ShoppingIntention_report

Exploratory Data Analysis:

There are no missing values in the data.

Goal: To guage whether a shopper is intending to buy or not. This can help marketeers to strategize for increasing the sales revenue.

#1. How many different 'Month' are there?
10 months only (Two months are missing from the data i.e. Jan & April)

#2. Which is the most common 'Month'?

May

#3. How many special days are there in the data? (i.e. marked by value equal to 1)

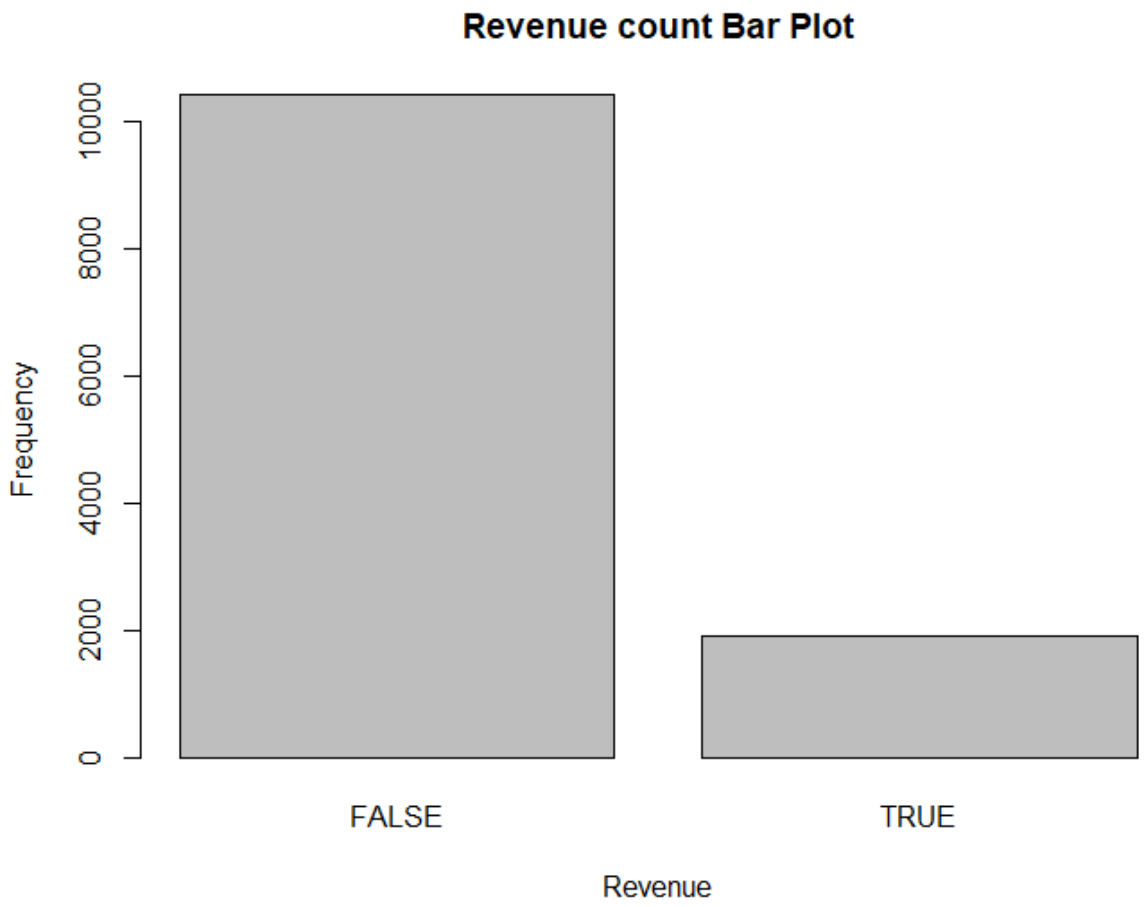
154 special day sessions from the special days marked by 1.

#4.Which months has sessions corresponding to special days?

Feb and May

Interpretation for Fig 1: ***There are more people who only surf the site but lesser people actually buy*** designated by Revenue value to be True.

Fig 1:



Interpretation for Fig 3: ***There is much more traffic on weekdays compared to weekends.***

Interpretation for Fig 2: And yet ***the conversion rate is more on weekend*** (almost 18%) than on weekdays (almost 15%).

Fig 2:

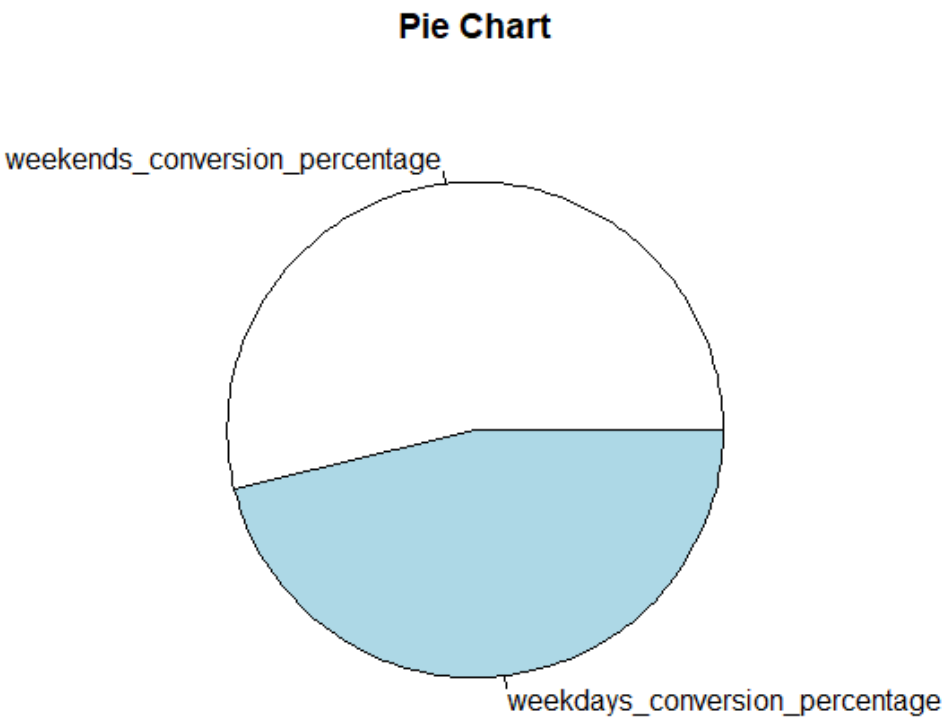

Pie Chart
weekends_conversion_percentage
weekdays_conversion_percentage

Fig 3:

Interpretation for Fig 4: Box-plot helps you understand the distribution of a numeric variable. It helps you visualize the instances which are above or below the median value. These points indicate the spread.

