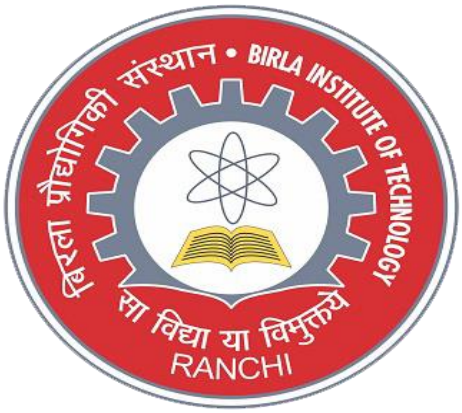
BIRLA INSTITUTE OF TECHNOLOGY

MESRA,RANCHI



ASSIGNMENT ON

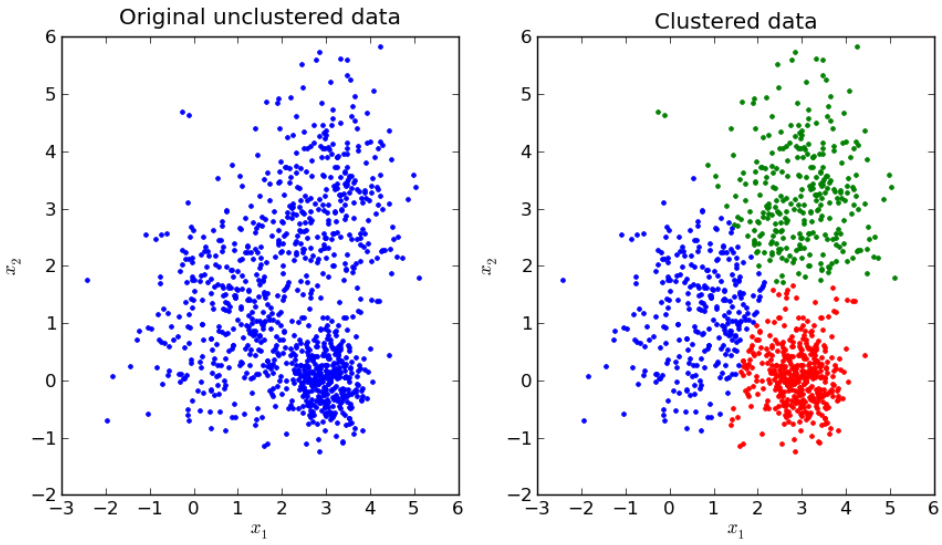
HIERARCHICAL CLUSTERING

PRESENTED BY:
RITESH KUMAR SINGH
(MCA/10029/18)
GROUB ASSIGNMENT WITH NIHARIKA RAJ
SEMESTER: 4TH
SUB: DATA ANALYTICS WITH PYTHON

