EXPERMINT: 08

● <u>Aim</u>: Implementation of a hierarchical clustering method.

● <u>Theory</u>:

● <u>Hierarchical clustering</u>

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabelled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.

Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

The hierarchical clustering technique has two approaches:

- **Agglomerative**: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
- **Divisive**: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

**<u>How the Agglomerative Hierarchical clustering Work</u>?**

Step-1: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.
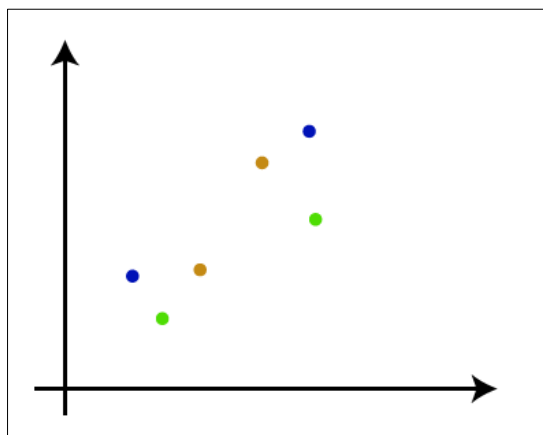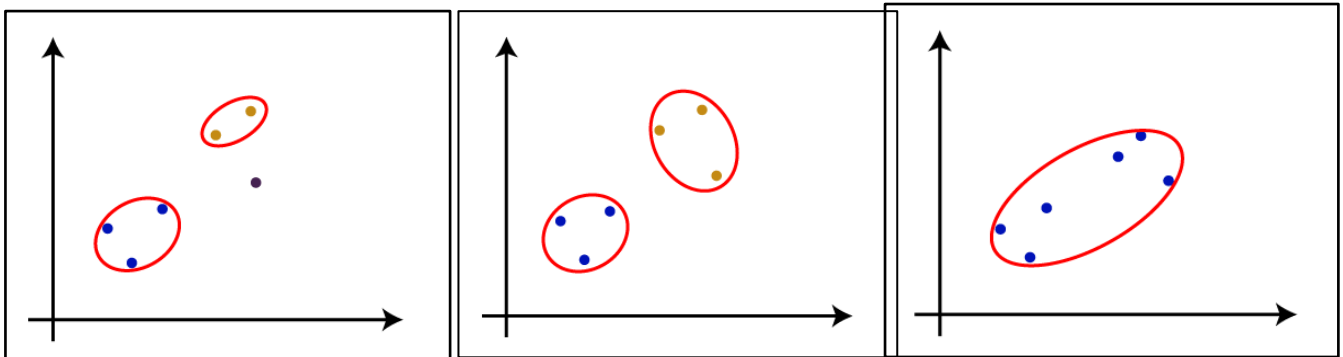
Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.
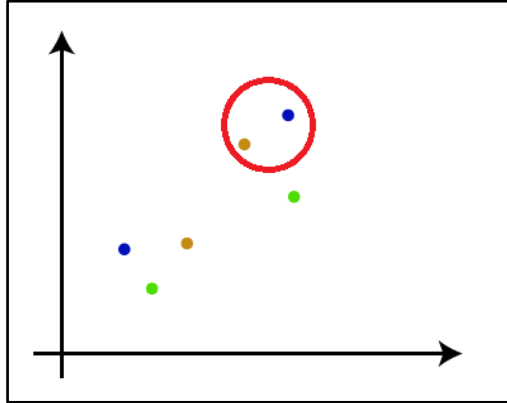
Step-3: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.
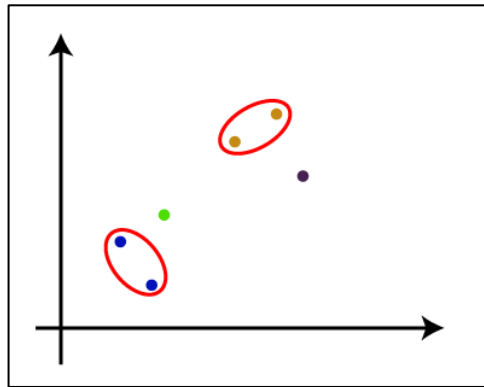
Step-4: Repeat Step 3 until only one cluster left. So, we will get the following clusters. Consider the below images:
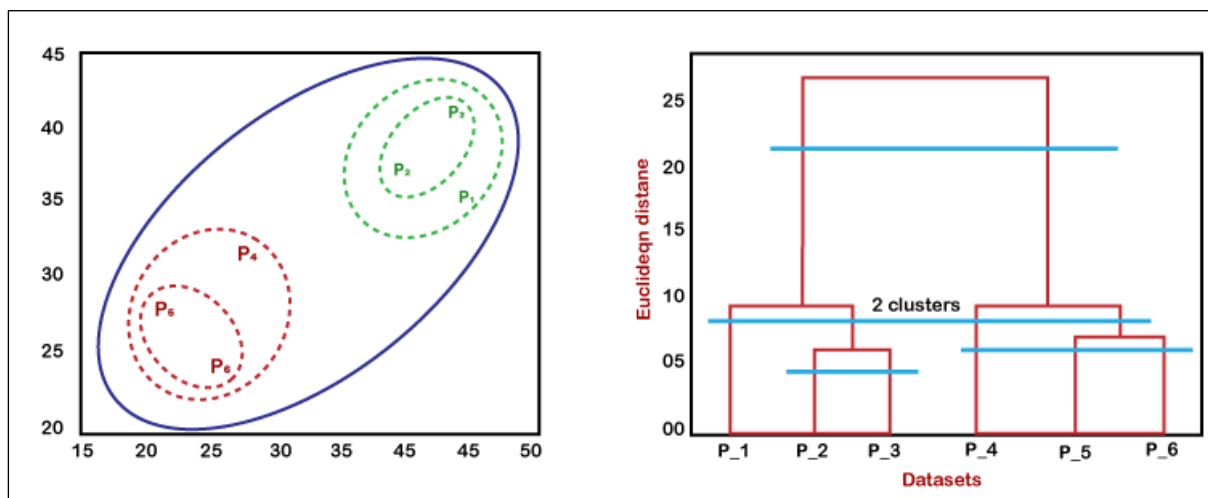

