Abstract:

I worked with two data sets Vatten Fall and Online Shoppers Intention with both Python and R. In **Vatten Fall** I used Linear regression to predict the next failure of various components. For **Online Shoppers Intention** to make a decision on purchase a product(Will buy or not), so I used clustering and classification.

1 Vatten Fall

1.1 Analysis Report of vattenfall predictive maintenance

This project focuses on predicting the failure of various components of turbines located across 5 countries (abbreviated as UK, GE, DK, NL, SE) including on shore as well as off shore turbines.

The components that are analysed in the data are blade, gearbox, generator, Blade and main bearing.

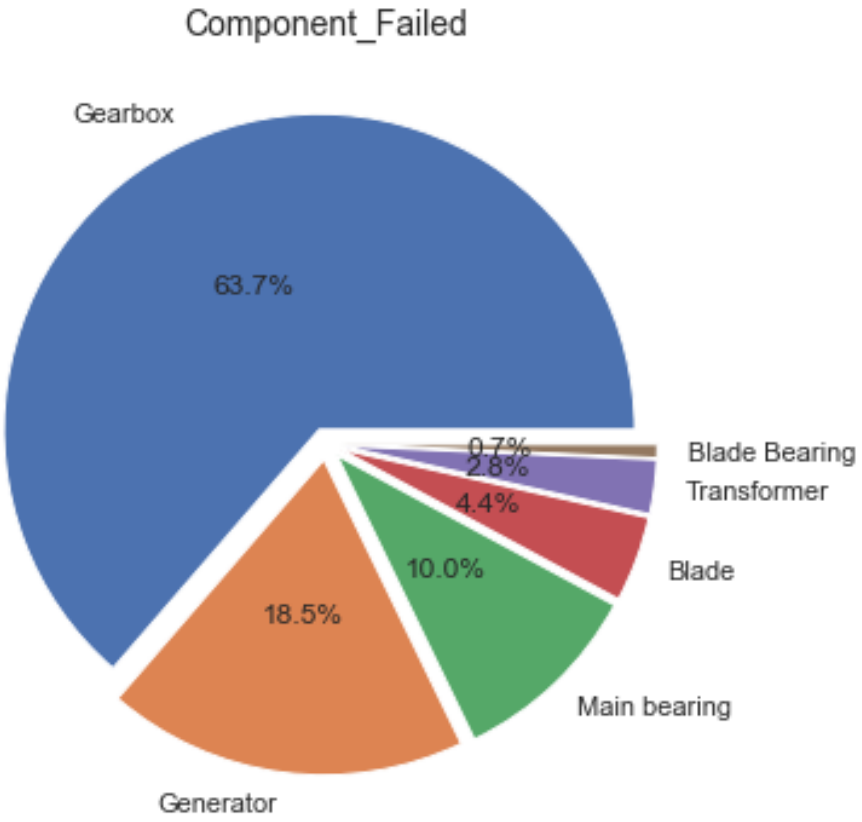

Figure 1: It's show how many failure of every component in percent sign

This data poses a Predictive Maintenance use-case. Every machine or components come with limited life. There can be various reasons for a failure of a component or reduced life of a component. Knowing beforehand when a given component would fail or in other words knowing remaining useful life of components in production line is of great importance.

Remaining Useful Life (RUL) states the duration of a component to reach its failure [1]. By taking RUL into account, engineers can schedule predictive maintenance, optimize operating efficiency, and avoid unplanned downtime. For this reason, estimating RUL is a top priority in the predictive maintenance program.

There are total 450 failure events recorded in the Vattenfall data.

Project commissioning is the process of assuring that all systems and components of a building or industrial plant are designed, installed, tested, operated, and maintained according to the operational requirements of the owner or final client[2]. For simplicity, we take commissioning date as the date the turbine is operational.

Component Exchange Date signifies the date till which the component fails, since exact date of failure is not available, assuming the components are replaced immediately after failure.

The difference between the Component Commissioning Date and the Component Exchange Date gives us approximately the total number of days the component worked before failing i.e. RUL of each component.

Limitations of the Data:

1. The data has too many missing values.
2. There is no value indicating the performance deterioration of the components over time. Alarm Code is one indicator close to gauging the performance of the components, however alarm code is absent in majority of the data points.
3. The Avg RUL is significantly different by country, manufacturer, component. All are used as inputs to predicting the RUL of the component.

The data is disproportional by number of training examples from different manufacturer. The manufacturer is one of the values analysed to predict the life of the component. This negatively impacts the generalizing the learning when there are more examples of only one kind and very less examples of the other.