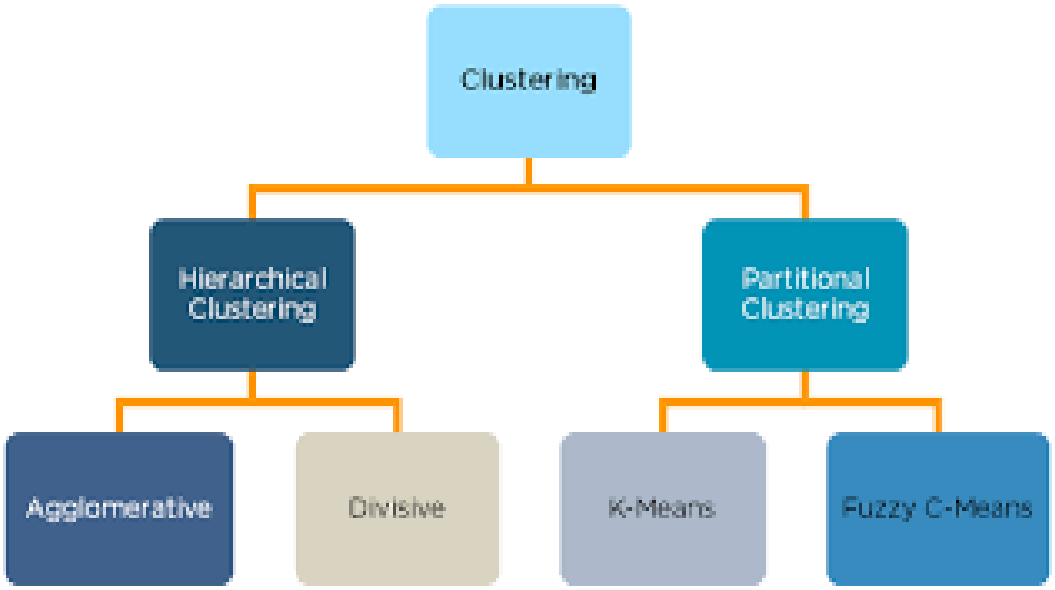
SUBMITTED TO:
PROF. RASHMI RATHI

CLUSTERING

The technique to group similar data points such that points within the same group are more similar than with points in other groups, is called clustering. Cluster is the group of similar data points.



CLASSIFICATION OF CLUSTERING



HIERARCHICAL CLUSTERING

Hierarchical clustering groups similar objects into clusters(goups).Hierarchical clustering is also known as Hierarchical cluster analysis.

Hierarchical decomposition of set of data objects is done by hierarchical clustering.

It groups data objects into tree structures.

One drawback of hierarchical method is that once a step is done, it can never be undone.

This leads to smaller computational cost which is useful.

Hierarchical clustering can be visualized using a dendogram.

Hierarchical clustering is classified into two types based on formation of hierarchical decomposition:

1. Agglomerative
2. Divisive

AGGLOMERATIVE HIERARCHICAL CLUSTERING TECHNIQUE

It is also called bottom up approach and starts with each object forms its own separate group.

In this each data point is considered as an individual cluster at the initial stage.

The similar clusters merge with other clusters in each iteration until k clusters are formed.

This type is used in most of the hierarchical clusterings.

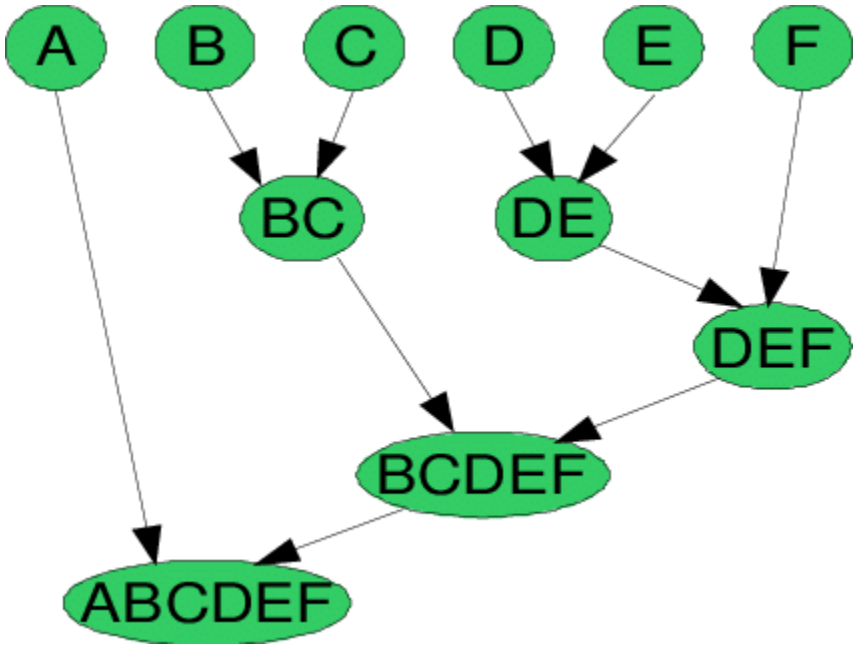
Algorithm for agglomerative is straight forward:

- Formulate the proximity matrix.
- Form cluster at each data point.
- Repeat: two closest clusters are merged and the matrix is updated.

- Until only a single cluster is left.

We will see a pictorial representation to see the agglomerative hierarchical clustering:

Let us assume that we have six data points(A, B, C, D, E, F).



Step 1: In the first step, proximity of individual step is calculated and all data points are considered as individual clusters.

Step 2: This step involve merging of similar clusters and their formation as a single cluster is done. For example B,C and D,E are similar and are merged in step 2.

Step 3: Proximity of new clusters is calculated again and the similar clusters are merged again to form new clusters i.e A, BC, DEF.

