IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – VII
Clustering

Student's Name: Sarthak Saptami Kumar Jha **Mobile No: 8825319259**

Roll Number: B20317 **Branch: CSE**

1

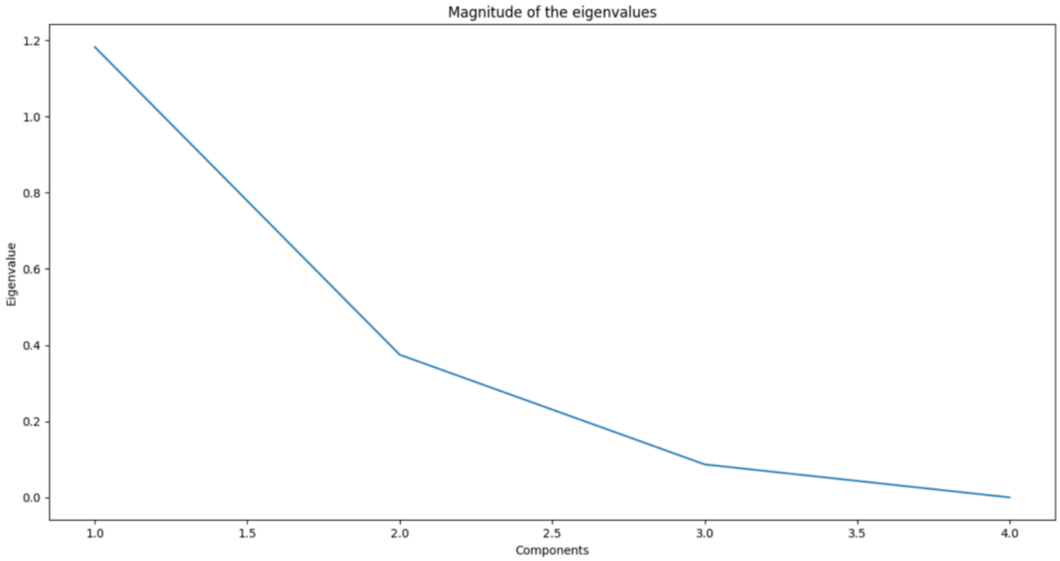

Figure 1 Eigenvalue vs. components

Inferences:

1. The magnitude of eigenvalues decrease with increase in number of components
2. With more components, we are including more less correlated components.

2 a.

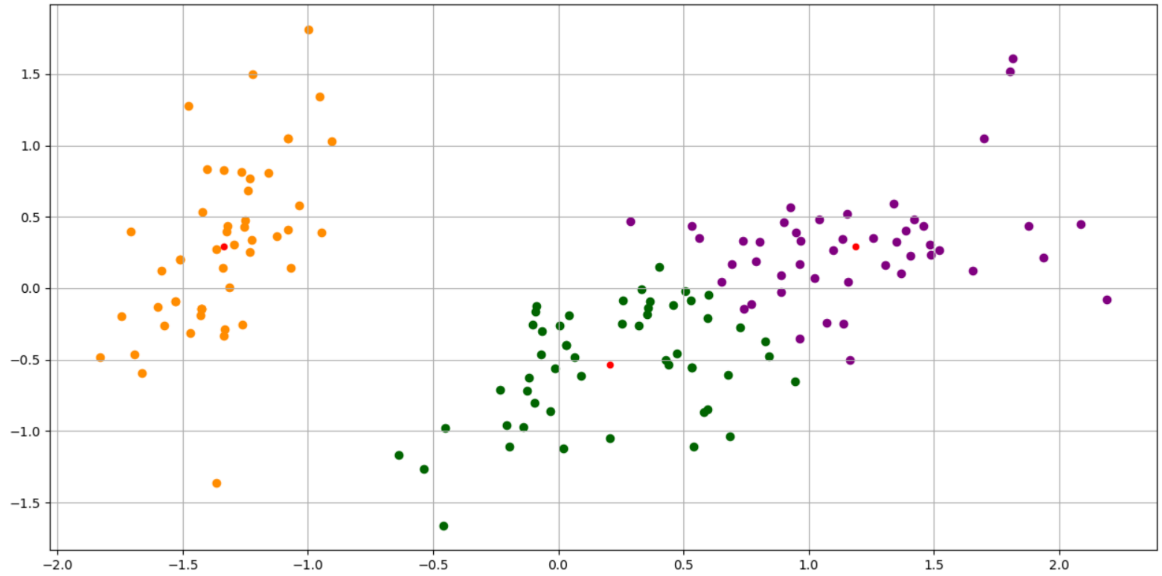

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering algorithm, assigns data points according to their distance from the mean of various clusters.
2. Yes the boundaries of the clusters appear to be quite circular, though some clusters can be said to be elliptical.