Fig 4:

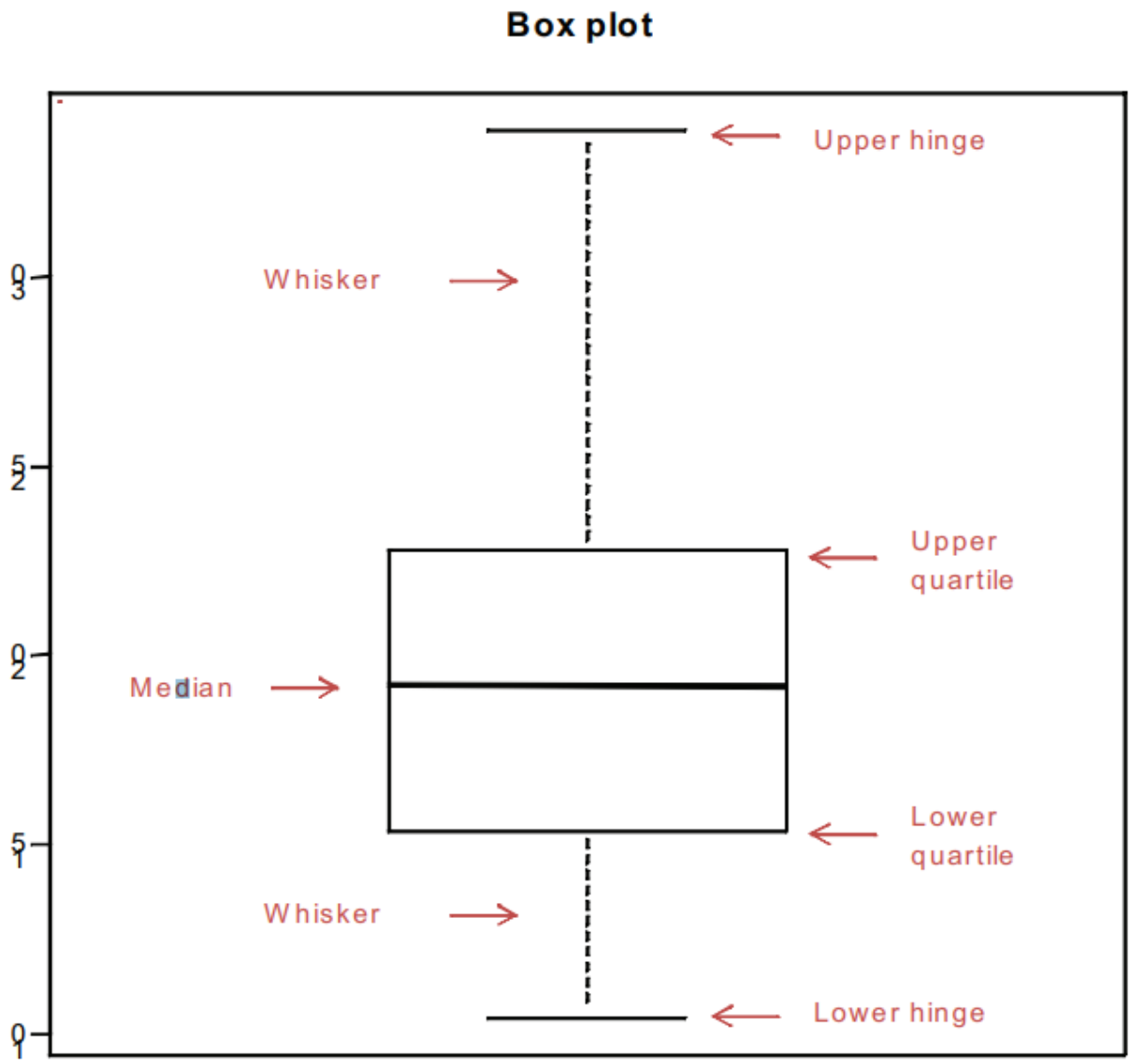
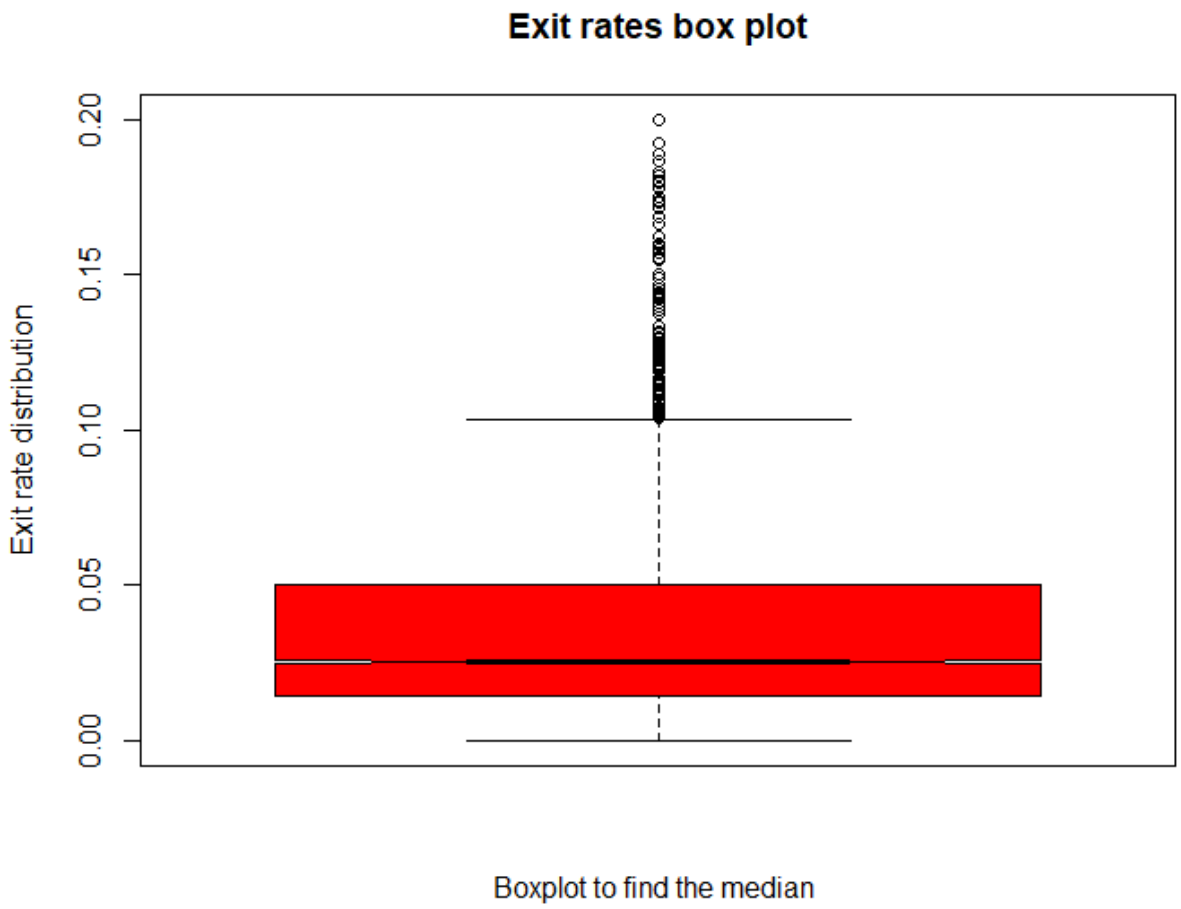

