DATA MINING AND MACHINE LEARNING

END ASSESMENT- A00279933

1. DECISION TREE ALGORITHM

1.1 BUSINESS UNDERSTANDING

This dataset is referenced from the UCI Machine Learning Repository, named Car Evaluation which is about evaluating the car based upon the overall Price, Technical characteristics and the overall Comfort.

Link of the dataset: https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

1.2.DATA UNDERSTANDING AND PREPARATION

1.2.1 Data Understanding

This dataset has a categorical data of ordinal type, with 7 attributes of 1729 instances. Attributes such as Buying, Maintaining, Doors, Persons, Lug_Boot(Luggage Boot), Safety and the class of the car evaluation dataset, in this variable, where Buying variable has 4 types of character instances such as (Very high, High, Medium, Less), Maint Variable has 3 types of character instances (High, Medium, Low), Door has continuous integers and character instances(2,3,4,5-more), Person variable has the (2,4, More) as its instances, Lug_Boot variable has (Small, Medium, High) instances, Safety Variable has (High, Medium, Low) and finally the class which has the (Unacc-Unaccounted, Acc-Accounted, Very Good, Good)

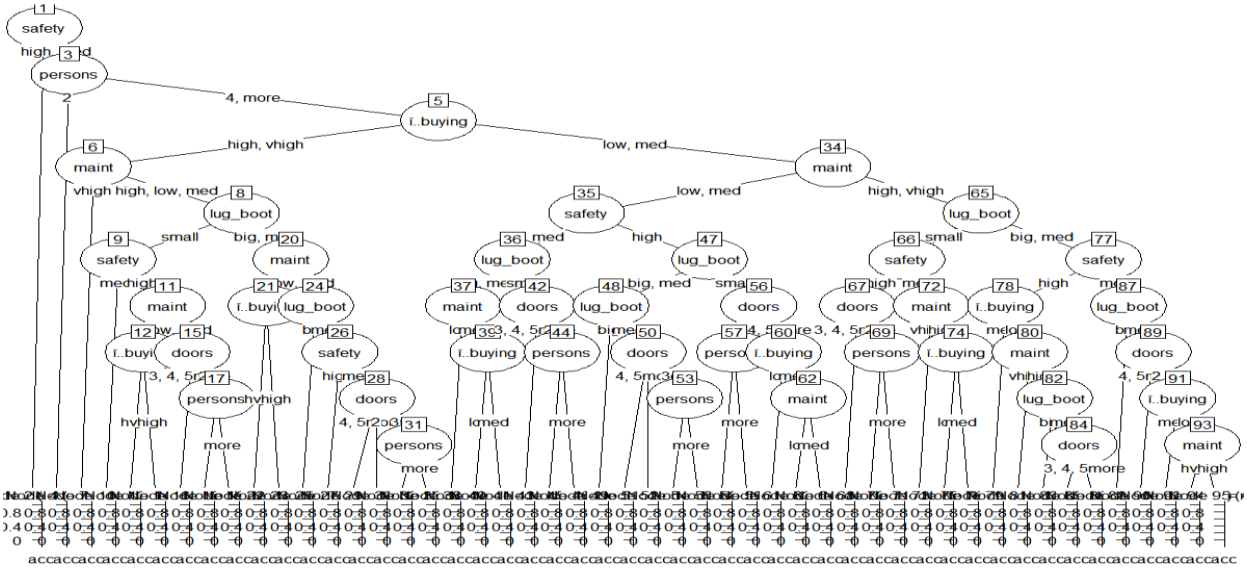
1.2.2 Preparation

As part of the dataset preparation, which needs to be pre-processed before we are importing to the R-studio, this dataset was not available in direct csv format instead as. data file and which is to be processed in spreadsheet and data processing work and cleaning the csv file has done. Importing the c50 Algorithm packages and then Splitting the dataset into 80 percentage for the train dataset and 20 percentage for the test data set and shuffling the entire dataset.