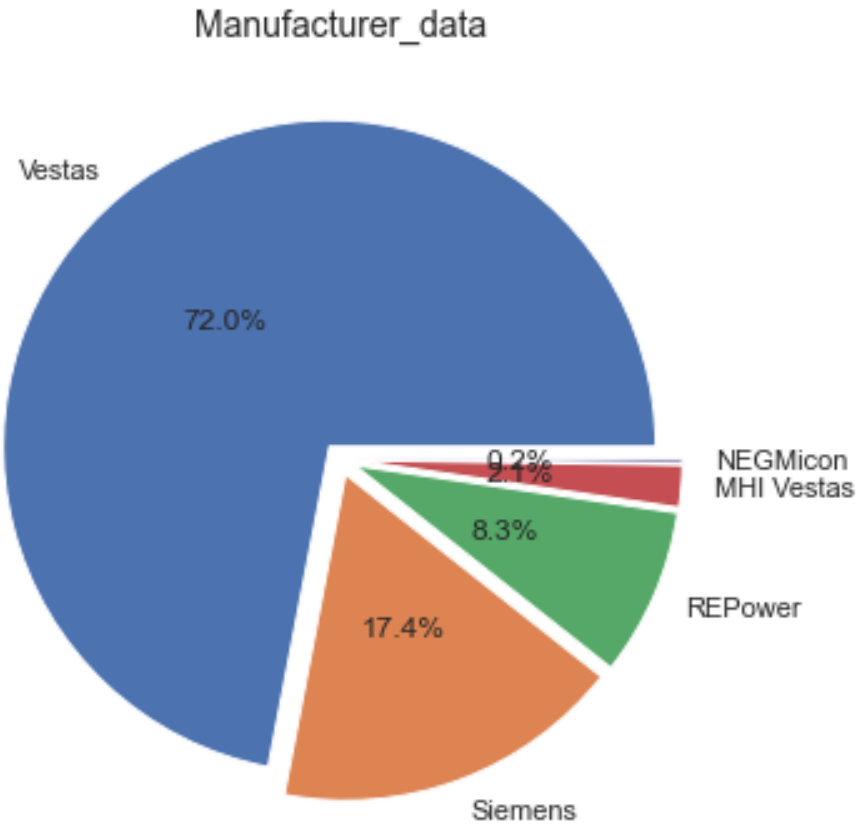

Figure 2: It's show the count of Manufacturer in percent sign

1.2 Missing data mappings

The available data has large missing attributes. Here is an overview of missing values in the data.

```
Your selected dataframe has 42 columns.
There are 33 columns that have missing values.
```

	Missing Values	% of Total Values
Alarm code 20	449	100.0
Alarm code 19	449	100.0
Alarm code 18	448	99.8
Alarm code 17	447	99.6
Alarm code 16	447	99.6
Alarm code 15	447	99.6
Alarm code 14	447	99.6
Alarm code 13	445	99.1
Alarm code 12	445	99.1
Alarm code 11	445	99.1
Alarm code 10	444	98.9
Alarm code 9	443	98.7
Alarm code 8	443	98.7
Alarm code 7	442	98.4
Alarm code 6	440	98.0
Alarm code 5	439	97.8
Alarm code 4	435	96.9
Alarm code 3	431	96.0
Alarm code 2	395	88.0
Jackup Call-off Date	387	86.2
Work Order	376	83.7
Outage cause	364	81.1
Jackup vessel	361	80.4
Component Commisioning Date	360	80.2
Turbine start date	350	78.0
Alarm code 1	338	75.3
Material number	304	67.7
Failure Mode	301	67.0
Turbine Stopp Date	268	59.7
Failed Component Serial number	234	52.1
New Component Serial number	114	25.4
Turbine	13	2.9
Estimated Lost Revenue	8	1.8

Figure 3: It's show the number of missing values

The data is spread into two sheets:

1. DATABASE i.e. Sheet_1
2. Turbine Data i.e. Sheet_2

The missing component commissioning date of Sheet_1 can be mapped to the corresponding commission year of the turbines provided in Sheet_2. Since only commissioning year is mentioned, the day and month is assumed to be 01-January-YEAR.

All the turbines belonging to the farms Nørrekær Enge, Kentish Flats, Princess Alexia, Stor Rotliden have same rest of the attributes (such as 'Manufacturer', 'Turbine Type', 'Rotor Diameter', 'Hub Height', 'Installed Power', 'Latitude', 'Longitude'). Hence these turbines can be given a dummy name to map with corresponding values in Sheet_2.

```
# Average RUL by Manufacturer
Avg_RUL_Manufacture= data_prepared.groupby('Manufacturer').mean()['RUL']
Avg_RUL_Manufacture
```

```
Manufacturer
MHI Vestas     896.111111
NEGMicon      2983.000000
REPower       1753.611111
Siemens       2294.853333
Vestas        2574.000000
Name: RUL, dtype: float64
```

Figure 4: It's show

The empty columns (100% missing) or beyond possibility of imputing (say 85% missing) are dropped.

The cost involved due to the component failure have no role in analysis. Those info are also kept away from the model building.

1.3 Visualization via Clustering

To visualize the data, reducing the dimensionality of the data can help.

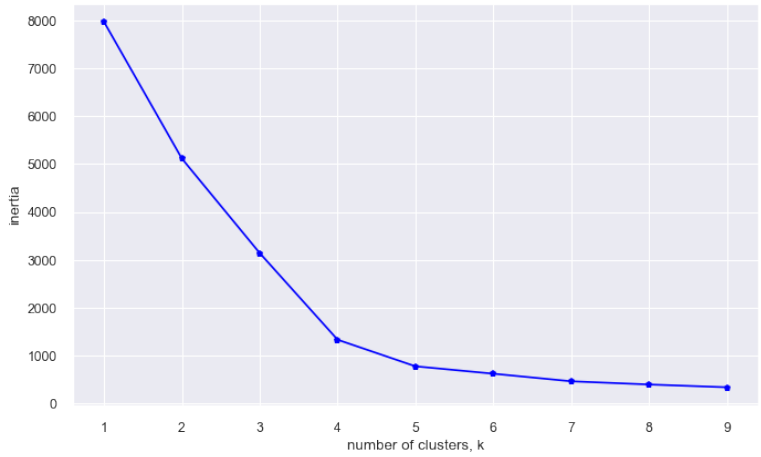

Figure 5: It's show how the optimal number of clusters via Elbow Method is 4.