Box plot
Upper hinge
Whisker
Upper
quartile
Median
Lower
quartile
Whisker
Lower hinge

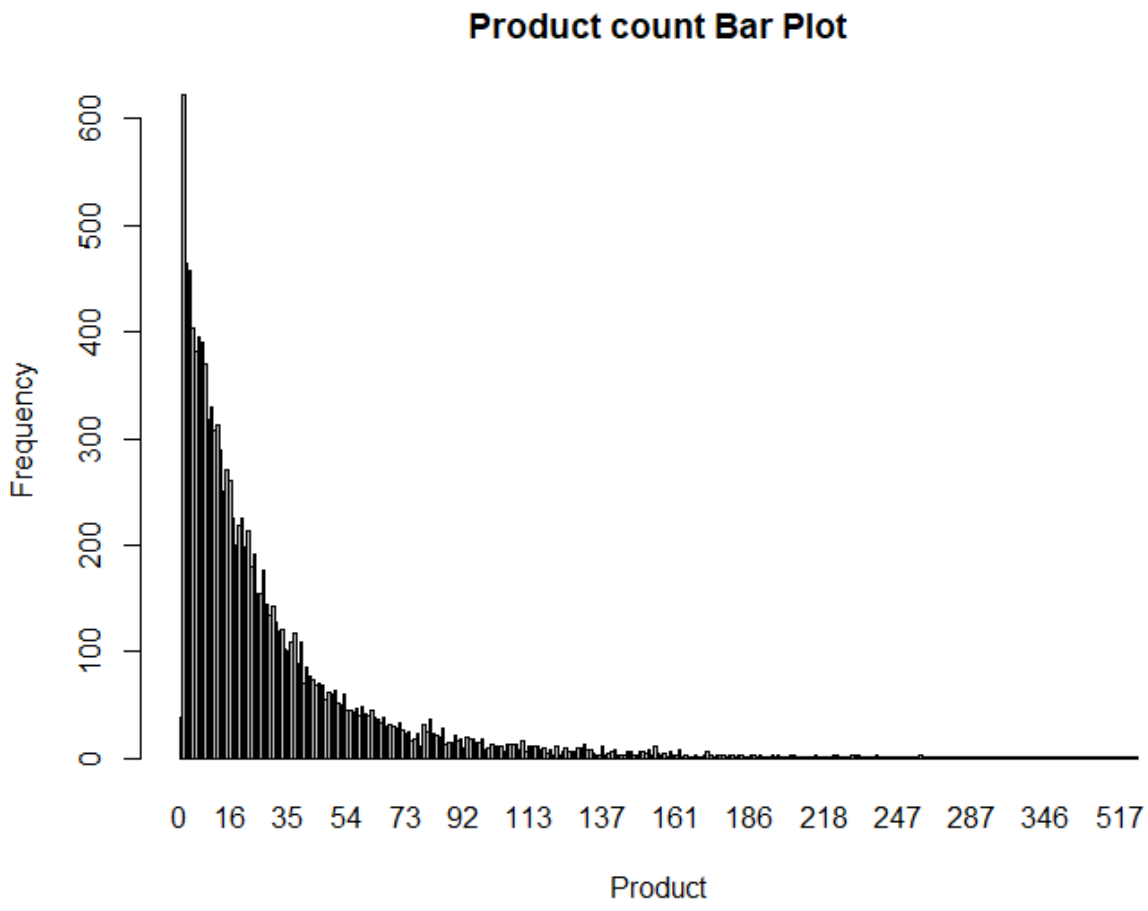
Fig 5:



Interpretation for Fig 5: As you can see in the box-plot visualization, there are many points in the whisker and above the Upper Hinge. These represent the spread. These can be treated as outliers and excluded from the analysis. However deleting from the model building process or training process is recommended only when the data is so large ex: Big Data systems.

Interpretation for Fig 6: ***Certain Products are way more popular*** than others.

Fig 6:



Interpretation for Fig 7:

Classification using Logistic Regression:

As we can observe below, the attempt to classify the online shoppers using Logistic Regression results is good accuracy classifier of 82.13 %. That's the classification performance on un-seen data.

Identifying the shoppers which don't really intent to buy (i.e. FALSE revenue) are important to business because Marketeers can follow up with these visitors to expand their businesses.