1.3. MODELLING

1.3.1 Building the Decision Tree - model 1:

Using the c50 algorithm decision tree has been generated, here is the decision tree with all node



In the training dataset of the first model, from the summary we can learn the evaluation of the training set of the model such as the total size and its class details and individual attribute usage has been clearly given.

```
        Decision Tree
      ----------------
      Size      Errors

        49   17( 1.2%)   <<


       (a)   (b)   (c)   (d)    <-classified as
      ----  ----  ----  ----
       287    10     2          (a): class acc
              56                (b): class good
         5         967          (c): class unacc
                          55    (d): class vgood


    Attribute usage:

    100.00% safety
     66.43% persons
     43.99% ï..buying
     43.99% maint
     38.57% lug_boot
     13.39% doors
```

1.3.2 Boosting the decision tree model 2:

Here the model is boosted with the **10 Trials**, where the model has improved

```
                                        Trial       Decision Tree
                                        -----      ----------------
  (a)   (b)   (c)   (d)    <-classified as          Size      Errors

  299                     (a): class acc
         56               (b): class good      0      49   17( 1.2%)
    1         971         (c): class unacc     1      12  157(11.4%)
                    55    (d): class vgood     2      24  101( 7.3%)
                                               3      29   84( 6.1%)
                                               4      27  104( 7.5%)
Attribute usage:                               5      35   91( 6.6%)
                                               6      40  127( 9.2%)
100.00% ï..buying                              7      36   54( 3.9%)
100.00% persons                                8      35   81( 5.9%)
100.00% safety                                 9      43   49( 3.5%)
 94.50% maint                              boost           1( 0.1%)   <<
 55.43% lug_boot
 42.76% doors
```

The usage of the attribute has been increased when the trial boost has been increased and the best model has been generated by adding the improvements.

1.4 EVALUATION

For the evaluation of the test data and prediction are performed using the cross-table matrix.by cross-classifying the factors and to build a contingency table of the counts at each combination of factor levels.

1.4.1 Model 1 Prediction and Confusion Matrix

In the model 1 of the decision tree of the test data the predictions are very much accurate compared with the training data and the

	actual				
predicted	acc	good	unacc	vgood	Row Total
acc	80	0	0	0	80
	0.231	0.000	0.000	0.000	
good	3	13	0	0	16
	0.009	0.038	0.000	0.000	
unacc	1	0	238	0	239
	0.003	0.000	0.688	0.000	
vgood	1	0	0	10	11
	0.003	0.000	0.000	0.029	
Column Total	85	13	238	10	346

The Accuracy of the Model Decision Trees has been found using the sum of Diagonal of confusion matrix to the sum of confusion matrix which gave the accuracy to 1

```
> cm=table(predictions)
> sum(diag(cm))/sum(cm) #  1
[1] 1
```

1.4.2 Model 2 Prediction and Confusion Matrix with boost

When the model is boosted with trials up to 10, here is the Actual and predicted with improved and much better. And the Accuracy of the Model 1 Decision Trees has been found using the sum of Diagonal of confusion matrix to the sum of confusion matrix which gave the accuracy to 1

	actual				
predicted	acc	good	unacc	vgood	Row Total
acc	83	0	1	1	85
	0.240	0.000	0.003	0.003	
good	0	12	0	0	12
	0.000	0.035	0.000	0.000	
unacc	1	0	237	0	238
	0.003	0.000	0.685	0.000	
vgood	1	1	0	9	11
	0.003	0.003	0.000	0.026	
Column Total	85	13	238	10	346

```
> cm=table(predictions1)
> sum(diag(cm))/sum(cm)
[1] 1
```

So, the Both Model Combinedly works well to do the prediction of the price, safety and Comfort of this Car Evaluation Dataset

2. K-NEAREST NEIGHBOUR ALGORITHM- KNN

2.1 BUSINESS UNDERSTANDING

This dataset is referenced from the UCI Machine Learning Repository, named Abalone data which is to predict the sex of the shell in their physical measurements.

Link of the dataset http://archive.ics.uci.edu/ml/datasets/Abalone

2.2.DATA UNDERSTANDING AND PREPARATION

2.2.1 Data Understanding

This dataset has 9 variables and 2732 Instances. Variables such as Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight and Shell weight are continuous values and Rings is the integer. Sex is nominal with (Male, Female, Infant), Length is the longest shell measurement in Millimetre (Mm), diameter is the perpendicular to length in Mm, Height is the total height of shell with meat in it -measured in Mm, Whole weight measured in grams of the abalone, Shucked weight is the weight of the meat in the grams, Viscera weight is the gut weight (after bleeding) in grams and Shell weight is the weight of the shell after its being dried- measured in grams. And Rings which is additionally added 1.5 years to the age shell.