b. The value for distortion measure is 50.38

c. The purity score after examples are assigned to the clusters is 0.82

3

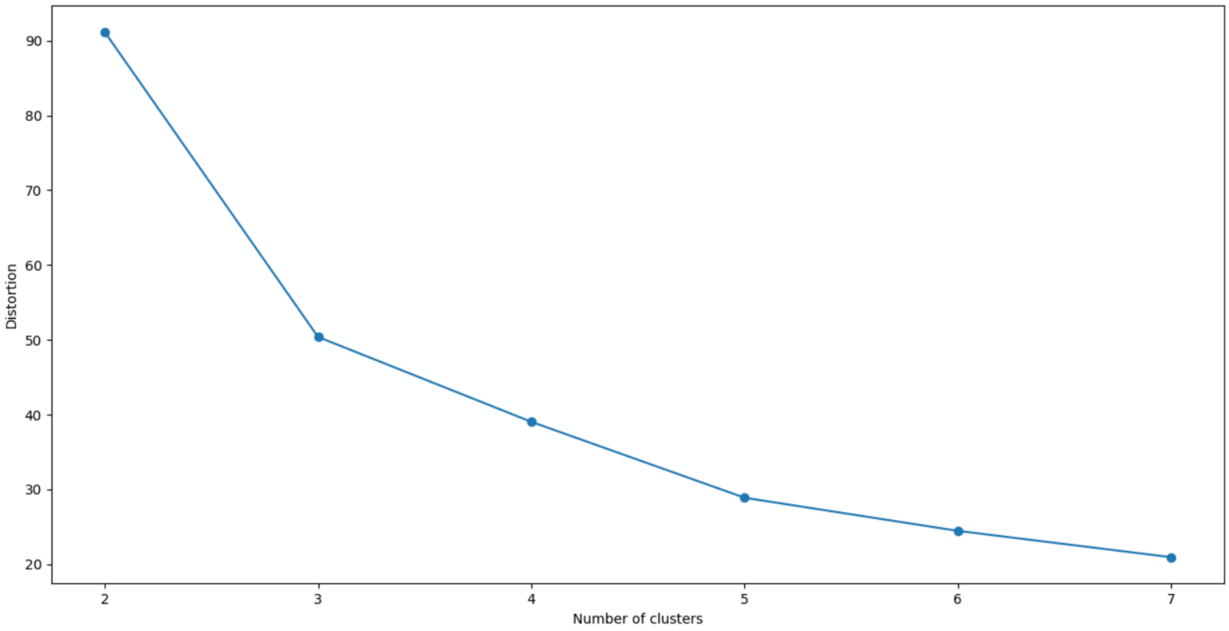

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases with increase in number of clusters
2. More number of clusters implies more cluster centers, thus the distance of each data point from its respective cluster center has a decreasing trend.
3. From the elbow method we can check that the optimum number of clusters will be 3.

Table 1 Purity score for K value = 2,3,4,5,6 & 7

K value	Purity score
2	0.667
3	0.820
4	0.667
5	0.540
6	0.467
7	0.433

Inferences:

1. The highest purity score is obtained with K = 3
2. The value of K increase to achieve a peak at K = 3 and then has a downward trend.

3. As we increase the value of K, we may be assigning data points belonging to a larger cluster into various small clusters thus decreasing the purity score.
4. If we ignore the values below the optimal number of clusters, rest of clusters have similar trend for distortion measure and purity score.

4 a.

