Lloyd's algorithm (k-means) building and comparison to Hierarchical agglomerative clustering

Santiago Velasco

This report details the implementation of Loyd's and Hierarchical Agglomerative Clustering algorithm and its performance over two different datasets. The first one contains 3500 two-dimensional examples generated by a Gaussian mixture model, the second consists of 14,801 three-dimensional examples.

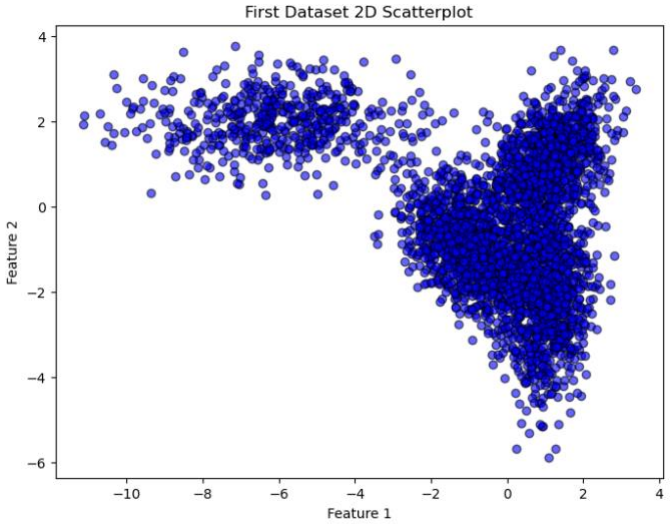

Figure 01. Data Points in the First Data Set.

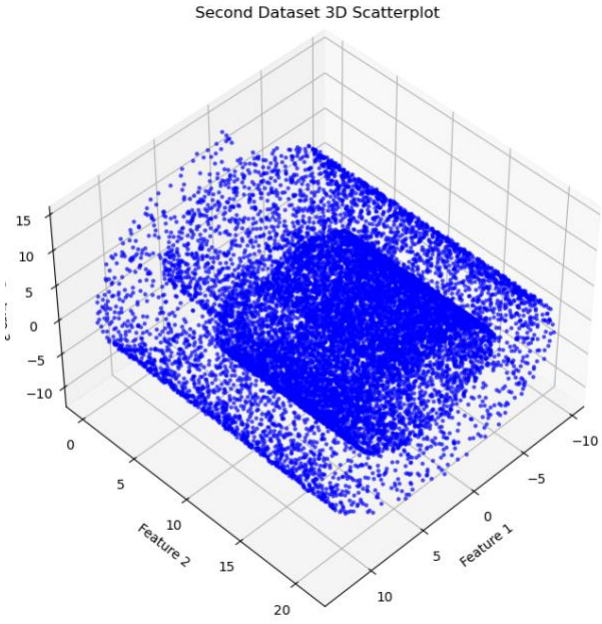

Figure 02. Data Points in the Second Data Set (First Perspective).

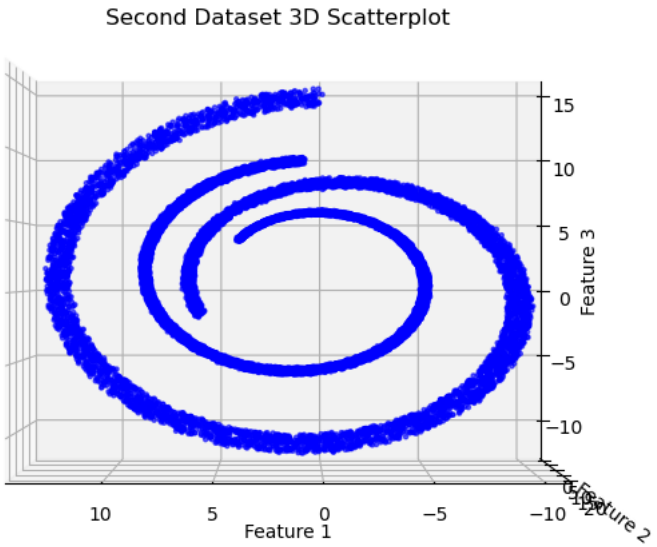

Figure 03. Data Points in the Second Data Set (Second Perspective).

The figures above show the datapoints distributions of the datasets along their features spaces, dataset one can be represented in a plane since it only contains two features. Data set two is more complex, since it contains three features it needs to be represented in a 3D space as it is shown in Figure 02, however the data follows a two spiral pattern when is viewed from the second perspective as in Figure 03.

The implemented function for Loyd's (k-means) algorithm follows the following process:

Initialization.

With ++ Initialization:

1. Choose the first centroid randomly.
2. Compute the initial squared distances from the first centroid.
3. Choose a new centroid by choosing from the remaining datapoints, where each datapoint have a probability to be chosen equal to the squared distance to their closest centroid over the sum of all squared distances.
4. Repeat step 2 and 3 until k centroids have been chosen.