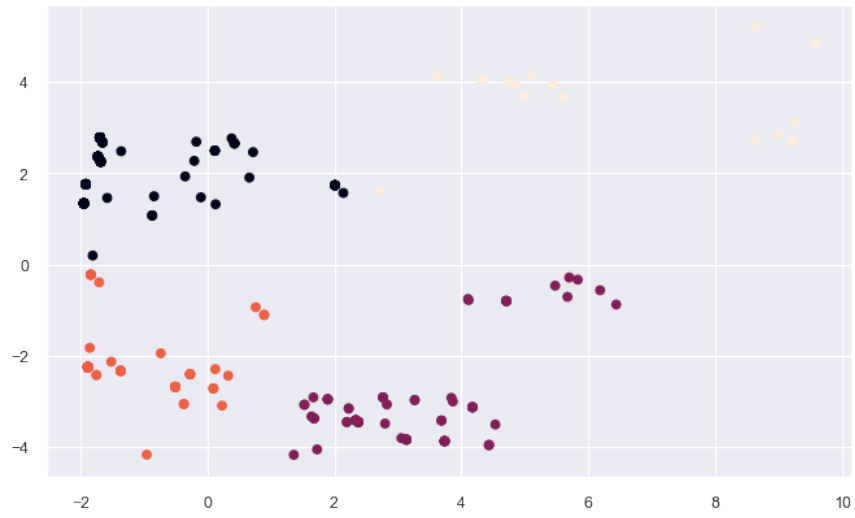

Figure 6: It's show how the 4 clusters are recognized.

Model highlights: Performance measurement benchmark: R2_score

R^2 (coefficient of determination) regression score function.

Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R^2 score of 0.0.

1.4 Regression Models

Decision tree resulted in performance is just 5%. Hence we can discard the model.

KNN regressor fits with 79% R2_score (0.79), which could correctly predict RUL values for most of the data. Here is sample plotting for 20 data points.

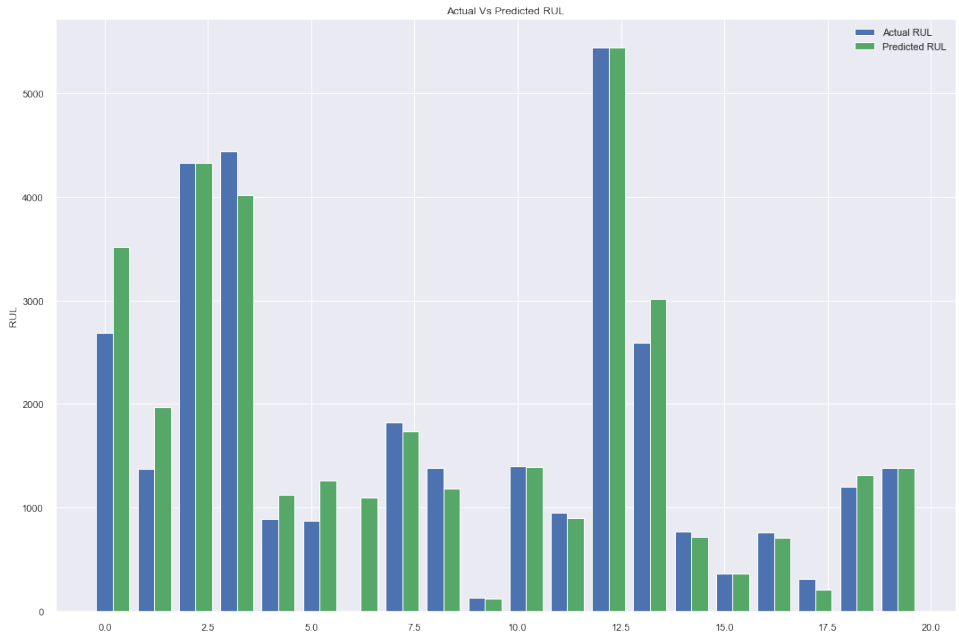

Figure 7: It's show how the predicted values are close to the actual values.

