

### EXPERIMENT: 08

- **Aim:** Implementation of a hierarchical clustering method.

- **Theory:**

- **Hierarchical clustering**

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.

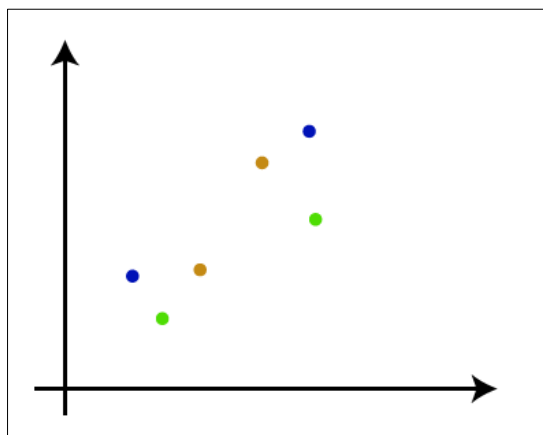
Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

The hierarchical clustering technique has two approaches:

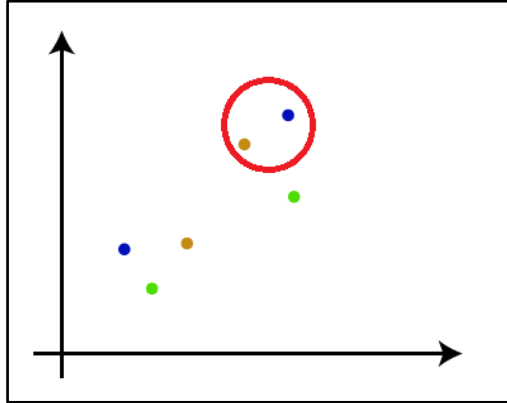
- **Agglomerative:** Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
  - **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

#### **How the Agglomerative Hierarchical clustering Work?**

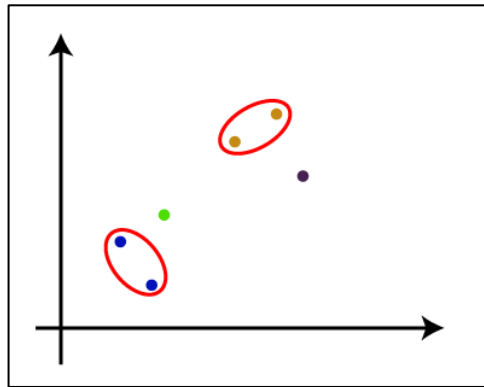
**Step-1:** Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.



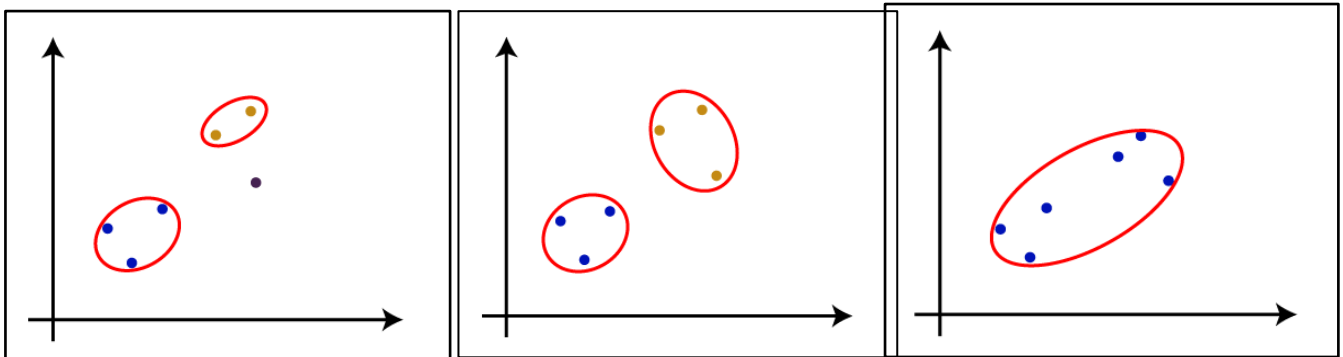
**Step-2:** Take two closest data points or clusters and merge them to form one cluster. So, there will now be  $N-1$  clusters.



