HW4-1

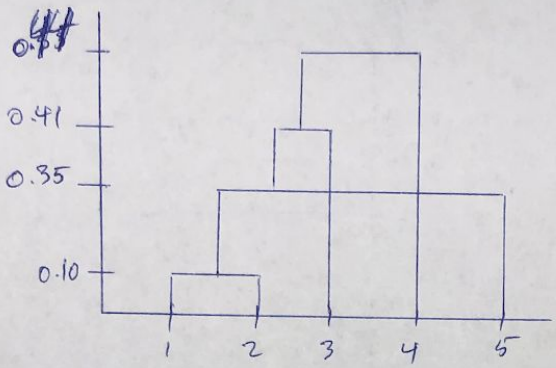
a) I studied the sklearn.cluster library and used some of its functions for this homework.
b) For optimal number of clusters, I first plotted the data set and looked at the visualization of data. Visual inspection suggests that there are 15 clusters in the data set. Then, I used elbow method by using the means algorithm to find the approximate/optimum number of clusters. The plot shows that the plot breaks and the elbow starts at 10, so from the elbow method and visual inspection we can estimate the optimal number of clusters is in the range of 10-15.
c) I applied KMeans algorithm to the dataset and plotted for each k= 5,10,15,20,25. I calculated the number of iterations until convergence and mentioned it on top of each plot. Number of iterations is how many iterations that algorithm does to converge. Convergence in K-means algorithm means the number of centroids do not change. This is the stopping criteria for K-means algorithm. About correlation, based on the SSE and K (number of clusters) by increasing the K the SSE decreases. If we increase K to the number of all points, SSE gets zero because at K=N, all points will become the ceiteroids and SSE becomes zero.
   I have attached my code to this file for your consideration.

HW4-2

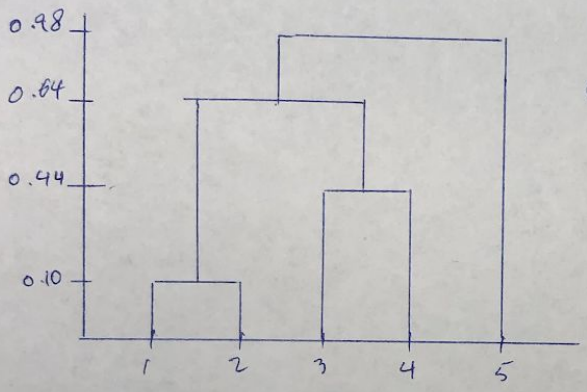
Based on the distance matrix:
A- based on single link scheme, we apply the scheme and we pick the points with the min distance between them. Then, we pick the min distance among all distances.

For complete link, we apply the scheme and we pick the max distance among all to fill out the tables. Then, we pick the min distance among all distances.

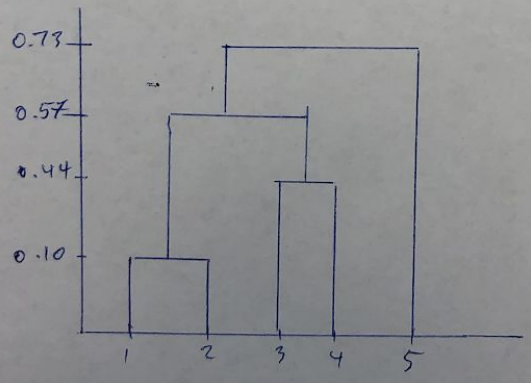
For group average, first we pick 2 points with the shortest distance and then we apply the scheme. The schee here is the average of any 2 clusters. Then we will pick the points with the shortest distance and cluster them.
The following is the Dendrograms of the 3 mentioned clustering schemes.