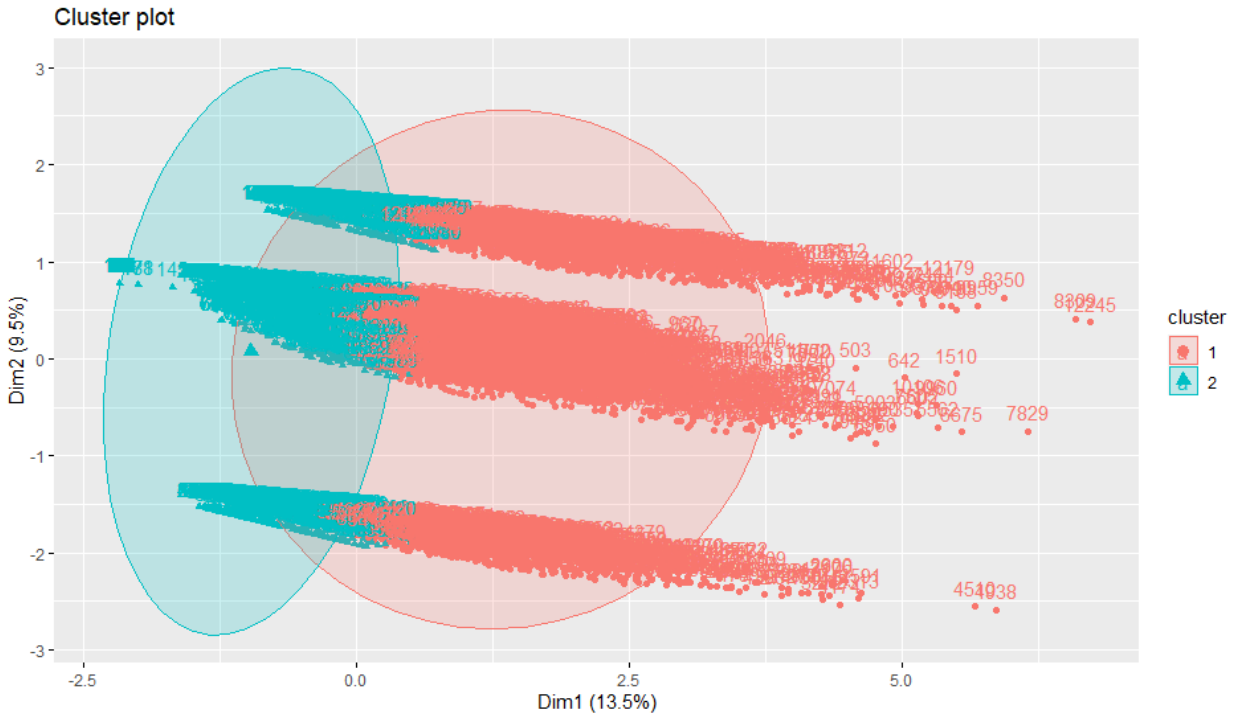
Fig 7:

```
Confusion Matrix and Statistics

       y_pred
        FALSE TRUE
  FALSE  3030   95
  TRUE    566    8

               Accuracy : 0.8213
                 95% CI : (0.8086, 0.8335)
    No Information Rate : 0.9722
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0248

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.84260
            Specificity : 0.07767
         Pos Pred Value : 0.96960
         Neg Pred Value : 0.01394
             Prevalence : 0.97215
         Detection Rate : 0.81914
   Detection Prevalence : 0.84482
      Balanced Accuracy : 0.46014

       'Positive' Class : FALSE
```

Fig 8:

Classification using K-means clustering:

```
Confusion Matrix and Statistics

       y_pred
        FALSE TRUE
  FALSE  6503 3919
  TRUE    599 1309

               Accuracy : 0.6336
                 95% CI : (0.625, 0.6421)
    No Information Rate : 0.576
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1812

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9157
            Specificity : 0.2504
         Pos Pred Value : 0.6240
         Neg Pred Value : 0.6861
             Prevalence : 0.5760
         Detection Rate : 0.5274
   Detection Prevalence : 0.8453
      Balanced Accuracy : 0.5830

       'Positive' Class : FALSE
```



Interpretation for Fig 9 & 10:

Why clustering is not a good fit?

As you can see from the following there is no discrete groups that can be used to be used as a classifier.

PCA scatter plot visualizations:

Fig 9:

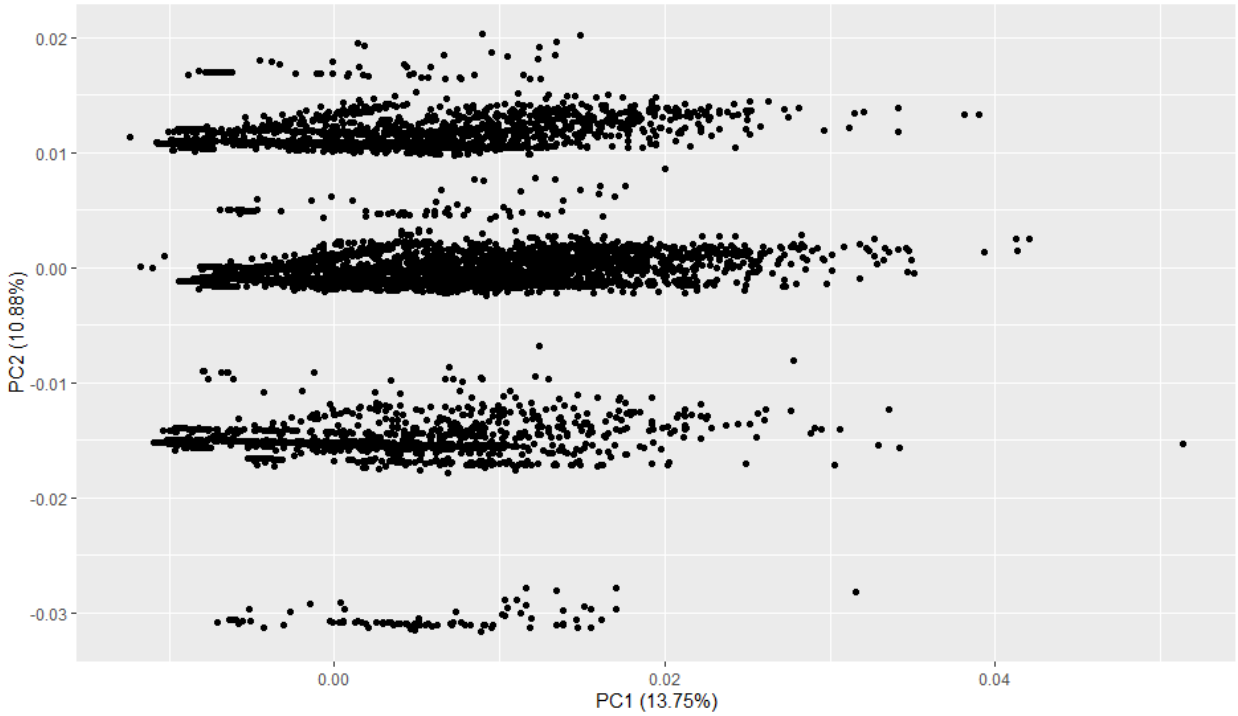
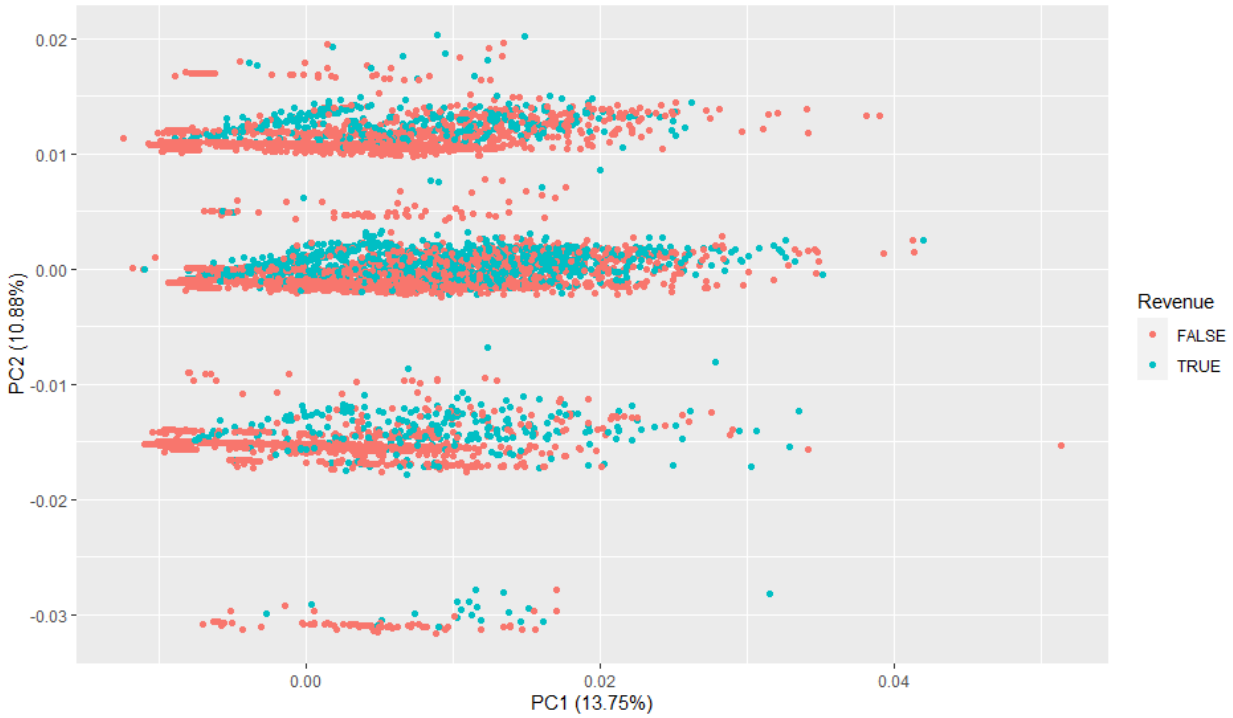


