

**Professional Master’s in Artificial Intelligence**

**Fundamentals for Applied Data Science (DTI 5126)**

Subject: Assignment 2 (Classification & Clustering)

By

Mohamed Sayed Abdelwahab Hussein

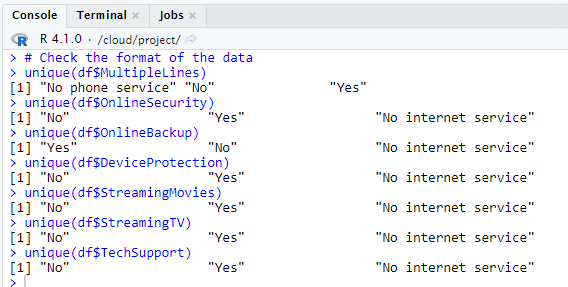
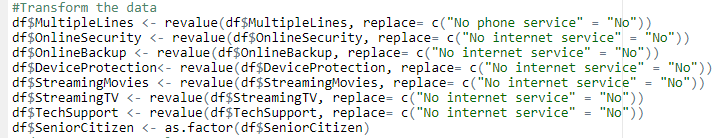
(300273145)

[Mhuss073@uottawa.ca](mailto:Mhuss073@uottawa.ca)

Under Supervision

Dr. Olubisi Runsewe

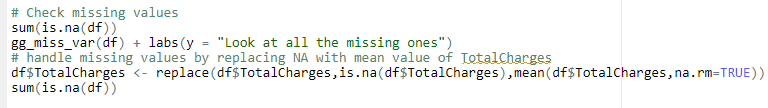
**Part A (Classification):**

1. Ensure data is in the correct format for downstream processes and address missing data

After investigating the data, we have an assumption here: the values of “No Internet service” has equal weight to the value of “No”, so we will convert them to “No” so the feature values will be binary “Yes” and “No”.

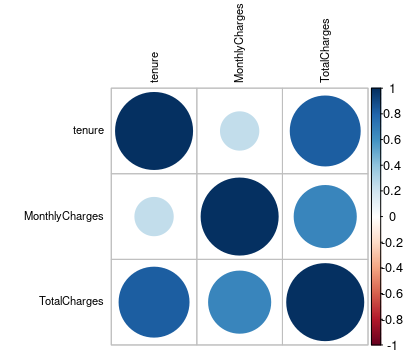
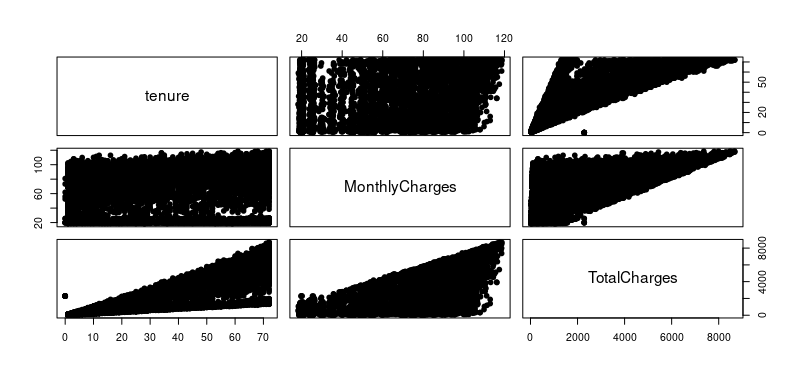
Then we will drop the customerID column: -

(df<- select(df, -customerID)).

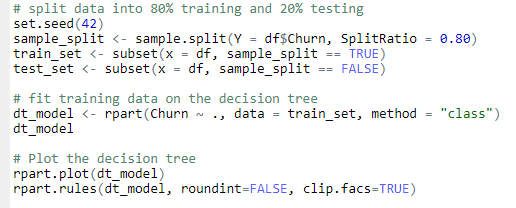
Let`s check and address the missing values

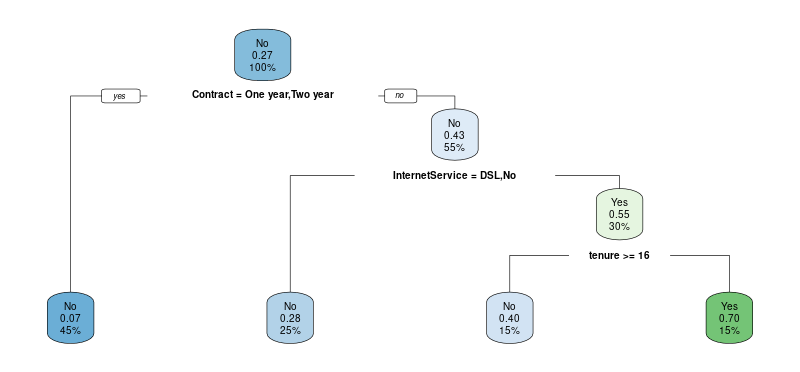
We noticed that “TotalChanges” column has 11 missing values, so we replaced them by the mean of the total charges column.