2.2.2 Preparation

As part of the dataset preparation, which needs to be pre-processed before we are importing to the R-studio, this dataset was not available in direct csv format instead as. data file and which is to be processed in spreadsheet and data processing work and cleaning the csv file has done. Importing the Class function for Knn and then setting factor for Sex attribute, Normalizing the data and Splitting the dataset into 80 percentage for the train dataset and 20 percentage for the test data set, Shuffling the dataset and labelling the dataset

2.3 MODELLING

In the KNN Algorithm, there is no model compared with the other algorithm as the model is training data here, we are going to improve based upon the K – Finding the nearest neighbour points (based upon majority of the neighbour datapoint we will be setting up the class)

2.3.3 when K is 5:

Here the k value when is equal to 5, it gives the prediction accuracy of **0.5311355** when using sex as the attribute

```
> cm=table(predictions,test_labels)
> m=sum(diag(cm))/sum(cm)
> m       #0.5311355
[1] 0.5311355
```

2.3.4 when K is 10:

Here the k value when is equal to 10, it gives the prediction accuracy of **0.5586081** when using the sex as the class attribute**, which is the best fitted**

```
> cm=table(predictions1,test_labels)
> m1=sum(diag(cm))/sum(cm)
> m1      # 0.5586081
[1] 0.5586081
```

2.3.5 when K is 20:

Here the k value when is equal to 20, it gives the prediction accuracy of **0.532967** when using sex as the class attribute**, here the accuracy starts to reduce**

```
> cm=table(predictions2,test_labels)
> m2=sum(diag(cm))/sum(cm)
> m2      #00.532967
[1] 0.532967
```

2.4 EVALUATION

In the evaluation phase of the KNN Algorithm which is based upon the predictions and which is evaluated by the cross-table matrix (Actual-Predicted) from this we found the best fitted model with the highest k values.

2.4.1 Cross-Table Evaluation:

Cross table evaluation talks about the actual minus the predicted value, here is the Confusion matrix of the **best prediction of the KNN when the k value is 20**

```
Total Observations in Table:  546


             | test_labels
predictions1 |           F |           I |           M | Row Total |
-------------|-------------|-------------|-------------|-----------|
           F |          75 |          15 |          65 |       155 |
             |       0.137 |       0.027 |       0.119 |           |
-------------|-------------|-------------|-------------|-----------|
           I |          24 |         135 |          45 |       204 |
             |       0.044 |       0.247 |       0.082 |           |
-------------|-------------|-------------|-------------|-----------|
           M |          69 |          23 |          95 |       187 |
             |       0.126 |       0.042 |       0.174 |           |
-------------|-------------|-------------|-------------|-----------|
Column Total |         168 |         173 |         205 |       546 |
-------------|-------------|-------------|-------------|-----------|
```

So, from the above table we can see that, how accurate and the predictions are matched over the cross-table matrix. F-(Female) row it has 75 datapoints in F, it has 15 Datapoints in the I-(Infant) and 65 Datapoints in the Male. -Which is like actual and the predicted ones.

For the I row which has 24 datapoints for actual in F and 135 datapoints in I and it has 45 Datapoints in the M (Both are similar for Actual and Predicted evaluation)

For the M –(Male) row it has 69 datapoints in F-(Female), 23 Datapoints in I(Infant) and 95 Datapoints in M-(Male): Which is same for the actual and predicted. **This is the best predicted model for the KNN by means of Sex**

4. K-MEANS ALGORITHM

4.1 BUSINESS UNDERSTANDING

This dataset is referenced from the UCI Machine Learning Repository, named Abalone data which is to predict the age of the shell by means of Rings Attribute dataset by means of their physical measurements -Further by cutting the shell through the cone and looking the number of rings through microscope, which is a time-consuming and daunting task and easiest way of prediction is by weather patterns and location and its food availability -Best approach to find the age using clustering algorithms.

Link of the dataset http://archive.ics.uci.edu/ml/datasets/Abalone

4.2.DATA UNDERSTANDING AND PREPARATION

4.2.1 Data Understanding

This dataset has 9 variables and 2732 Instances. Variables such as Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight and Shell weight are continuous values and Rings is the integer. Sex is nominal with (Male, Female, Infant), Length is the longest shell measurement in Millimetre (Mm), diameter is the perpendicular to length in Mm, Height is the total height of shell with meat in it -measured in Mm, Whole weight measured in grams of the abalone, Shucked weight is the weight of the meat in the grams, Viscera weight is the gut weight (after bleeding) in grams and Shell weight is the weight of the shell after its being dried- measured in grams. And Rings which is additionally added 1.5 years to the age shell.