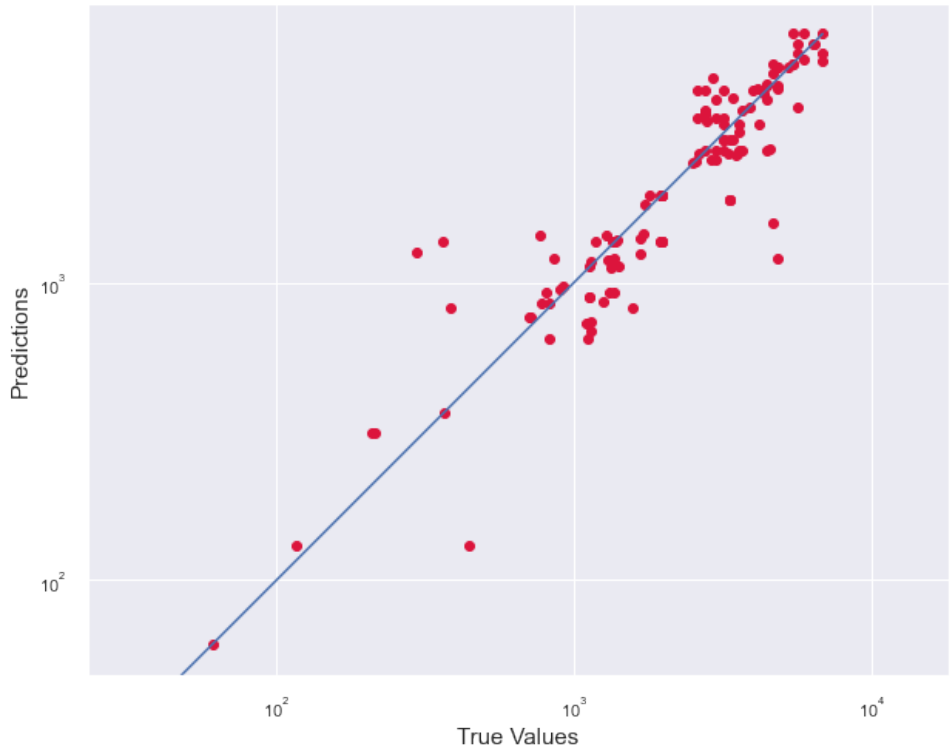

Figure 8: It's show scatter plot which see how our predicted data is different from our True Values.

2 Online Shoppers Purchasing Intention

2.1 Abstract

We are using the "Online Shoppers Purchasing Intention" UCI dataset related to visits of customers to an online store and their decision on purchase a product(s). The goal of the analysis is use clustering algorithms and classification algorithms.

The dataset used is based on "Online Shoppers Purchasing Intention" UCI dataset , (detailed description at: `https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset`)

2.2 Introduction

we will analyze the activity of users who vist a specific product offered online through a website. The objective is to predict if the user will buy or not.

2.3 Data Set Information

The dataset consists of feature vectors belonging to 12,330 sessions.

The dataset was formed so that each session would belong to a different user in a 1-year period to avoid any tendency to a specific campaign, special day, user profile, or period.

2.3.1 Attribute Information

The dataset consists of 10 numerical and 8 categorical attributes. The 'Revenue' attribute can be used as the class label.

"Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

FEATURES:

- **Administrative Duration, Informational, Informational Duration, Product Related and Product Related Duration**: represent the number of pages visited by the user in that session and total time spent in each page.
- **Bounce Rate**: percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytic server during that session.
- **Exit Rate**: the percentage that were the last in the session
- **Page Value**: feature represents the average value for a web page that a user visited before completing an e-commerce transaction
- **Special Day**: indicates the closeness of the site visiting time to a specific special day (e.g. Mother's Day, Valentine's Day)
- **Operating system,browser, region, traffic type**: Different types of operating systems, browser, region and traffic type used to visit the website

- **Visitor type**: Whether the customer is a returning or new visitor
- **Weekend**: A Boolean value indicating whether the date of the visit is weekend
- **Month**: Month of the year
- **Revenue**: class whether it can make a revenue or not

2.4 Exploratory Data Analysis

There are no missing values in the data.

Goal: To gauge whether a shopper is intending to buy or not. This can help marketeers to strategies for increasing the sales revenue, and we can start by asking a few questions to achieve our goals.

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 -> may could be high special days in this month

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 -> as we said before about May where it was the common month.