Single link ≡ MIN scheme



Complete link ≡ MAX scheme



Group Average

HW4-3

In this exercise I use DBscan clustering algorithm to find the parameter values that would result in the optimal number of clusters. The optimal number of clusters is provided in the site by as K for each data set. I used the optimal number of clusters (K) from that site as the reference for finding the optimal parameters. The parameters of DBscan algorithm are:

1- radius of neighborhood (Eps)

2- minimum number of samples (MinPts)

I changed these parameters by trial and error to find the optimal K for each data set. I found the following parameters for the data sets:

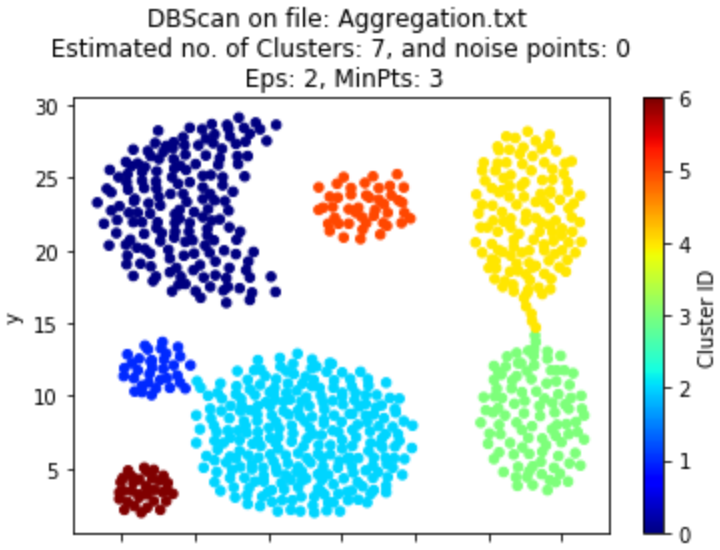
file_name_list=["Aggregation.txt","Compound.txt","pathbased.txt","spiral.txt","D31.txt","R15.txt","jain.txt","flame.txt"]

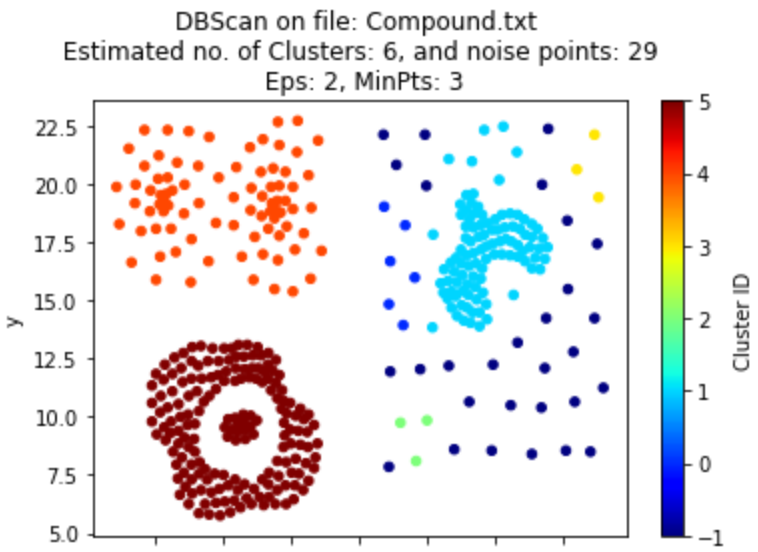
DBScan parameters

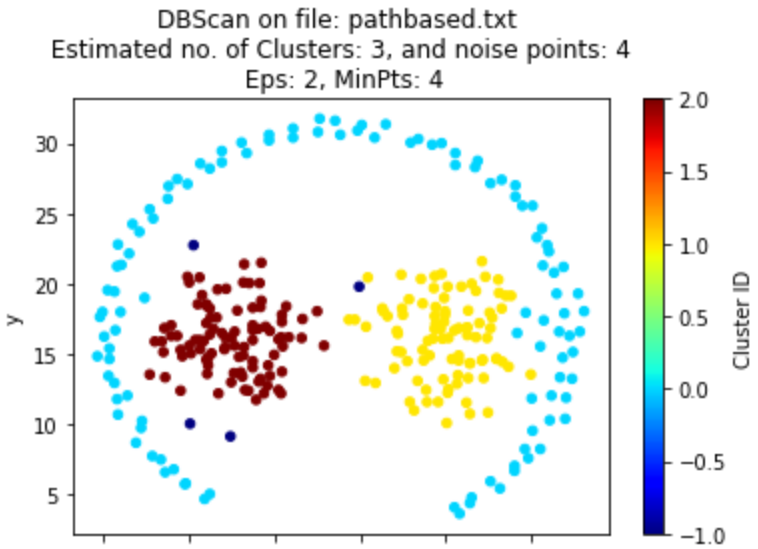
eps_list=[2,2,2,1.5,0.6,0.4,2.8,0.9]

