**Assignment No:-4**

**Problem Statement:-**

Write a program to do following:

We have given a collection of 8 points. P1=[0.1,0.6] P2=[0.15,0.71] P3=[0.08,0.9] P4=[0.16,

0.85] P5=[0.2,0.3] P6=[0.25,0.5] P7=[0.24,0.1] P8=[0.3,0.2]. Perform the k-mean clustering

with initial centroids as m1=P1=Cluster#1=C1 and m2=P8=cluster#2=C2.

Answer the following:

a) Which cluster does P6 belong to?

b) What is the population of a cluster around m2?

c) What is the updated value of m1 and m2?

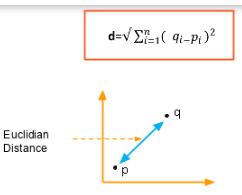
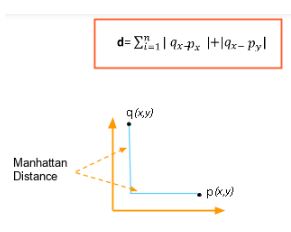
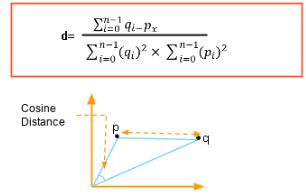
**Theory**:-

K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

The term ‘K’ is a number. You need to tell the system how many clusters you need to create. For example, K = 2 refers to two clusters. There is a way of finding out what is the best or optimum value of K for a given data.

Distance measure determines the similarity between two elements and influences the shape of clusters.

K-Means clustering supports various kinds of distance measures, such as:

* Euclidean distance measure
* 
* Manhattan distance measure
* 
* A squared euclidean distance measure
* squared eu
* Cosine distance measure
* 

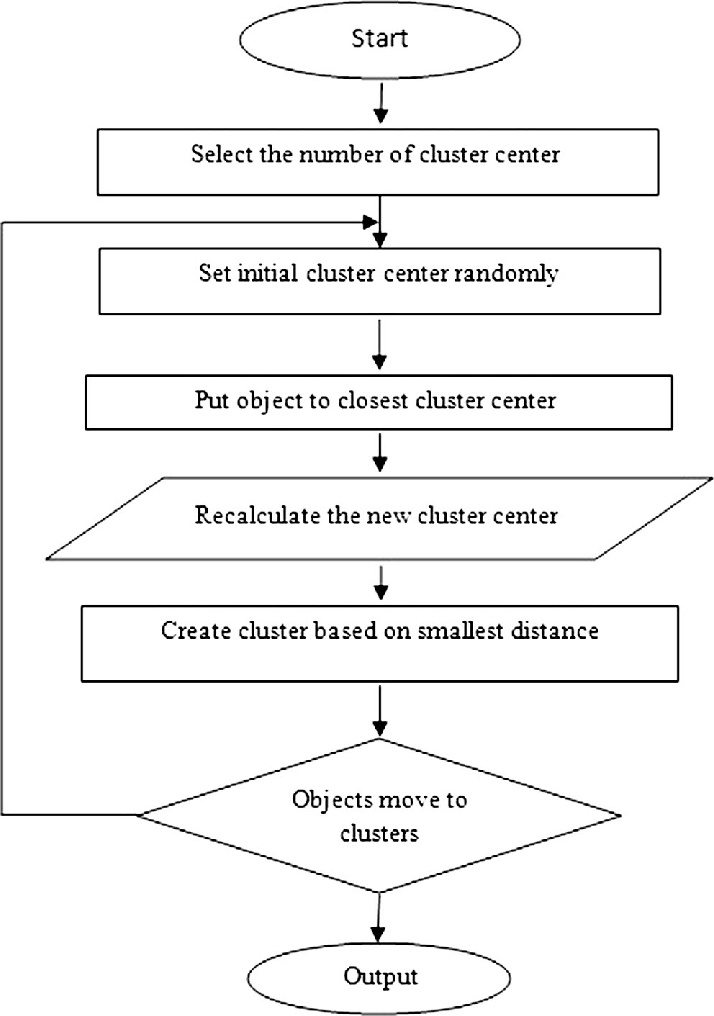
Advantages and Disadvantages & Limitation/Example:

1. Advantages:
   * Simple and Intuitive: K-means clustering is easy to understand and implement.
   * Efficient: It works well for large datasets and can handle high-dimensional data efficiently.
   * Scalability: K-means scales well with increasing dataset sizes.
   * Interpretability: Results are straightforward and easy to interpret.
2. Disadvantages & Limitations/Example:
   * Sensitivity to Initial Centroids: Results can vary depending on the initial centroid selection.
   * Assumption of Spherical Clusters: K-means assumes that clusters are spherical, which may not always be the case.
   * Impact of Outliers: Outliers can significantly affect the cluster centroids and result in suboptimal clustering.
   * Determining Number of Clusters: The number of clusters needs to be specified beforehand, which can be subjective and challenging to determine.