Step 4: Again the proximity of new cluster is calculated. Clusters DEF and BC are merged to form a new cluster as they are similar.

Step 5: In the final step, all clusters are merged and form a single cluster.

ALORITHM OF AGGLOMERATIVE CLUSTERING

Dist{c1,c2} is the distance function.

i=1;

do

ci={xi}

while i<=n

Given: Object set X {x1,x2,x3,…,xn}

end do

C={c1,…,cn}

l=n+1

while C.size > 1 do

- (cmin1, cmin2)=minimum dist(ci,cj) for all ci,cj in C
- remove cmin1, cmin2 from C
- add {cmin1, cmin2} to C
- l=l+1

end while

DIVISIVE CLUSTERING

- It is exactly opposite than agglomerative clustering and not much used in the real world.

- It is top-down approach and begins with all objects in the same cluster.
- The cluster is subdivided into smaller pieces. All the data points are considered as single cluster and the data point which is not similar is separated from the cluster in each iteration.
- The data point which separated is considered as individual cluster and at the last there are n number of clusters left.
- It is called as divisive since a single cluster is divided into n-clusters.

AGGLOMERATIVE VERSUS DIVISIVE HIERARCHICAL CLUSTERING

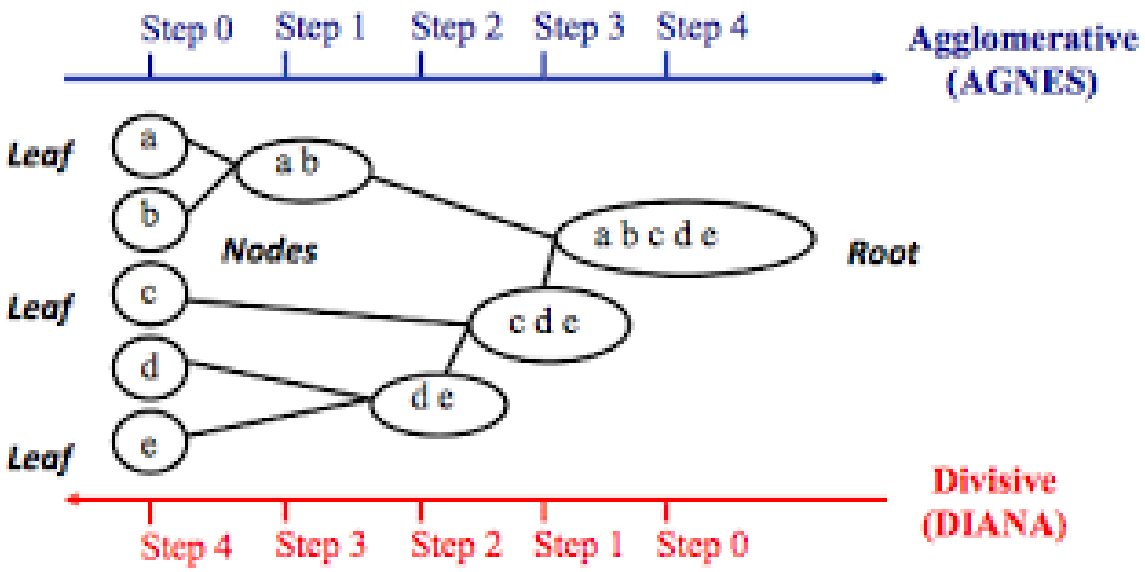


Fig: Agglomerative versus Divisive Hierarchical Clustering on data objects {a,b,c,d,e}

In the above figure, in the first step, AGNES puts each object in a cluster of its own.

Then merging of clusters takes place based on some criteria.

For example, c1 and c2 are merged if distance between them forms minimum Euclidean distance.

This is a single linkage approach in which each cluster is represented by all the objects in the clusters and similarity is measured among two clusters.

This process continues to merge all the data into one cluster.

In DIANA, all objects form one cluster initially.

Based on maximum Euclidean distance between neighboring objects, the cluster is split.

The splitting of clusters repeats till all clusters have single object.

The user can specify the termination condition i.e no of clusters in both agglomerative and divisive clustering.