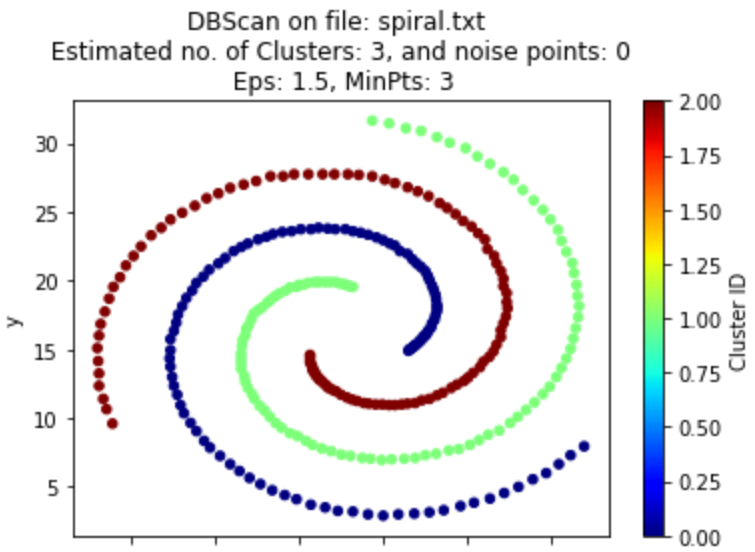
min_sample_list=[3,3,4,3,3,5,5,3]

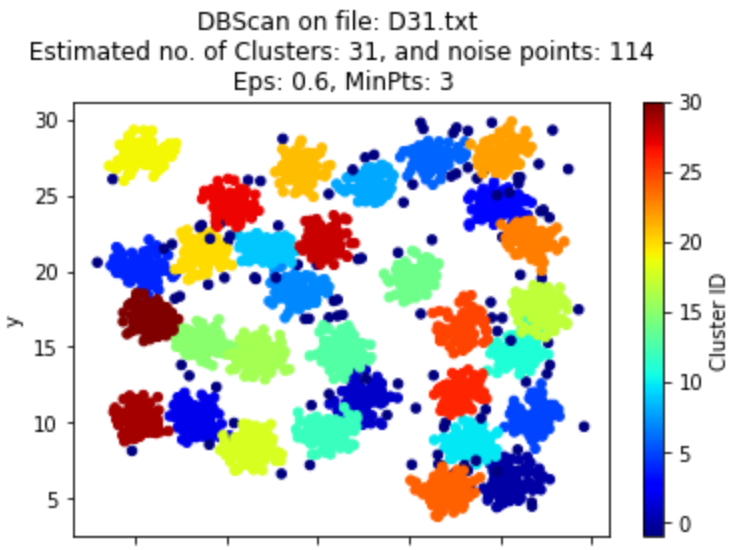
I have uploaded my code with this problem. The code clusters and plots the data sets based on the optimum radius and min number of points for each cluster.