4.2.2 Preparation

As part of the dataset preparation, which needs to be pre-processed before we are importing to the R-studio, this dataset was not available in direct csv format instead as. data file and which is to be processed in spreadsheet and data processing work and cleaning the csv file has done. Importing the Class function for Knn and the foreign function, then Normalizing the data and labelling the dataset

4.3 MODELLLING

When creating the k means model for the dataset, we used the k-Means function and inserted the dataset and the number of class instances (number of classes in the class column). We then found the model summary, cluster, tot. wittiness and centres. This helped in analysing the data.

The below is the Model Cluster

```
> model$cluster
   [1] 26  9 14 19  9 18 25  4  2 25 26 19 20 14 19 20  9 19  9 18 10 10  3 14 22  3  3  3 23  6
  [31]  3  7 24  7  7  2 23  2  3 10  2  4 21 13 13  9  1  2  9  6  1 18 20 19 18  1  2  1 21  1
  [61]  2  1 14 18  1  2  3 11 20 21  4 18 24 22 23 23  3  3  4 23  6 11 25 24 23 23 23  6  2  3
  [91]  4  4 23 15  7  7 14  2  2 20  9  4  4 14 22  4  4 20  1  2  1  2 18  2 14 20 20 19  6 18
 [121]  2  9  4  9  9 21  9 10  7 12 24  2  9  9 21 18 21 10 18 10  6 14 11  3 20 20 19 21 13 13
 [151] 23 22 11 22  3 14 23  7  3 23 23 22 15 12 12  8 12 12 12  7 12 14  5 14 13  9 21  9 13  3
 [181] 23 24  3 22 16 11 22 15 15  3 23 22  4 10 14  4 20 11  4  3  2  4 20  4 18  2 10 18  4  9
 [211] 20  9 19  1 26 14  2 18 19 18  2 19  1 19 19  2 18  9  3  4  4  4 24 21 19 21 13 13 13 13
 [241] 25 21 13 21  9 10 10 10 21  9  9 23  3 23 23 23 25  3 11 23  3  3  3 21 21 20  4  9  1 19
 [271] 24 23 23  7 11 12 11  7 24 14  4  9  2 26  4 14 14 26  2 14 25 22 23 25 24 21 21 21 20 10
 [301] 18  4 10  9 19 13 13  7  4  3 24  3 24 24 24  2 11 19 25  9 21 13 10 21  9  9  9 20 10  9
 [331] 20 18 21 13 12 15 23 23 15 23 23 11 15 20 14 20 14  9 10 23  5 25  3  6 11 12 16  7 12  7
 [361] 24  3 24 23 22 24  3  5 22  7  7  7 12 11 11 12 22  5 23 22 23 14 20 20  6  2  4 19 26  2
 [391] 19 20 20 10  9 10 14  1  4  3  1 14  2 14 18  4  2 23  3  6 23 23  3  4 11 23 24 24 23  4
 [421]  7 20 20 21 21 23 23 25 25 25 25 25 25 25 19 20  9 10 10  4 10  9 25 10 19 14  4 25 22 25
 [451] 12  7  7  3 11  3 11 10 10 14 10  3 21 13 21 13 11 11  7 24 14 20 18 20 14 25  2 24 12 11
 [481] 24 25  4 25 22  3  3 11 14 23 23  3 11 11 24 23 11 25 22  3  3 24 23 23 23 23  3 23  3 25
 [511] 22 19 20 21 21 21 10 21 21  9 13  9 10 13 13 13 13  4  3 10 25 19 19  2 19 19 20 21 21 10
 [541] 26 19 19  2 10 10 21 13  4  3 23 23  4 20  1 20  4 11 11 20 18 20 19 14 26 10 20 19 21 19
 [571] 26 19 25 25  3 22  6  3  3 24 22  4  7 19 19 20 14  4 19 19  4 10 25 24  4  4 14  3 23 23
 [601] 25 10 10 19 20 19 10 19 10 10 19 13  9 26 26 19 19  9 21 21 10 26 20 26 26 20 19 18 25 10
 [631]  2 18 19  2 10  9  9  9 19 20  9 11 25  9  2 19 21  1 19  2 21  9 19 21 10 21 25 23 24  3
 [661]  7  4 19 10 26 10 19 26  4 19 26 26 25 26 25 25  4 20 25  9  9 20 18  4 20  4  4 25 14  3
 [691] 20 21 19 10 13 21 13 21 19 19 10 26  2 10 18 19 10 10 10 21  9 10 21  9 10 21 21 21 13 13
 [721] 13  3 23  4 19 26 10 20  4 26  4  4  4 20 25 20 20 20  4 19  4  9  4 25 25 25  7  1  4  7
 [751] 20 20 23 23 24 16 24 11  3  4 23 25 25  7 23  3 23 10  4 25  3  1 10  2 20  4 26 20 26 26
 [781]  4 20 14  9 13  4  1 20  1  3  3 20  2  3  3 26 14  2  4 19 19 19  2  9 20 18 21 26 19  4
 [811]  4 26 21 21  9  9  9  9  9  9  9  9 18 18 18 18 18 18 18 18 18  2  2 18  2  1  2  1  2
 [841]  1 14  1 14 14  6 14  6  6  3  6  6  5  6 14  6  5  3 22  5  5 22 15  5 15 15 22 15 15 15
 [871] 11 15 15 22 15 11 15 15 15 15 16 15 12 11 16 15 17 16 17 17 17 12 13 21 13 21 21 21 13 21
 [901] 21 21 21 21 21 21  9  9  9  9  9  9  9  9  9  9  9 18 18 18 18 18 18 18 18 18 18 18 18  2
 [931] 18  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  1  2  2  1  1  1  2  1  2  1  1  1  1  1
 [961]  1  1  1  1  1 14  1 14 14 14 14  1 14 14 14 14  1 14 14  6  6  6  6 14  5  5  5  5  6  5
 [991]  5  5  5  5  5  5 22  5  5  6
```