DENDOGRAM

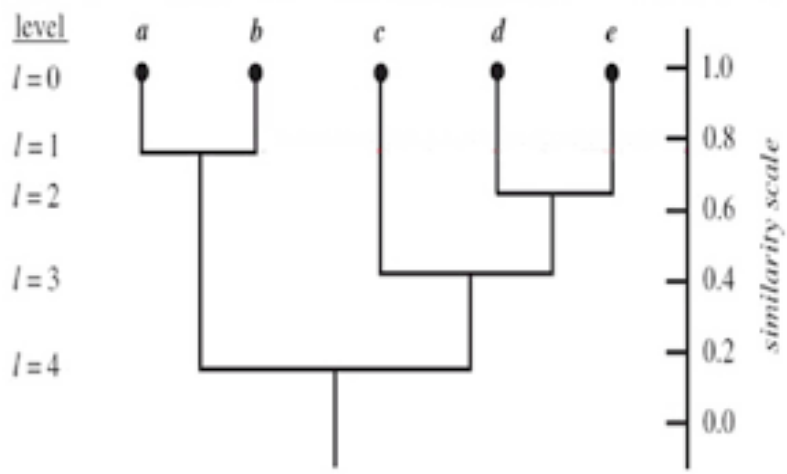


Fig2: Dendogram representing Hierarchical clustering of data objects{a, b, c, d, e}

Dendogram is a tree structure used to represent hierarchical clustering process.

It shows step by step grouping of objects.

In figure2, at l=0 all five objects are clustered as singleton object.

At l=0, first cluster is formed by grouping a and b objects.

Vertical axis is used to show similarity between the clusters.

HOW TO MEASURE SIMILARITY BETWEEN TWO CLUSTERS:

To merge or divide the clusters, it is important to measure similarity between two clusters. To calculate similarity following measures are considered:

1. MAX
2. MIN
3. Group Average
4. Distance between centroids
5. Wards method

MEASURES FOR DISTANCE BETWEEN CLUSTERS

Following are the widely used measures to find the distance between clusters where distance between two objects p and p' is |p-p'|, mi is mean of clustering for Ci and no of objects in Ci is ni.

Minimum distance: $d_{min}(C_i,C_j) = \min_{p \in Ci,\, p' \in Cj} |p-p'|$

Maximum distance: $d_{max}(C_i,C_j) = \max_{p \in Ci,\, p' \in Cj} |p-p'|$

Mean distance: $d_{mean}(C_i,C_j) = |m_i - m_j|$

Average distance: $d_{avg}(C_i,C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p-p'|$

When an algorithm uses min distance measure to calculate the distance, it is called nearest neighbor clustering algorithm.

It is called singleton clustering if the process is terminated when the distance between nearest clusters exceeds the threshold value.

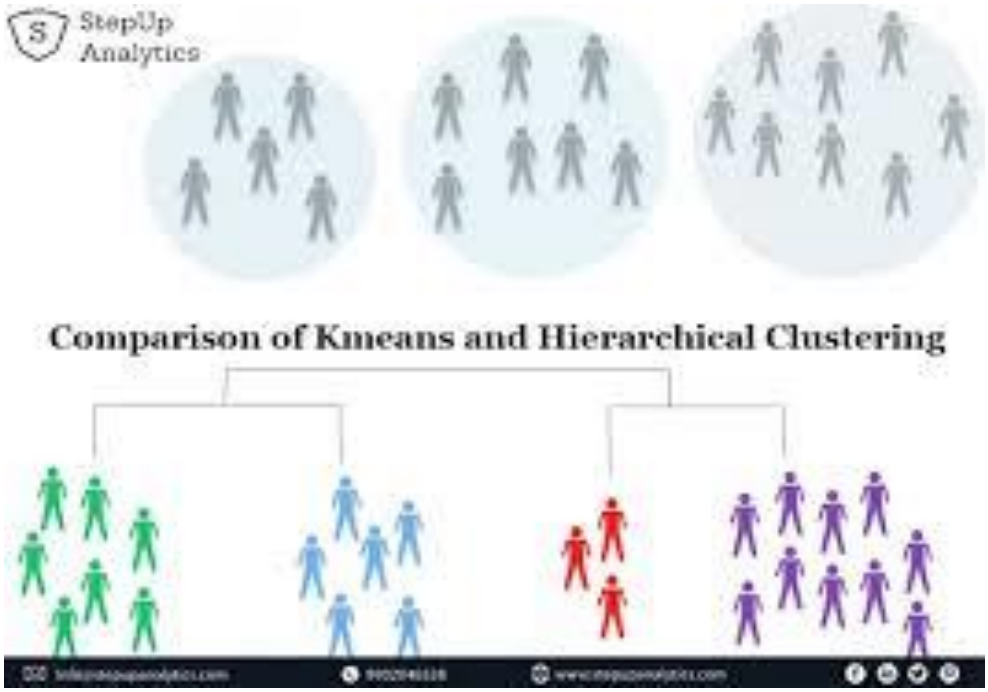
An agglomerative clustering algorithm is also called minimum spanning tree algorithm if it also uses minimum distance measure.

