2. Clustering

2.1. First dataset

Scatter Plot

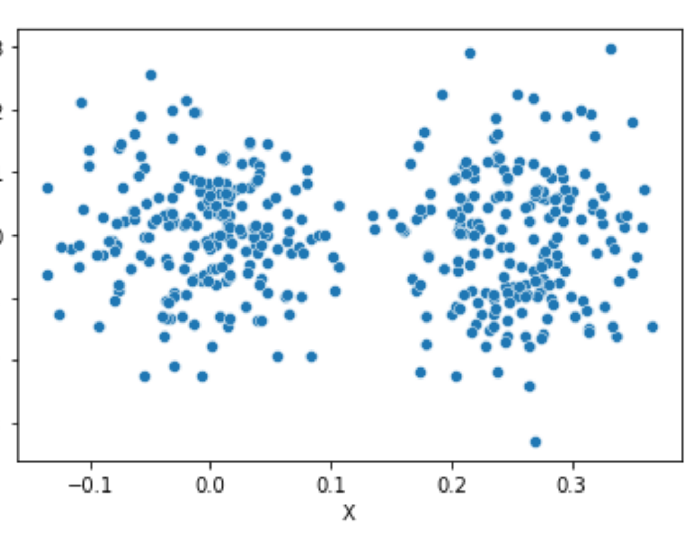


Figure 1: Scatter plot for the First dataset

Type of clusters: Center-based Clusters

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster.
- The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster.

Predicted number of clusters: $K = 2$

Lloyd's k-means Clustering

Clusters are separated by colors and the **centroids** are colored **red**.