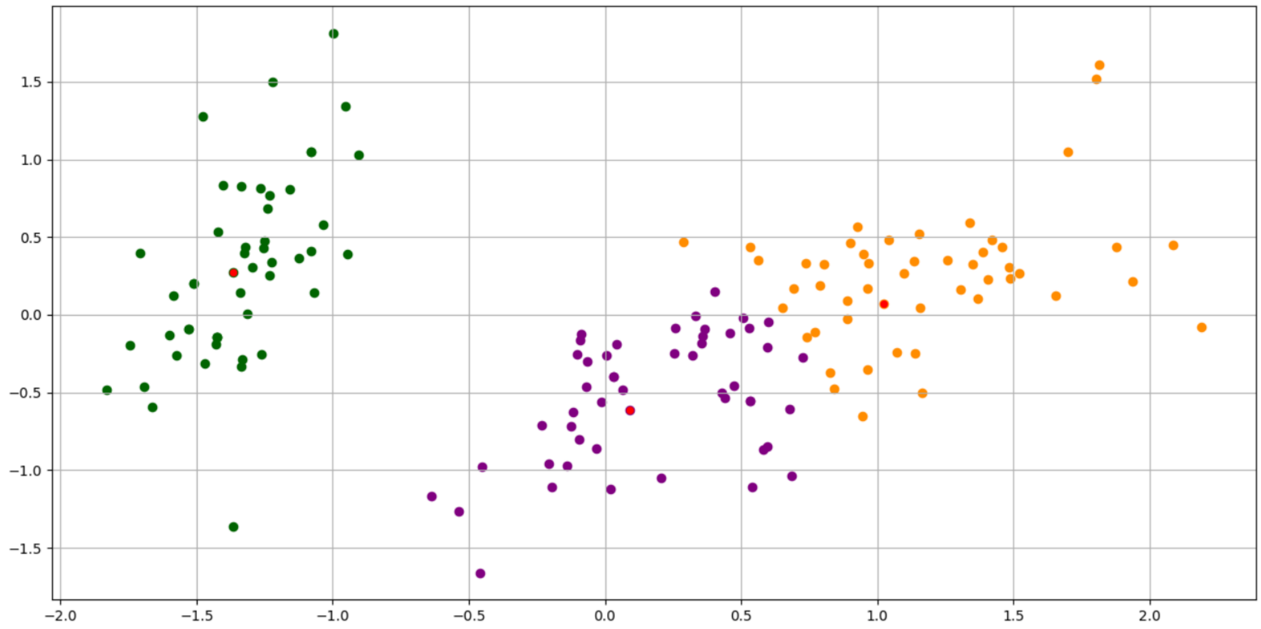

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. The data points are assigned according to their probabilities in each of the Gaussian Models.
2. Yes the cluster boundaries are elliptical in nature.
3. There

b. The value for distortion measure is -1.767.

c. The purity score after examples are assigned to the clusters is 0.84.

