Assignment 2

DMW-2

▷▷ PROBLEM STATEMENT: Consider a suitable dataset. For clustering of data instances in different groups, apply different clustering techniques (minimum 2). Visualize the clusters using suitable tool.

▷▷ OBJECTIVE: 1) Understand various clustering types and how to implement the same.
2) Use python libraries and appropriate datasets to perform clustering and visualize the same.

▷▷ ~~PROBLEM~~ OUTCOME: Understood K-Means, Hierarchical (Agglomerative) clustering and performed it on dataset ~~'Market~~ cars data by brand.

▷▷ SOFTWARE & HARDWARE REQUIREMENT: Any CPU with i3 processor or higher, 8 GB RAM or higher, 1 GB HDD; 64 bit LINUX / UNIX OS;

▷▷ THEORY:

- Clustering, or cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more ~~responsible~~ similar to each other than to those in other groups.

• Clustering is the main task of exploratory data mining, and a common technique for statistical data analysis, and is used in many fields.

• K-means clustering is a type of unsupervised learning, which is used when you unlabeled data (i.e. data without defined categories or groups). The goal of this algorithm ~~works~~ is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively ~~to~~ assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

1. The centroids of the K ~~features~~ clusters, which can be used to label new data
2. Labels for the training data (each data point is assigned to a single cluster)

• Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically.

• Hierarchical clustering is an algorithm that groups similar objects into groups called clusters. The endpoints is a set of clusters, where each cluster is distinct from each other, and the objects within each cluster are ~~wide~~ broadly similar to one another.

- Given a set of N items to be clustered, and an NxN distance matrix, the basic process of hierarchical clustering is:
  1) ~~find~~ assign each item to its own cluster, so now there are N clusters; let the distance between the clusters equal the distances (similarities) between the items they contain.
  2) find the closest (most similar) pair of clusters and merge them into a single cluster.
  3) compute the distances (similarities) between the new cluster and each of the old clusters
  4) Repeat steps 2 and 3 until all clusters are clustered into a single cluster of size N.

- The default distance measure is the Euclidean distance, which is the square root of the sum of the square differences.

- In the agglomerative clustering approach, there are four possible methods, ward's method being one of them. It says that the distance between two clusters, A and B, is how much the sum of squares will increase when they are merged.

- The dataset used was the 'cars' which contained the parameters 'mileage', 'hp', 'weight', 'year', and 'brand'.

- The entries were clustered by brand vs. weight.

CONCLUSION

K-means and hierarchical (agglomerative) clustering techniques were understood, successfully implemented; and the required output was obtained.