ASSIGNMENT - 02

Date: 08/08/2020

Title: Clustering

Problem definition: Consider a suitable dataset for clustering of data instances in differentgroups, apply different clustering techniques (min. 2). Visualize clusters using suitable tools.

S/W & H/w requirements : · R tool / Anaconda Python
· PIV, 2GB RAM, 500 GB HDD.

Learning objectives: Use R functions / scikit-learn functions to create K-Means clustering models and heirarchial clustering models.

Learning outcomes: Visualize the effects of K-means and heirarchial clustering using graphic capabilities.

Theory:

1] K-means Clustering.

- It is a type of unsupervised learning, which is used when you have unlabelled data.
- The goal of this algorithm is to find groups in the data, with the number of groups

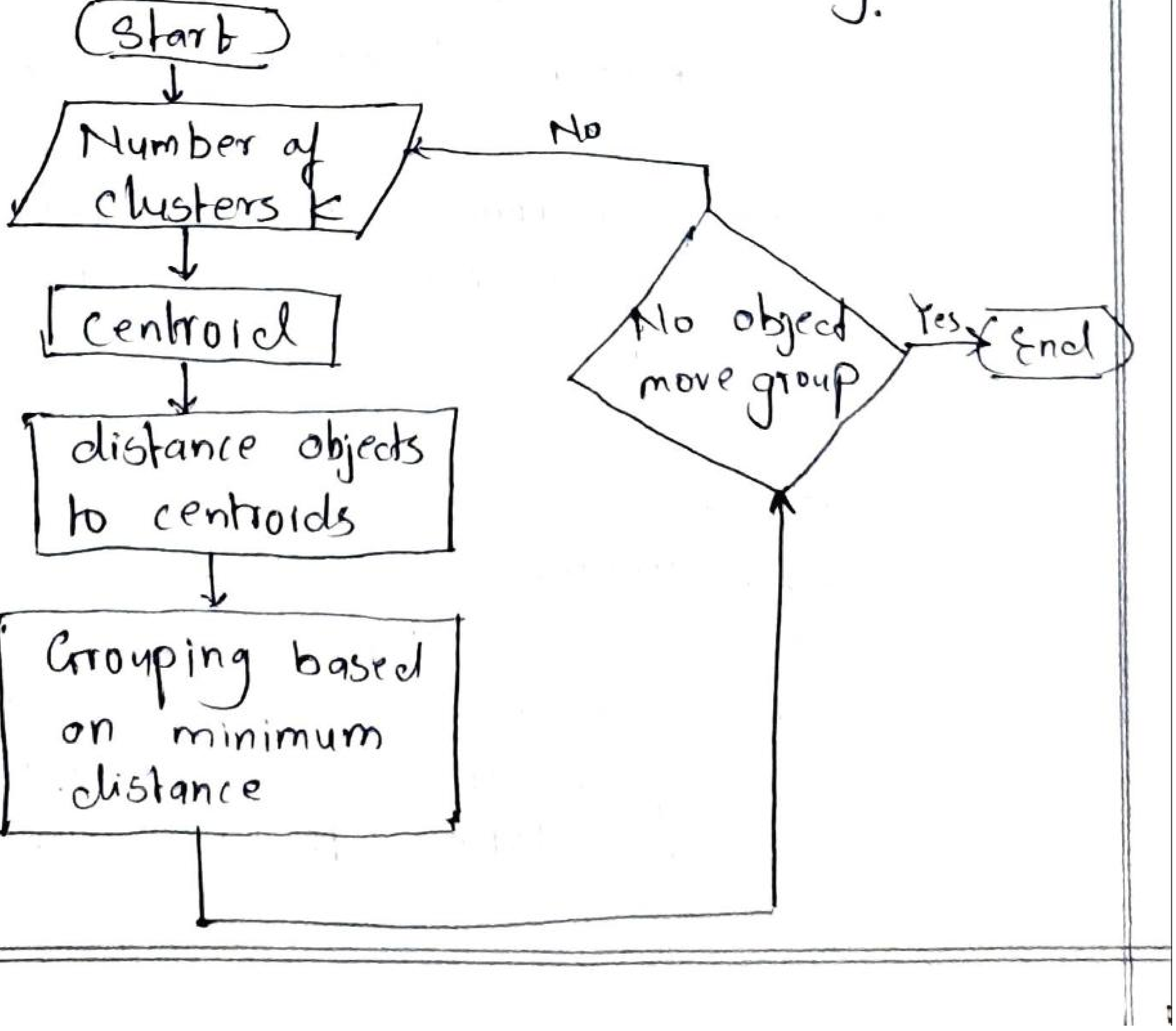
represented by the variable K.