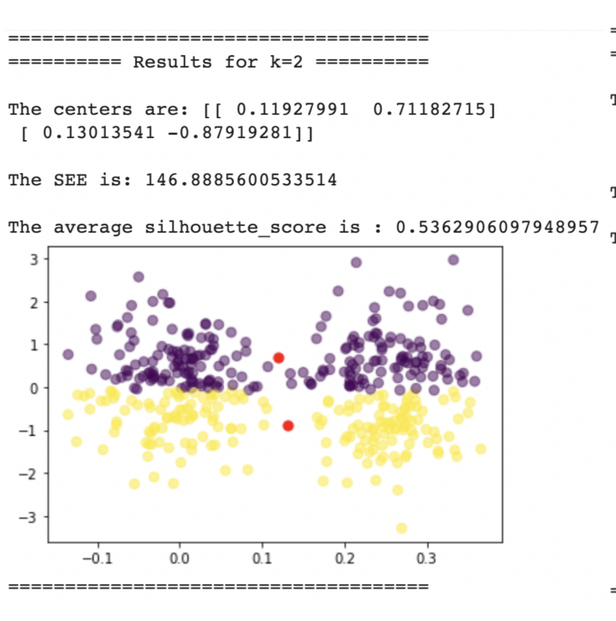


Figure 2: Number of clusters = 2

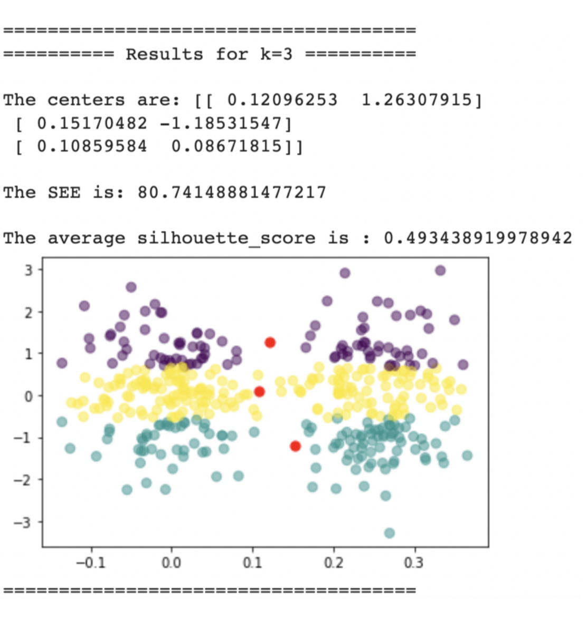


Figure 3: Number of clusters = 3

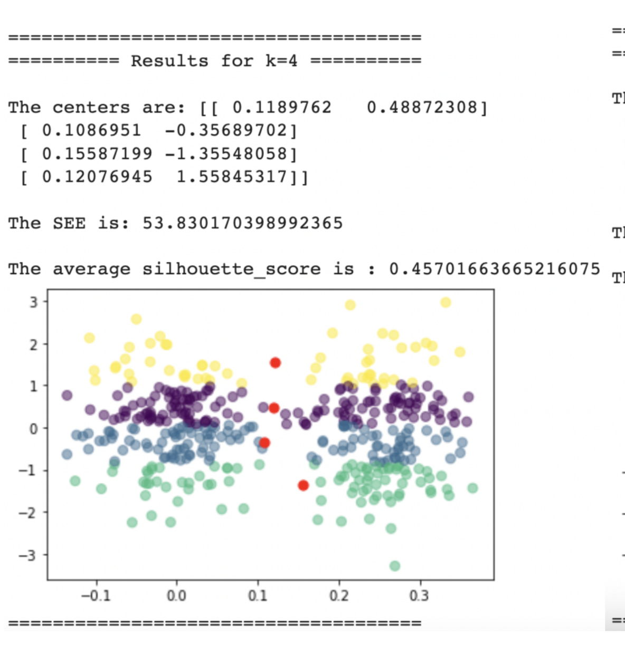


Figure 4: Number of clusters = 4

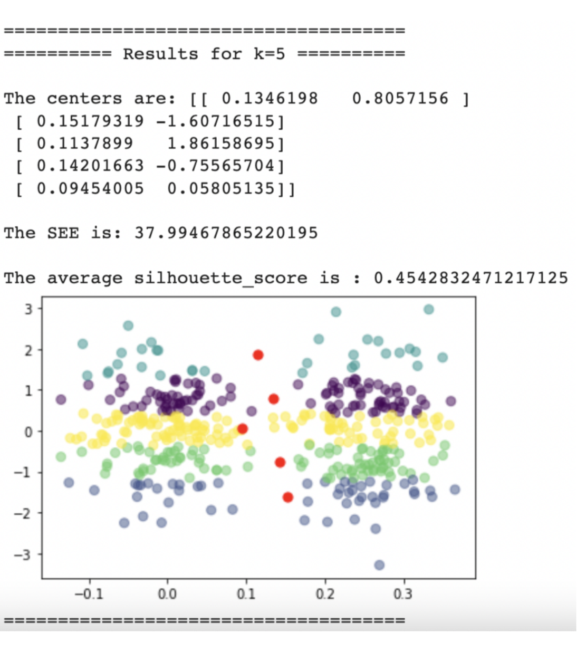


Figure 5: Number of clusters = 5

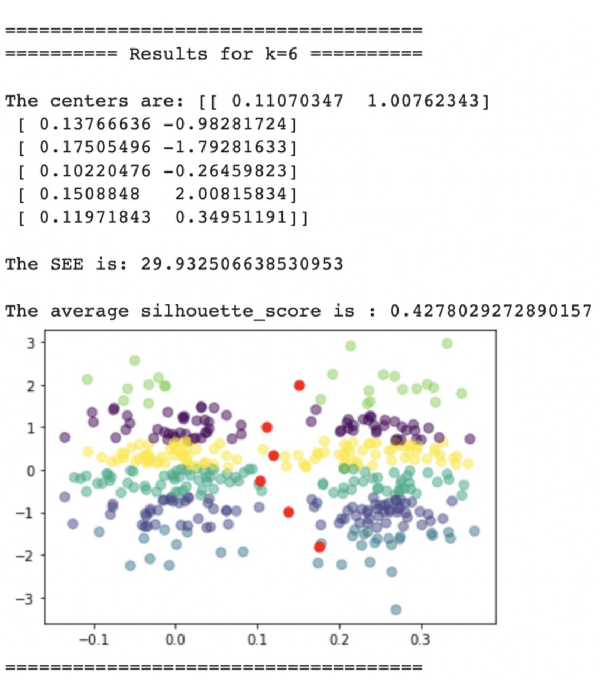


Figure 6: Number of clusters = 6

Minimum SSE: 29
Number of clusters with minimum SSE: 6