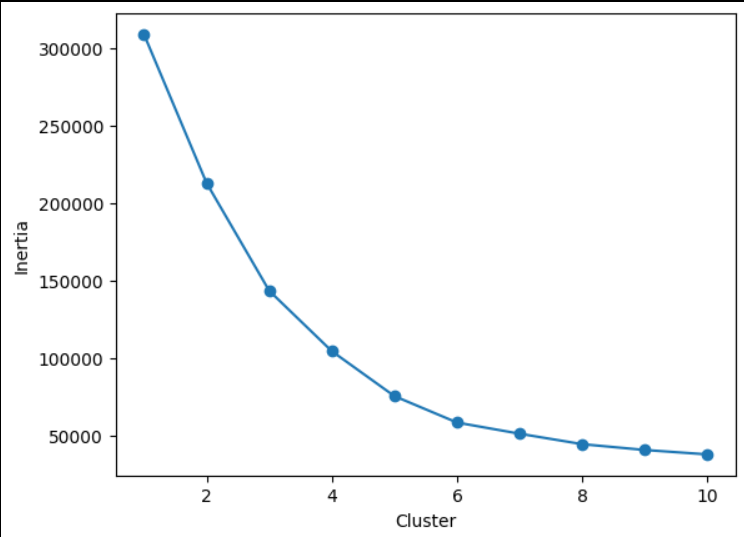
**Methodology:**

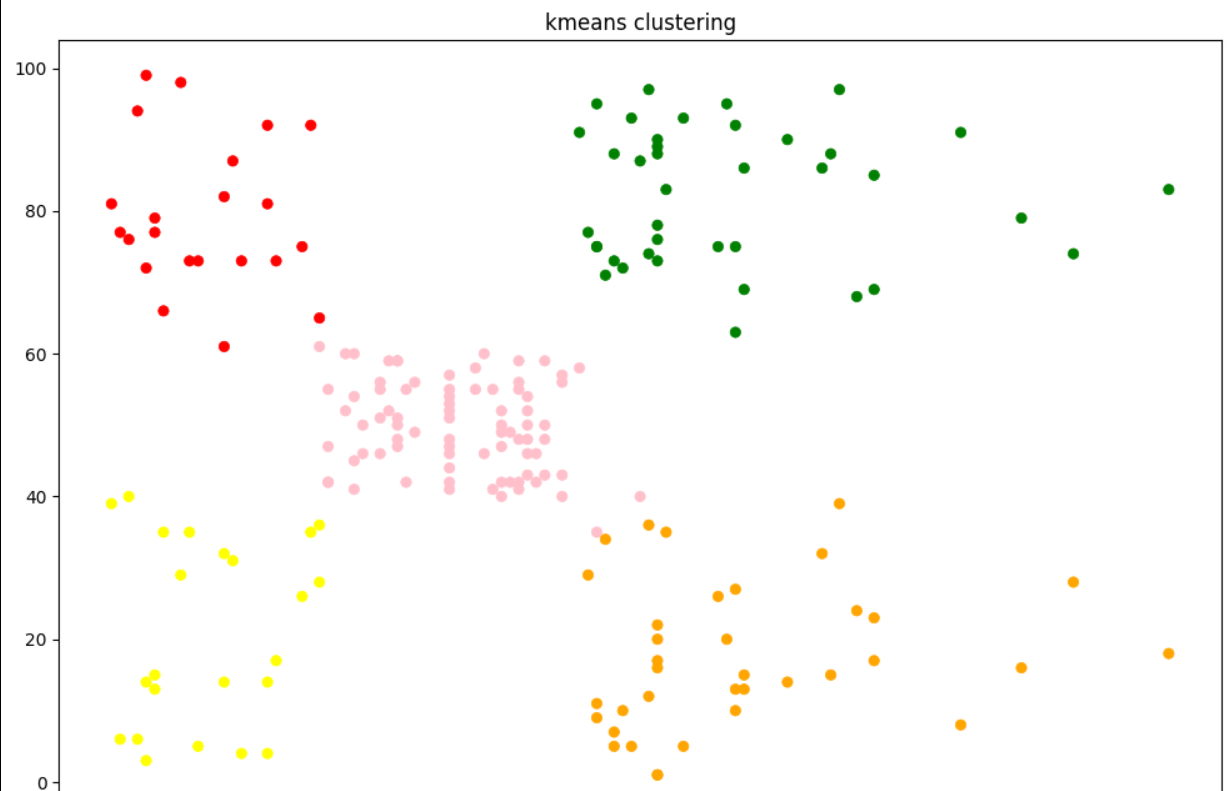
1. Initialization: Start by defining the initial centroids, in this case, m1=P1 and m2=P8.
2. Assign Points to Clusters: Calculate the Euclidean distance between each point and both centroids. Assign each point to the cluster corresponding to the nearest centroid.
3. Update Centroids: Calculate the mean of all points belonging to each cluster. Update the centroids to the new mean values.
4. Repeat: Iterate steps 2 and 3 until the centroids no longer change significantly or until a specified number of iterations is reached.
5. Answering Questions:
   * To determine which cluster P6 belongs to, calculate its distance from both centroids and assign it to the cluster with the nearest centroid.
   * To find the population of the cluster around m2, count the number of points assigned to cluster C2.
   * To update the values of m1 and m2, calculate the mean of points in each cluster and set m1 and m2 to these new mean values.