- The algorithm works iteratively to assign each data point to one of K groups based on features that are provided.
- Data points are clustered based on feature similarity.
- The results of K-means clustering algorithm are :

    1. The centroids of the K-clusters, which can be used to label new data.
    2. Labels for 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 been formed organically.

→ steps to perform K-means Clustering.


Start
Number of clusters K
Centroid
distance objects to centroids
Grouping based on minimum distance
No object move group
No
Yes
End

Scanned by TapScanner

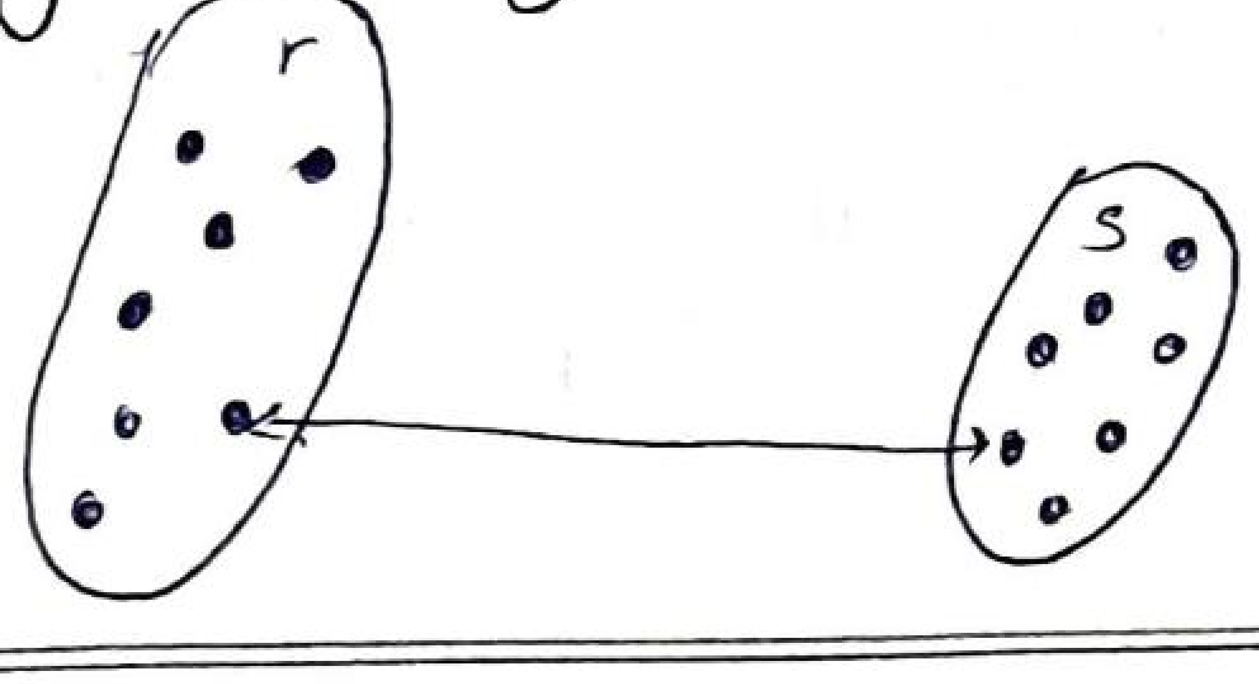
B] Hierarchial Clustering

Hierarchial clustering involves creating clusters that have pre-determined ordering from top to bottom.