An algorithm is called farthest neighbor clustering algorithm if it uses maximum distance to measure the distance between clusters.

It is called complete linkage algorithm if the clustering is terminated if the distance between the clusters reaches the threshold value.

K-MEANS VERSUS HIERARCHICAL CLUSTERING

K-Means Clustering	Hierarchical Clustering
These are less computationally intensive and preferred with very large data sets.	It is useful when clusters are required to be arranged in natural hierarchy
It divides the set of data into non overlapping subsets so that each data object is in one subset.	It is a set of nested structure that are organized as a tree.
It assumes a particular value of k.	It does not assume a particular value of k.
Running the algorithm multiple times result in different data since random choice of clusters.	In hierarchical clustering results are reproducible.
It works well with hyper spherical shape of clusters.	It doesn't work well with hyper spherical shape of clusters.



ADVANTAGES OF HIERARCHICAL CLUSTERING:

- Any form of similarity and distance can be easily handled.
- It is applicable to any attribute types.

LIMITATIONS OF HIERARCHICAL CLUSTERING:

- For very large data sets, comparison and storage of n x n distance matrix is expensive and slow.
- The records placed incorrectly in the early process cannot be reallocated subsequently since only one pass through data is done.
- It has low stability.

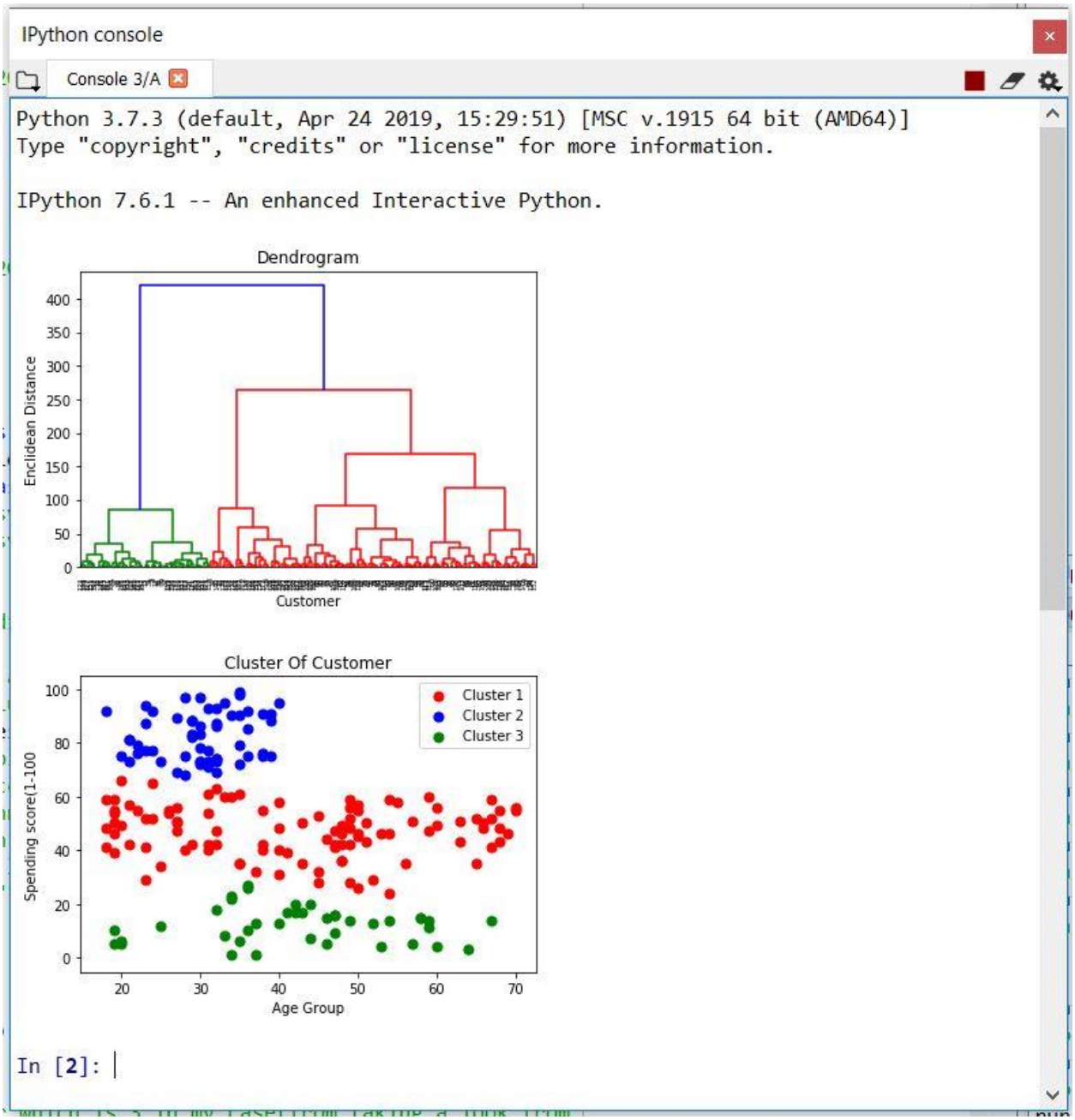
CODE TO IMPLEMENT HIERARCHICAL CLUSTERING:

```
# -*- coding: utf-8 -*-
"""
Created on Thu Jun 11 22:58:59 2020

@author: singh
"""

import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
from sklearn.cluster import AgglomerativeClustering
import scipy.cluster.hierarchy as sch
"""Loading data set from data.csv file into dataset variable"""
dataset = pd.read_csv('./data.csv')
