**Lab 4: Hierarchical clustering**

**Objectives:** To Implement Hierarchical clustering and Visualizing the clusters.

**Description:**

* At first a sample named **Daraz\_Customers.csv** file is taken as input . (Fig.01)
* On the right the data are visualized.

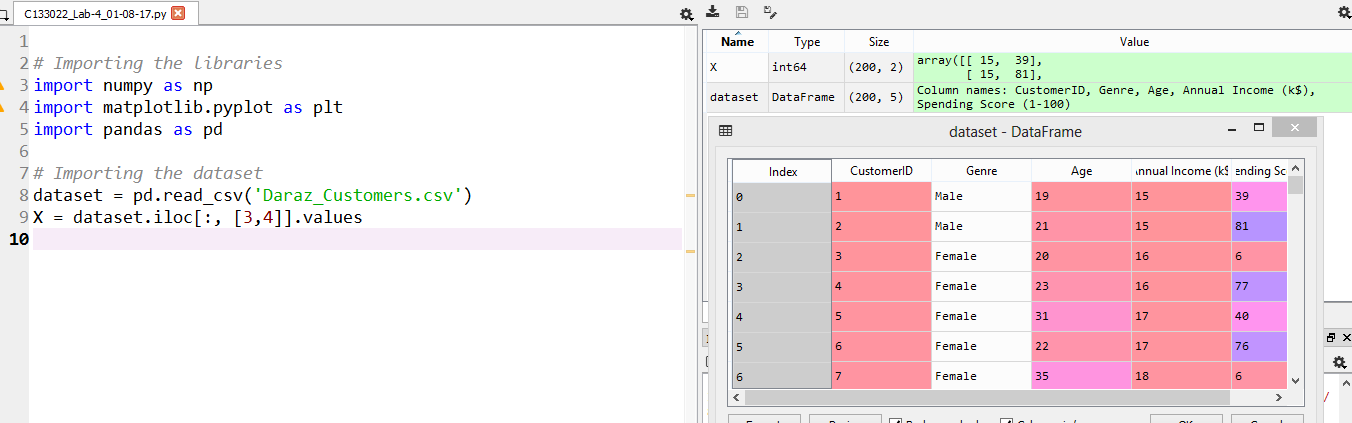


Fig.01

* Then **scipy.cluster.hierarchy** is imported as sch (Fig.02 - line 11)
* **scipy.cluster.hierarchy :** To cut hierarchical clusterings into flat clusterings or to find the roots of the forest formed by a cut by providing the flat cluster ids of each observation.
* In line number 12: (Fig.02)

|  |  |
| --- | --- |
| [**dendrogram**](https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html#scipy.cluster.hierarchy.dendrogram)(Z[, p, truncate\_mode, …]) | Plots the hierarchical clustering as a dendrogram. |
| [**linkage**](https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html#scipy.cluster.hierarchy.linkage)(y[, method, metric]) | Performs hierarchical/agglomerative clustering. |
| [**ward**](https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.ward.html#scipy.cluster.hierarchy.ward)(y) | Performs Ward’s linkage on a condensed distance matrix. |

* After running the code, the dendogram is visualized on the right.

