CIS6930 Fall 2017: Introduction to Data Mining
Project II: Clustering

Debarshi Mitra
UFID: 33813136

Brief description of clustering techniques –

Hierarchical Clustering: The hierarchical clustering method is a type of cluster analysis where a cluster is built hierarchically and can have a bottom up or top down approach called agglomerative and divisive approach respectively. A greedy approach is used for the merges and splits in this clustering and results are expressed in a dendrogram.

K-means Clustering: In the K-means clustering, if there are n observations, they are partitioned to k number of clusters and each of the observations belong to cluster having the nearest mean to the observation point. The number of clusters must be defined as an input parameter to get that many number of clusters. Euclidean distance, correlation and other similar methods measure the closeness of the observations. The most common way to evaluate K-means clusters is by the sum of the squared error.

Density-based Clustering: In density based clustering, the clusters are set as areas with higher density compared to the remainder of the data set. DBSCAN is the most popular type of density based clustering method. Density is defined as the number of points within the specified radius called eps. MinPts is defined as the core points inside the clusters within the eps. The noise points are defined as any other point other than the core points and border points.

Graph-based Clustering: The proximity graph is used in the graph based clustering. Each of the points are considered as nodes in the graph and they are connected by the edges having weight and defined as the proximity between the points. sNNclust method is used for the graph based clustering. For a given value of neighborhood size, a shared nearest neighbor graph is constructed.

Detailed analysis of the required tasks for Dataset1 –
Three script files are created in total for the complete project. The clustering_methods.R script file contains all the method definitions for the four clustering methods. The clustering1.R file contains the script to run the four clustering techniques over the given dataset1 with 1000 entries. Initially in the clustering_methods.R file, the path is set correctly using the setwd command. The required packages, “dbscan” and “plotly” are installed and loaded. The dataset1.csv file is read and stored as data matrix in the variable “data”. The detailed analysis of the four clustering methods are done below for the dataset1.
Hierarchical Clustering: In the h_cluster_method, the hclust method is used for the hierarchical clustering. The hclust method takes in a method argument, depending on which the final accuracy computation varies. Also, the dist method is used inside the hclust method as a parameter and the dist method also takes in method argument, depending on which the final accuracy of the computation varies. The different hclust methods are "ward.D", "ward.D2", "single", "complete", "average", "mcquitty", "median", and "centroid". The different dist methods are "euclidean", "maximum", "manhattan", "canberra", "binary", and "minkowski". The first three column of the dataset1.csv is taken as input for the hclust method. By running for loops over the hclust function, the different accuracies are

generated using different combination of the hclust and the dist functions. It is seen that the maximum accuracy is obtained with dist method maximum and hclust method ward.D2. The different accuracies achieved are shown in the results section below. The result is then divided into 8 clusters using the cutree method. The derived cluster computed now is compared row by row with the given value of cluster in the dataset1 and accuracy is measured by finding the sum of the diagonal of the comparison table var1. The 3D plot is generated using the plot_ly method and printed. The 8x8 table is returned to be used in the clustering1.R script.

K-means Clustering: In the K-means clustering, in the k_cluster_method, with different set.seed value and different values of nstart parameter in the kmeans method, the accuracies vary. A for loop is run over set.seed and compared with different values of nstart parameter to get the highest accuracy possible. It is found that for set.seed (178) and with nstart = 2, the accuracy achieved is maximum for dataset1. In the kmeans method, the first three columns of the dataset1.csv file is taken as input parameter along with number of clusters being 8 and value of iter.max set to 10. A table is created, var2, with the derived cluster and given cluster values and sum of the diagonal of var2 is computed and printed to get the accuracy. The 3D plot is generated using the plot_ly method and printed. The var2 is returned to be used in the clustering1.R script.

Density-based Clustering: The function d_cluster_method is defined for the density based clustering. The dbscan package is called and dbscan method is used with the parameter eps set to 1.313 and minPts set to 4. The value of the eps is set using trial and error to get the best value generating highest accuracy. A table with the derived cluster values and the given cluster values in the dataset1.csv file is created and stored in variable var3 and the table is shown with the rows two to nine. The density_result contains the data frame with the given input values and the derived cluster values. The density_result is plotted in the 3D graph and it is printed. To determine the optimum eps value, the value of k is specified to 4 and kNNdistplot method is used. The k distances are plotted in ascending order and the knee is determined which gives the optimal eps parameter.