"""print(dataset)"""

""" Finding the cluster for finding the age group which spends more money"""

"""In table index 2 contain age and index 4 contain spending
now from dataset it take all values into X variable"""
```

```
X = dataset.iloc[:, [2,4]].values
"""dendrogram function help to plot dendrogram
linkage function helps to connecct ward
and method define for how to connect the data so in this case its ward
ward is nothing but the lines in dendrogram"""
dendrogram = sch.dendrogram(sch.linkage(X, method='ward'))
"""defining title of dendrogram"""
plt.title('Dendrogram')
"""giving name to x axis"""
plt.xlabel('Customer')
"""giving name to y axis"""
plt.ylabel('Enclidean Distance')
""" drawing dendrogram"""
plt.show()
""" till now we find the cluster which is 3 in my case(from taking a look from
dendrogram, now we make clusters"""

""" model variable contain cluster and for that we have use
AgglomerativeClustering method which is heirarchical method of clustering """

"""this method take no of clustor affinity say the technique through which we
are finding cluster and next is linkage which difine which linkage method
we are using """
```

```
model = AgglomerativeClustering(n_clusters=3, affinity='euclidean',
linkage='ward')

Y_model = model.fit_predict(X);

"""scatter is use to visualize the clustering"""

plt.scatter(X[Y_model==0, 0], X[Y_model==0, 1], s=50, marker='o', color='red',
label='Cluster 1')

plt.scatter(X[Y_model==1, 0], X[Y_model==1, 1], s=50, marker='o', color='blue',
label='Cluster 2')

plt.scatter(X[Y_model==2, 0], X[Y_model==2, 1], s=50, marker='o', color='green',
label='Cluster 3')

plt.title('Cluster Of Customer')

plt.xlabel('Age Group')

plt.ylabel('Spending score(1-100')

plt.legend()

plt.show()


""" form the graph we can find which age group spends the most"""
```

OUTPUT



-*- coding: utf-8 -*-

"""

