.5512495  0.8924093 -0.1673278  0.5245451
## Individual silhouette widths:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.6772  0.7647  0.9375  0.7975  0.9497  0.9595
```

```
#Plot
plot(sil, nmax= 10, cex.names=0.9)
```



k-means clustering with scaled data

```
Scaled_DF_1 <- as.data.frame(scale(DF2[1:17]))
set.seed(123)

# function to compute total within-cluster sum of square
wss <- function(k) {
  kmeans(Scaled_DF_1, k, nstart = 10 )$tot.withinss
}

# Compute and plot wss for k = 1 to k = 15
k.values <- 1:15

# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, wss)

plot(k.values, wss_values,
       type="b", pch = 19, frame = FALSE,
       xlab="Number of clusters K",
       ylab="Total within-clusters sum of squares")
```




```
# As we saw above 4 was the optimal number of clusters.

set.seed(240)
clusters_1 <- kmeans(Scaled_DF_1[, 1:17],4 , nstart = 10)

# Save the cluster number in the dataset as column 'Borough'
DF2$KMEAN_SCALED <- as.factor(clusters$cluster)
```

```
str(clusters_1)
```

```
## List of 9
##  $ cluster     : int [1:12330] 2 1 2 2 1 1 2 2 1 1 ...
##  $ centers     : num [1:4, 1:17] -0.1735 -0.684 0.0836 1.5991 -0.1622 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:4] "1" "2" "3" "4"
##   .. ..$ : chr [1:17] "Administrative" "Administrative_Duration" "Informational" "Informational_Dura
##  $ totss       : num 209593
##  $ withinss    : num [1:4] 76708 9607 25062 40550
##  $ tot.withinss: num 151927
##  $ betweenss   : num 57666
##  $ size        : int [1:4] 8347 1047 1670 1266
##  $ iter        : int 4
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
CrossTable(DF2$KMEAN_SCALED,DF2$Revenue)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##                  | DF2$Revenue
## DF2$KMEAN_SCALED |         0 |         1 | Row Total |
## -----------------|-----------|-----------|-----------|
##                1 |      2048 |       561 |      2609 |
##                  |    11.216 |    61.265 |           |
##                  |     0.785 |     0.215 |     0.212 |
##                  |     0.197 |     0.294 |           |
##                  |     0.166 |     0.045 |           |
## -----------------|-----------|-----------|-----------|
##                2 |      7951 |      1133 |      9084 |
##                  |     9.685 |    52.902 |           |
##                  |     0.875 |     0.125 |     0.737 |
##                  |     0.763 |     0.594 |           |
##                  |     0.645 |     0.092 |           |
## -----------------|-----------|-----------|-----------|
##                3 |        41 |        25 |        66 |
##                  |     3.919 |    21.409 |           |
##                  |     0.621 |     0.379 |     0.005 |
##                  |     0.004 |     0.013 |           |
##                  |     0.003 |     0.002 |           |
## -----------------|-----------|-----------|-----------|
##                4 |       382 |       189 |       571 |
##                  |    20.986 |   114.630 |           |
##                  |     0.669 |     0.331 |     0.046 |
##                  |     0.037 |     0.099 |           |
##                  |     0.031 |     0.015 |           |
## -----------------|-----------|-----------|-----------|
##     Column Total |     10422 |      1908 |     12330 |
##                  |     0.845 |     0.155 |           |
## -----------------|-----------|-----------|-----------|
##
##
```

```
CrossTable(DF2$KMEAN,DF2$Revenue)
```

```
##
```

```
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  12330
##
##
##                | DF2$Revenue
##    DF2$KMEAN |         0 |         1 | Row Total |
## -------------|-----------|-----------|-----------|
##            1 |      2048 |       561 |      2609 |
##              |    11.216 |    61.265 |           |
##              |     0.785 |     0.215 |     0.212 |
##              |     0.197 |     0.294 |           |
##              |     0.166 |     0.045 |           |
## -------------|-----------|-----------|-----------|
##            2 |      7951 |      1133 |      9084 |
##              |     9.685 |    52.902 |           |
##              |     0.875 |     0.125 |     0.737 |
##              |     0.763 |     0.594 |           |
##              |     0.645 |     0.092 |           |
## -------------|-----------|-----------|-----------|
##            3 |        41 |        25 |        66 |
##              |     3.919 |    21.409 |           |
##              |     0.621 |     0.379 |     0.005 |
##              |     0.004 |     0.013 |           |
##              |     0.003 |     0.002 |           |
## -------------|-----------|-----------|-----------|
##            4 |       382 |       189 |       571 |
##              |    20.986 |   114.630 |           |
##              |     0.669 |     0.331 |     0.046 |
##              |     0.037 |     0.099 |           |
##              |     0.031 |     0.015 |           |
## -------------|-----------|-----------|-----------|
## Column Total |     10422 |      1908 |     12330 |
##              |     0.845 |     0.155 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
clusters_1$withinss
```