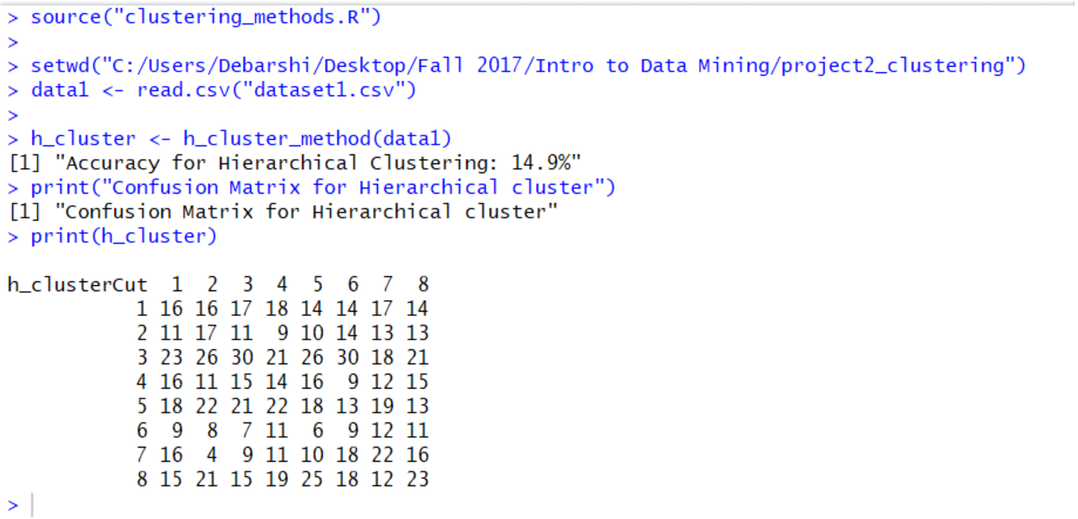
Graph-based Clustering: For the graph based clustering, the g_cluster_method is defined which uses the sNNclust method over the first three columns of dataset1.csv. The euclidean method is used for the dist function and value of parameters k set to 20, eps set to 10 and minPts set to 16 to get the best accuracy using trial and error method. The var4 stores the table with the derived cluster value to the given cluster value for rows 2 to 9. The sum of diagonal of var4 is printed to get the accuracy for the graph based clustering. The p4 stores the 3D plot using the plot_ly method and is printed. The var4 is returned to be used in the clustering1.R script.

Detailed analysis of the required tasks for Dataset2 – The third script file, clustering2.R is created to run the dataset2.csv file. The function for dataset2, k_cluster_method2 is defined in the clustering_methods.R script file. The k-means clustering is used here as the other clustering methods create data matrix which are nxn in size where n is the number of rows in the input file. For large datasets, this will give an out of bound exception as it fails to converge. The set.seed is taken as 178 like in dataset1 and kmeans is run over the first four column of the dataset2.csv file with nstart equals to 2 and iter.max equals to 10. The number of clusters is passed as an argument and in the clustering2.R a for loop is run for different values of number of clustering ranging from 2 to 25 and the k_cluster_method2 is run over data2 containing dataset2.csv data and the output is stored in the variable k_cluster. Two lists are initialized where the k_cluster$tot.withinss and k_cluster$betweenss are stored for all values of cluster from 2 to 25. The lists are converted to a data frame l1 and l2 respectively and the transpose are plotted for those two lists. In k-means, the tot.withinss is defined as the total within cluster sum of squares and betweenss is defined as the between cluster sum of squares. The lower the value of tot.withinss, the dense the cluster will be and higher the betweenss value, the dense the clusters will be. From the plots, it is found that the value of k_cluster$tot.withinss reduces for increase in the values of number of clusters and maintains a straight line after number of cluster reaches 20. The value of k_cluster$betweenss increases for increase in value of the number of clusters and maintains a straight line after the number of cluster reaches 20. Hence the 20 clusters defines the first point in the graphs where the tot.withinss is least and the betweenss is highest. Therefore, it can be concluded that the value of number of clusters can be taken as 20 for dataset2. The similarity matrix is not possible to be produced for dataset this large.