**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:

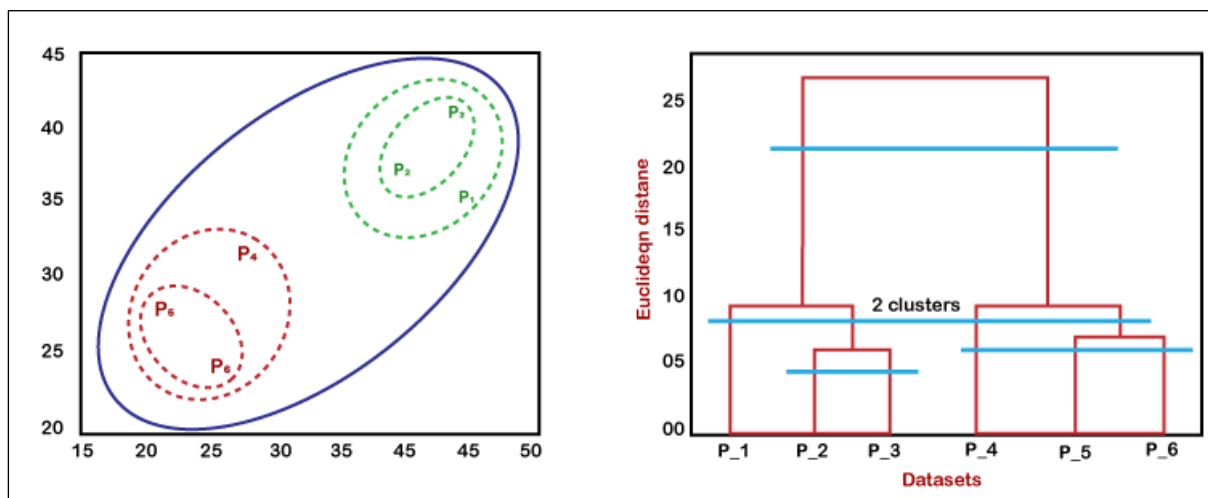


**Step-5:** Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

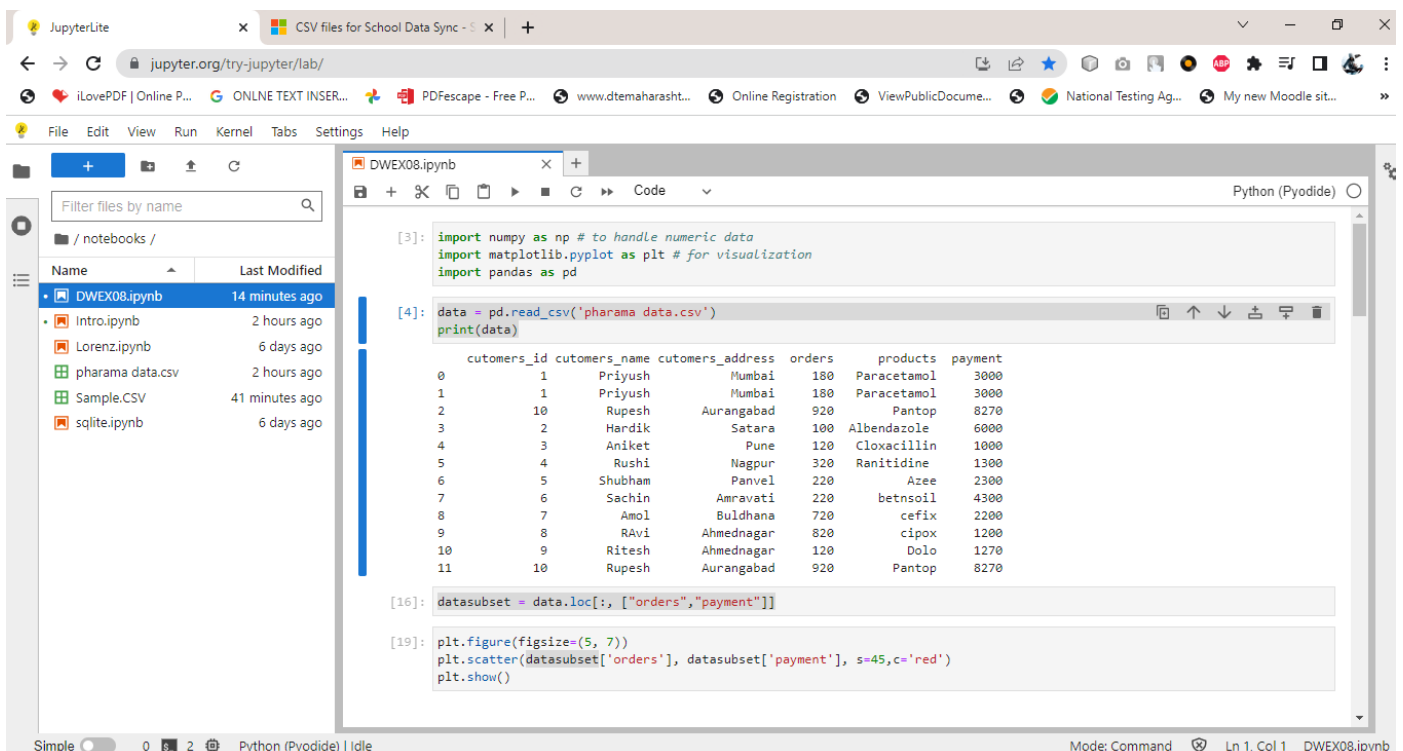
### Working of Dendrogram in Hierarchical clustering