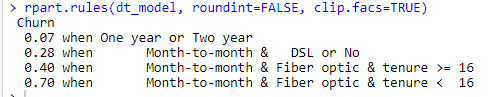
1. Generate a scatterplot matrix to show the relationships between the variables and a correlation matrix to determine correlated attributes



1. Split the dataset into 80 training /20 test set and fit a decision tree to the training data. Plot the tree, and interpret the results.
2. Describe the first few splits in the decision tree. Extract some rules.





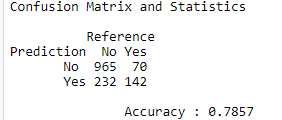
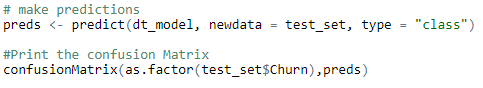
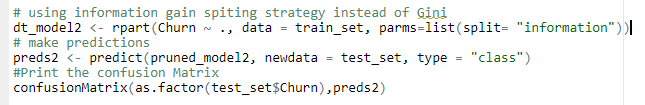
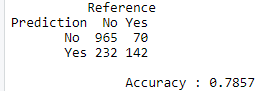


From the above figure we can tell that: -

* root node is the overall probability of Churn. It shows the proportion of customers that churned. 27% of customers churned. This node asks whether the customer’s contract is one year or two-year. If yes, 45% from customers are with one or two-year contract with a churn probability of 7%. If no 55% from customers are without one or two-year contract with a churn probability of 43%.
* in the second node, you ask if the customer’s internetService is DSL or No. If yes, then the chance of churned is 28%.
* in the third node, you ask if the customer’s tenure greater than or equal 16. If yes, then the chance of churned is 40%.

1. Try different ways to improve your decision tree algorithm (e.g., use different splitting strategies, prune tree after splitting). Does pruning the tree improves the accuracy?

Use different splitting strategies

1. With default splitting parameter (Gini): -
2. With information gain splitting parameter: -

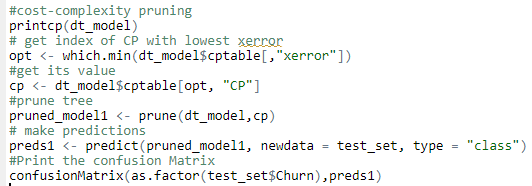
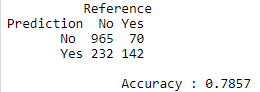
We tried Gini and Information Gain strategies for splitting, there is no different in accuracy.

Prune tree after splitting

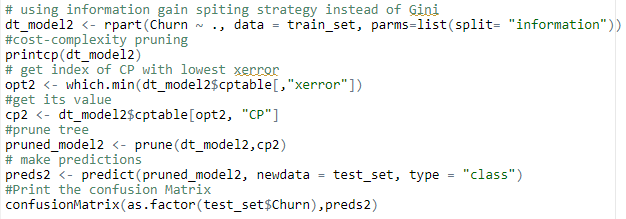
1. With default splitting parameter (Gini): -

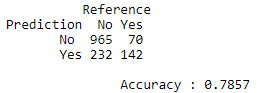
Printcp(dt\_model)



It is clear from the above, that the lowest cross-validation error (xerror in the table) occurs for alpha =0.0100. we will get this value programmatically, then we will prune the tree based on the value of cp, make predications and get accuracy.

It’s the same accuracy of the default model.