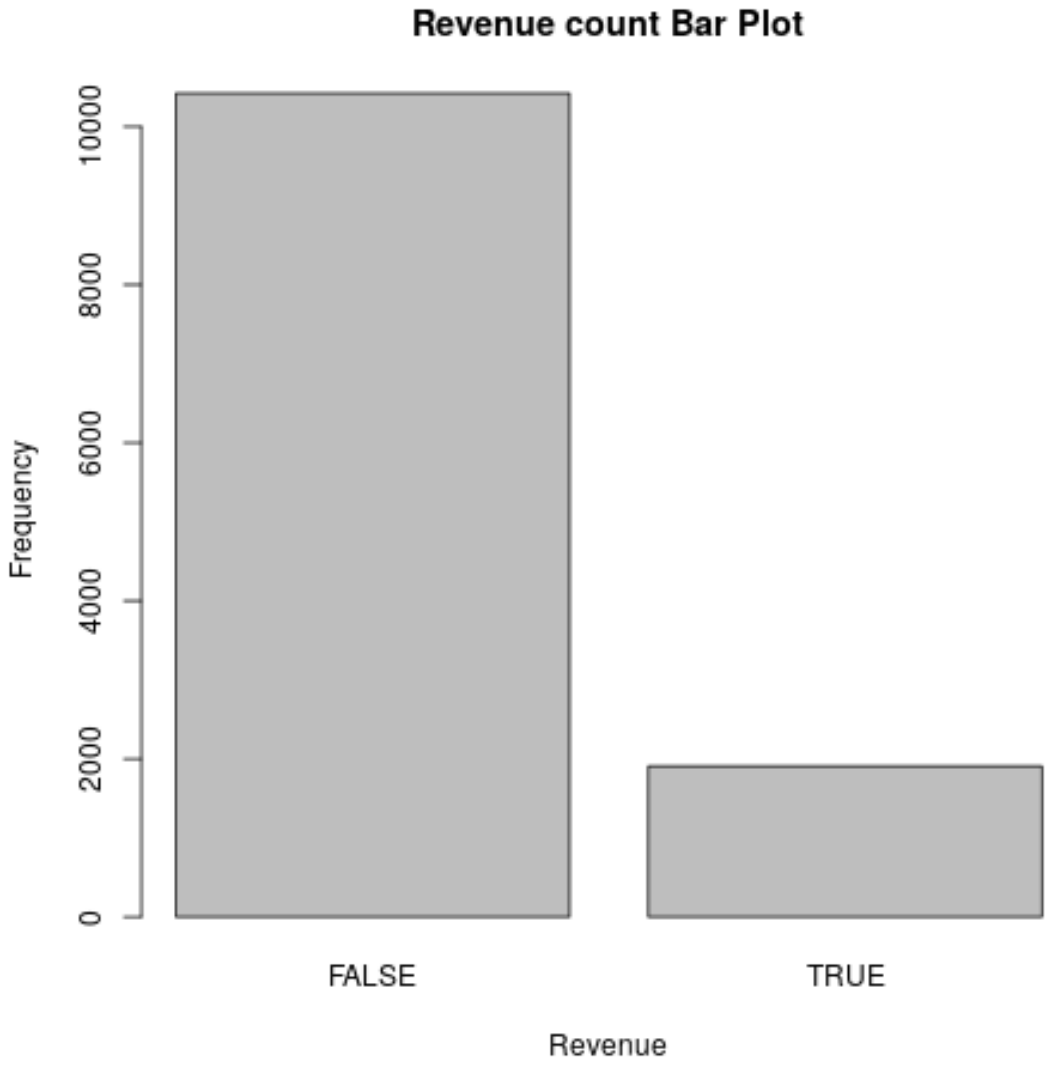

Figure 9: There are more people who only surf the site but lesser people actually buy designated by Revenue value to be True.

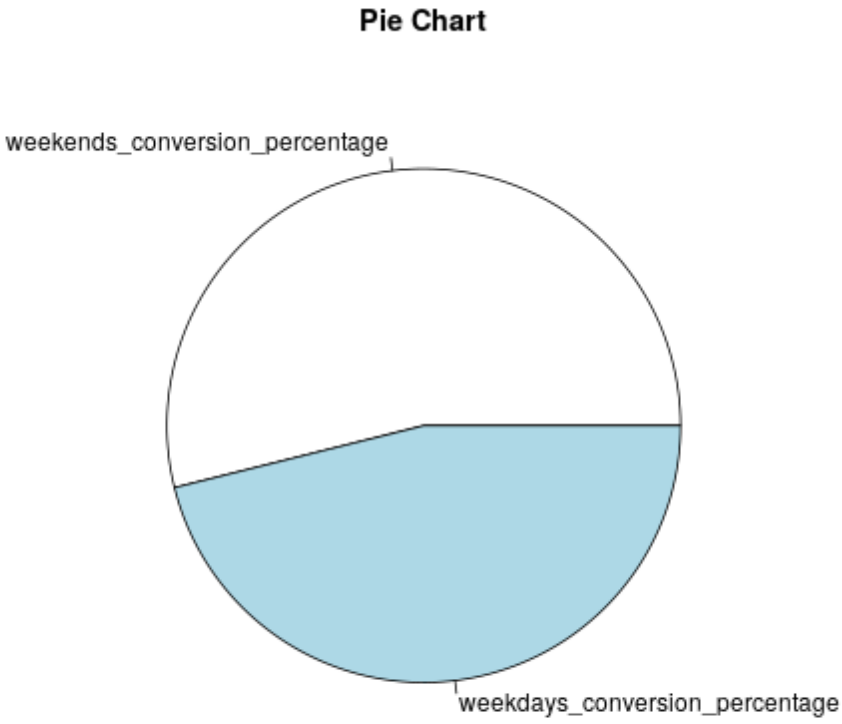

Figure 10: the conversion rate is more on weekend (almost 18%) than on weekdays (almost 15%).

Figure 11: There is much more traffic on weekdays compared to weekends.

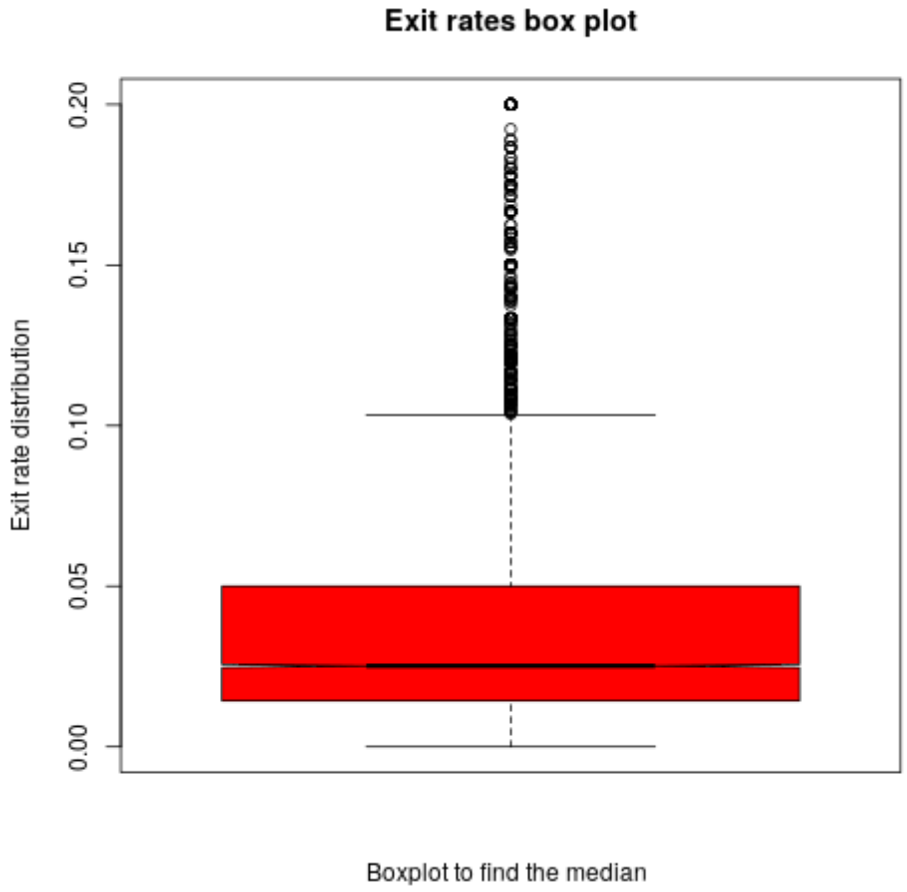

Figure 12: As you can see in the box-plot visualization, there are many points in the whisker and above the Upper Hinge. These represent the spread. These can be treated as outliers and excluded from the analysis. However deleting from the model building process or training process is recommended only when the data is so large ex: Big Data systems.