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:



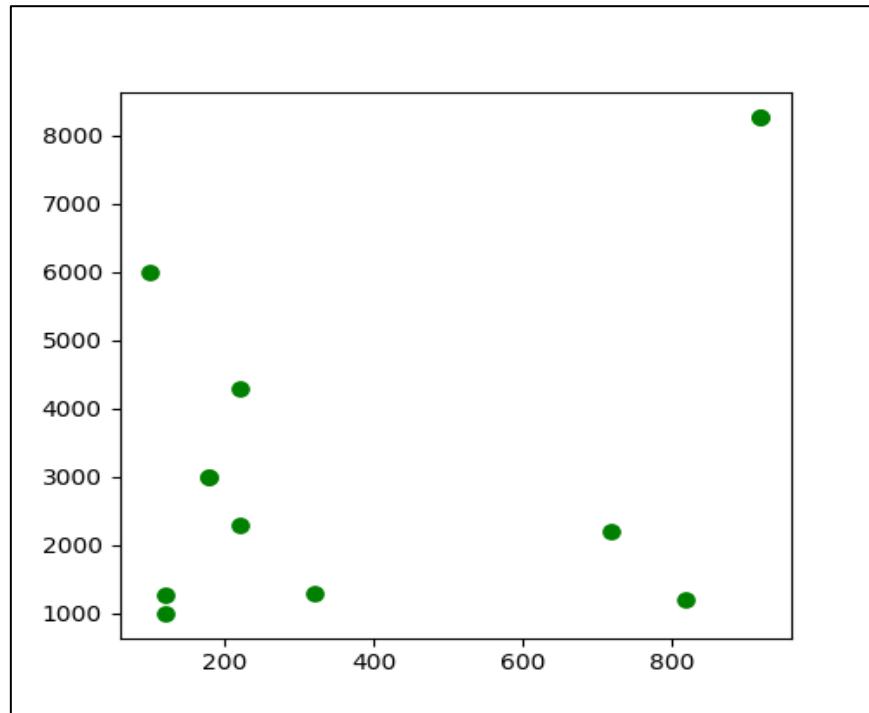
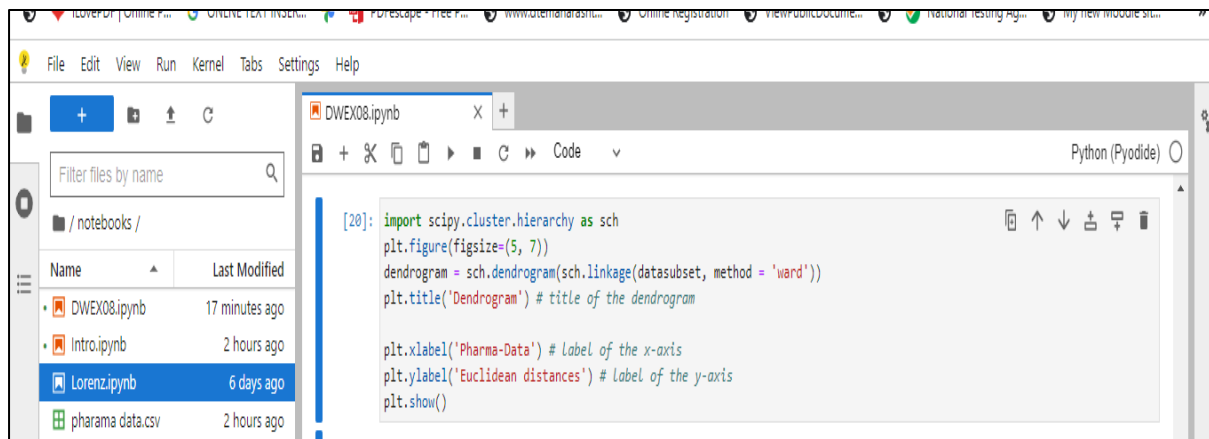
```
[3]: import numpy as np # to handle numeric data
import matplotlib.pyplot as plt # for visualization
import pandas as pd

[4]: data = pd.read_csv('pharama data.csv')
print(data)

customers_id customers_name customers_address orders products payment
0 1 Priyush Mumbai 180 Paracetamol 3000
1 1 Priyush Mumbai 180 Paracetamol 3000
2 10 Rupesh Aurangabad 920 Pantop 8270
3 2 Hardik Satara 100 Albendazole 6000
4 3 Aniket Pune 120 Cloxacillin 1000
5 4 Rushi Nagpur 320 Ranitidine 1300
6 5 Shubham Panvel 220 Azee 2300
7 6 Sachin Amravati 220 betnsoil 4300
8 7 Amol Buldhana 720 cefix 2200
9 8 Ravi Ahmednagar 820 cipox 1200
10 9 Ritesh Ahmednagar 120 Dolo 1270
11 10 Rupesh Aurangabad 920 Pantop 8270

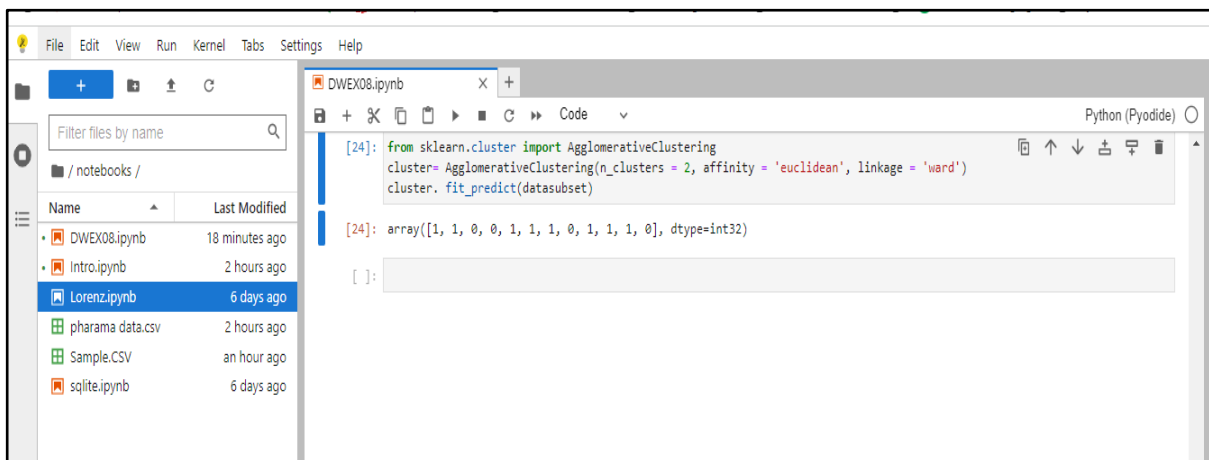
[16]: datasubset = data.loc[:, ["orders","payment"]]

[19]: plt.figure(figsize=(5, 7))
plt.scatter(datasubset['orders'], datasubset['payment'], s=45,c='red')
plt.show()
```