Step-5: Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.
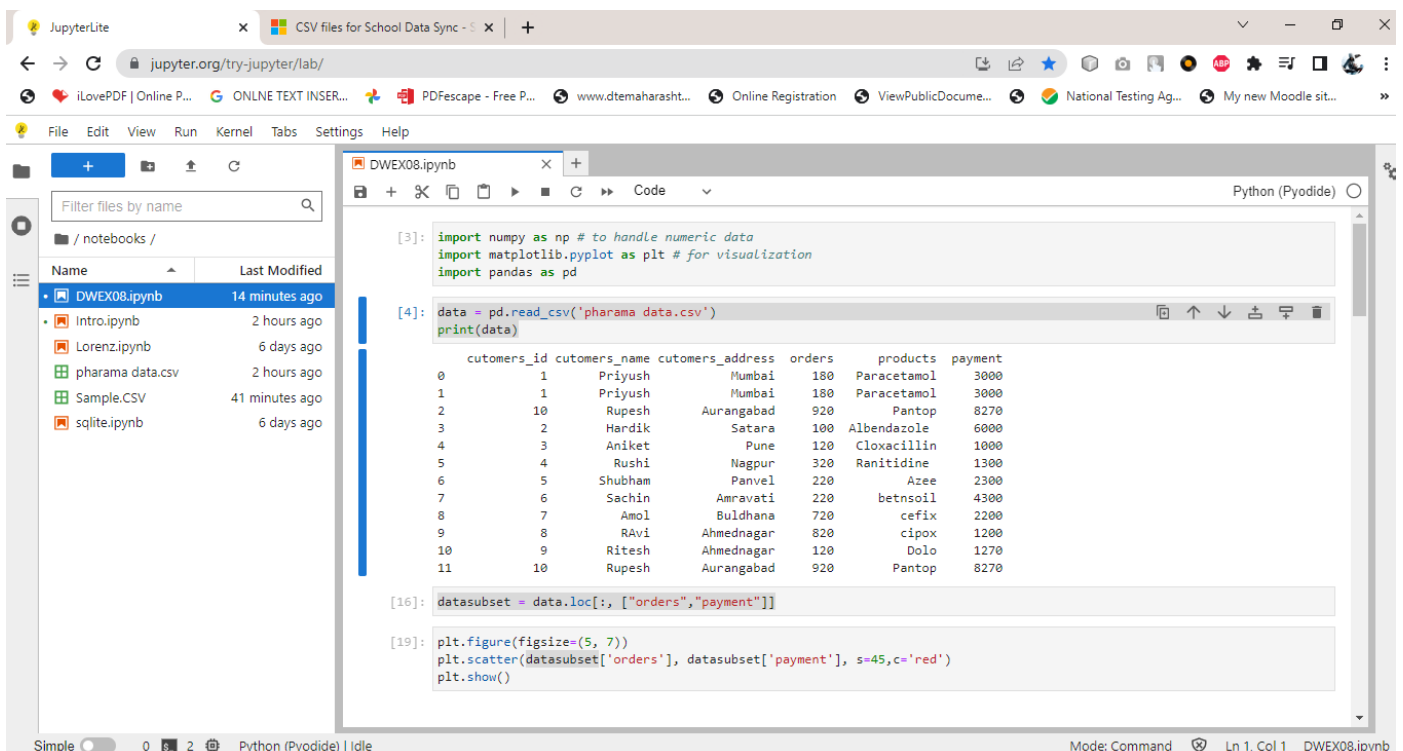
Woking of Dendrogram in Hierarchical clustering