Fig 10:



Decision Tree:

Fig

```
Confusion Matrix and Statistics

          predict_unseen
            FALSE TRUE
  FALSE  2917  208
  TRUE    253  321

                Accuracy : 0.8754
                  95% CI : (0.8643, 0.8859)
    No Information Rate : 0.857
    P-Value [Acc > NIR] : 0.0006319

                   Kappa : 0.509

  Mcnemar's Test P-Value : 0.0404343

             Sensitivity : 0.9202
             Specificity : 0.6068
          Pos Pred Value : 0.9334
          Neg Pred Value : 0.5592
              Prevalence : 0.8570
          Detection Rate : 0.7886
    Detection Prevalence : 0.8448
       Balanced Accuracy : 0.7635

        'Positive' Class : FALSE
```

Interpretation for fig :

3 components displayed in visualization:

1. The higher probabilty class of either of the label FALSE or TRUE.
2. The actual probability of class TRUE is mentioned.
3. The percentage of data left at each node.

The root node indicated by 1) : The higher Probability class is FALSE as we know the 15% of total number of instances in the data. The FALSE number of instances are 8631 & TRUE instances is 1334.

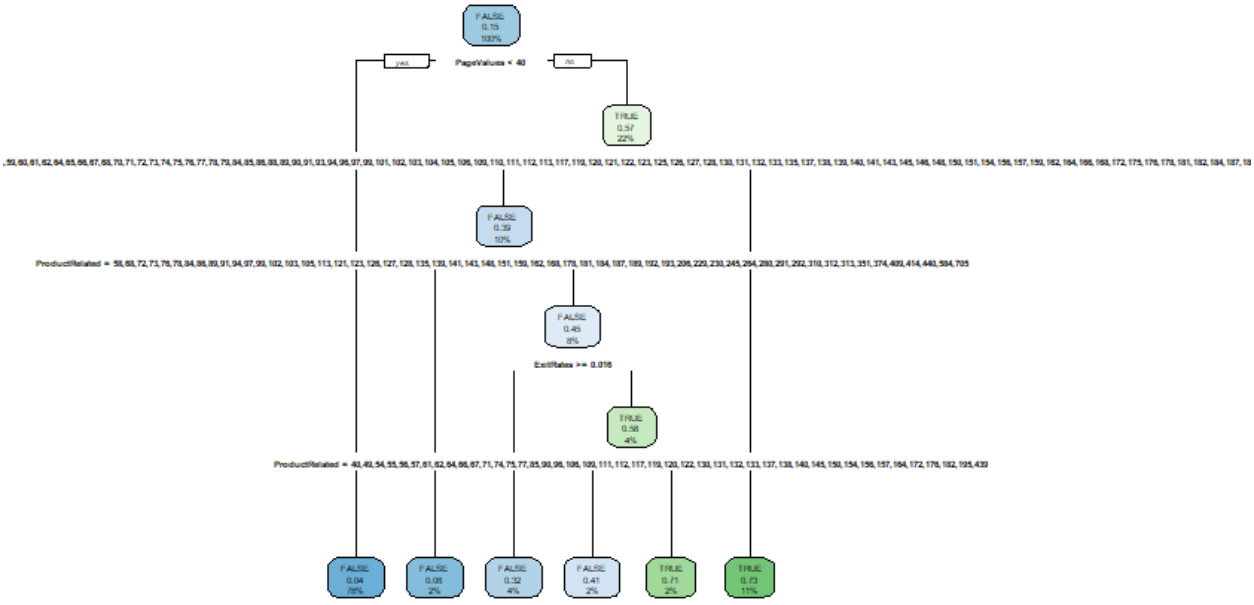
If ***PageValues*** is less than 39.5, its marked as FALSE, with probability of being FALSE is 96%.

If not, a further question is posed on ***ProductRelated*** variable equal to the list as shown below.