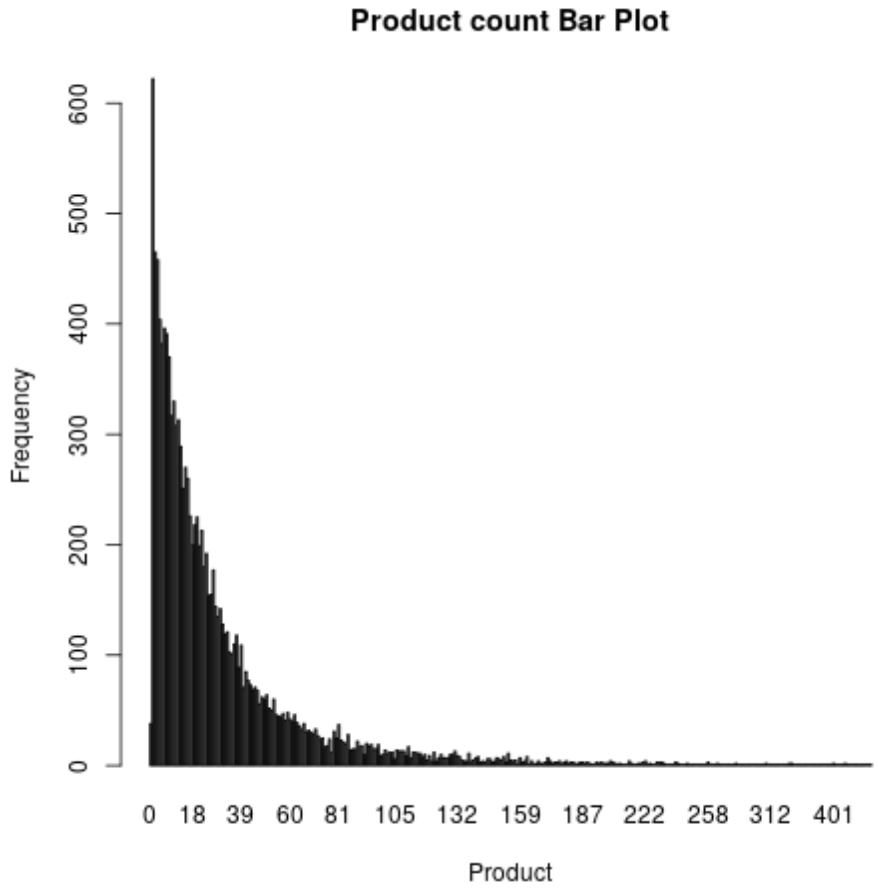

Figure 13: Certain Products are way more popular than others.

2.5 Classification

2.5.1 Classification by 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.

```
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
```

Figure 14: Confusion matrix and Statistic.

2.5.2 Decision Tree

```
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
```

Figure 15: Confusion matrix and Statistic.

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.

We include the decision tree values because it's hard to see by zooming in decision tree plot.

```
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) *
```

Figure 16: Decision tree values.

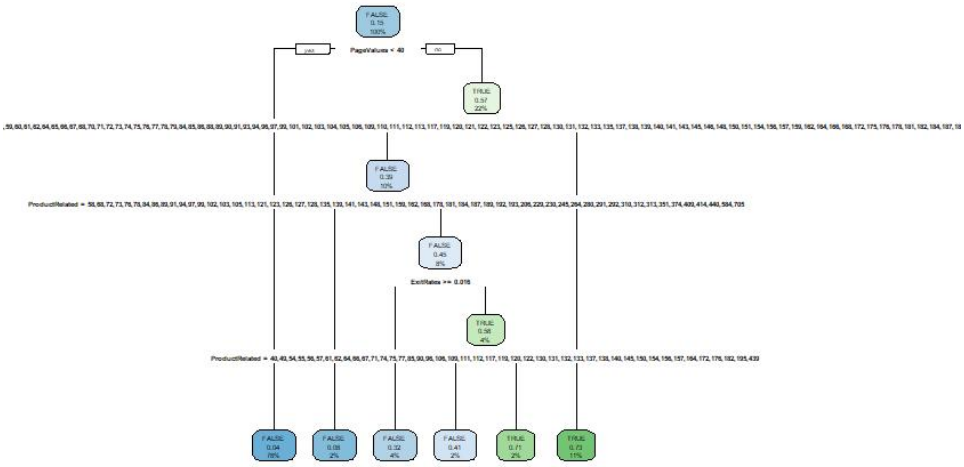

Figure 17: Decision tree plot.

2.5.3 Support Vector Machine

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.

```
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
```

Figure 18: Confusion matrix and Statistic.