```
## [1] 76707.52  9607.17 25062.15 40550.06
```

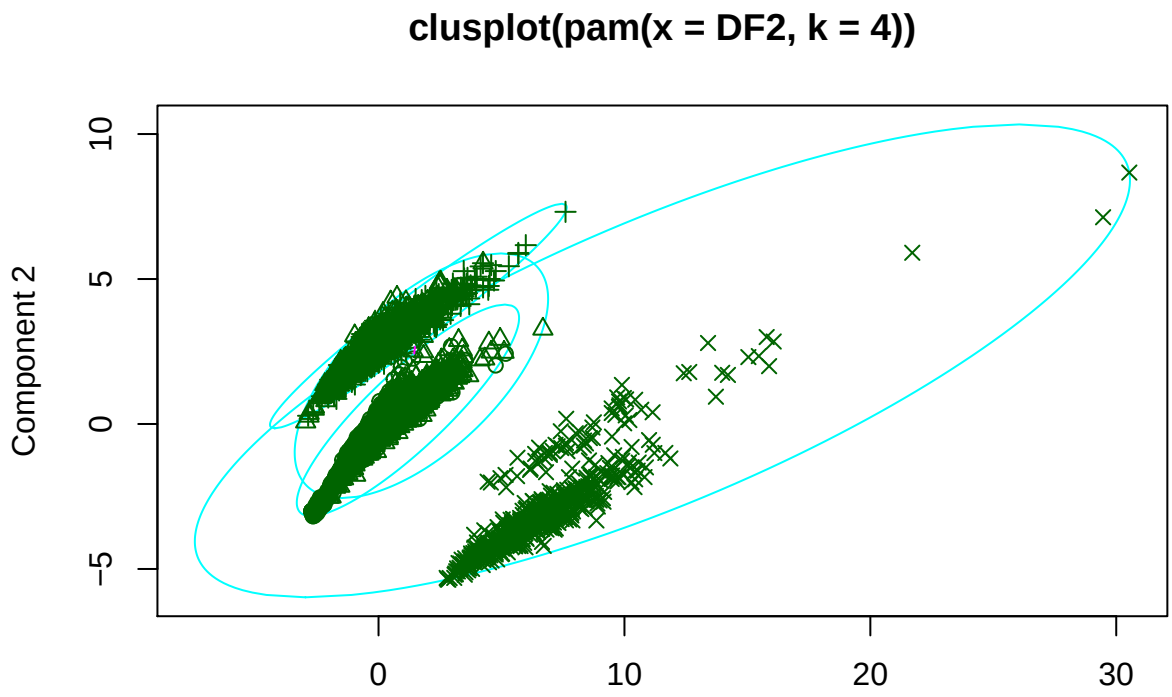
```
clusters_1$tot.withinss/clusters_1$betweenss
```

```
## [1] 2.634596
```

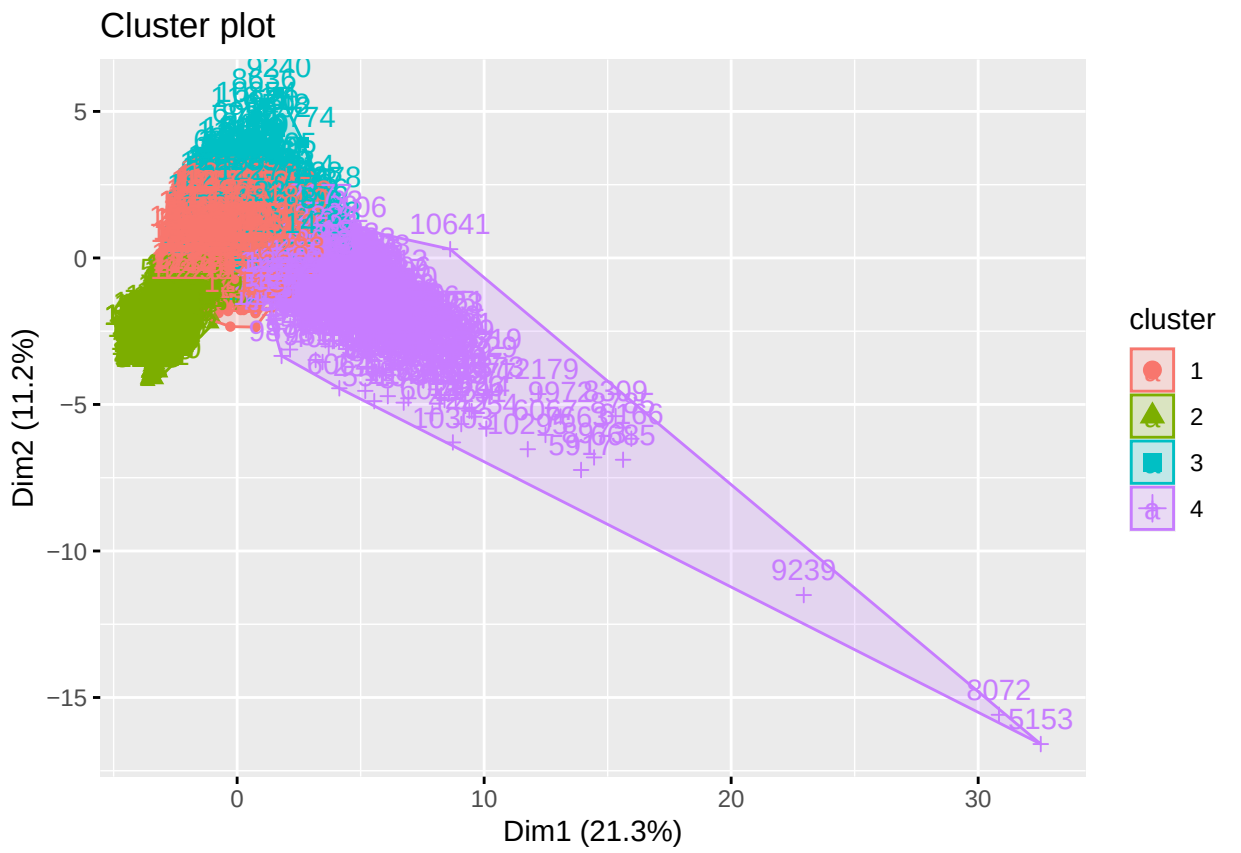
```
data.frame(clusters_1$centers)
```