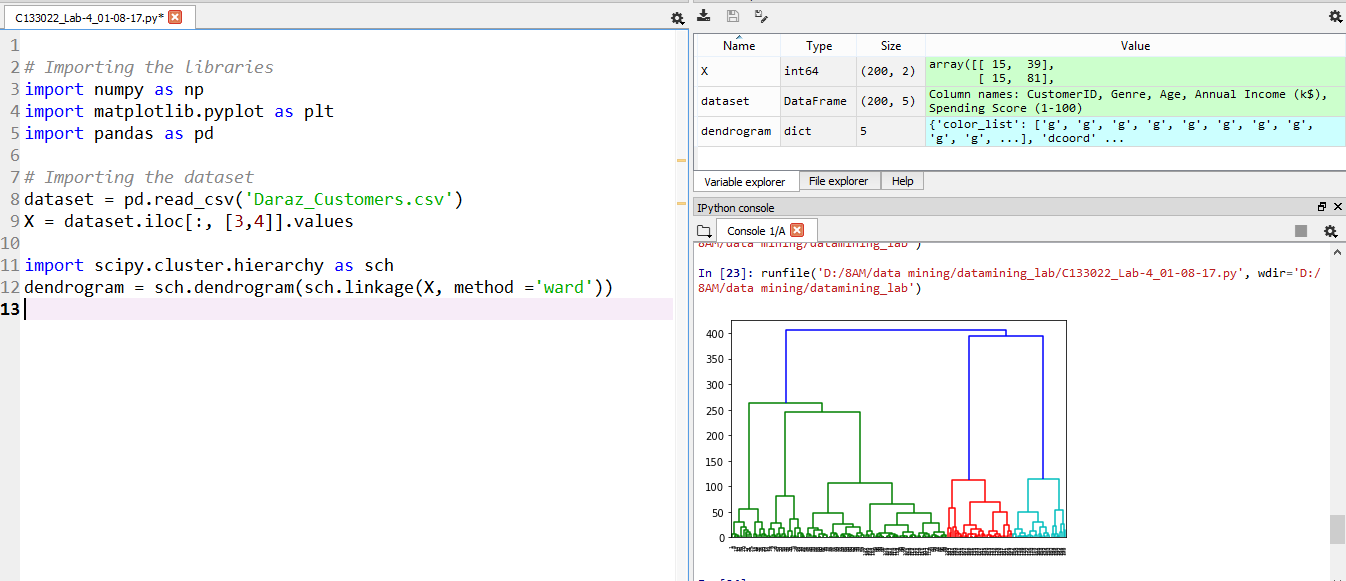


Fig.02

* In line 13-16: the dendogram is titled as ‘Dendogram’ and it’s x and y-axis are labeled as ‘Customers’ and ‘Eucledean Distance’ respectively. (Fif. 03)

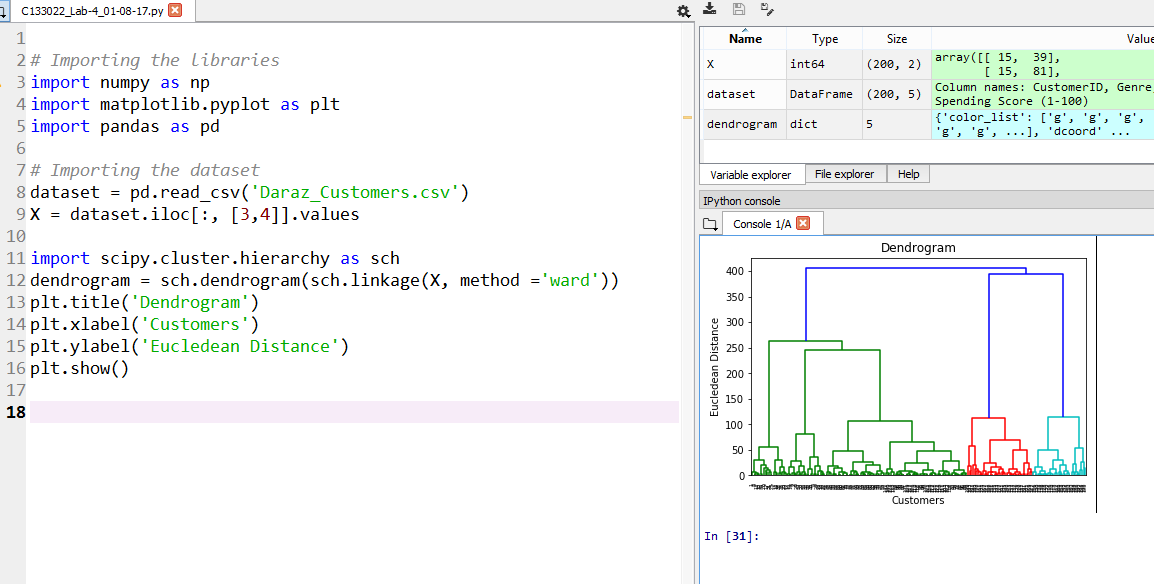


Fig.03

* sklearn.cluster.AgglomerativeClustering : Recursively merges the pair of clusters that minimally increases a given linkage distance.
* In line 19: (Fig.04) :

**Parameters:**

* + **n\_clusters** : The number of clusters to find.
  + **affinity** : Metric used to compute the linkage. Can be “euclidean”, “l1”, “l2”, “manhattan”, “cosine”, or ‘precomputed’. If linkage is “ward”, only “euclidean” is accepted.
  + **linkage** : {“ward”, “complete”, “average”}, optional, default: “ward”
  + Which linkage criterion to use. The linkage criterion determines which distance to use between sets of observation. The algorithm will merge the pairs of cluster that minimize this criterion.
    - ward minimizes the variance of the clusters being merged.
* **In line 20: fit\_predict**(X, y=None) : Performs clustering on X and returns cluster labels