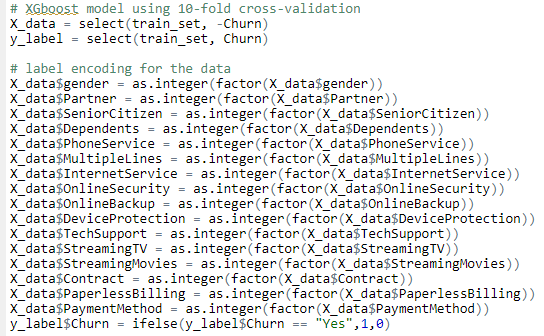
1. With spiriting parameter (Information Gain): -

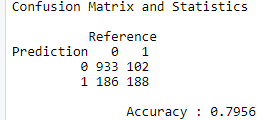


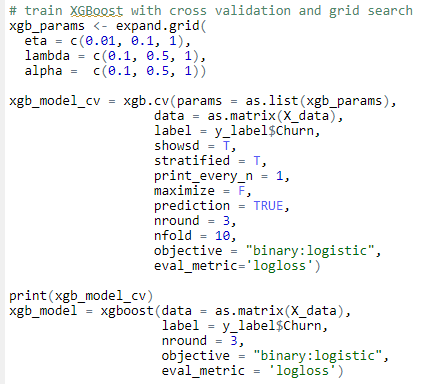
It’s the same accuracy of the default model.

1. Train an XGboost model using 10-fold cross-validation repeated 3 times and a hyperparameter grid search to train the optimal model. Evaluate the performance.

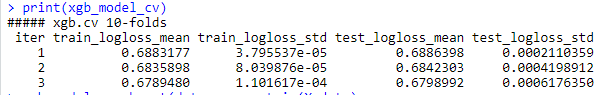
First, we will split the original data to features and labels, then convert all categorical data to numeric.



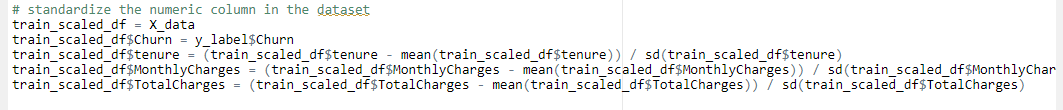
Fit XGBoost model using cross validation and grid search

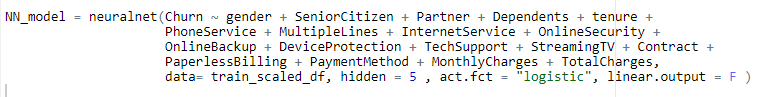
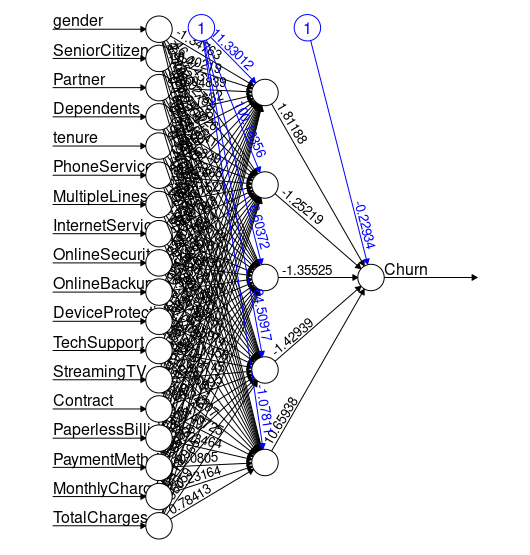


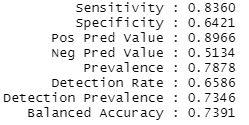
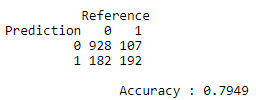
the accuracy increased by 0.1 in XGBoost.

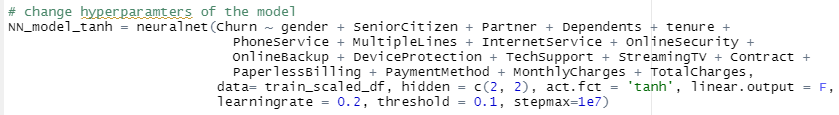
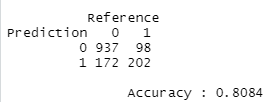
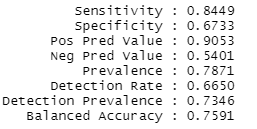


1. Build a multilayer perceptron with 5 nodes at the hidden layer. Use a standard or normalization to scale the variables.