There are two types of hierarchial clustering.

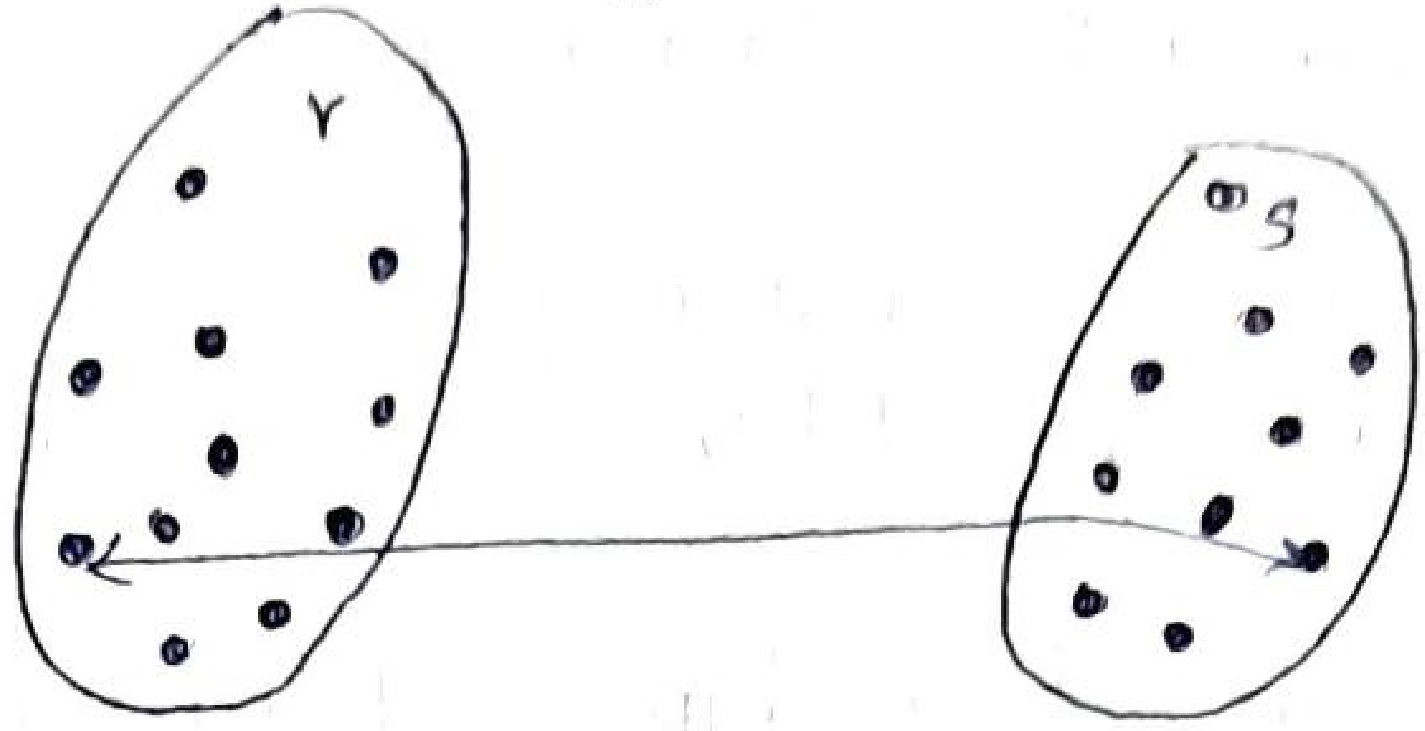
1. Divisive method
   - It is also known as top-down clustering. We assign all of observations to a single cluster and the partition cluster to two least similar clusters.
   - Finally, we proceed recursively on each cluster until there is one cluster for each observation.