```
##   Administrative Administrative_Duration Informational Informational_Duration
## 1    -0.17346646             -0.16221091   -0.21328876             -0.1808159
## 2    -0.68402589             -0.44940105   -0.38367846             -0.2442610
## 3     0.08358403              0.02763157   -0.09144022             -0.1160928
## 4     1.59914244              1.40470195    1.84418469              1.5473037
##   ProductRelated ProductRelated_Duration BounceRates  ExitRates PageValues
## 1    -0.18201192             -0.17791130  -0.2530359 -0.1856218 -0.2496910
## 2    -0.64516906             -0.59543937   2.9799023  2.8304976 -0.3171650
## 3     0.04067971              0.05117628  -0.3605570 -0.4862000  1.3992553
## 4     1.67994500              1.59793543  -0.3204872 -0.4756648  0.0627852
##    SpecialDay OperatingSystems       Browser      Region  TrafficType
## 1  0.02834645      -0.00158833  0.0006498772  0.01950263 -0.012644692
## 2  0.24528922       0.06410392 -0.0333039075 -0.03590833  0.239928473
## 3 -0.18056969      -0.02042879  0.0779847845  0.01942371  0.006514511
## 4 -0.15155945      -0.01559472 -0.0796128938 -0.12451033 -0.123648577
##   VisitorType      Weekend    Revenue
## 1 -0.01541764 -0.006107908 -0.4278544
## 2  0.30545037 -0.159450956 -0.4172913
## 3 -0.34987973  0.065976489  2.2410272
## 4  0.31057159  0.085108309  0.2098653
```

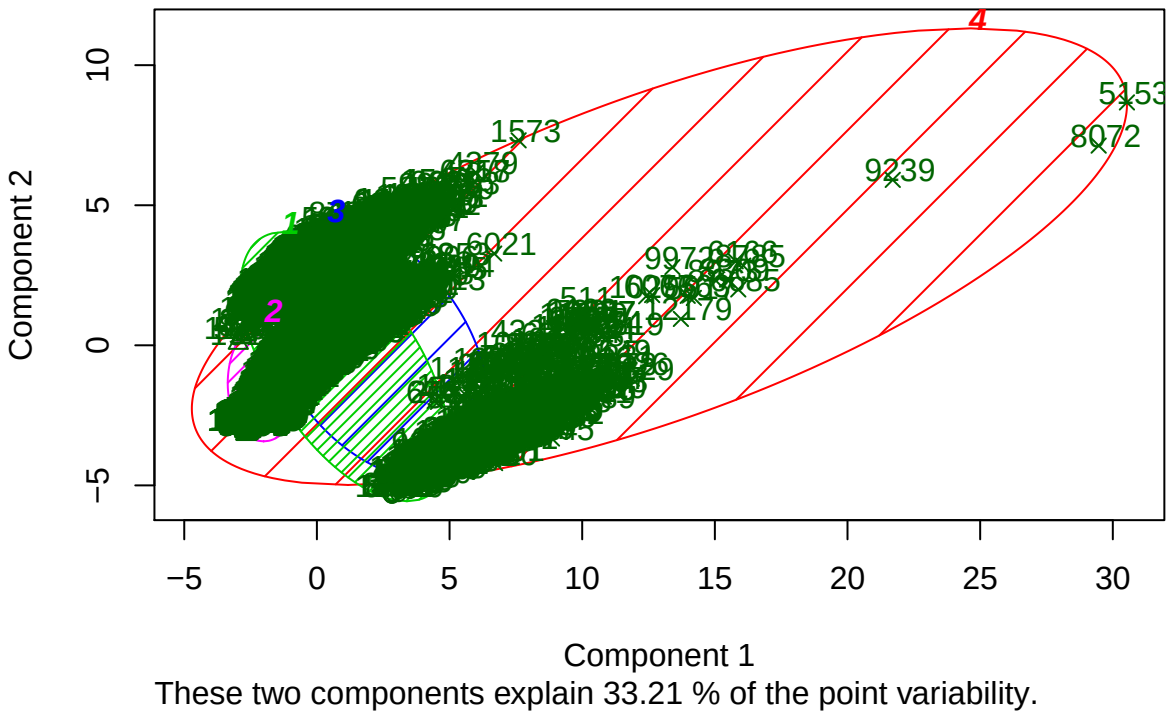
```
clusplot(pam(DF2,4))
```



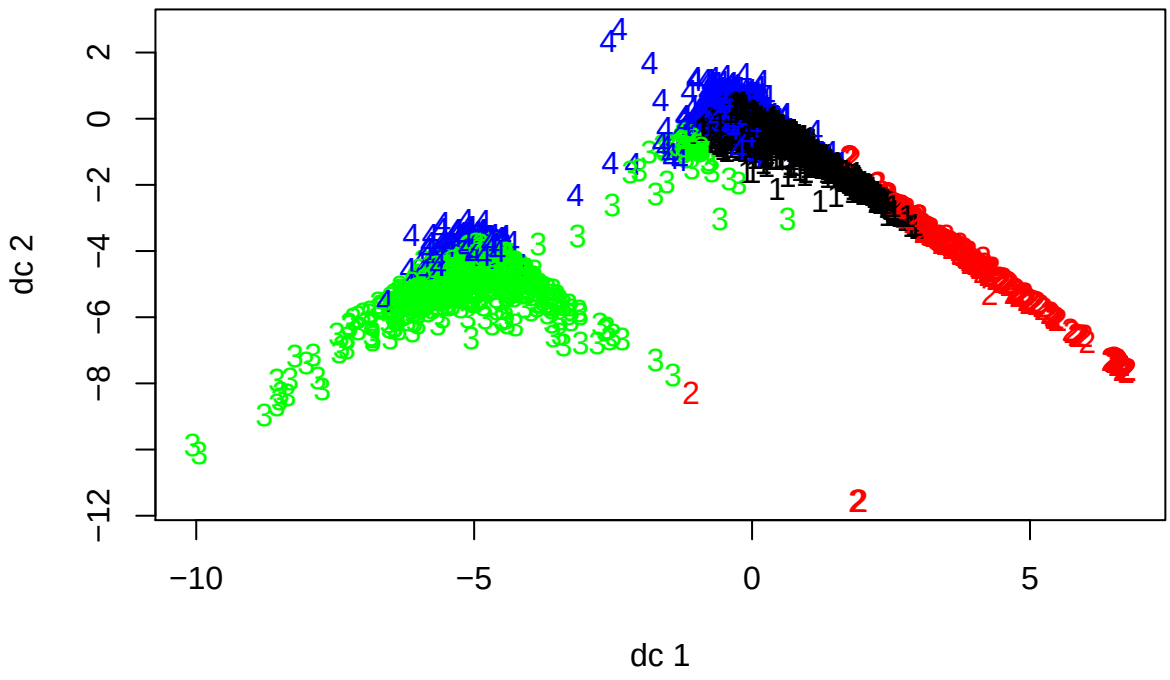
```
fviz_cluster(clusters_1, data = DF2[1:16],ellipse=TRUE)
```