Fig :

```
n= 8631

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 8631 1334 FALSE (0.84544085 0.15455915)
   2) PageValues< 39.5 6769  269 FALSE (0.96026001 0.03973999) *
   3) PageValues>=39.5 1862  797 TRUE (0.42803437 0.57196563)
     6) ProductRelated=6,26,27,33,35,36,38,39,40,41,42,43,45,46,47,48,49,50,51,52,54,55,56,57,58,59,60,
61,62,64,65,66,67,68,70,71,72,73,74,75,76,77,78,79,84,85,86,88,89,90,91,93,94,96,97,99,101,102,103,104,
105,106,109,110,111,112,113,117,119,120,121,122,123,125,126,127,128,130,131,132,133,135,137,138,139,14
0,141,143,145,146,148,150,151,154,156,157,159,162,164,166,168,172,175,176,178,181,182,184,187,189,192,1
93,195,206,229,230,245,264,280,291,292,310,312,313,351,374,409,414,439,440,584,705 873  341 FALSE (0.60
939290 0.39060710)
      12) ProductRelated=58,68,72,73,76,78,84,86,89,91,94,97,99,102,103,105,113,121,123,126,127,128,13
5,139,141,143,148,151,159,162,168,178,181,184,187,189,192,193,206,229,230,245,264,280,291,292,310,312,3
13,351,374,409,414,440,584,705 140   11 FALSE (0.92142857 0.07857143) *
      13) ProductRelated=6,26,27,33,35,36,38,39,40,41,42,43,45,46,47,48,49,50,51,52,54,55,56,57,59,60,6
1,62,64,65,66,67,70,71,74,75,77,79,85,88,90,93,96,101,104,106,109,110,111,112,117,119,120,122,125,130,1
31,132,133,137,138,140,145,146,150,154,156,157,164,166,172,175,176,182,195,439 733  330 FALSE (0.549795
36 0.45020464)
        26) ExitRates>=0.016419 367  117 FALSE (0.68119891 0.31880109) *
        27) ExitRates< 0.016419 366  153 TRUE (0.41803279 0.58196721)
          54) ProductRelated=40,49,54,55,56,57,61,62,64,66,67,71,74,75,77,85,90,96,106,109,111,112,117,
119,120,122,130,131,132,133,137,138,140,145,150,154,156,157,164,172,176,182,195,439 157   64 FALSE (0.5
9235669 0.40764331) *
          55) ProductRelated=6,26,27,33,35,36,38,39,41,42,43,45,46,47,48,50,51,52,59,60,65,70,79,88,93,
101,104,146,166 209   60 TRUE (0.28708134 0.71291866) *
     7) ProductRelated=1,2,3,4,5,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,28,29,30,31,32,3
4,37,44,53,63,69,80,81,82,83,87,92,95,98,100,107,108,114,115,116,118,124,129,134,136,142,149,152,153,15
5,160,161,163,165,170,171,173,183,194,198,200,202,216,218,219,221,225,233,237,238,243,248,261,276,318,3
24,346,357,359,397,401,470,501,517 989  265 TRUE (0.26794742 0.73205258) *
```

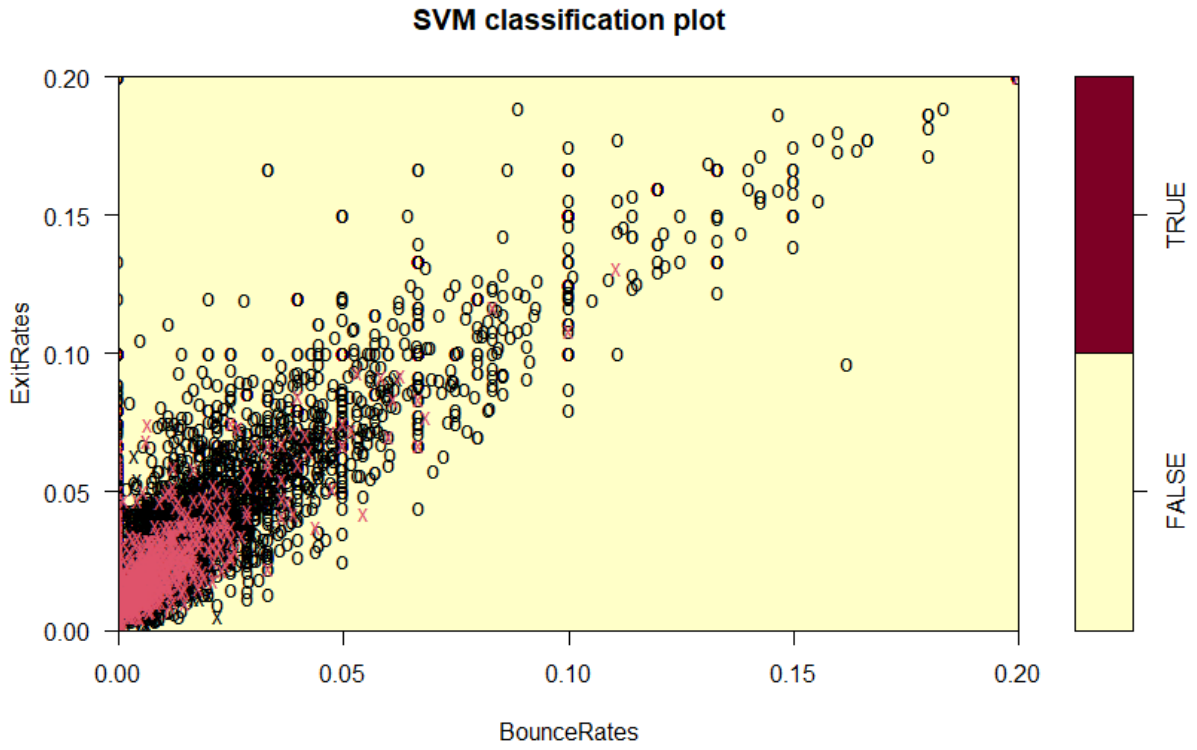
Fig



SVM:

Interpretation for fig : SVM develops support vectors in order to classify the label variable i.e. Revenue. The figure visualizes the support vectors developed across any two numerical variables. The major premise is FALSE label as it is a dominant class.