** Be careful! a small number of SSE does not guarantee a good clustering! Elbow method is used to solve this in the next part.**

Maximum silhouette score: 0.54
Number of clusters with maximum silhouette score: 2

K-Means Elbow Method

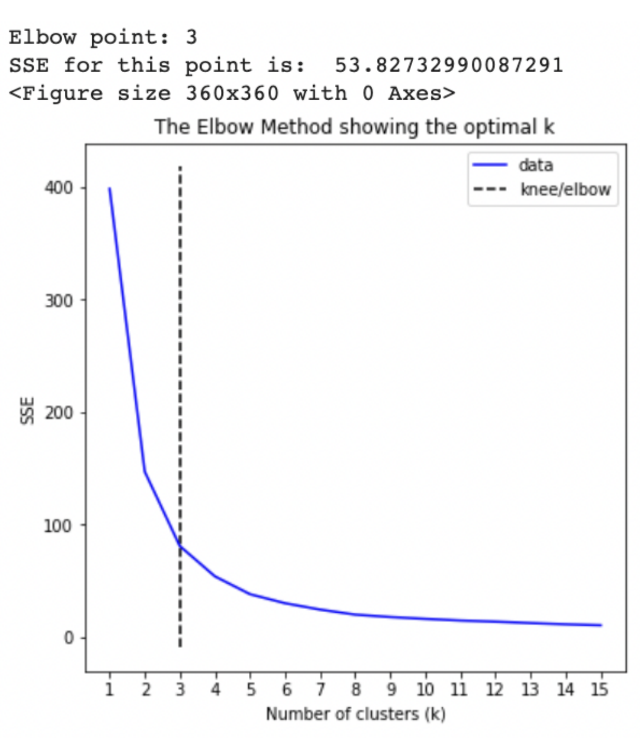


Figure 7: Elbow Method for the first dataset

Thus we conclude that the **best clustering using k-means algorithm and validating by SSE and elbow method** is:

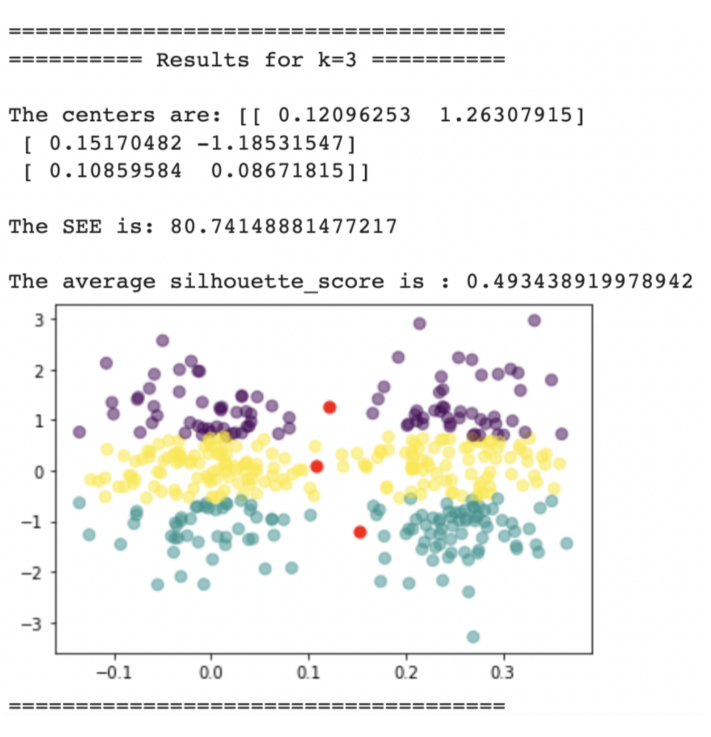


Figure 8: Best clustering for the first dataset using SSE and elbow method

best clustering using k-means algorithm and validating by silhouette score is:

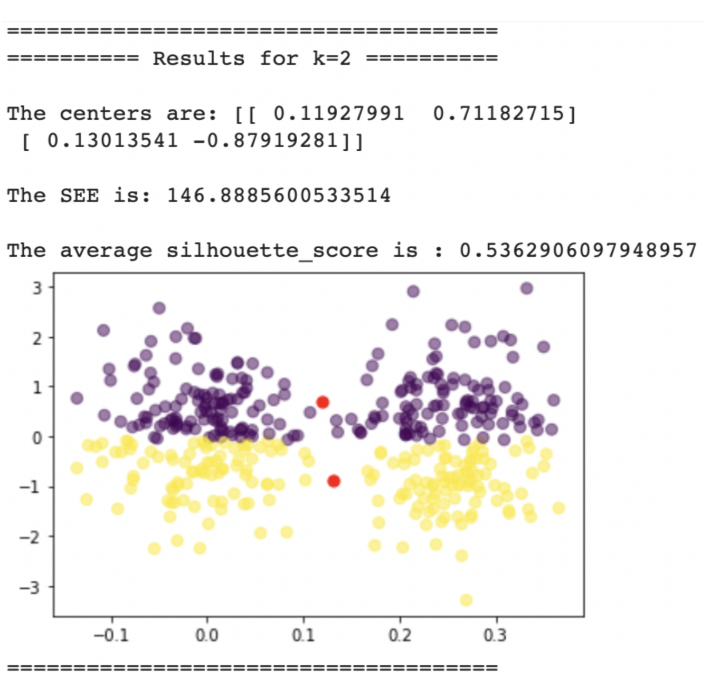


Figure 9: Best clustering for the first dataset using silhouette score

Fuzzy C-Means Clustering

Clusters are separated by colors, the **centroids** are colored **red** and the data points that are a member of **more than two clusters** are colored **purple**. (except for the first one)