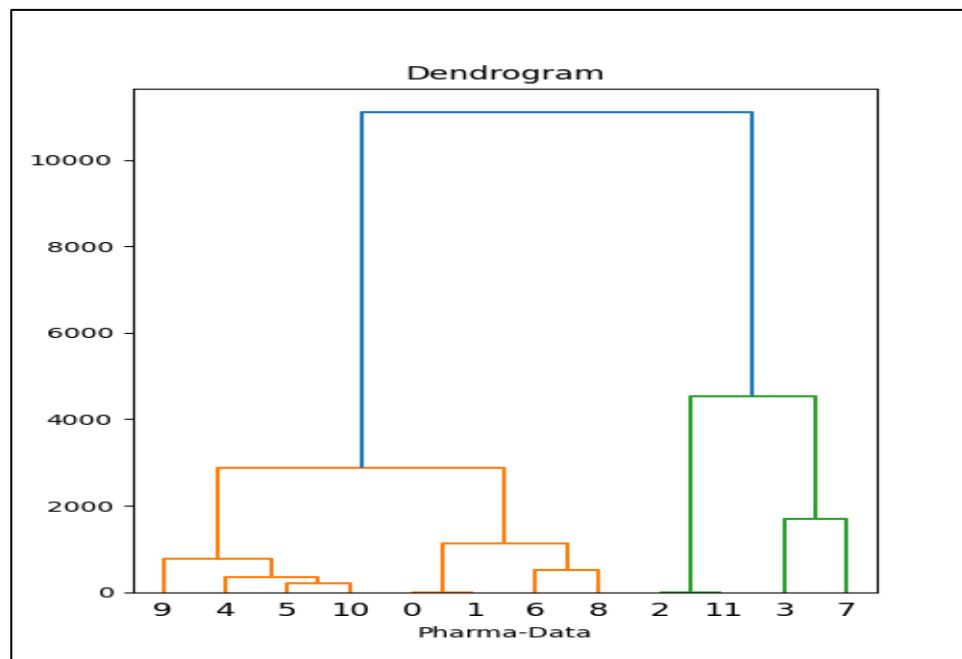
```
[20]: import scipy.cluster.hierarchy as sch
plt.figure(figsize=(5, 7))
dendrogram = sch.dendrogram(sch.linkage(datasubset, method = 'ward'))
plt.title('Dendrogram') # title of the dendrogram

plt.xlabel('Pharma-Data') # Label of the x-axis
plt.ylabel('Euclidean distances') # Label of the y-axis
plt.show()
```



```
[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