```
# Cluster Plot against 1st 2 principal components

# vary parameters for most readable graph
clusplot(DF2, clusters_1$cluster, color=TRUE, shade=TRUE,
   labels=2, lines=0)
```




```
# Centroid Plot against 1st 2 discriminant functions
plotcluster(DF2[1:17], clusters_1$cluster)
```

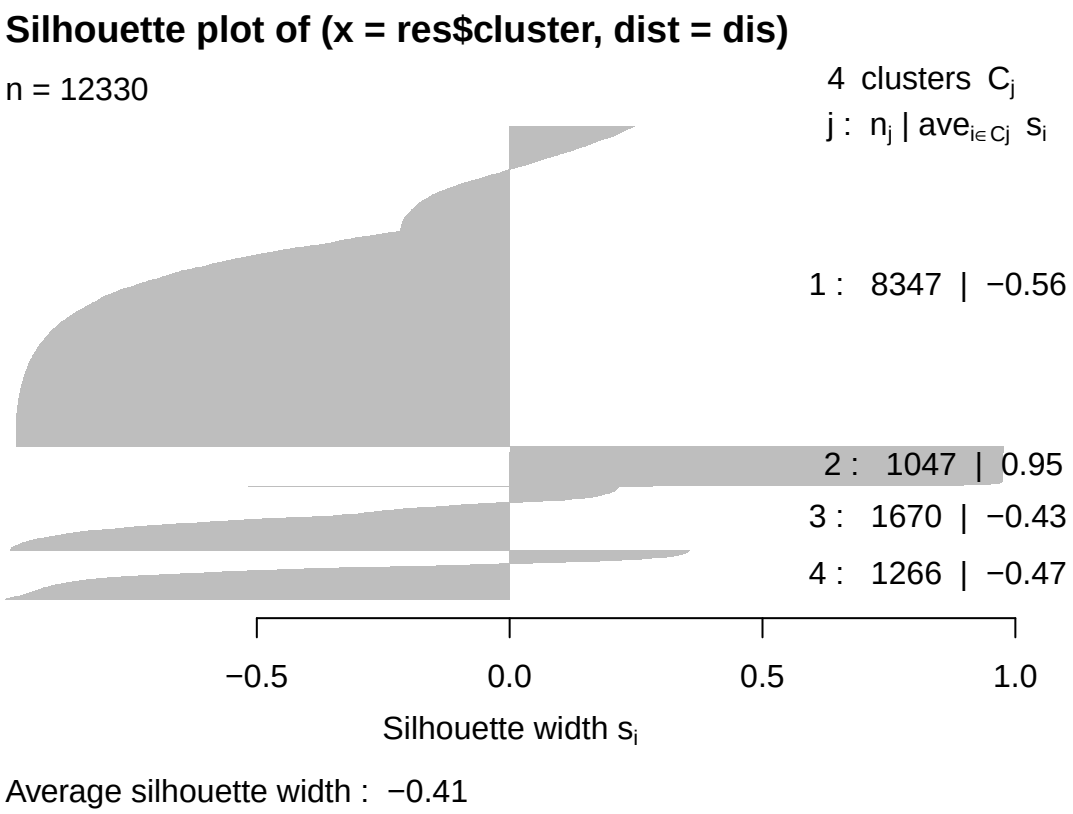



```
#Preperation
dis = dist(DF2)^2
res = clusters_1
sil = silhouette (res$cluster, dis)
```

```
summary(sil)
```

```
## Silhouette of 12330 units in 4 clusters from silhouette.default(x = res$cluster, dist = dis) :
##  Cluster sizes and average silhouette widths:
##       8347       1047       1670       1266
## -0.5626146  0.9516690 -0.4345101 -0.4672120
## Individual silhouette widths:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.99709 -0.93421 -0.59450 -0.40688 -0.03401  0.97685
```

```
#Plot
plot(sil, nmax= 10, cex.names=0.9)
```



Model Based Clustering

```
fit <- Mclust(DF2)
```

```
summary(fit)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust XXX (ellipsoidal multivariate normal) model with 1 component:
##
##  log-likelihood     n  df     BIC     ICL
##        686012.1 12330 230 1369858 1369858
##
## Clustering table:
##     1
## 12330
```

Hierarchical Clustering

##Scaling

