

Assignment No. 2

* Problem Statement :

Consider a Suitable dataset, for clustering of data instances in diff. groups, apply diff. clustering techniques (min. 2) - Visualize the clusters using Suitable tools.

* Objective : i) understand various clustering types & how to implement the same. ii) use Python libraries & appropriate data sets to perform clustering & visualize the same.

* Outcome : Understood K-means, Hierarchical clustering & perform it on dataset cars data by brand.

* Theory :

- ① Clustering or cluster analysis is a task of grouping a set of objects in such a way that objects in the same group (cluster) are more similar to each other than to those in each other.
- ② Clustering is the main task of exploratory data mining & a common

technique for statistical data analysis, and is used in many fields.

- ① 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 is to find groups in the data, with the number of groups represented by variable k . The algorithm works iteratively to assign each data point to one of k groups based on the feature provided. Data points are clustered based on feature similarity. The result of this algorithm is -
 - ⇒ the centroid of k -clusters, can be used to label new data
 - ⇒ labels from the training data (each data point is assigned to single cluster)

- ② Rather than defining groups, before looking at the data, clustering allows you to find & analyze the groups that have formed organically.

- ③ Hierarchical clustering is an algorithm that groups similar objects into groups called cluster. The end point is a set of cluster, where each cluster is distinct from each other, and the object

within each cluster are broadly similar to each other.

① Given a set of N items to be clustered and an $N \times N$ distance matrix, the basic process of hierarchical clustering is:

→ assign each item to its own cluster, so now there are N clusters; let the dist. between the cluster equal the dist.

(Similarities) between the items they contain.

→ find the closest (most similar) pair of clusters and merge them into a single cluster

→ compute the distances (similarities) between the new cluster & each of the old clusters

→ Repeat above 2 steps until all clusters are clustered into a single cluster of size N .

② The default distance measured in the Euclidean dist. which is the Sq. root of the Sum of the differences.

③ In the agglomerative clustering approach, there are 4 possible methods, Ward's method being one of them. It says that the dist. between two clusters, A & B , is how much the Sum of Squares will increase when they are merged



- ① The dataset used was 'mall-customers' which contains various parameters.
- ② these entries were clustered using the algorithm.

* Conclusion :

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