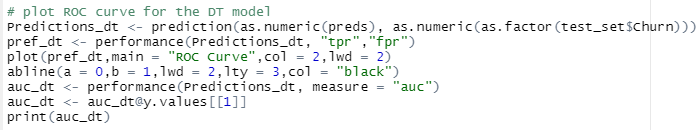
Here we will start by standardize the three numeric features in the dataset

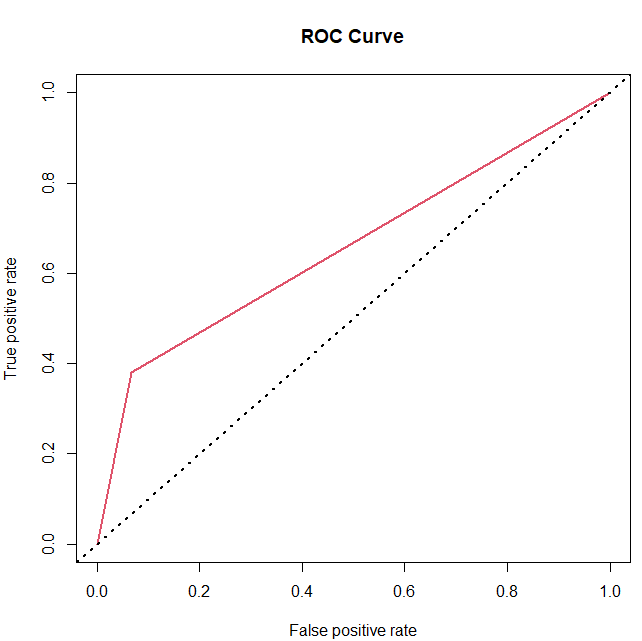
Fit NN model on the training data

Making prediction using NN model and calculate accuracy

When we change some parameters like (activation function = “thanh”, learning rate = “0.2”, neurons=2, no\_hidden layers= 2, threshold= 0.1, stepmax= 1e7)

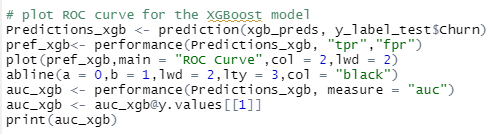
the accuracy increased by a little bit when we changed some parameters.

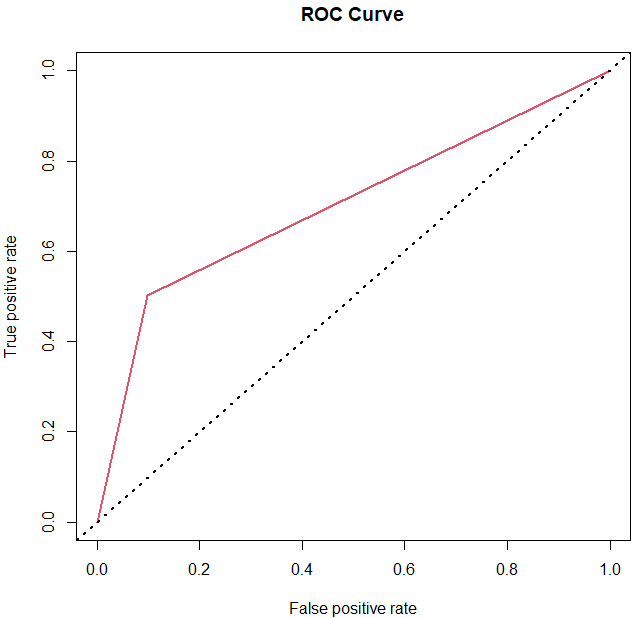
1. Carry out a ROC analysis to compare the performance of the DT, XGboost & NN techniques. Plot the ROC graph of the models.
   1. Decision Tree ROC Curve: -