```
Scaled_DF <- as.data.frame(scale(DF2[1:17]))
summary(Scaled_DF)
```

```
##  Administrative    Administrative_Duration Informational
##  Min.   :-0.6970   Min.   :-0.45717        Min.   :-0.3965
##  1st Qu.:-0.6970   1st Qu.:-0.45717        1st Qu.:-0.3965
##  Median :-0.3959   Median :-0.41475        Median :-0.3965
##  Mean   : 0.0000   Mean   : 0.00000        Mean   : 0.0000
##  3rd Qu.: 0.5072   3rd Qu.: 0.07036        3rd Qu.:-0.3965
##  Max.   : 7.4312   Max.   :18.76880        Max.   :18.4988
##  Informational_Duration ProductRelated    ProductRelated_Duration
##  Min.   :-0.2449        Min.   :-0.7135   Min.   :-0.6243
##  1st Qu.:-0.2449        1st Qu.:-0.5561   1st Qu.:-0.5281
##  Median :-0.2449        Median :-0.3087   Median :-0.3113
##  Mean   : 0.0000        Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.2449        3rd Qu.: 0.1409   3rd Qu.: 0.1408
##  Max.   :17.8680        Max.   :15.1380   Max.   :32.8055
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :-0.4577   Min.   :-0.8863   Min.   :-0.3172   Min.   :-0.3088
##  1st Qu.:-0.4577   1st Qu.:-0.5924   1st Qu.:-0.3172   1st Qu.:-0.3088
##  Median :-0.3935   Median :-0.3687   Median :-0.3172   Median :-0.3088
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.:-0.1109   3rd Qu.: 0.1425   3rd Qu.:-0.3172   3rd Qu.:-0.3088
##  Max.   : 3.6670   Max.   : 3.2292   Max.   :19.1656   Max.   : 4.7184
##  OperatingSystems     Browser            Region          TrafficType
##  Min.   :-1.2334   Min.   :-0.7903   Min.   :-0.89414   Min.   :-0.76260
##  1st Qu.:-0.1361   1st Qu.:-0.2079   1st Qu.:-0.89414   1st Qu.:-0.51416
##  Median :-0.1361   Median :-0.2079   Median :-0.06136   Median :-0.51416
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000
##  3rd Qu.: 0.9612   3rd Qu.:-0.2079   3rd Qu.: 0.35503   3rd Qu.:-0.01729
##  Max.   : 6.4477   Max.   : 6.1975   Max.   : 2.43698   Max.   : 3.95770
##   VisitorType         Weekend           Revenue
##  Min.   :-2.4876   Min.   :-0.5505   Min.   :-0.4279
##  1st Qu.: 0.4078   1st Qu.:-0.5505   1st Qu.:-0.4279
##  Median : 0.4078   Median :-0.5505   Median :-0.4279
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.4078   3rd Qu.:-0.5505   3rd Qu.:-0.4279
##  Max.   : 0.4078   Max.   : 1.8163   Max.   : 2.3371
```

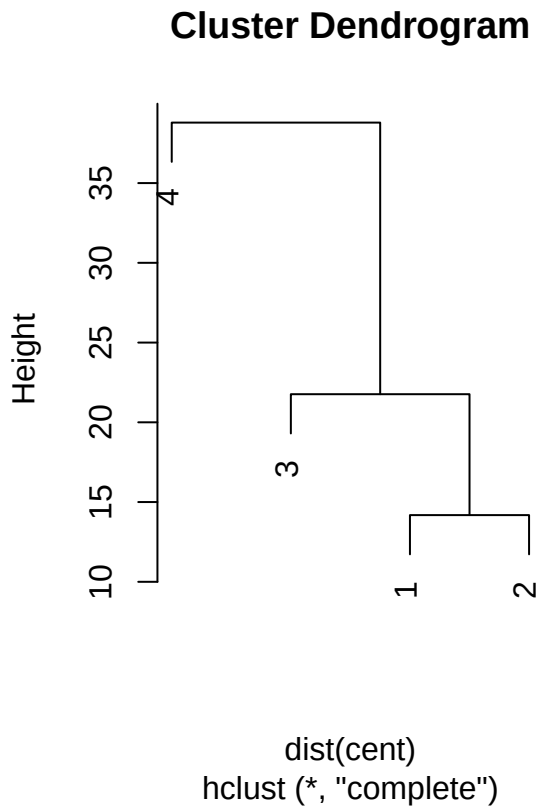
```
#To remove any missing value that might be present in the data.

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

##hclust

```
dist_mat <- dist(Scaled_DF, method = 'euclidean')
hclust_avg <- hclust(dist_mat, method = 'average')
hc1 <- hclust(dist_mat, method = "complete" )
```

```
memb<- cutree(hc1, k=4)
cent<- NULL
for(k in 1:4){
```

```
  cent<-rbind(cent, colMeans(Scaled_DF[memb == k, ,drop=FALSE]))}
hc2<-hclust(dist(cent), method="complete", members=table(memb))
opar<-par(mfrow=c(1,2))
plot(hc2)
```