5

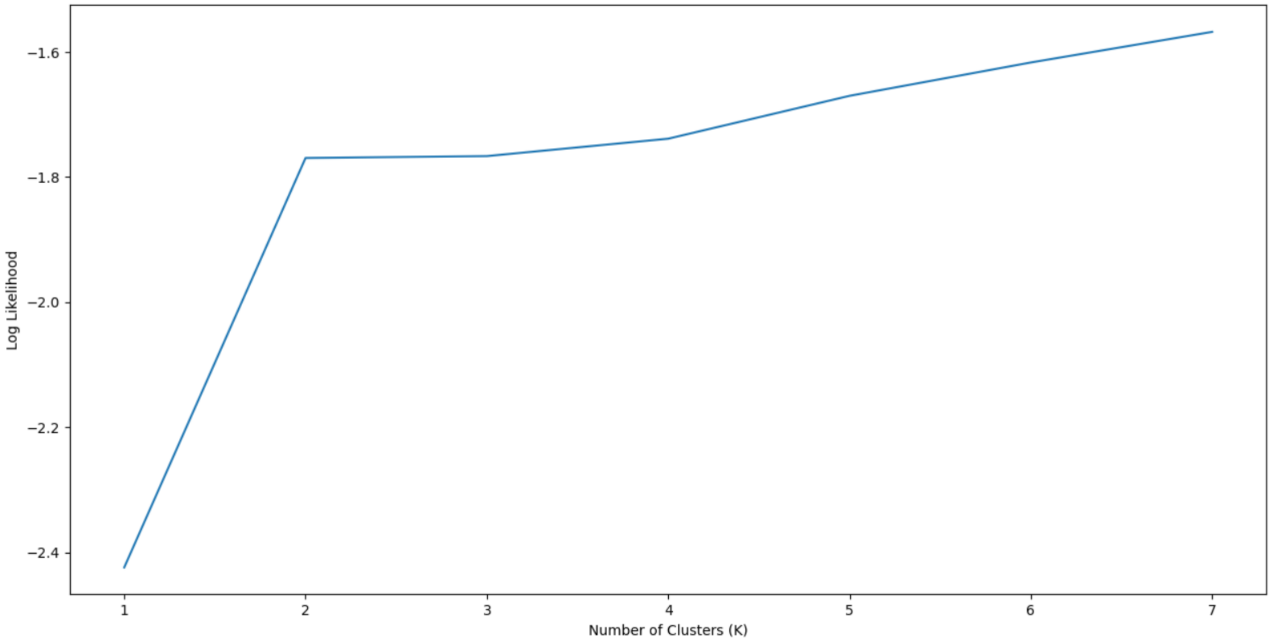

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion increases heavily at first, but then almost saturates.
2. As we increase, we can see from the data as well that there are 2 major clusters, the same is reflected in the log likelihood curve.
3. The number of clusters will optimally be 2 in case of GMM Clustering, as we have a clear elbow at K=2.

Table 2 Purity score for K value = 2,3,4,5,6 & 7

K value	Purity score
2	0.66
3	0.873
4	0.7
5	0.533
6	0.547
7	0.527

Inferences:

1. The highest purity score is obtained with K = 3

2. The value of purity score, increase to a peak and then decreases.
3. With more number of clusters, data points will be assigned to same cluster.
4. Yes, there is some similarities in the general trend of distortion measure and purity score
5. GMM Clustering algorithm is much more complex and can identify complex patterns, same can be observed in the purity score, as GMM has higher purity score compared to K-Means.

6

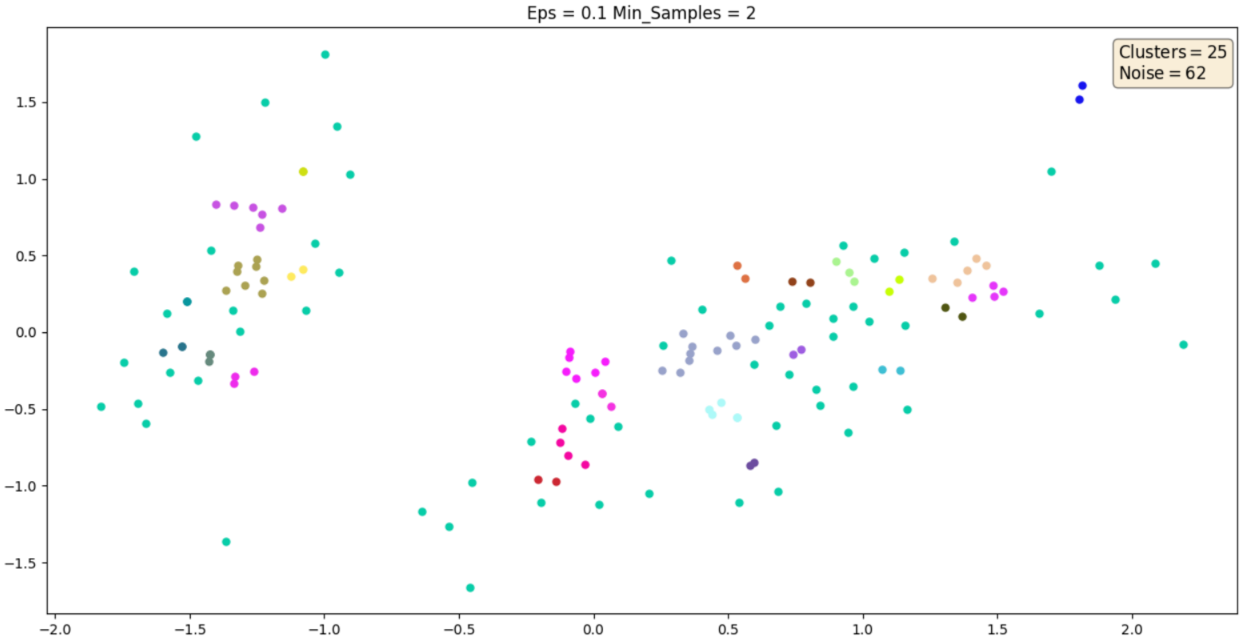

Figure 6 DBSCAN clustering on Iris flower dataset with eps = 0.1, min_samples = 2