Results:

Dataset1:

1. **Hierarchical Clustering:**



```
> source("clustering_methods.R")
>
> setwd("C:/Users/Debarshi/Desktop/Fall 2017/Intro to Data Mining/project2_clustering")
> data1 <- read.csv("dataset1.csv")
>
> h_cluster <- h_cluster_method(data1)
[1] "Accuracy for Hierarchical Clustering: 14.9%"
> print("Confusion Matrix for Hierarchical cluster")
[1] "Confusion Matrix for Hierarchical cluster"
> print(h_cluster)

h_clusterCut  1  2  3  4  5  6  7  8
           1 16 16 17 18 14 14 17 14
           2 11 17 11  9 10 14 13 13
           3 23 26 30 21 26 30 18 21
           4 16 11 15 14 16  9 12 15
           5 18 22 21 22 18 13 19 13
           6  9  8  7 11  6  9 12 11
           7 16  4  9 11 10 18 22 16
           8 15 21 15 19 25 18 12 23
>
```

Fig: Percentage Accuracy and Confusion Matrix

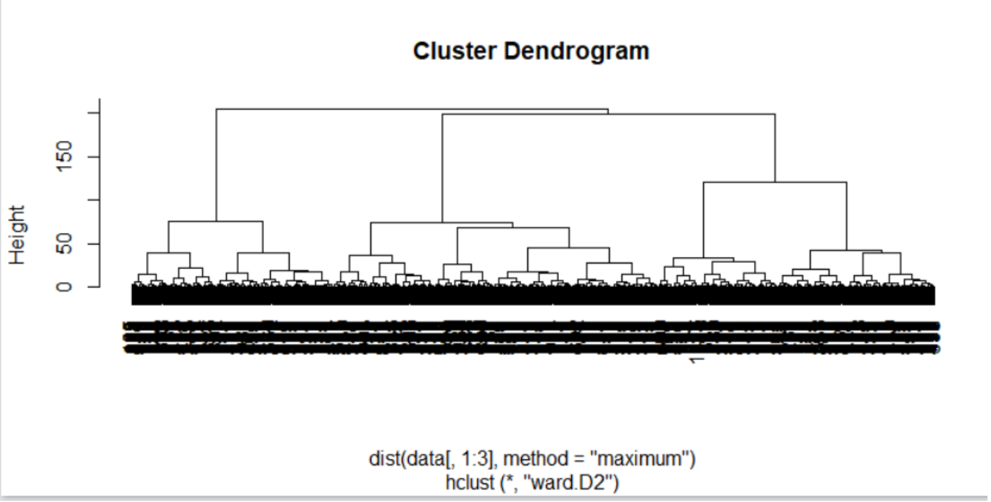


Fig: Cluster Dendrogram for Hierarchical Clustering

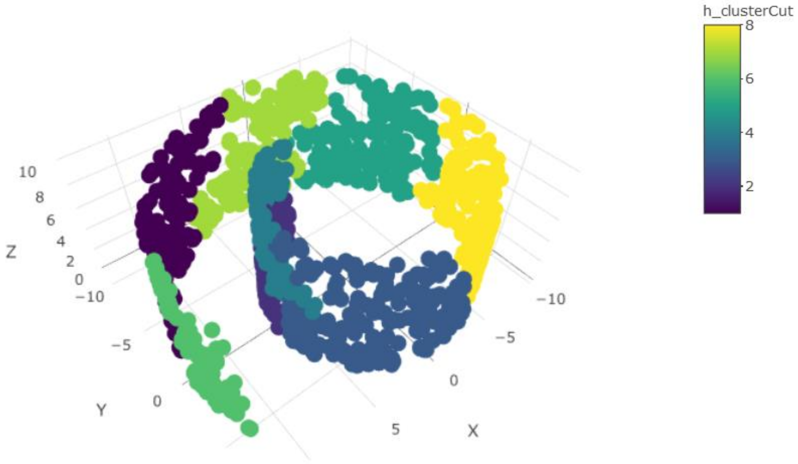
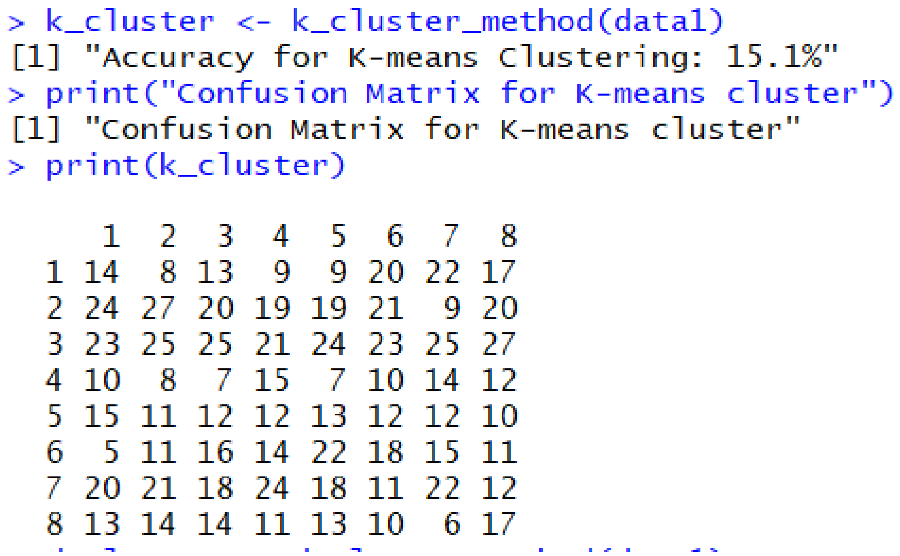


Fig: 3D Plot for Hierarchical Clustering

2. **K-means Clustering:**



```
> k_cluster <- k_cluster_method(data1)
[1] "Accuracy for K-means Clustering: 15.1%"
> print("Confusion Matrix for K-means cluster")
[1] "Confusion Matrix for K-means cluster"
> print(k_cluster)

    1  2  3  4  5  6  7  8
 1 14  8 13  9  9 20 22 17
 2 24 27 20 19 19 21  9 20
 3 23 25 25 21 24 23 25 27
 4 10  8  7 15  7 10 14 12
 5 15 11 12 12 13 12 12 10
 6  5 11 16 14 22 18 15 11
 7 20 21 18 24 18 11 22 12
 8 13 14 14 11 13 10  6 17
```

Fig: Percentage Accuracy and Confusion Matrix

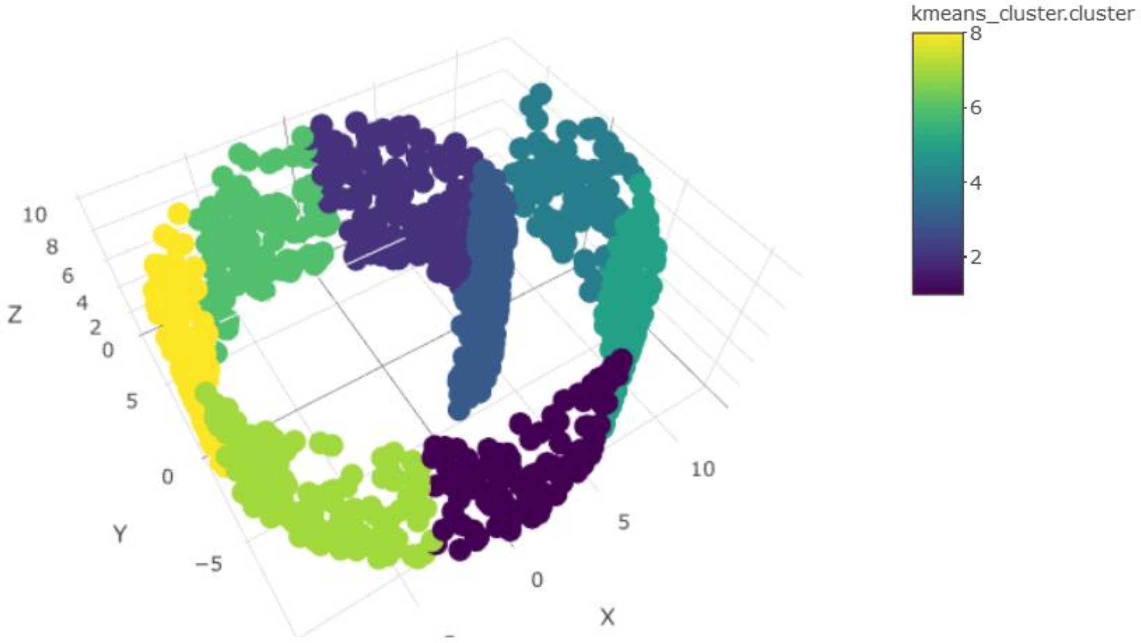
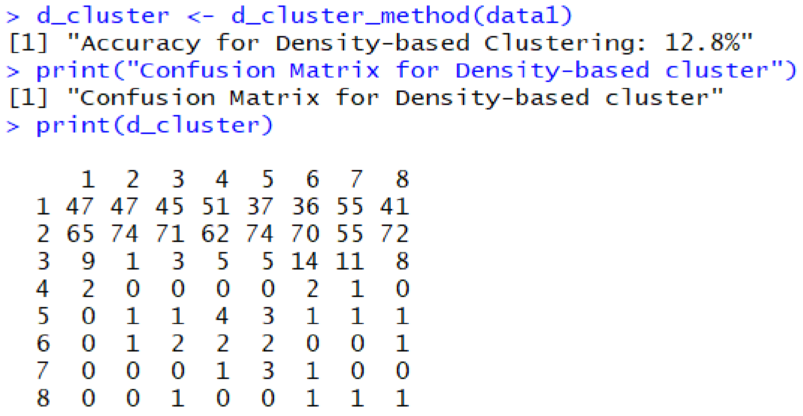