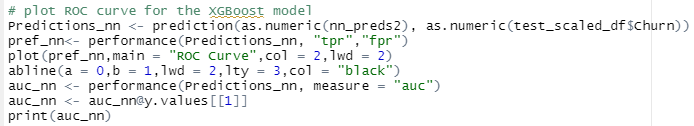
**AUC = 0.6560231**

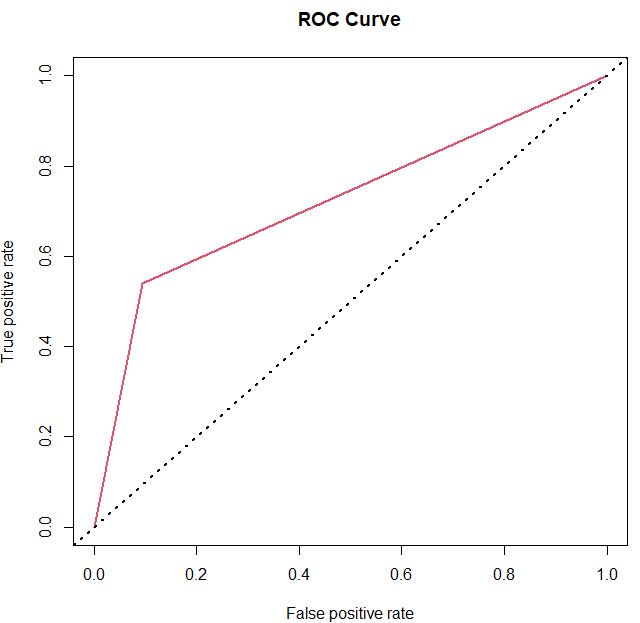
* 1. XGBoost ROC Curve: -





**AUC = 0.7020615**

* 1. NN ROC Curve: -



**AUC = 0.7227105**

**Part B (Clustering):**

1. a) Perform k‐means clustering, specifying k = 2 clusters and plot. Determine the attribute that is most correlated with the clusters.

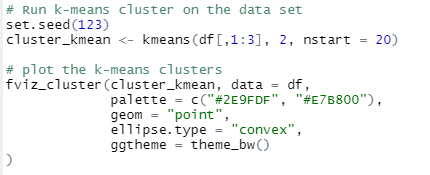
First, we remove the first two columns from the original dataset

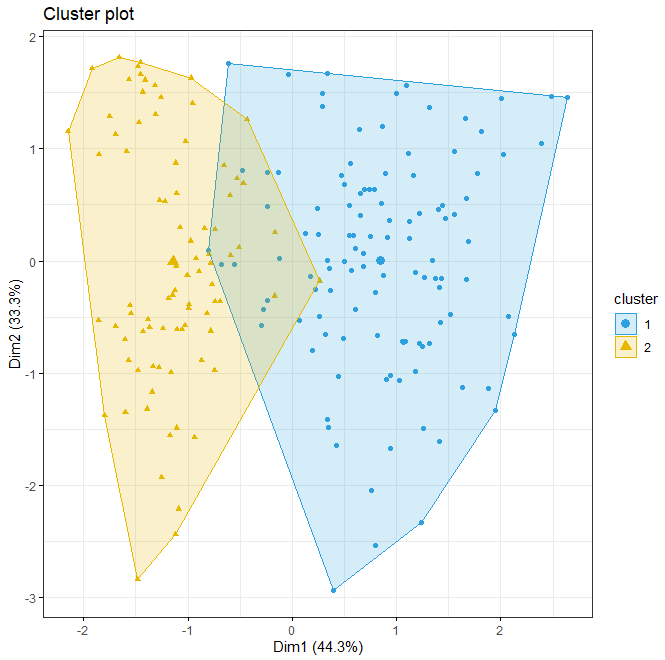


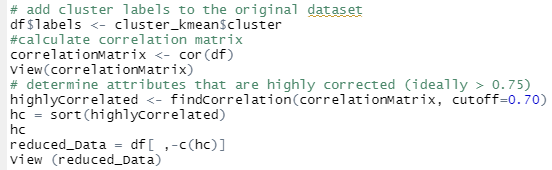
Then we will check the missing values

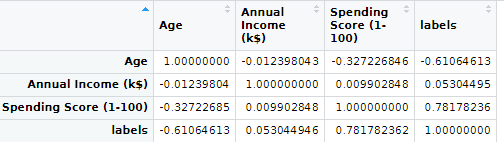
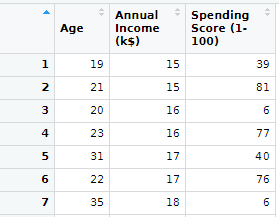


After assessing the data, we will run K-means cluster with 2 clusters.