Fig


SVM classification plot
ExitRates
BounceRates
0.00
0.05
0.10
0.15
0.20
TRUE
FALSE

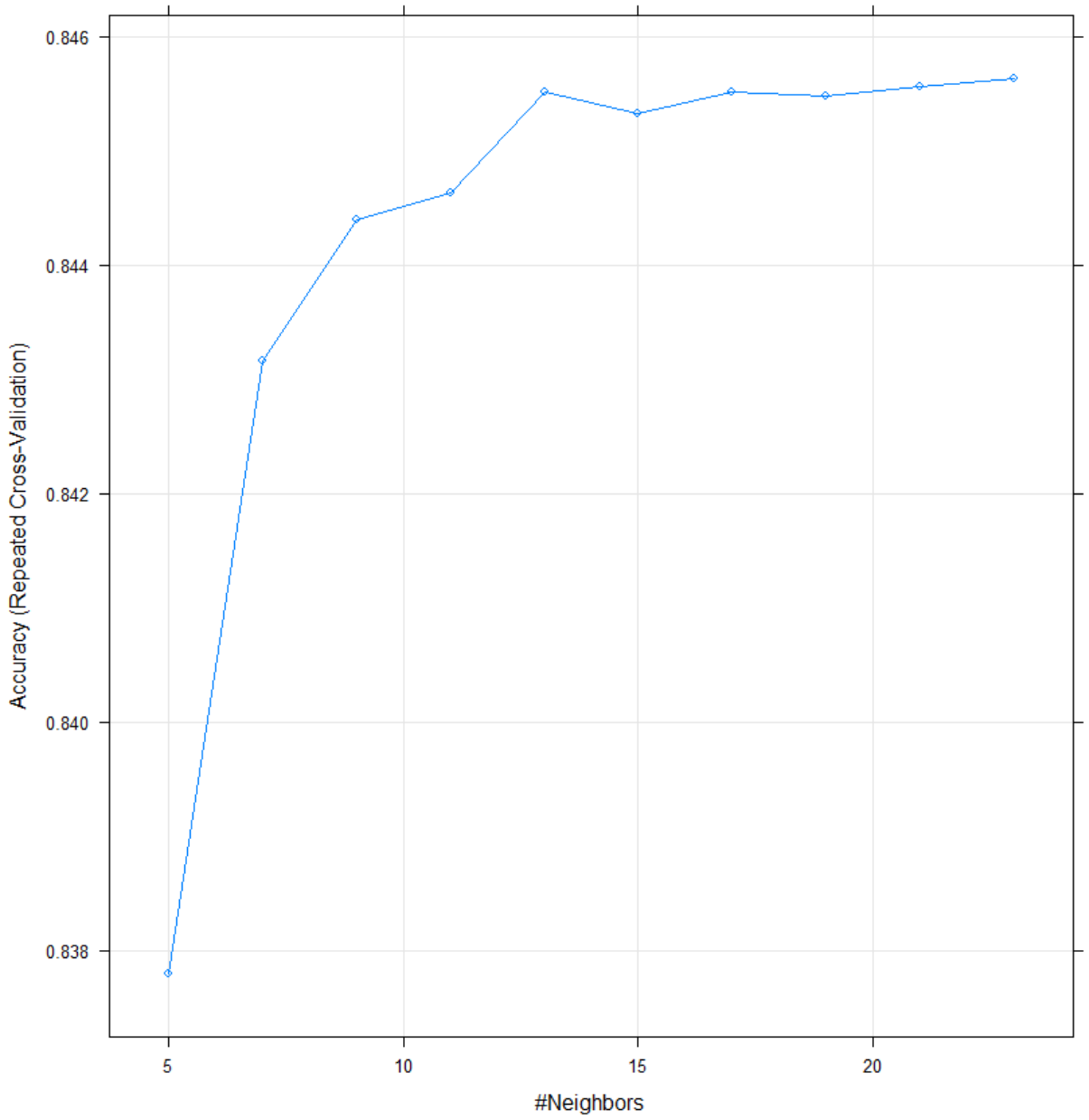
Fig

```
Confusion Matrix and Statistics

          Actual
Predicted FALSE TRUE
    FALSE  2936  313
    TRUE    189  261

               Accuracy : 0.8643
                 95% CI : (0.8528, 0.8752)
    No Information Rate : 0.8448
    P-Value [Acc > NIR] : 0.0004837

                  Kappa : 0.4323

 Mcnemar's Test P-Value : 4.025e-08

            Sensitivity : 0.9395
            Specificity : 0.4547
         Pos Pred Value : 0.9037
         Neg Pred Value : 0.5800
             Prevalence : 0.8448
         Detection Rate : 0.7937
   Detection Prevalence : 0.8783
      Balanced Accuracy : 0.6971

       'Positive' Class : FALSE
```

KNN

Accuracy is optimum at K = 23 of around 84.5% . This implies the 23 nearest neighbours from the training data and chooses the class that belongs to most of the 23 nearest neighbours identified.

Fig


0.846
0.844
0.842
0.840
0.838
Accuracy (Repeated Cross-Validation)
5
10
15
20
#Neighbors

Fig

```
Confusion Matrix and Statistics

          Reference
Prediction FALSE TRUE
     FALSE  3122  574
     TRUE      3    0

               Accuracy : 0.844
                 95% CI : (0.8319, 0.8556)
    No Information Rate : 0.8448
    P-Value [Acc > NIR] : 0.5651

                  Kappa : -0.0016

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9990
            Specificity : 0.0000
         Pos Pred Value : 0.8447
         Neg Pred Value : 0.0000
             Prevalence : 0.8448
         Detection Rate : 0.8440
   Detection Prevalence : 0.9992
      Balanced Accuracy : 0.4995

       'Positive' Class : FALSE
```