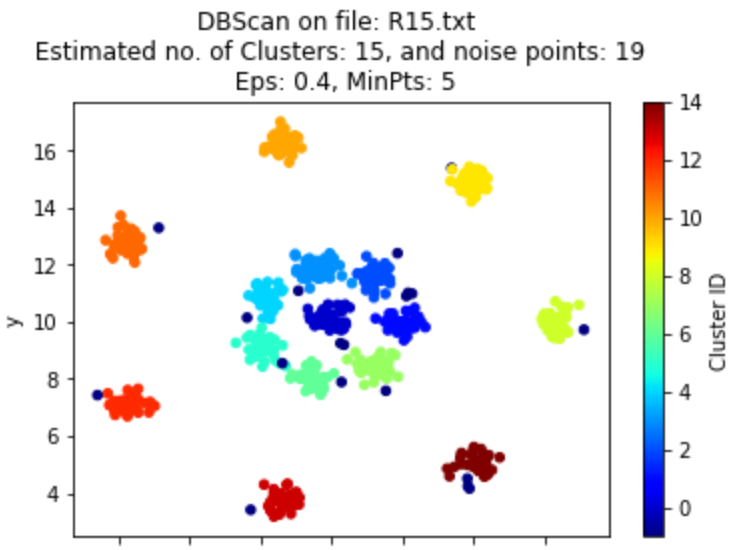


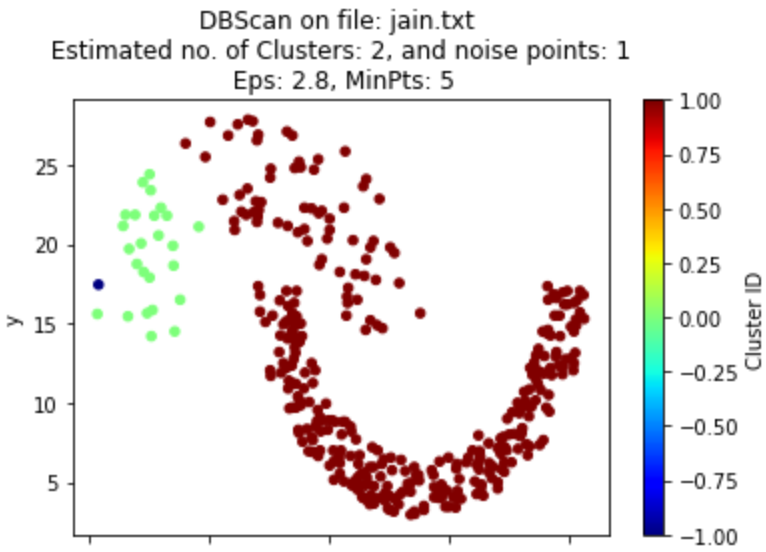

DBScan on file: Compound.txt
Estimated no. of Clusters: 6, and noise points: 29
Eps: 2, MinPts: 3
y
Cluster ID

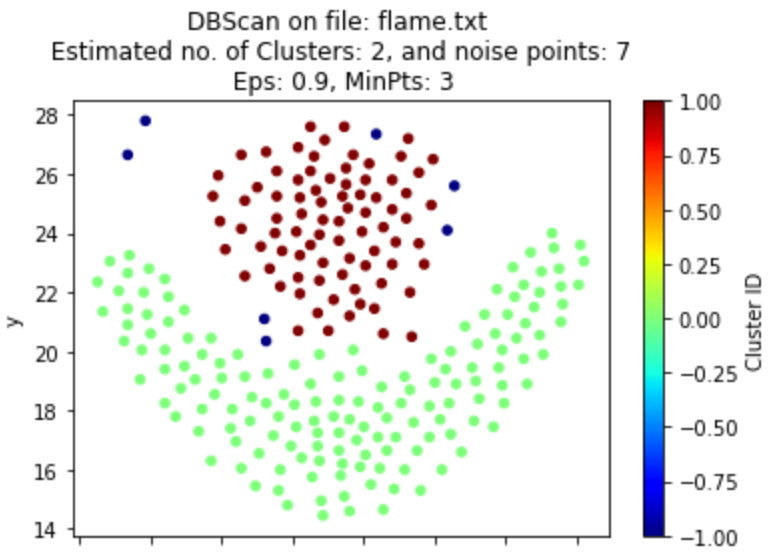

DBScan on file: pathbased.txt
Estimated no. of Clusters: 3, and noise points: 4
Eps: 2, MinPts: 4
y
Cluster ID


DBScan on file: spiral.txt
Estimated no. of Clusters: 3, and noise points: 0
Eps: 1.5, MinPts: 3
y
Cluster ID


DBScan on file: D31.txt
Estimated no. of Clusters: 31, and noise points: 114
Eps: 0.6, MinPts: 3
y
Cluster ID


DBScan on file: R15.txt
Estimated no. of Clusters: 15, and noise points: 19
Eps: 0.4, MinPts: 5
y
Cluster ID


DBScan on file: jain.txt
Estimated no. of Clusters: 2, and noise points: 1
Eps: 2.8, MinPts: 5
y
Cluster ID


DBScan on file: flame.txt
Estimated no. of Clusters: 2, and noise points: 7
Eps: 0.9, MinPts: 3
y
28
26
24
22
20
18
16
14
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Cluster ID