The dendrogram is a tree-like structure that is mainly used to store each step as a memory that the HC algorithm performs. In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset.

The working of the dendrogram can be explained using the below diagram:
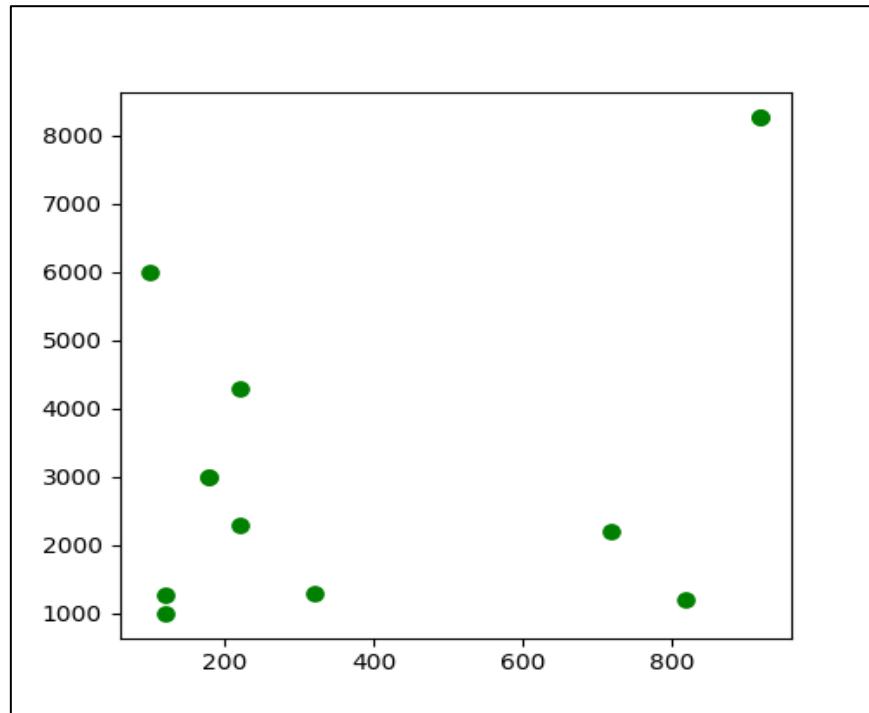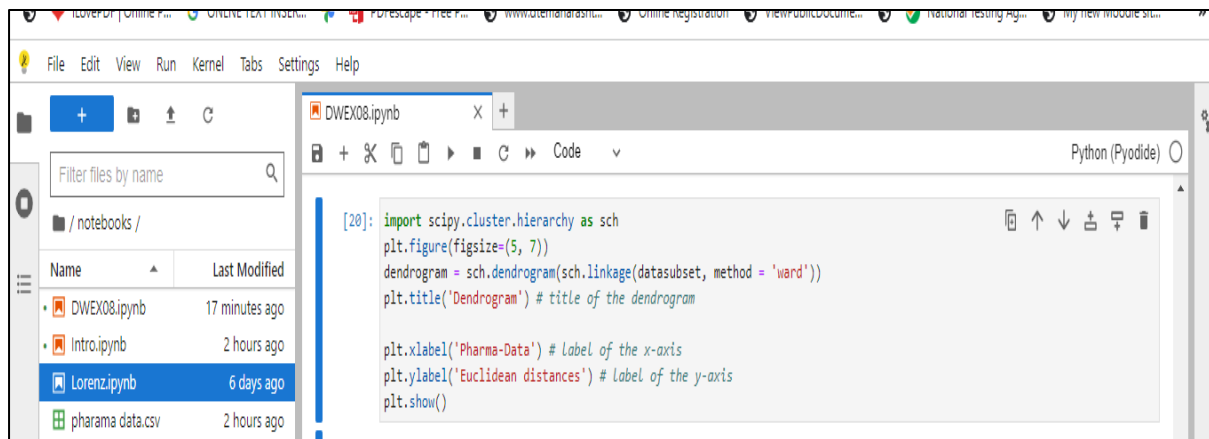