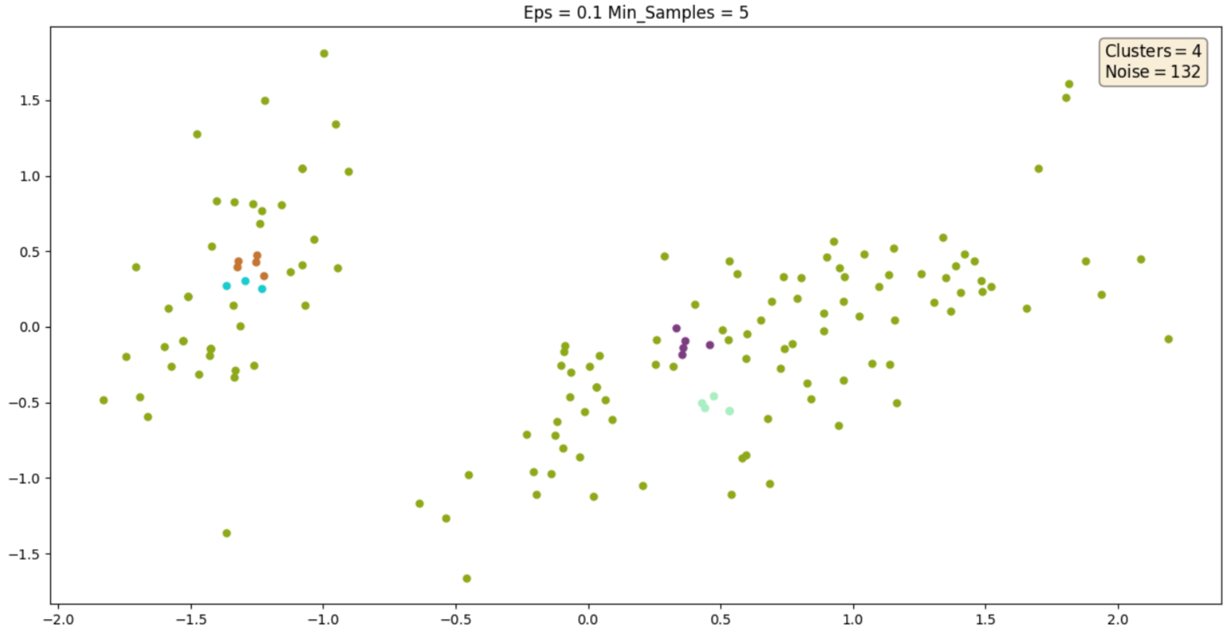

Figure 7 DBSCAN clustering on Iris flower dataset with eps = 0.1, min_samples = 5