2. Agglomerative method.
   - It is also known as bottom-up clustering.
   - We assign each observation to its own cluster.
   - Computation algorithim :-
     1. Compute the proximity matrix.
     2. Let each data point be a cluster.
     3. Repeat : Merge the two closest clusters and update the proximity matrix.
     4. Until only a single cluster remains
   - Following are the methods to determine proximity matrix.
     - Single Linkage



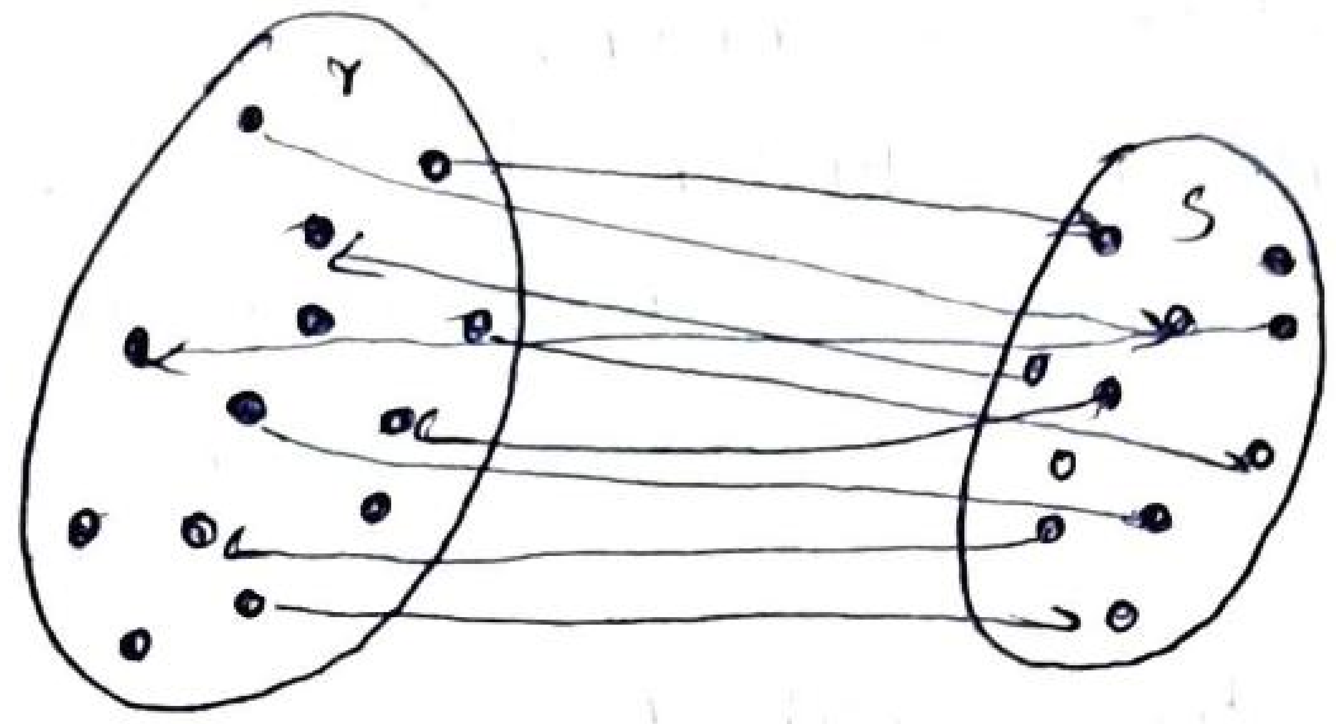
$L(r,s) = \min(D(x_{ri}, x_{sj}))$

2. Complete linkage.



$$L(r,s) = max(D(x_{ri}, x_{sj}))$$

3. Average Linkage.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Test cases:

Sr. no.	Description.	Expected o/p	Actual o/p
1.	In hierarchial clustering construct a dendrogram using "ward laverge" method.	No. of clusters rendered = 5	success
2.	Visuale cluster using single, complete and average linkages.	~~A~~ clusters are displayed by means of scatter plot	Success
3.	While fitting k-means to dataset, put random_state = 42.	Success	success.

Conclusion: Hence, we have successfully implemented hierarchial clustering and K-means clustering algorithm in python using jupyter notebooks.