cutoff_coefficient = 0.001

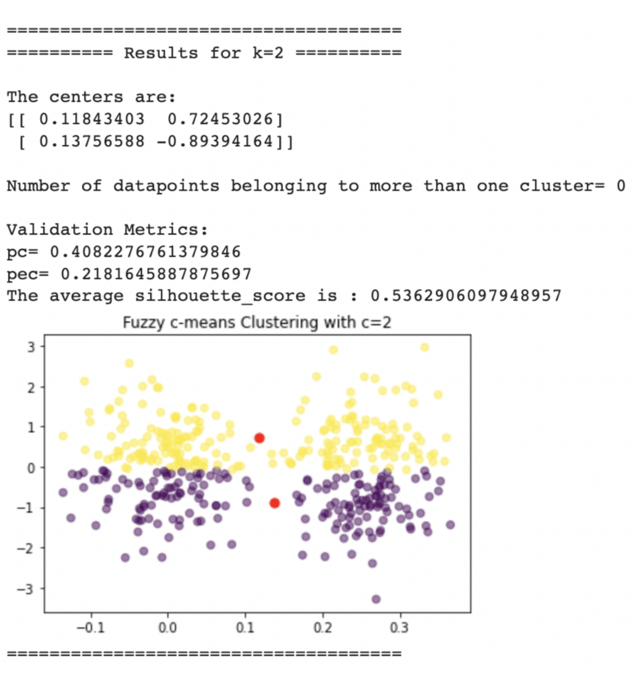


Figure 10: Number of clusters = 2

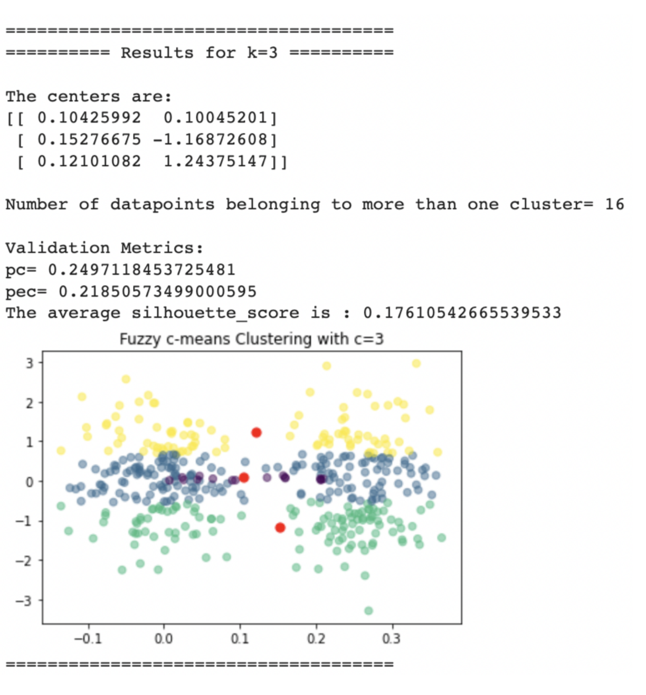


Figure 11: Number of clusters = 3

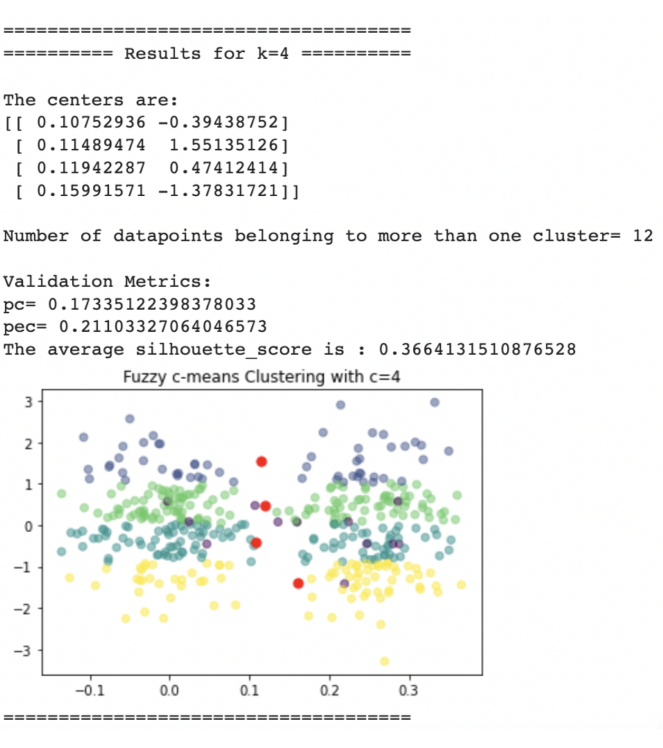