Below which is the Model Centre

```
> model$centers
      Length  Diameter     Height Whole.weight Shucked.weight Viscera.weight Shell.weight
1  0.5779689 0.5584925 0.11319389   0.21319875     0.18379083     0.17482395   0.16496203
2  0.5218948 0.5031914 0.10360752   0.16267838     0.13734883     0.13126266   0.13441207
3  0.6791470 0.6764320 0.13741171   0.33461950     0.26104218     0.27336913   0.28574759
4  0.6255737 0.6197083 0.13424612   0.28873070     0.20587862     0.23355112   0.26044223
5  0.7003406 0.6849764 0.13788762   0.38528552     0.35073619     0.31007242   0.27214910
6  0.6699614 0.6530132 0.13031606   0.31089276     0.27066001     0.25212828   0.24062923
7  0.7999173 0.8024353 0.17283728   0.57163818     0.41131301     0.45983528   0.50413349
8  0.8901594 0.8808446 0.19083277   0.78899166     0.71232735     0.63687311   0.55767634
9  0.3777294 0.3560272 0.07844608   0.07433576     0.06213170     0.05923091   0.06336579
10 0.3982647 0.3837853 0.08708769   0.08689046     0.06544217     0.07116674   0.07977193
11 0.7630977 0.7688429 0.15670524   0.46551879     0.34909730     0.36018636   0.41580622
12 0.8567102 0.8640974 0.18523039   0.72016267     0.51047005     0.54566298   0.68812604
13 0.1573057 0.1402311 0.04023783   0.01239596     0.00938866     0.01067725   0.01007412
14 0.6193292 0.6081433 0.12439423   0.26361342     0.22194535     0.21192279   0.21505682
15 0.7569591 0.7448297 0.14985957   0.46616984     0.41013132     0.38123884   0.34115625
16 0.8011493 0.7915652 0.16280705   0.54828054     0.46700061     0.44937951   0.41198329
17 0.8457743 0.8342003 0.16926535   0.63758535     0.55808542     0.52179274   0.47217279
18 0.4614185 0.4396102 0.09103524   0.12084020     0.10267685     0.09526666   0.09854562
19 0.4989371 0.4889057 0.10664214   0.15267329     0.11283181     0.12670222   0.13726116
20 0.5695180 0.5553149 0.11855670   0.21640722     0.16385652     0.18377527   0.18918322
21 0.2806375 0.2581307 0.06360197   0.03902498     0.03272894     0.03289629   0.03253879
22 0.7348824 0.7244423 0.14461705   0.39760209     0.32430706     0.33390483   0.32550820
23 0.7001168 0.6970640 0.15093412   0.38408682     0.27918088     0.30552914   0.34193902
24 0.7509888 0.7556876 0.16263760   0.47112000     0.33013762     0.34995745   0.43344635
25 0.6463964 0.6386555 0.13938053   0.32473585     0.22475342     0.24084924   0.29946022
26 0.5506757 0.5403727 0.12052713   0.20456418     0.14060992     0.16737270   0.19102706
27 0.5135135 0.5042017 1.00000000   0.20966885     0.22259583     0.15207373   0.13153961

       Rings
1  0.2497835
2  0.2455113
3  0.3646789
4  0.4407008
5  0.2804878
6  0.2846939
7  0.5269679
8  0.3763736
9  0.1987490
10 0.3133117
11 0.4384615
12 0.6059113
13 0.1132812
14 0.2855017
15 0.3186484
16 0.3508011
17 0.3503401
18 0.2195689
19 0.3663724
20 0.3648748
21 0.1742094
22 0.3358171
23 0.4863316
24 0.6898955
25 0.6327381
26 0.5248447
27 0.2500000
```