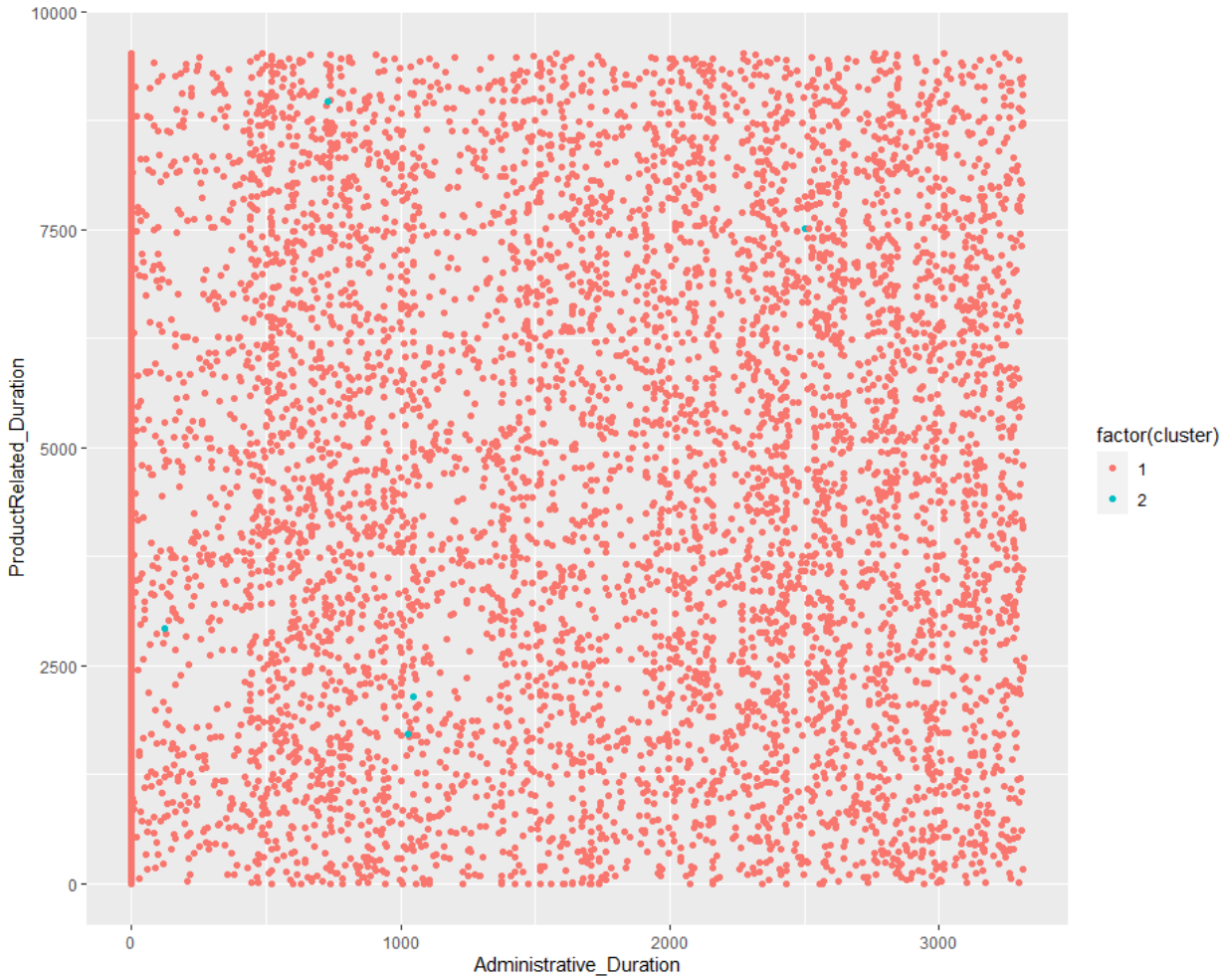
Hierarchical Clustering

```
Confusion Matrix and Statistics

       FALSE  TRUE
FALSE 10415  1905
TRUE      7     3

               Accuracy : 0.8449
                 95% CI : (0.8384, 0.8513)
    No Information Rate : 0.8453
    P-Value [Acc > NIR] : 0.5457

                  Kappa : 0.0015

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.999328
            Specificity : 0.001572
         Pos Pred Value : 0.845373
         Neg Pred Value : 0.300000
             Prevalence : 0.845255
         Detection Rate : 0.844688
   Detection Prevalence : 0.999189
      Balanced Accuracy : 0.500450

       'Positive' Class : FALSE
```

The number of instances belonging to two clusters is as below:

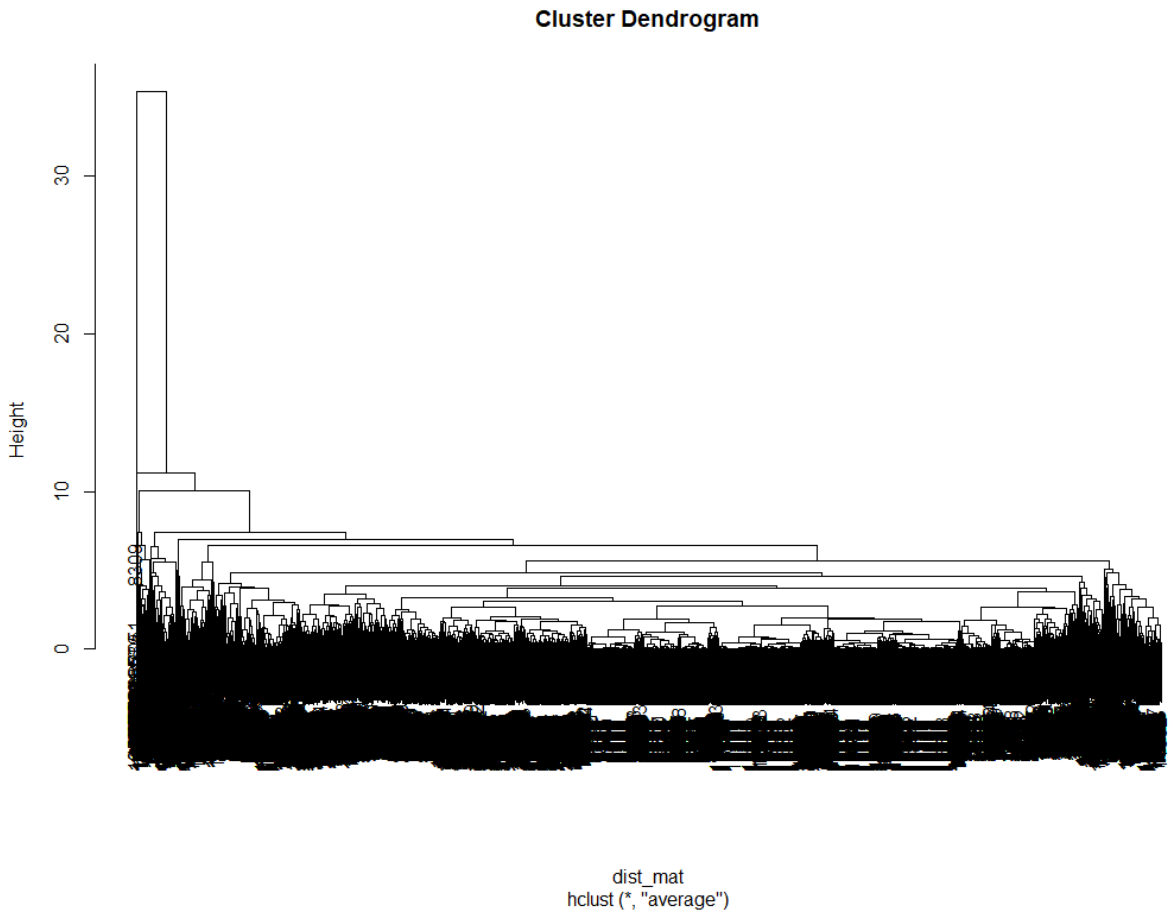
```
> count(seeds_df_cl,cluster)
   cluster     n
1:       1 12320
2:       2    10
```

No discrete groups are formed. One group is much more dominant than other.



Dendogram:

The y axis is the distance values (Ex Eucledian distance).


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