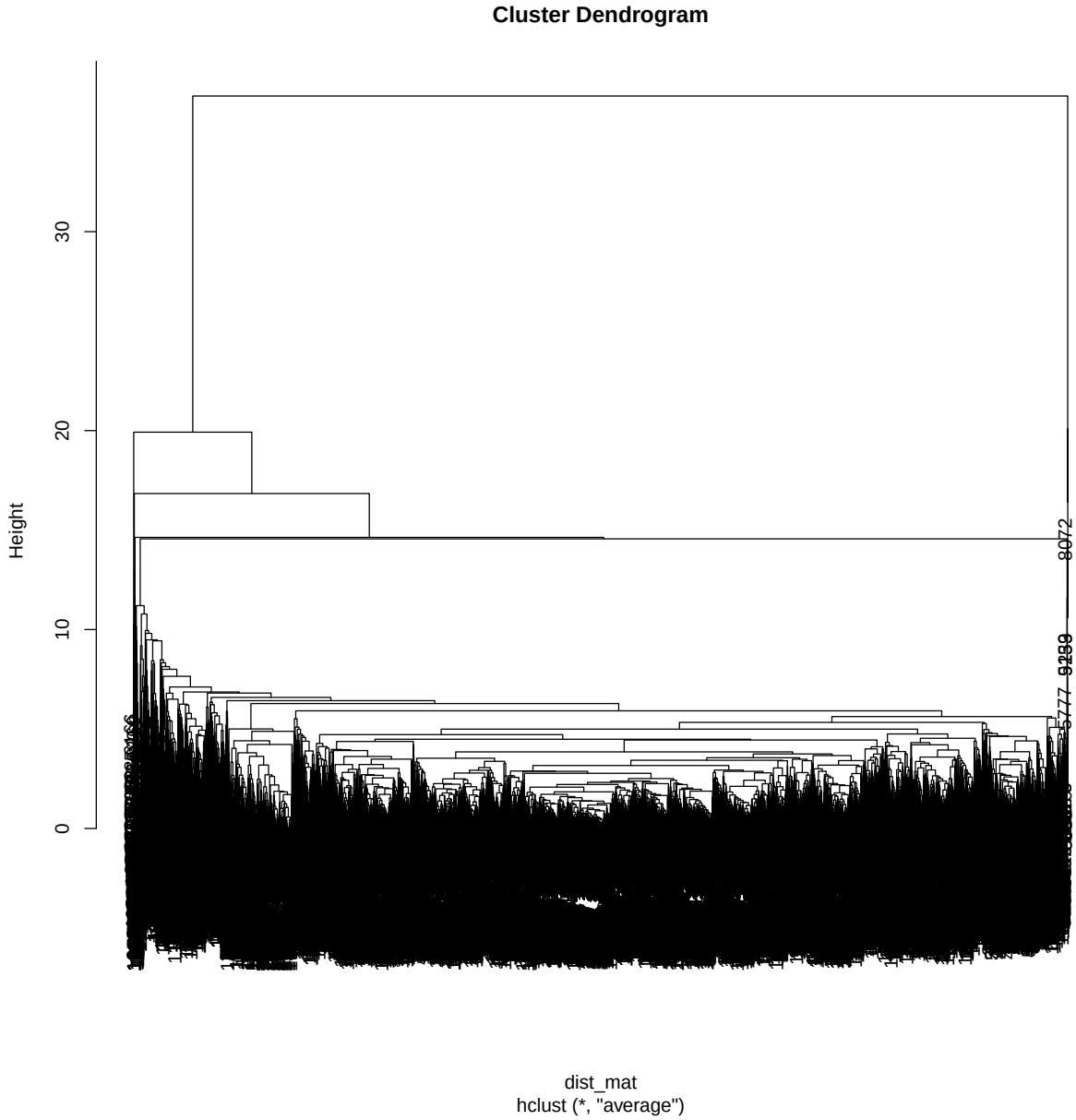



```
DF2$CUT_AVG=memb
```

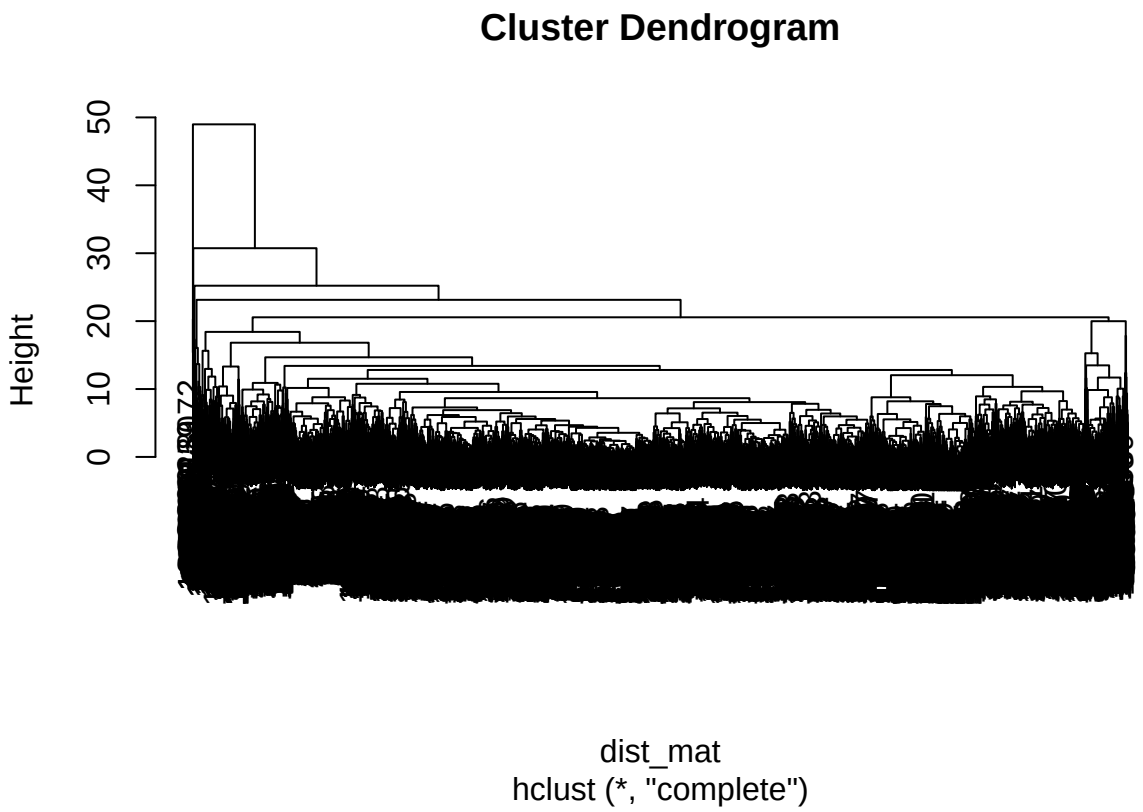
```
table(memb, DF2$Revenue)
```

```
##
## memb     0     1
##    1 10411  1892
##    2     2    12
##    3     6     4
##    4     3     0
```

```
plot(hclust_avg)
```

