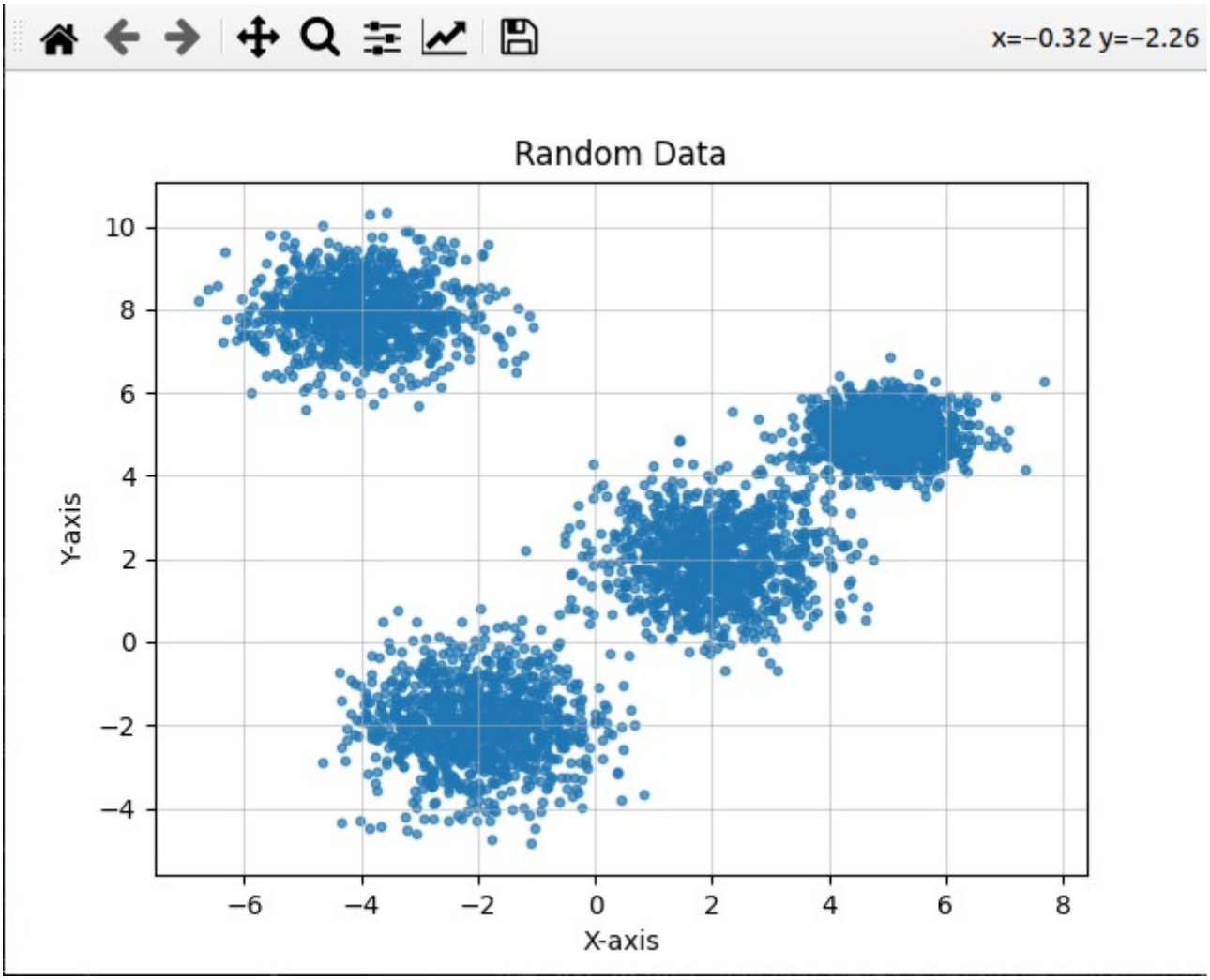
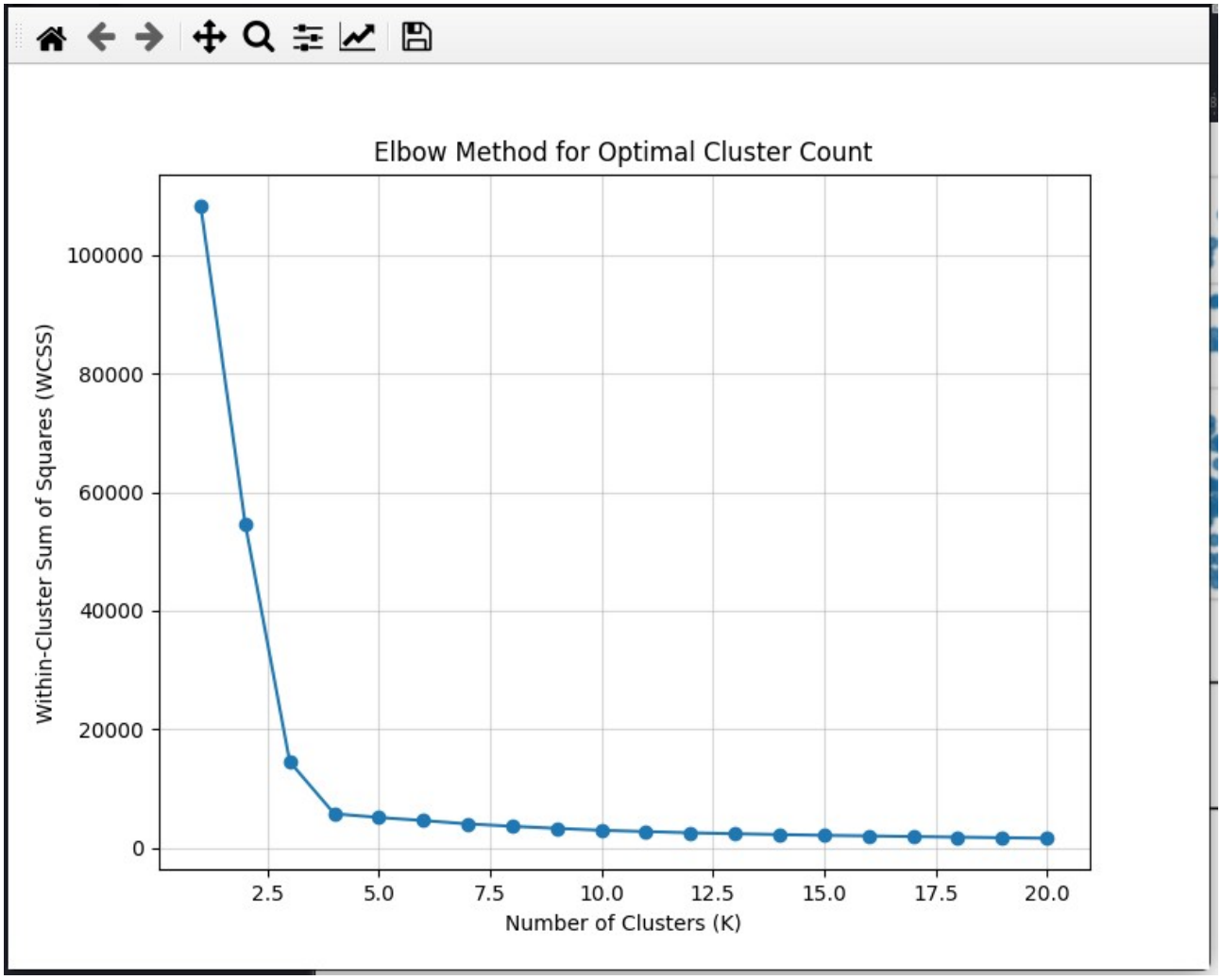
Usman Akinyemi
U20220090
MACHINE LEARNING AND PATTERN RECOGNITION
LAB2 REPORT