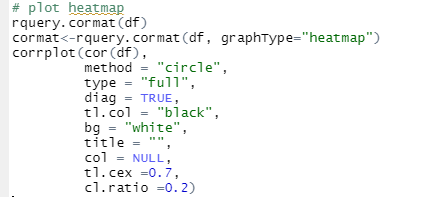


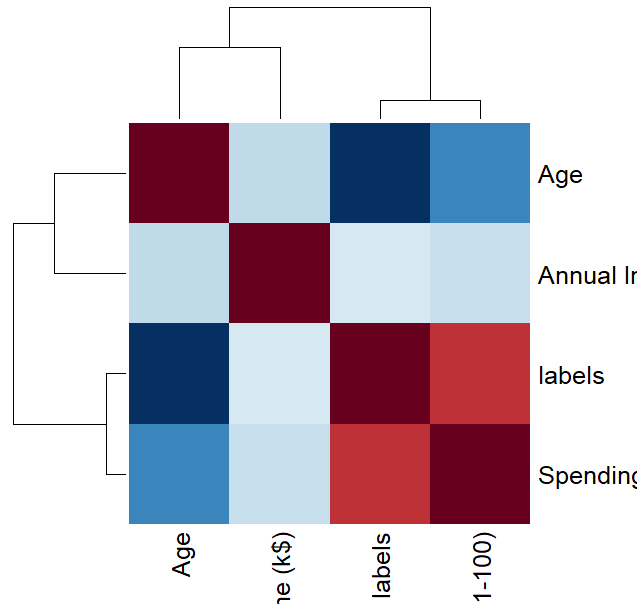
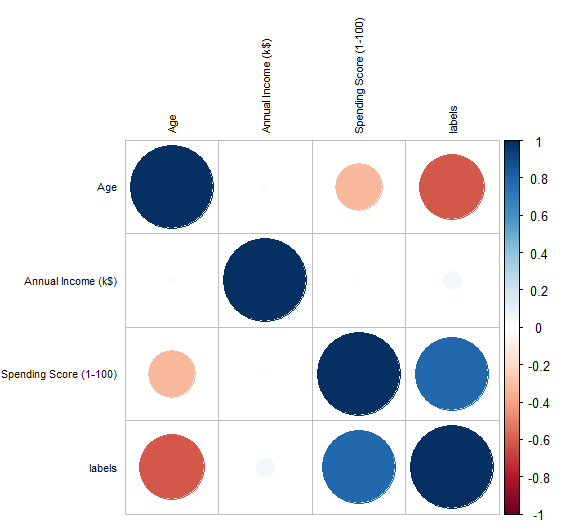


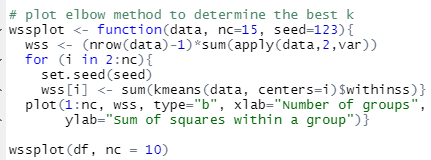
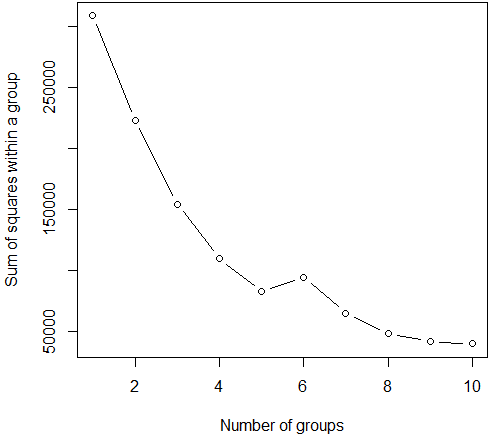
Add cluster labels to the original dataset and calculate correlation and determine the attributes that most correlated with the cluster.