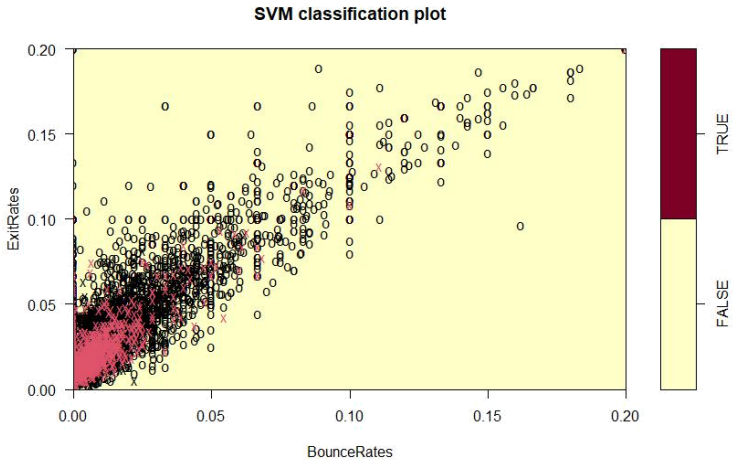

Figure 19: SVM plot.

2.5.4 K Nearest Neighbor

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.

```
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
```

Figure 20: Confusion matrix and Statistic.

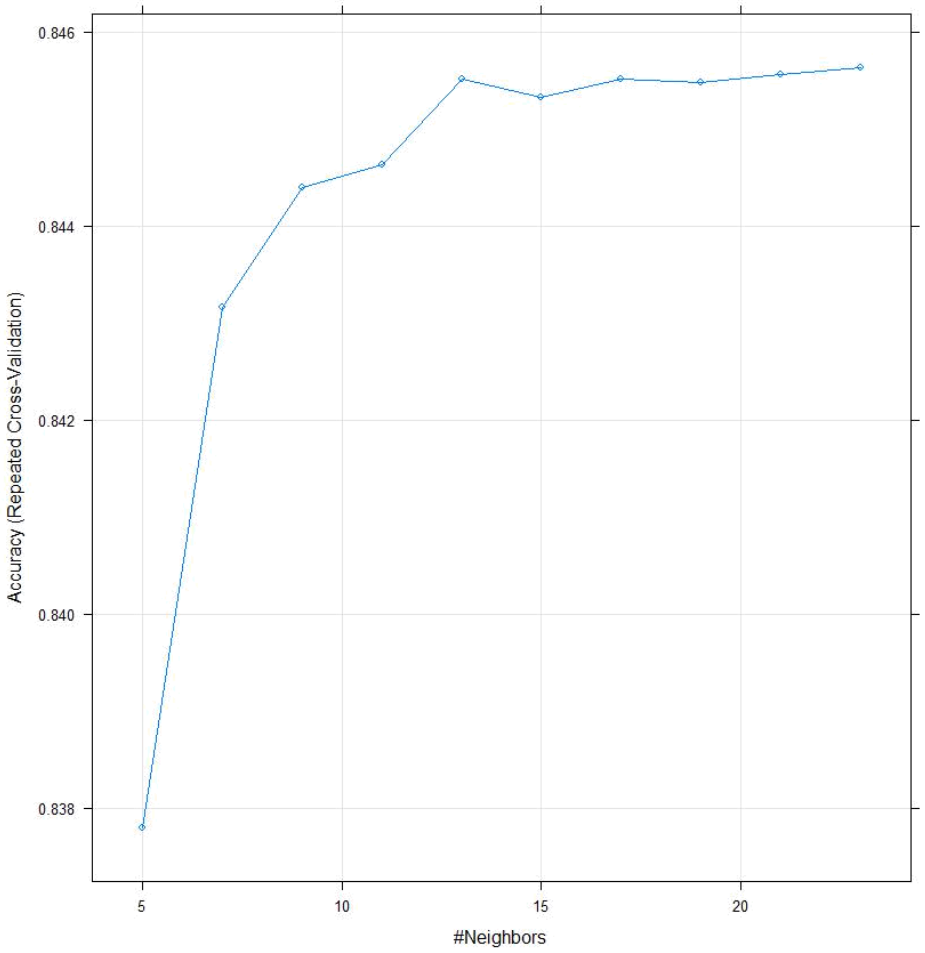

Figure 21: Accuracy(RepeatedCross-Validation.

2.6 Clustering

2.6.1 K-means clustering

Why clustering is not a good fit?

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

```
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
```

Figure 22: Confusion matrix and Statistic.

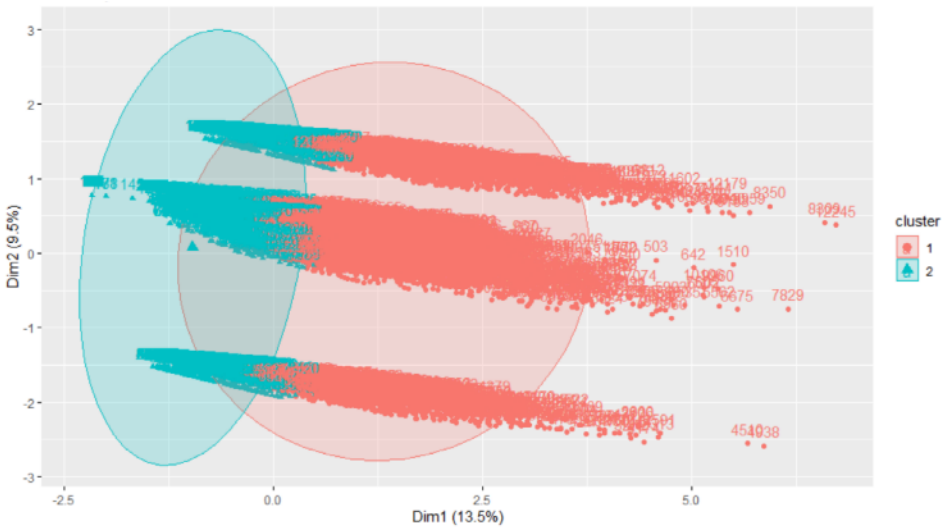

Figure 23: Cluster plot have two clusters TRUE or FALSE.