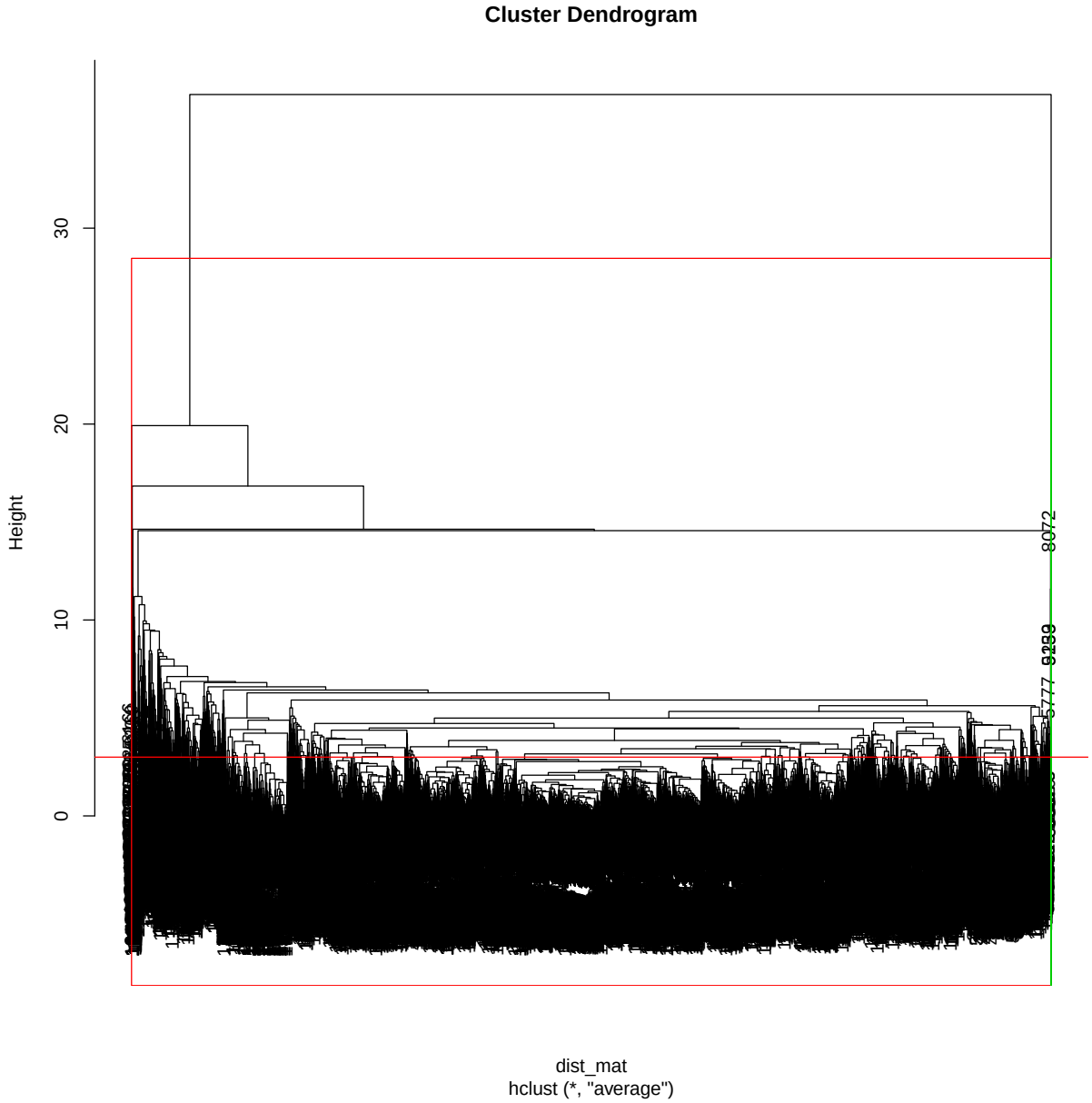

Cluster Dendrogram
Height
0
10
20
30
8072
dist_mat
hclust (*, "average")

```
plot(hc1)
```




```
cut_avg <- cutree(hclust_avg, k = 2)
```

```
plot(hclust_avg)
rect.hclust(hclust_avg , k = 2, border = 2:6)
abline(h = 3, col = 'red')
```


Cluster Dendrogram
Height
0
10
20
30
8072
dist_mat
hclust (*, "average")