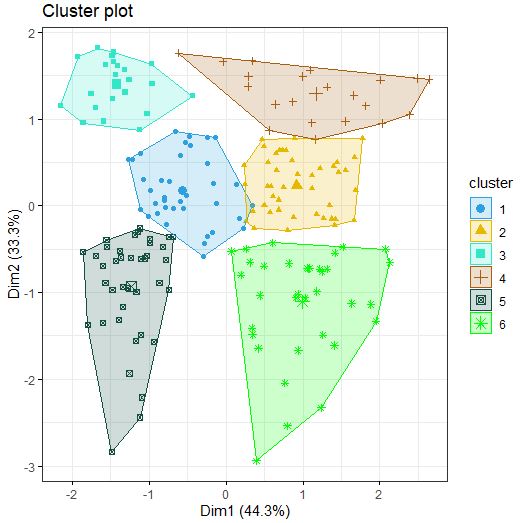
Plot heatmap

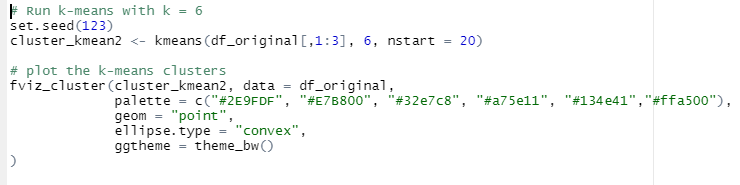


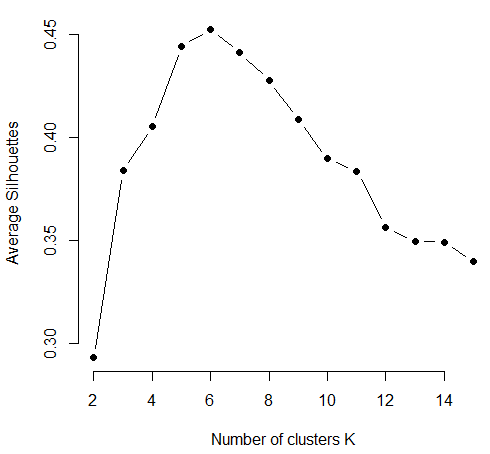
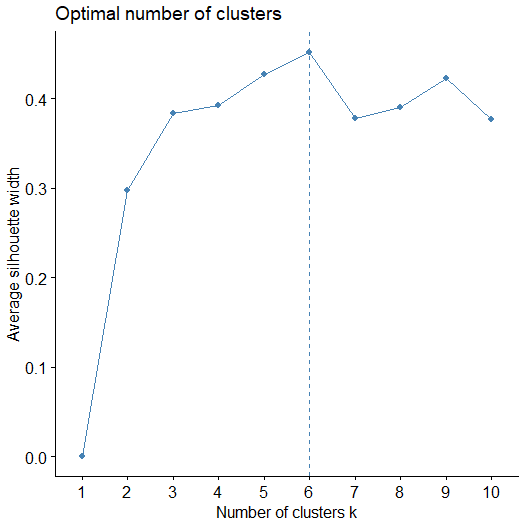
 

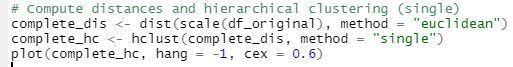
B) Apply the elbow method to determine the best k and plot.

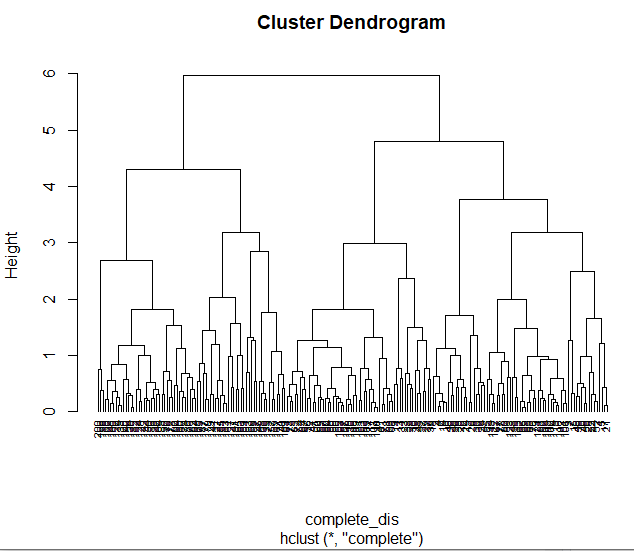
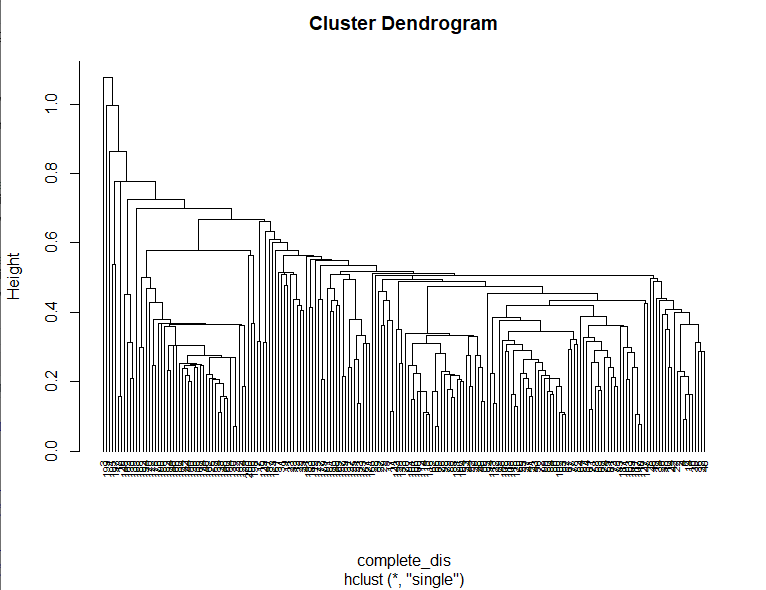
From above figure

the optimal number of clusters = 6



c) Evaluate the quality of the clusters using Silhouette Coefficient method.

d) Apply hierarchical clustering (single & complete linkage) to the dataset using Euclidean-based distance, and plot the dendrogram. Do your results depend on the type of linkage used.