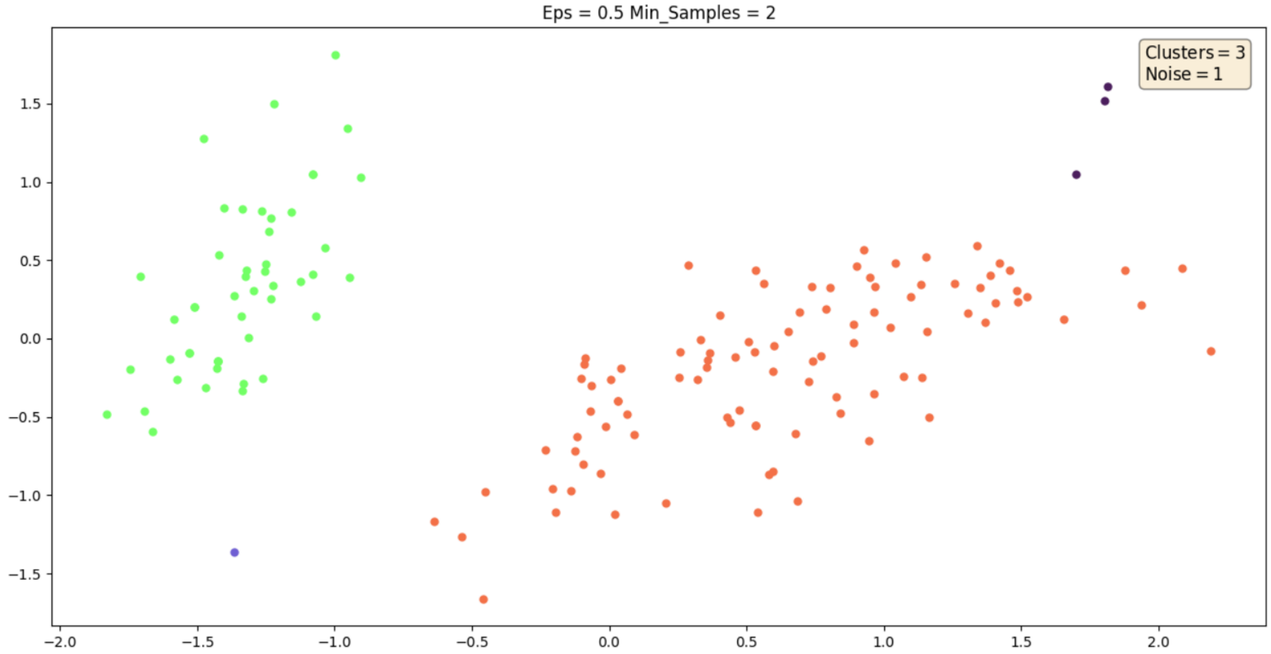

Figure 8 DBSCAN clustering on Iris flower dataset with eps = 0.5, min_samples = 2

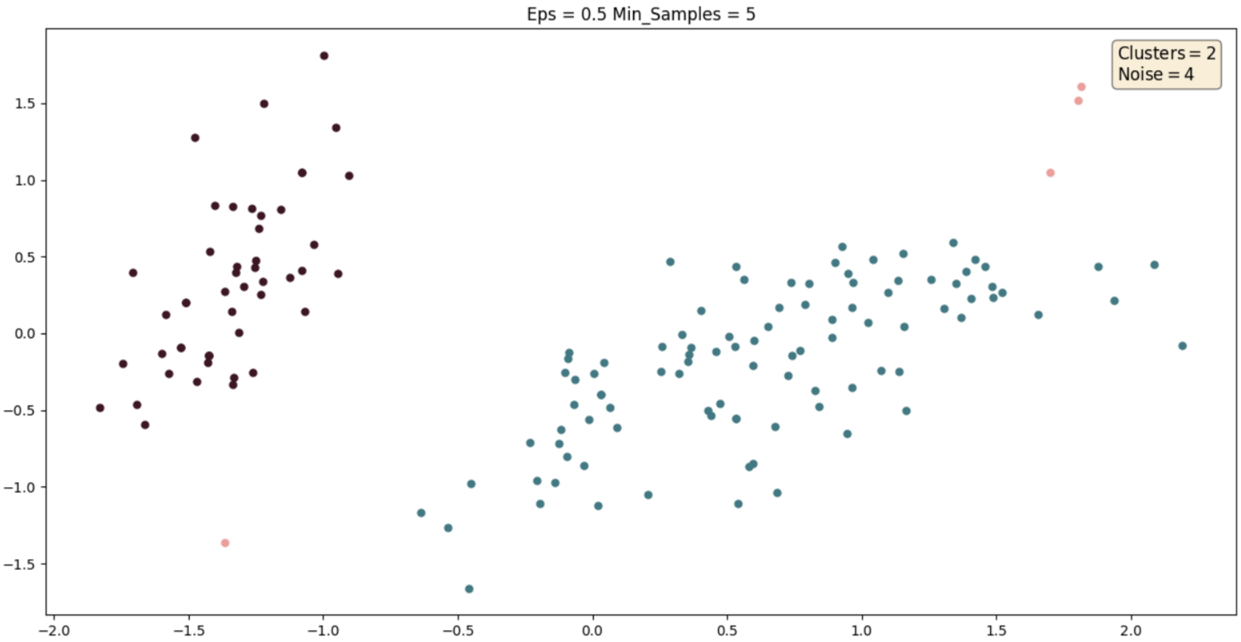

Figure 9 DBSCAN clustering on Iris flower dataset with eps = 0.5, min_samples = 5

Inferences:

1. DBSCAN clustering algorithm works on the basis of absence of data points between two clusters. Thus, it works best for data where the inherent clusters are separated by noise.
2. K-means forms circular clusters, while GMM forms Elliptical clusters in data, DBSCAN on the other hand forms clusters based on noise in between clusters.

b.

Eps	Min_samples	Purity Score
0.1	2	0.247
	5	0.36
	10	0.333
0.4	2	0.68
	5	0.68
	10	0.673

Inferences:

1. There is a general upward trend in purity score on increasing number of Min_Samples.
2. The purity score increase on increasing eps value for same number of Min_Samples.