Figure 12: Number of clusters = 4

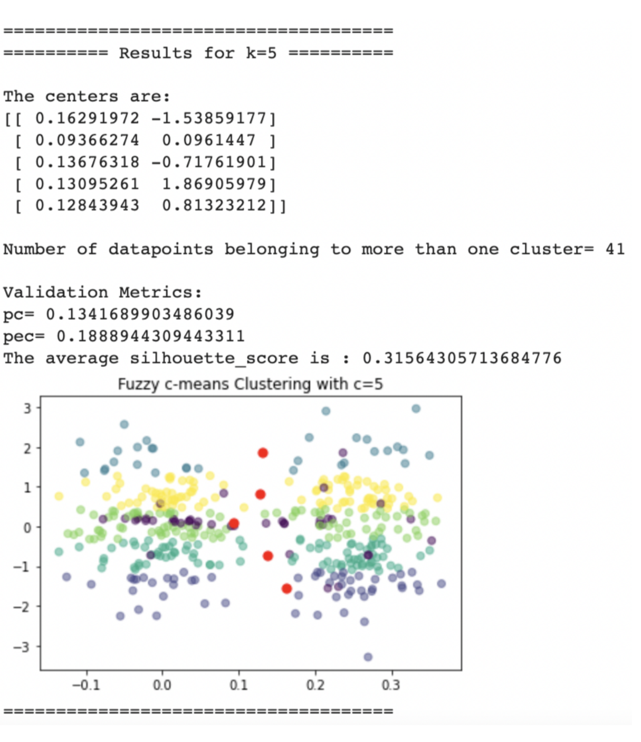


Figure 13: Number of clusters = 5

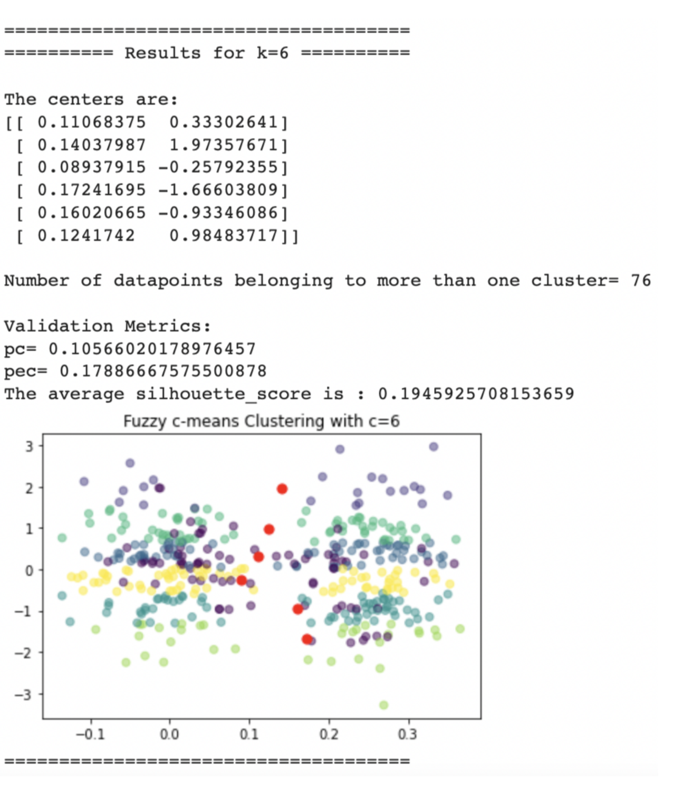
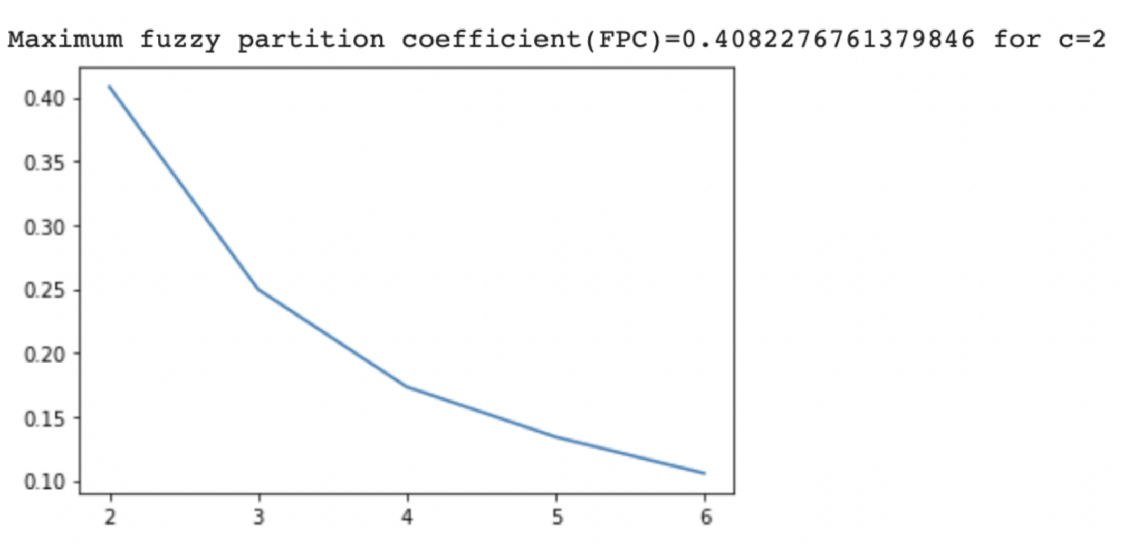