Figure 24: PCA scatter plot visualizations.

2.6.2 Hierarchical Clustering

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

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

Figure 25: Which shows the numbers of observations in each cluster.

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

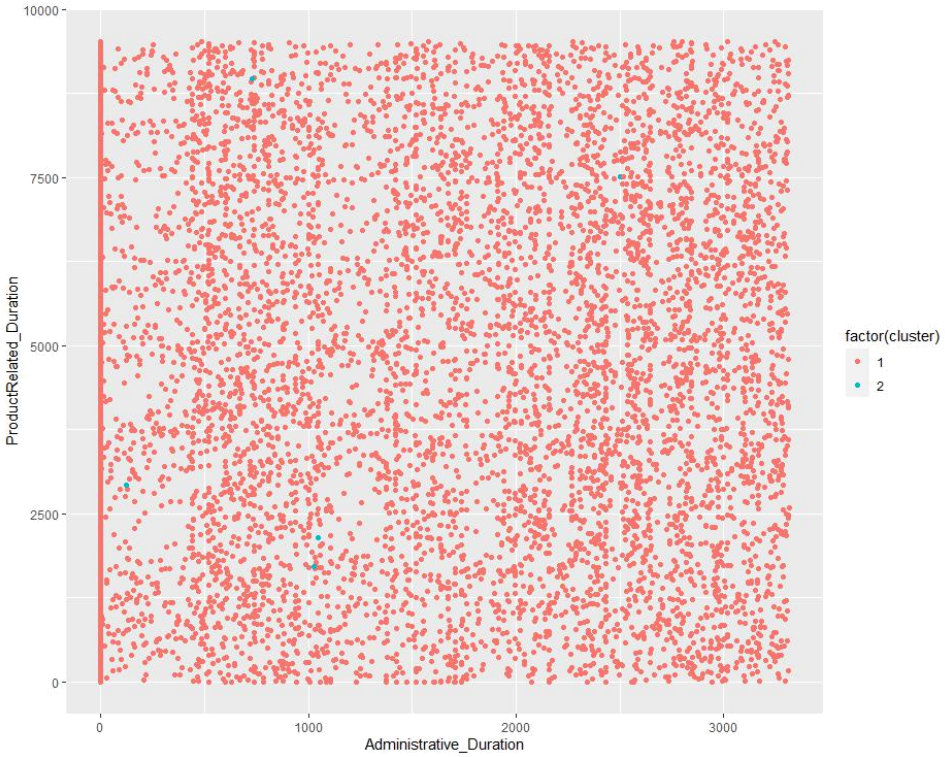

Figure 26: It's a little bit hard to see the blue points because of the huge numbers of another cluster class(red points).

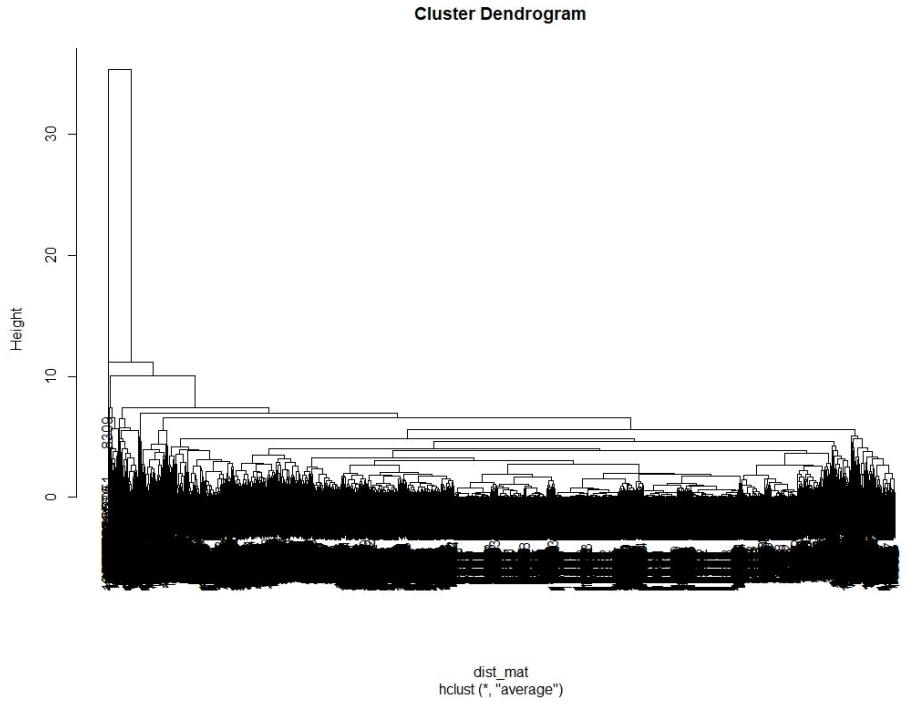

Figure 27: It's hard to see the values, because of the amount of data is enormous.