Fig: 3D Plot for K-Means Clustering

3. Density-Based Clustering:



```
> d_cluster <- d_cluster_method(data1)
[1] "Accuracy for Density-based Clustering: 12.8%"
> print("Confusion Matrix for Density-based cluster")
[1] "Confusion Matrix for Density-based cluster"
> print(d_cluster)

     1  2  3  4  5  6  7  8
  1 47 47 45 51 37 36 55 41
  2 65 74 71 62 74 70 55 72
  3  9  1  3  5  5 14 11  8
  4  2  0  0  0  0  2  1  0
  5  0  1  1  4  3  1  1  1
  6  0  1  2  2  2  0  0  1
  7  0  0  0  1  3  1  0  0
  8  0  0  1  0  0  1  1  1
```

Fig: Percentage Accuracy and Confusion Matrix

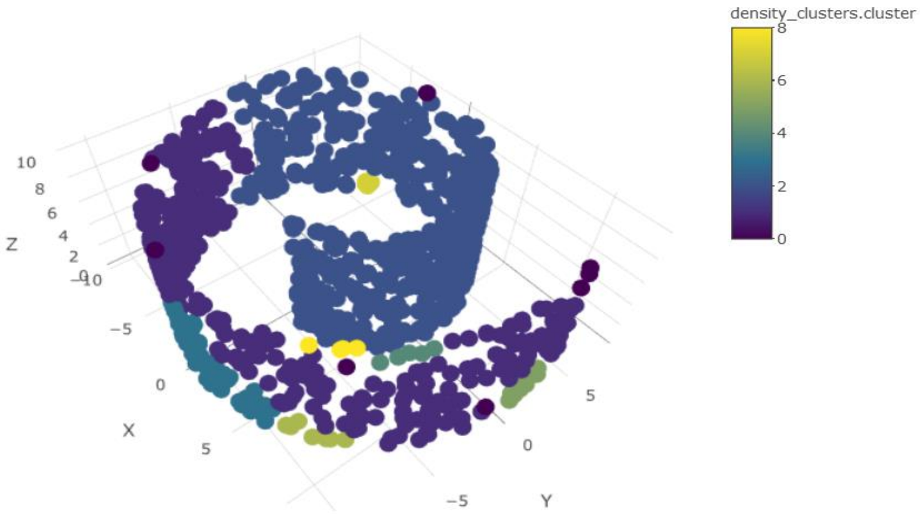


Fig: 3D Plot for Density Based Clustering

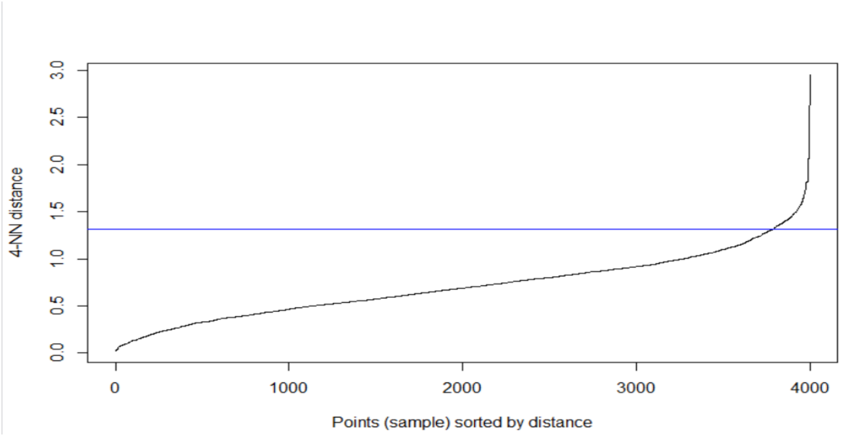
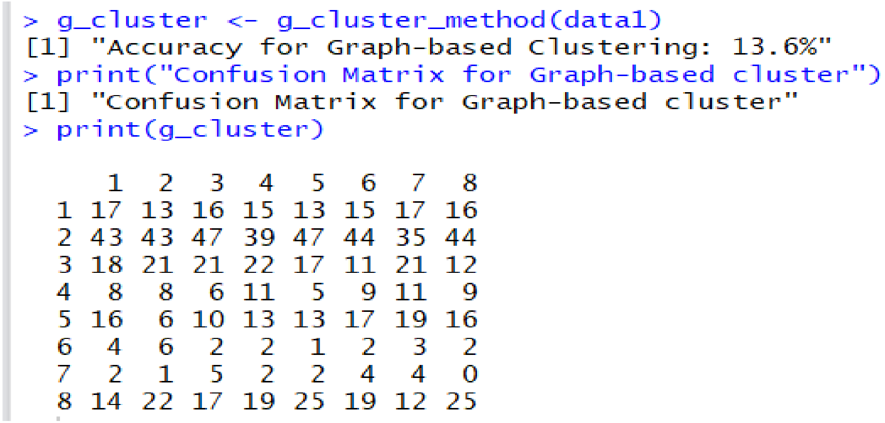


Fig: KNNdistplot for Density-Based Clustering

4. Graph-based Clustering:



```
> g_cluster <- g_cluster_method(data1)
[1] "Accuracy for Graph-based Clustering: 13.6%"
> print("Confusion Matrix for Graph-based cluster")
[1] "Confusion Matrix for Graph-based cluster"
> print(g_cluster)

     1  2  3  4  5  6  7  8
  1 17 13 16 15 13 15 17 16
  2 43 43 47 39 47 44 35 44
  3 18 21 21 22 17 11 21 12
  4  8  8  6 11  5  9 11  9
  5 16  6 10 13 13 17 19 16
  6  4  6  2  2  1  2  3  2
  7  2  1  5  2  2  4  4  0
  8 14 22 17 19 25 19 12 25
```