Figure 14: Number of clusters = 6



Minimum PEC: 0.17
Number of clusters with minimum PEC: 6
** Not a good cluster! **

Maximum PC: 0.4
Number of clusters with Maximum PC: 2

Maximum silhouette score: 0.54
Number of clusters with maximum silhouette score: 2

Thus we conclude that the **best clustering using fuzzy c-means algorithm and validating by both partition coefficient and silhouette score** is:

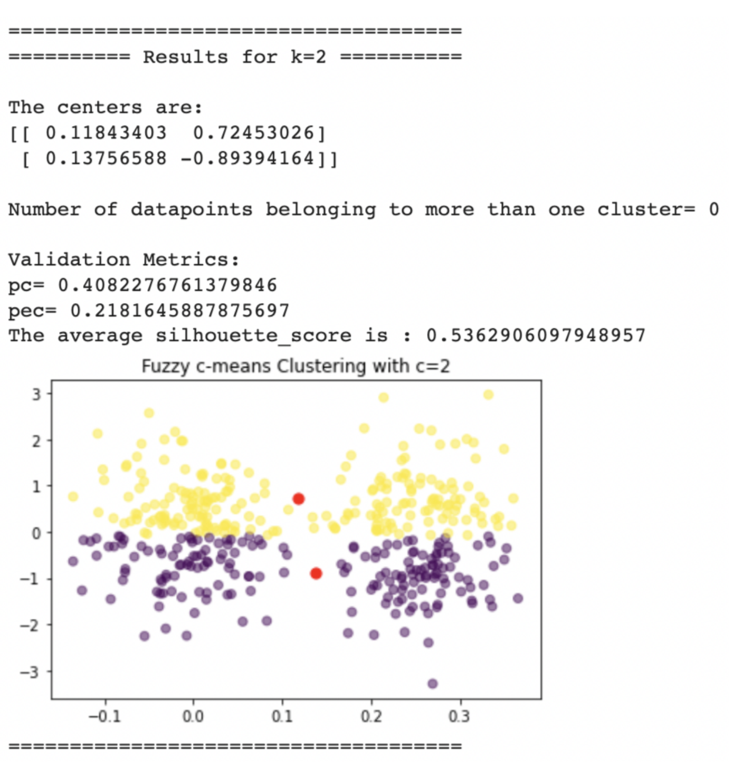


Figure 15: Best clustering for the first dataset using partition coefficient and silhouette score