* Single-linkage (nearest neighbor) is the shortest distance between a pair of observations in two clusters. It can sometimes produce clusters where observations in different clusters are closer together than to observations within their own clusters.
* Complete-linkage (farthest neighbor) is where distance is measured between the farthest pair of observations in two clusters. This method usually produces tighter clusters than single-linkage, but these tight clusters can end up very close together.

1. Consider the following “data” to be clustered: **10 20 40 80 85 121 160 168 195**.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80 | 85 | 121 | 160 | 168 | 195 |
| 10 | 0 |  |  |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |  |  |
| 80 | 70 | 60 | 40 | 0 |  |  |  |  |  |
| 85 | 75 | 65 | 45 | 5 | 0 |  |  |  |  |
| 121 | 111 | 101 | 81 | 41 | 36 | 0 |  |  |  |
| 160 | 150 | 140 | 120 | 80 | 75 | 39 | 0 |  |  |
| 168 | 158 | 148 | 128 | 88 | 83 | 47 | 8 | 0 |  |
| 195 | 185 | 175 | 155 | 115 | 110 | 74 | 35 | 27 | 0 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80, 85 | 121 | 160 | 168 | 195 |
| 10 | 0 |  |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |  |
| 80, 85 | 70 | 60 | 40 | 0 |  |  |  |  |
| 121 | 111 | 101 | 81 | 36 | 0 |  |  |  |
| 160 | 150 | 140 | 120 | 75 | 39 | 0 |  |  |
| 168 | 158 | 148 | 128 | 83 | 47 | 8 | 0 |  |
| 195 | 185 | 175 | 155 | 110 | 74 | 35 | 27 | 0 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10 | 20 | 40 | 80, 85 | 121 | 160, 168 | 195 |
| 10 | 0 |  |  |  |  |  |  |
| 20 | 10 | 0 |  |  |  |  |  |
| 40 | 30 | 20 | 0 |  |  |  |  |
| 80, 85 | 70 | 60 | 40 | 0 |  |  |  |
| 121 | 111 | 101 | 81 | 36 | 0 |  |  |
| 160, 168 | 150 | 140 | 120 | 75 | 39 | 0 |  |
| 195 | 158 | 175 | 155 | 110 | 74 | 27 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 10, 20 | 40 | 80, 85 | 121 | 160, 168 | 195 |
| 10, 20 | 0 |  |  |  |  |  |
| 40 | 20 | 0 |  |  |  |  |
| 80, 85 | 60 | 40 | 0 |  |  |  |
| 121 | 101 | 81 | 36 | 0 |  |  |
| 160, 168 | 40 | 120 | 80 | 39 | 0 |  |
| 195 | 175 | 155 | 110 | 74 | 27 | 0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 10, 20, 40 | 80, 85 | 121 | 160, 168 | 195 |
| 10, 20, 40 | 0 |  |  |  |  |
| 80, 85 | 40 | 0 |  |  |  |
| 121 | 81 | 36 | 0 |  |  |
| 160, 168 | 120 | 75 | 39 | 0 |  |
| 195 | 155 | 110 | 74 | 27 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 10, 20, 40 | 80, 85 | 121 | 160, 168, 195 |
| 10, 20, 40 | 0 |  |  |  |
| 80, 85 | 40 | 0 |  |  |
| 121 | 81 | 36 | 0 |  |
| 160, 168, 195 | 120 | 75 | 39 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 10, 20, 40 | 80, 85, 121 | 160, 168, 195 |
| 10, 20, 40 | 0 |  |  |
| 80, 85, 121 | 40 | 0 |  |
| 160, 168, 195 | 120 | 39 | 0 |

|  |  |  |
| --- | --- | --- |
|  | 10, 20, 40 | 80, 85, 121, 160, 168, 195 |
| 10, 20, 40 | 0 |  |
| 80, 85, 121, 160, 168, 195 | 40 | 0 |