Program:

```python
[20]: import scipy.cluster.hierarchy as sch
      plt.figure(figsize=(5, 7))
      dendrogram = sch.dendrogram(sch.linkage(datasubset, method = 'ward'))
      plt.title('Dendrogram') # title of the dendrogram

      plt.xlabel('Pharma-Data') # label of the x-axis
      plt.ylabel('Euclidean distances') # label of the y-axis
      plt.show()
```
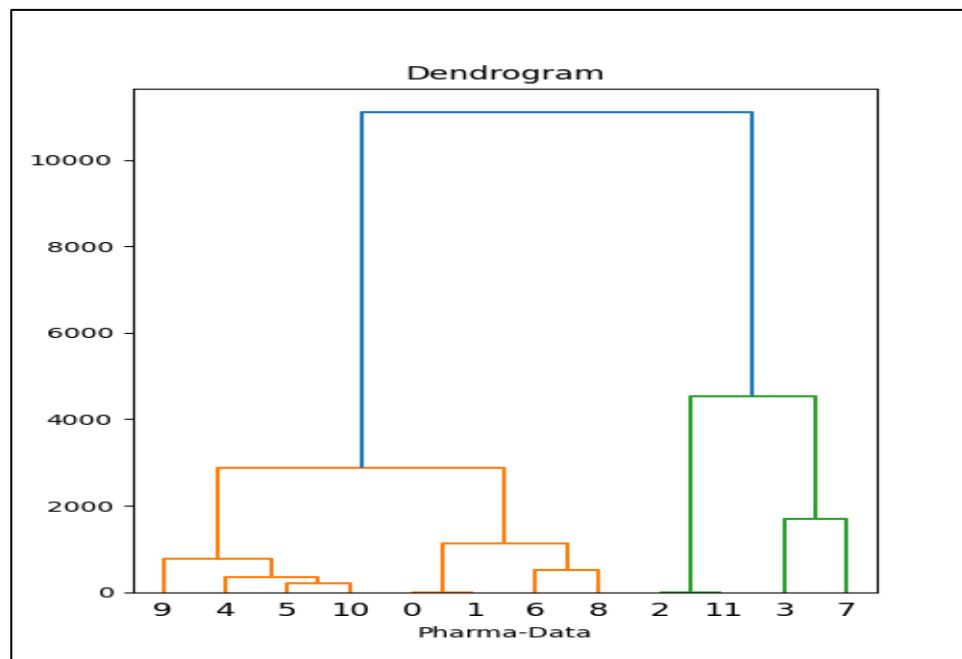
```python
[24]: from sklearn.cluster import AgglomerativeClustering
      cluster= AgglomerativeClustering(n_clusters = 2, affinity = 'euclidean', linkage = 'ward')
      cluster. fit_predict(datasubset)

[24]: array([1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0], dtype=int32)
```

- **Conclusion: -** Thus we learnt and implemented hierarchical clustering algorithm