DBSCAN Clustering

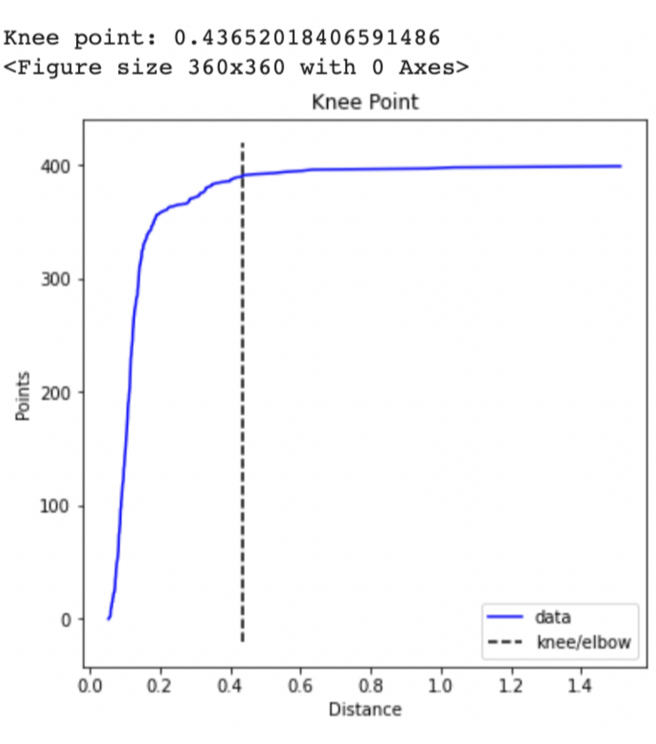
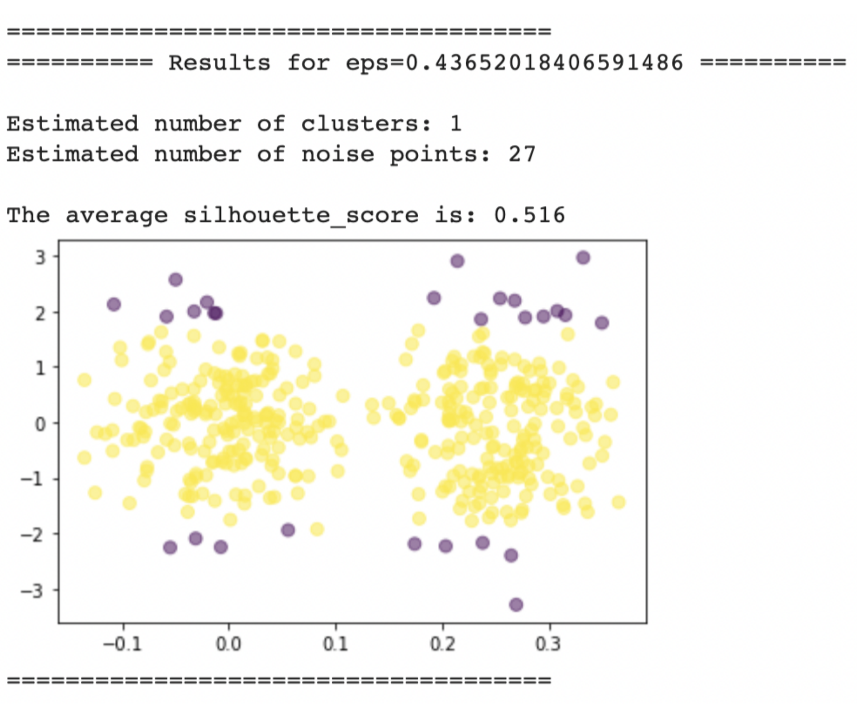


Figure 16: Best eps for DBSCAN using kNN algorithm

As explained in the previous section, we find the best *eps* parameter for the DBSCAN using the kNN algorithm. Thus the **best eps is 0.436**.

Then we try a range of numbers on the *min_ samples* parameter and maximize the silhouette score. Doing so we achieve the below clustering:
Purple dots are the **noise**.



Thus we conclude that the **DBSCAN** algorithm **fails** miserably to cluster this dataset. Which is probably because it is density-based.