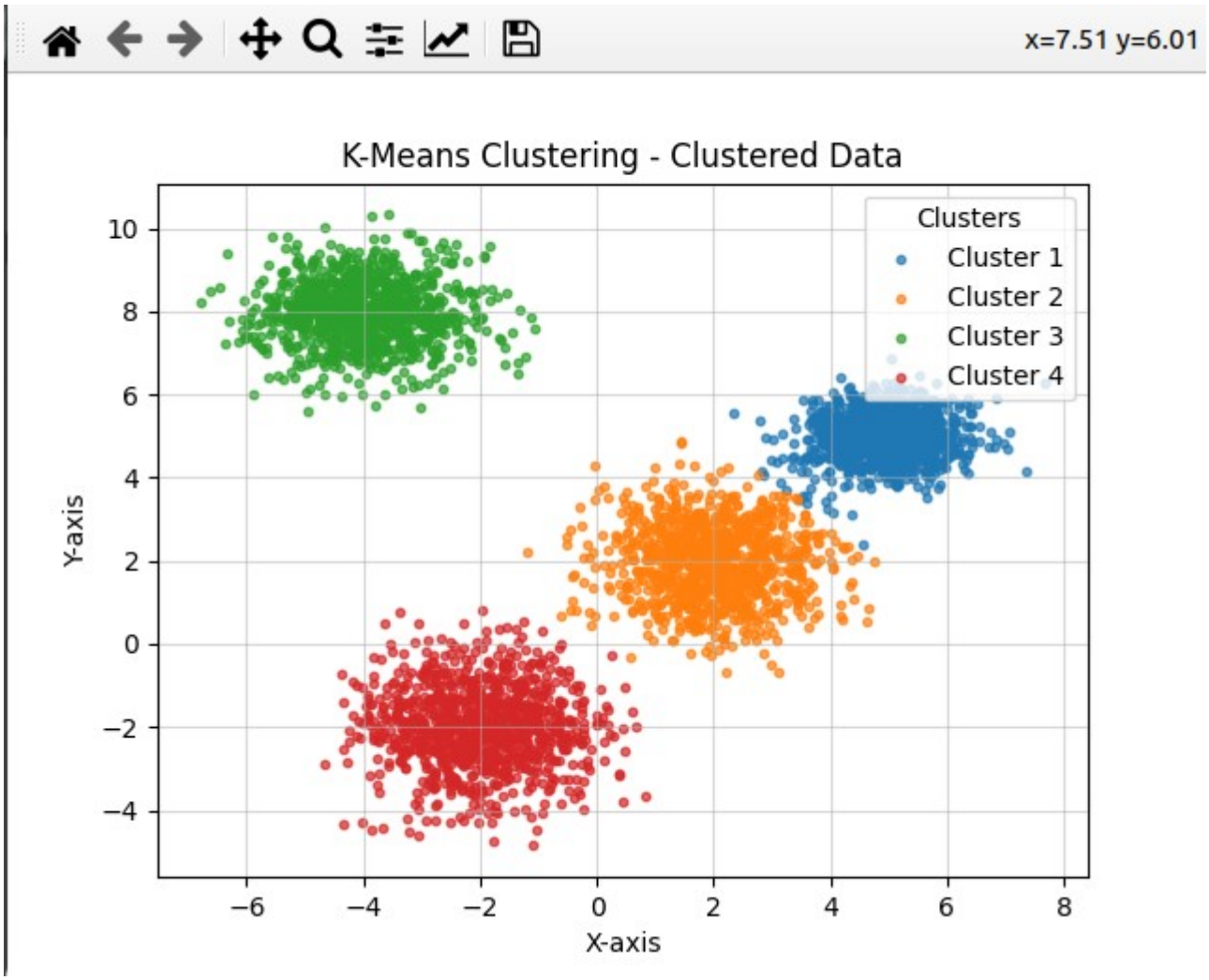
• Data Plot



Elbow Method plot



Clustered Data Plot



Questions:

1. What are the key factors to consider when choosing the value of k in kmeans clustering?

i. One thing to consider while choosing a Kmeans will be domain knowledge of the data someone is working with. The domain knowledge can give some intuition about what the value of K can be.

ii Another consideration is the WCSS(within-clusters sum of squares). One can take multiple values of K and then use WCSS to find the best value for K. We can get the best value by plotting the WCSS value by taken the K value where the graph tends to flatten out and forming an elbow.

iii. We can also consider using the Silhouette method to find the value of K which works by determining the degree of separation between clusters and higher value of silhouette score indicate better defined clusters.

iv. One other consideration might be to visualize the data clustering for different value of K as it might give insight into different natural grouping.

2. How does the initialization of centroids impact the performance of the kmeans algorithm? Are there any methods to improve initialization?

Initialization of centroids is very important in the performance of Kmeans as it can lead to a very slow convergence of the algorithm and might lead to incorrect clustering. One simple way to look at this is through understanding what Kmeans algorithm does. It basically looking for the shortest distance(euclidean) to the centroid from the other point. So, placing the centroids in a wrong position can make it hard to find the optimal configuration.

There are lot of methods of to improve initialization.

1. Placing the mean(centroid) at the edges of the datapoint far away from each other.
2. One way to is also to use pseudo-random algorithm for the initialization.

3. Why is it important to consider inter-cluster distance when determining the optimal number of clusters in a clustering algorithm?

Considering inter-cluster distance is crucial in clustering algorithms as it ensure distinct and well-separated clusters. It also enhance interpretability, avoids overlapping clusters, improves model generalization, and facilitates reliable decision-making.

4. Can the silhouette coefficient be negative? If so, what does a negative silhouette coefficient indicate about the clustering solution

Yes, the silhouette coefficient can be negative. The silhouette coefficient is a measure of how well-separated clusters are, ranging from -1 to 1. A negative silhouette coefficient indicates that the average distance between points in different clusters is greater than the average distance between points in the same cluster. In other words, it suggests that the clustering solution is less appropriate than a random assignment of points to clusters.

Interpretation of silhouette coefficients:

Near +1: Indicates a well-defined clustering solution.

Near 0: Suggests overlapping clusters.

Near -1: Implies a poor clustering solution, where points may be assigned to the wrong clusters.

5. What is the silhouette coefficient in the context of clustering, and how is it used to evaluate the quality of cluster

The silhouette coefficient is a metric (ranging from -1 to 1) used to assess the quality of clusters in clustering algorithms. It measures how well-separated clusters are, with higher values indicating better-defined clusters. The average silhouette coefficient is often used to select the optimal number of clusters, considering both cluster cohesion and separation.