Below which is the Model Witness

```
> model$tot.withinss
[1] 23.35264
```

4. EVALUATION

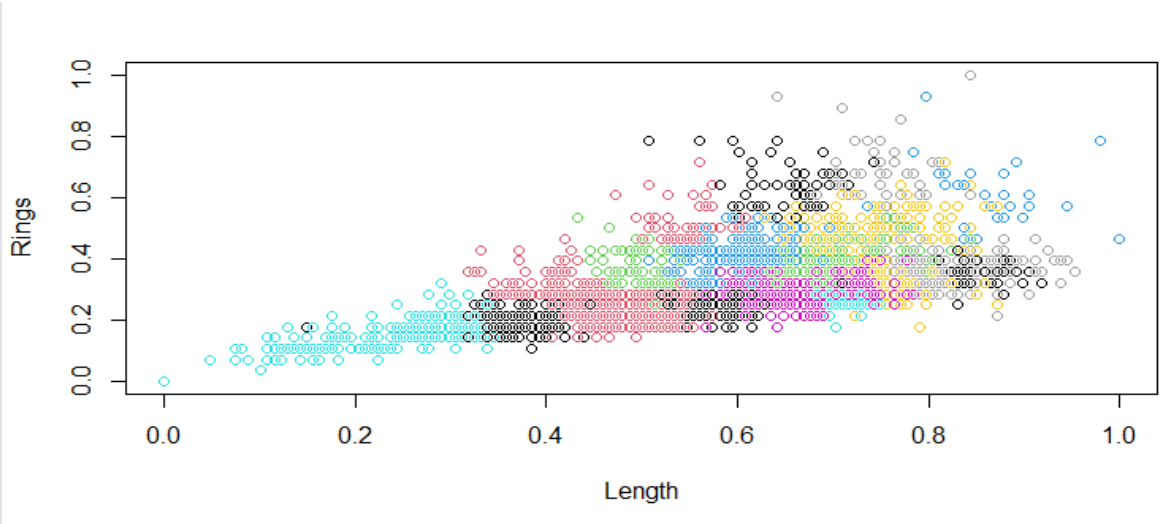
The model is evaluated by the means of verifying, how the clusters are created using the model to that of the actual classes in the dataset, here are the clusters of the attribute Ring

```
> table(dataset1$Rings, model$cluster)

     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
1    0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2    0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
3    0  0  0  0  0  0  0  0  0  0  0  0 12  0  0  0  0  0  0  0  0  0  0  0  0  0  0
4    0  0  0  0  0  0  0  0  1  0  0  0 28  0  0  0  0  0  0  0  9  0  0  0  0  0  0
5    0  0  0  0  0  0  0  0 16  0  0  0 17  0  0  0  0  5  0  0 45  0  0  0  0  0  0
6    8 18  0  0  1  0  0  0 63  0  0  0  4  3  1  0  0 35  0  0 45  0  0  0  0  0  0
7   37 46  0  0  6  5  0  0 77  0  0  0  1  8  1  1  0 74  0  0 22  0  0  0  0  0  0
8   68 62  0  0 35 46  0  0 19 11  0  0  0 43 12  1  2 40  0  0  6  4  0  0  0  0  1
9   52 55  3  0 50 78  0  2  1 29  0  0  0 62 48 14  4 15  0  0  3 22  0  0  0  0  0
10   0  2 25  0 30 41  0  5  0 30  1  0  0 37 54 29 17  0 34 34  1 53  0  0  0  0  0
11   0  0 42  8  1  5  0 12  0 10  6  0  0 15 37 32 26  0 24 22  0 36  0  0  0  0  0
12   0  0 26 22  0  0  1 13  0  4 12  0  0  0 13 18  9  0 13 27  0 22  2  0  0  0  0
13   0  0 11 30  0  0  6  4  0  3 20  0  0  0  1  9  5  0 13 14  0  2 17  0  0  0  0
14   0  0  2 24  0  0  6  3  0  1 12  4  0  0  0  3  0  0  3  0  0  0 20  0  0 13  0
15   0  0  0 16  0  0 10  0  0  0  9  1  0  0  0  0  0  0  1  0  0  0 23  0  0 16  0
16   0  0  0  6  0  0  8  0  0  0  5  4  0  0  0  0  0  0  1  0  0  0 11  0  5  5  0
17   0  0  0  0  0  0 10  0  0  0  0  6  0  0  0  0  0  0  0  0  0  0  6  5 16  4  0
18   0  0  0  0  0  0  4  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  2  8  9  4  0
19   0  0  0  0  0  0  3  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0  4 12  2  0
20   0  0  0  0  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  0  9  8  0  0
21   0  0  0  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  5  3  2  0
22   0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  2  3  0  0
23   0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  4  4  0  0
25   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
26   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
27   0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
29   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
```