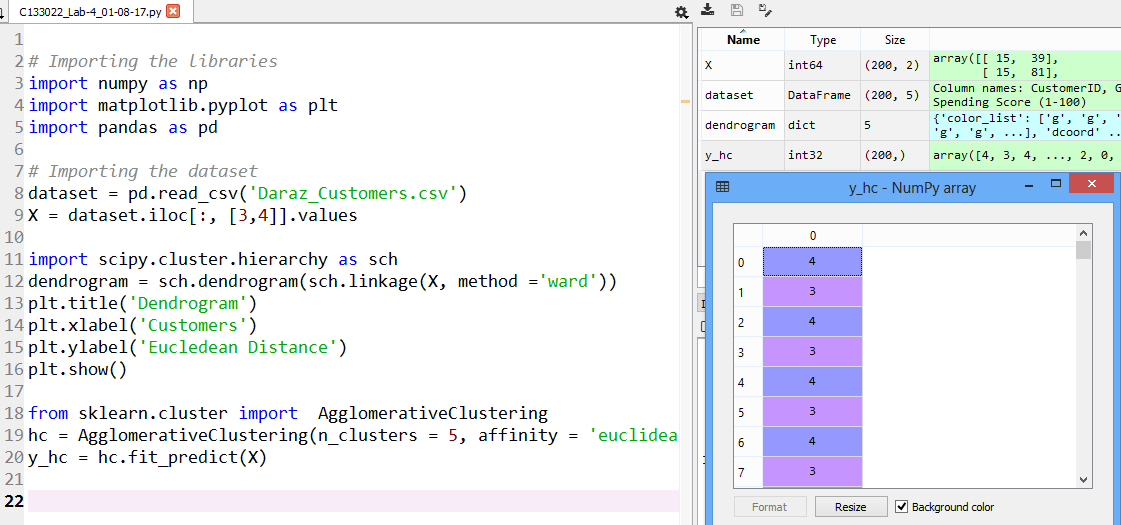


Fig.04

* Rest of the lines are to visualize the clusters.
* matplotlib.pyplot.scatter is used to make a scatter plot of x vs y. (Fig.05)

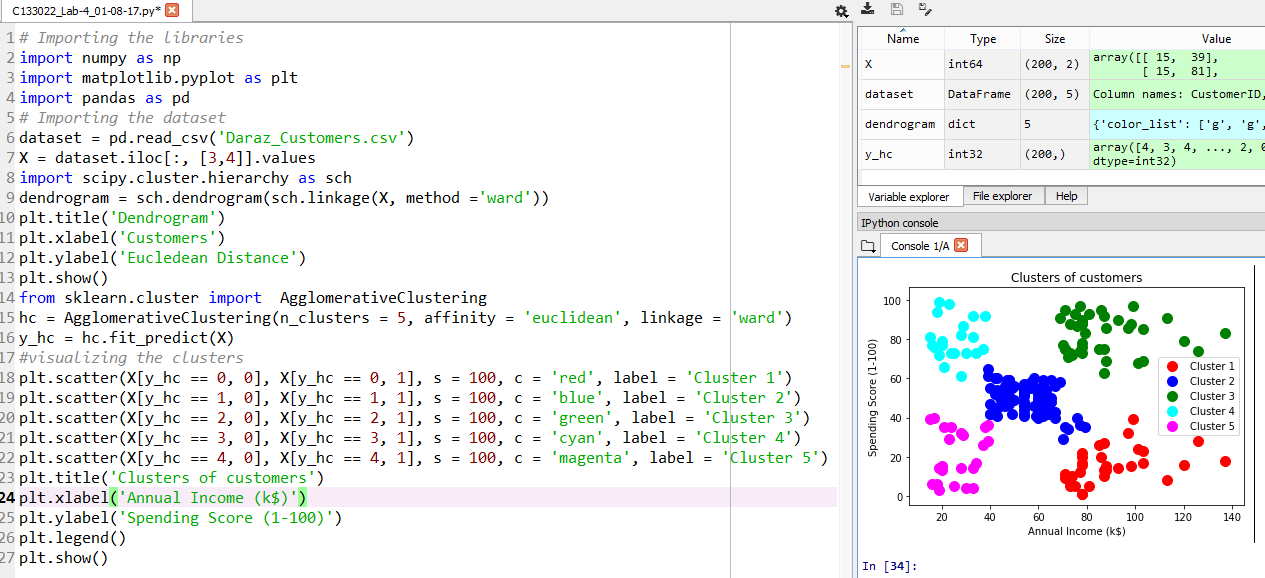


Fig.05

**Results/Conclusions:**

* The sample data were successfully clustered and visualized.
* The data are clustered into 5 clusters.