Created on Thu Jun 11 21:14:28 2020

@author: singh

```
"""
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt
from sklearn.cluster import AgglomerativeClustering
import scipy.cluster.hierarchy as sch
"""Loading data set from data.csv file into dataset variable"""
dataset = pd.read_csv('./data.csv')
"""print(dataset)"""

""" Finding the cluster for targeting the customer to sell cars with the help
of data which contain anual income and spending"""

"""In table index 3 contain anual income and index 4 contain spending
now from dataset it take all values into X variable"""
X = dataset.iloc[:, [3,4]].values
"""dendrogram function help to plot dendrogram
linkage function helps to connecct ward
and method define for how to connect the data so in this case its ward
ward is nothing but the lines in dendrogram"""
dendrogram = sch.dendrogram(sch.linkage(X, method='ward'))
"""defining title of dendrogram"""
plt.title('Dendrogram')
```

```
"""giving name to x axis"""
plt.xlabel('Customer')
"""giving name to y axis"""
plt.ylabel('Enclidean Distance')
""" drawing dendrogram"""
plt.show()
""" till now we find the cluster which is 5 in my case(from taking a look from
dendrogram, now we make clusters"""

""" model variable contain cluster and for that we have use
AgglomerativeClustering method which is heirarchical method of clustering """

"""this method take no of clustor affinity say the technique through which we
are finding cluster and next is linkage which difine which linkage method
we are using """

model = AgglomerativeClustering(n_clusters=5, affinity='euclidean',
linkage='ward')
Y_model = model.fit_predict(X);
"""scatter is use to visualize the clustering"""
plt.scatter(X[Y_model==0, 0], X[Y_model==0, 1], s=50, marker='o', color='red',
label='Cluster 1')
plt.scatter(X[Y_model==1, 0], X[Y_model==1, 1], s=50, marker='o', color='blue',
label='Cluster 2')
```

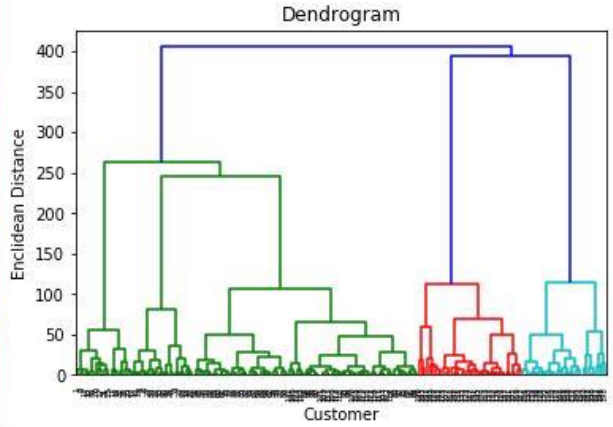
```
plt.scatter(X[Y_model==2, 0], X[Y_model==2, 1], s=50, marker='o', color='green', label='Cluster 3')

plt.scatter(X[Y_model==3, 0], X[Y_model==3, 1], s=50, marker='o', color='purple', label='Cluster 4')

plt.scatter(X[Y_model==4, 0], X[Y_model==4, 1], s=50, marker='o', color='orange', label='Cluster 5')

plt.title('Cluster Of Customer')

plt.xlabel('Anual Income')

plt.ylabel('Spendings')

plt.legend()

plt.show()


""" form the graph company can find to whom they have to target to

sell the cars"""
```

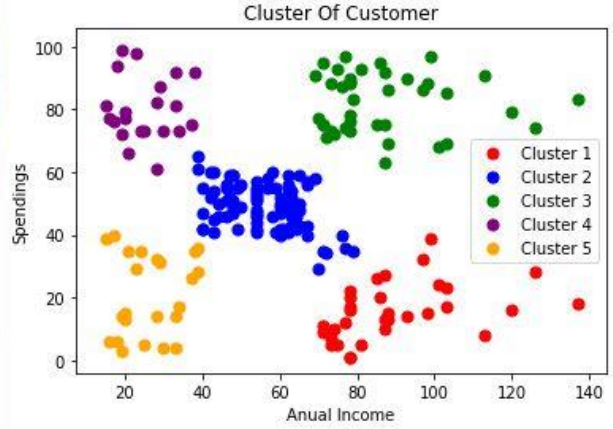
OUTPUT

IPython console

Console 3/A

```
In [2]: runfile('D:/mooc/MoocProject/danalytic.py', wdir='D:/mooc/MoocProject')
```






```
In [3]:
```