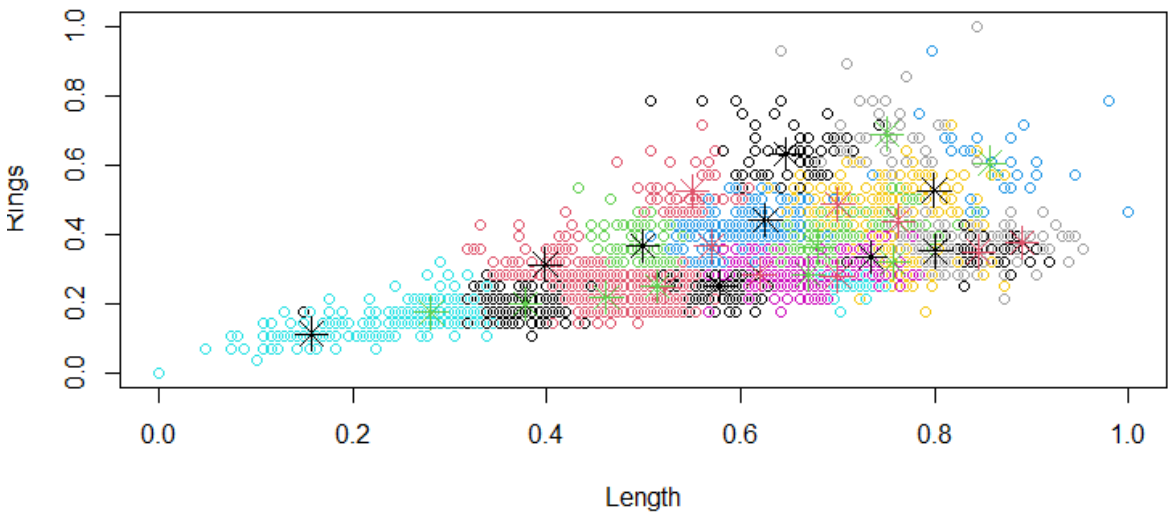
4.4.1 Plotting the Model Cluster

The model is first plotted using two elements from the data frame, which are the Length Attribute and the Ring attribute



4.4.2 Plotting the Model Cluster Centres

In this plot each cluster centre is plotted with the two attributes of the data frame (Length and Ring) attribute.



From the above graph we can understand how close the rings attribute and its clusters, and it is well predicted.