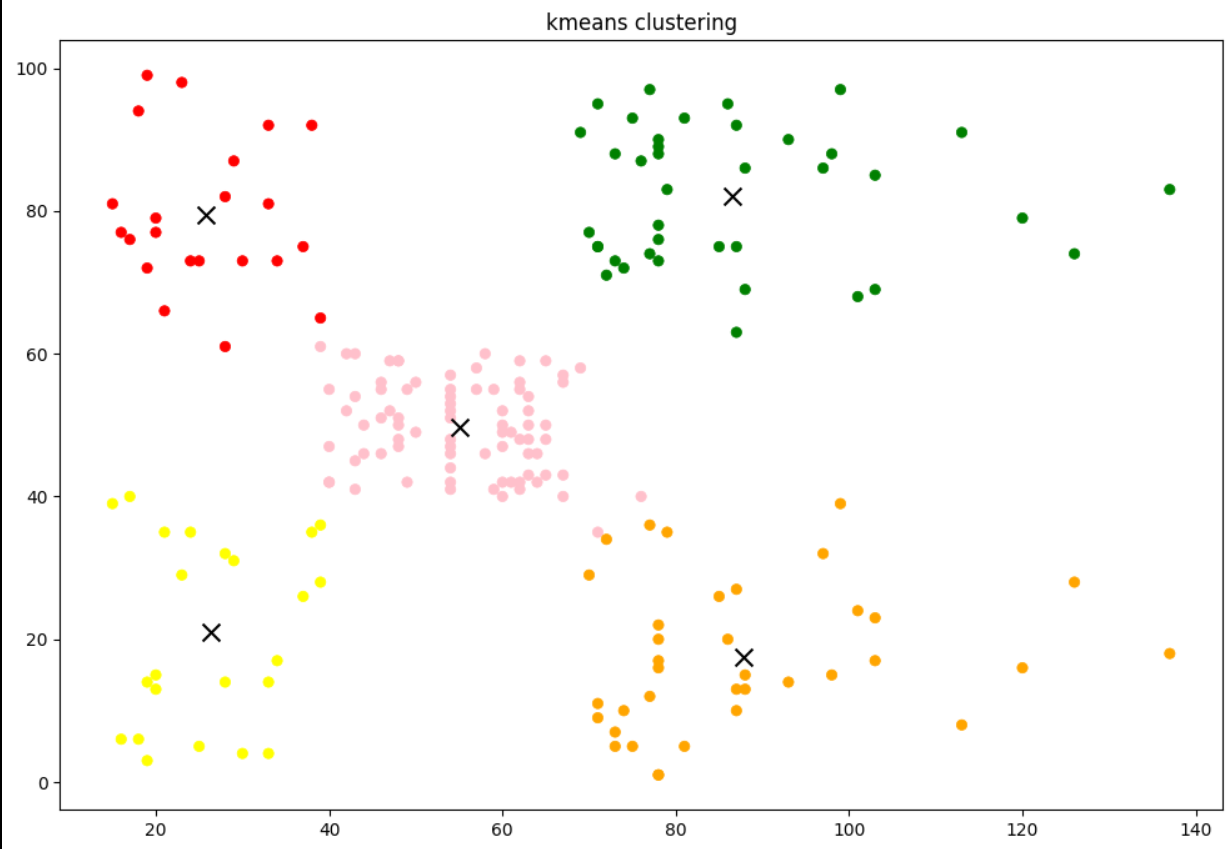
**OUTPUT**

****

**Diagram**

****

****

****

**Conclusion**

In conclusion, K-means clustering is a simple yet effective method for grouping data points into clusters. While it's straightforward to implement and scales well with large datasets, it requires careful consideration of initial centroids and assumes spherical clusters. Despite its limitations, K-means remains a popular choice for clustering tasks due to its efficiency and interpretability.