Fig: Percentage Accuracy and Confusion Matrix

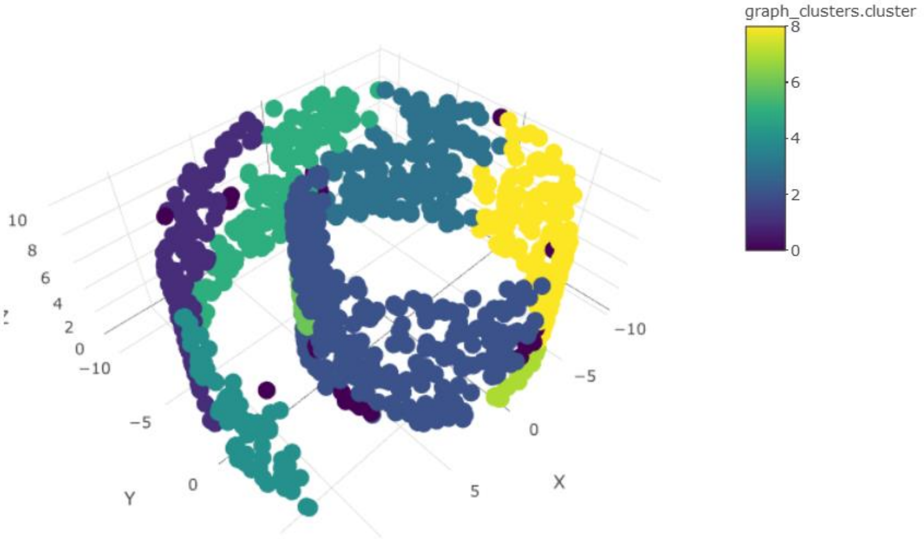


Fig: 3D Plot for Graph Based Clustering

Dataset2:
K-Means Clustering:

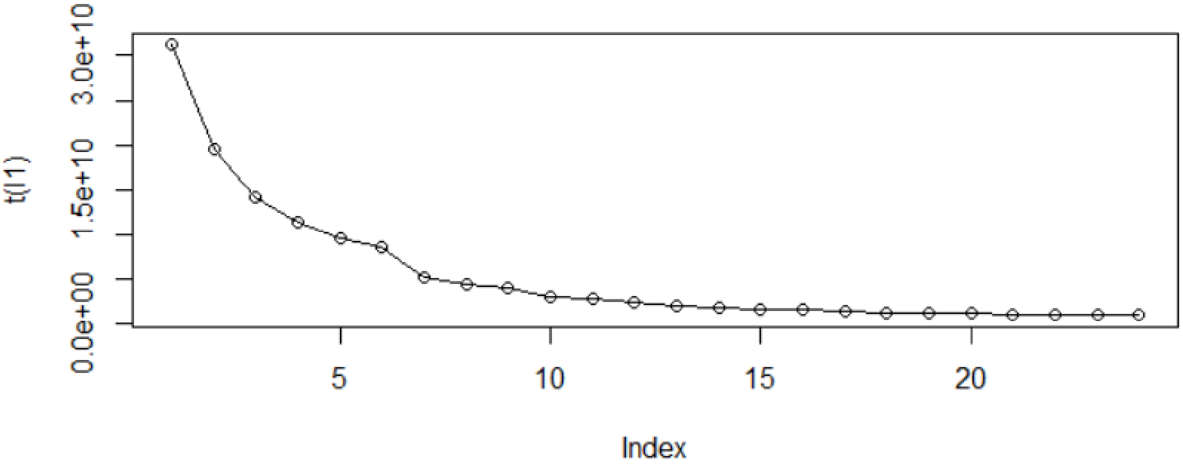


Fig: Plot for k_cluster$tot.withinss

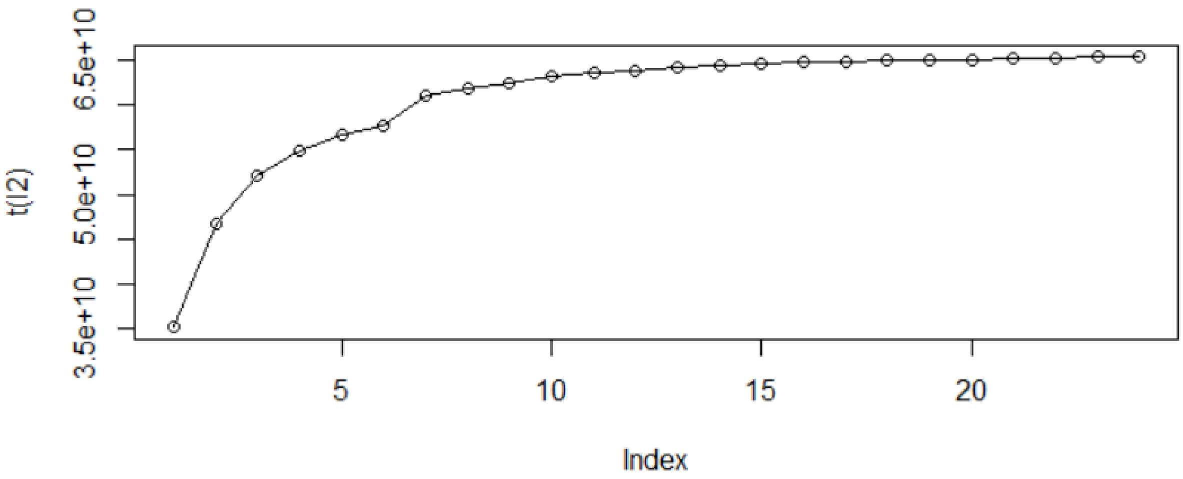
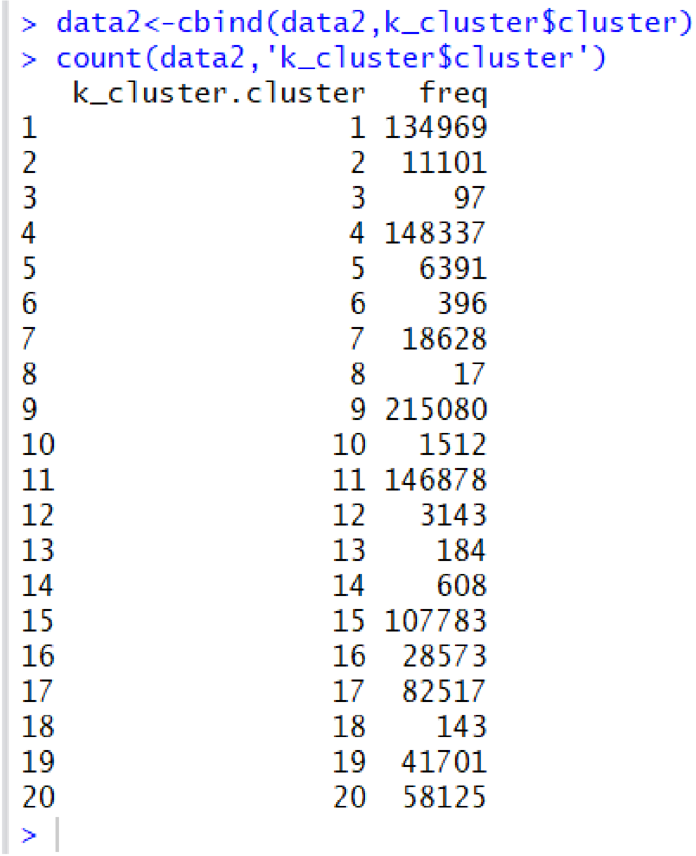


Fig: Plot for k_cluster$betweenss



```
> data2<-cbind(data2,k_cluster$cluster)
> count(data2,'k_cluster$cluster')
   k_cluster.cluster   freq
1                  1 134969
2                  2  11101
3                  3     97
4                  4 148337
5                  5   6391
6                  6    396
7                  7  18628
8                  8     17
9                  9 215080
10                10   1512
11                11 146878
12                12   3143
13                13    184
14                14    608
15                15 107783
16                16  28573
17                17  82517
18                18    143
19                19  41701
20                20  58125
>
```

Fig: Frequency of data for each clusters in dataset2

Conclusion:

For dataset1, it can be concluded that the K-Means Clustering gives the highest accuracy of 15.1%

Clustering Technique	**Percentage Accuracy**
Hierarchical Clustering	14.9%
K-Means Clustering	15.1%
Density Based Clustering	12.8%
Graph Based Clustering	13.6%

For dataset2, it can be concluded that using the K-Means clustering, the dataset can be divided into 20 clusters and the frequency in each cluster is given above in the results.

Reference list:

1. https://www.rdocumentation.org/packages/stats/versions/3.4.1/topics/hclust
2. https://www.rdocumentation.org/packages/dendextend/versions/1.4.0/topics/cutree
3. https://www.rdocumentation.org/packages/plotly/versions/4.7.1/topics/plot_ly
4. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html
5. https://cran.r-project.org/web/packages/dbscan/README.html
6. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/dist.html
7. https://www.rdocumentation.org/packages/dbscan/versions/1.1-1/topics/kNNdist
8. https://www.rdocumentation.org/packages/dbscan/versions/1.1-1/topics/sNNclust
9. http://www.r-tutor.com/r-introduction/list
10. https://www.rdocumentation.org/packages/graphics/versions/3.4.0/topics/plot
11. https://stat.ethz.ch/R-manual/R-devel/library/base/html/source.html