Uniform Random Initialization:

1. Draw k different centroids at random.

K-means clustering:

1. Compute the distances of each centroid with respect to all data points.
2. Label each datapoint with the closest centroid.
3. For every centroid update its value to be the average distance of the datapoints clustered by that centroid.

4. Repeat until the difference between the new and old centroids are less than a specified threshold also known as tolerance.

With the Loyd's algorithm implemented, the following values for k were chosen: 2, 4, 8, 12, 16, 20, 24, 28, 32.

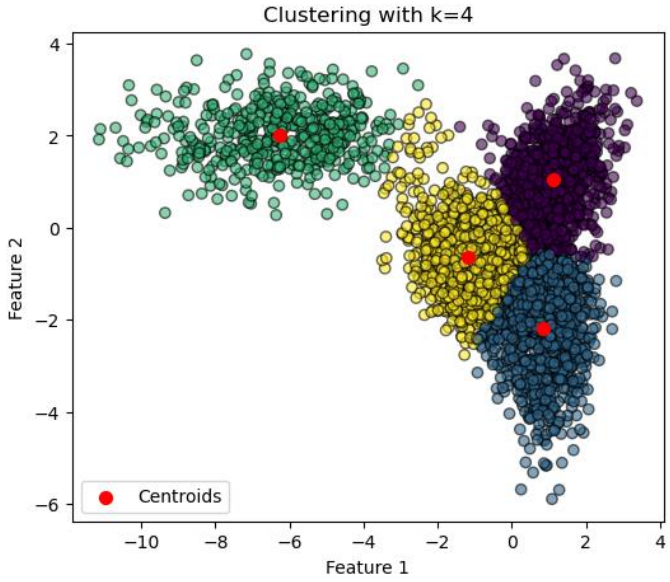

Figure 04. Clustering with four centroids using Uniform Random Initialization.

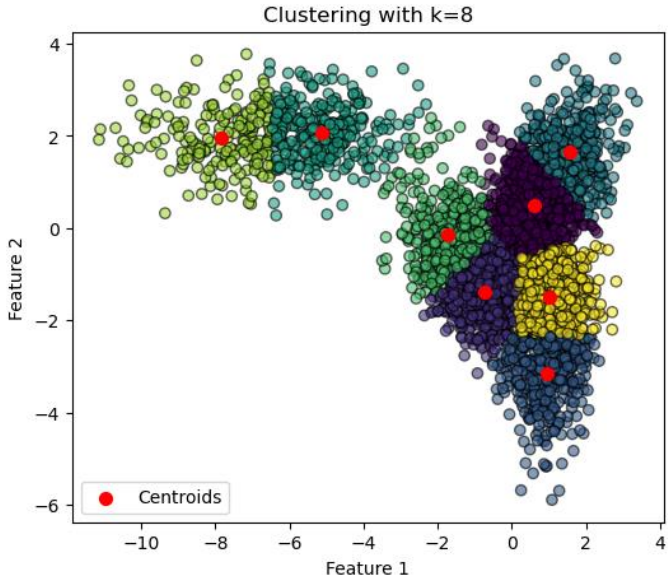

Figure 05. Clustering with eight centroids using Uniform Random Initialization.

Figures 04 and 05 shows the datapoints clustering with four and eight centroids respectively when Loyd's algorithm is used with Uniform Random Initialization.

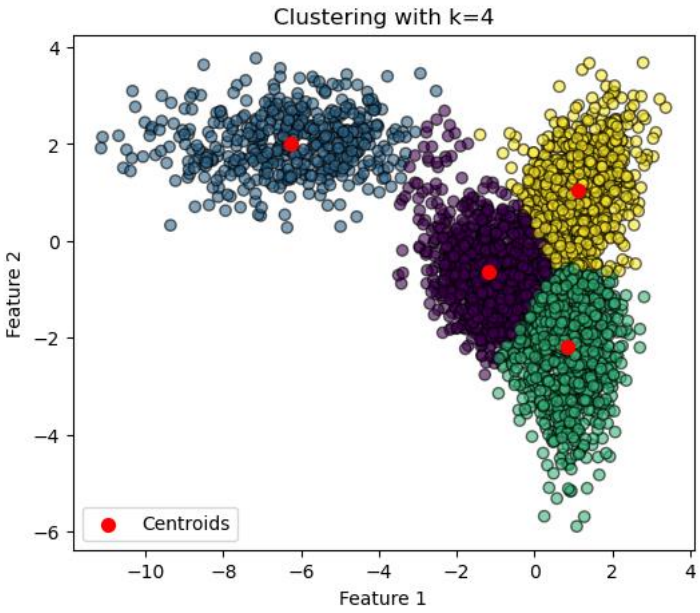

Figure 06. Clustering with four centroids using ++Initialization.

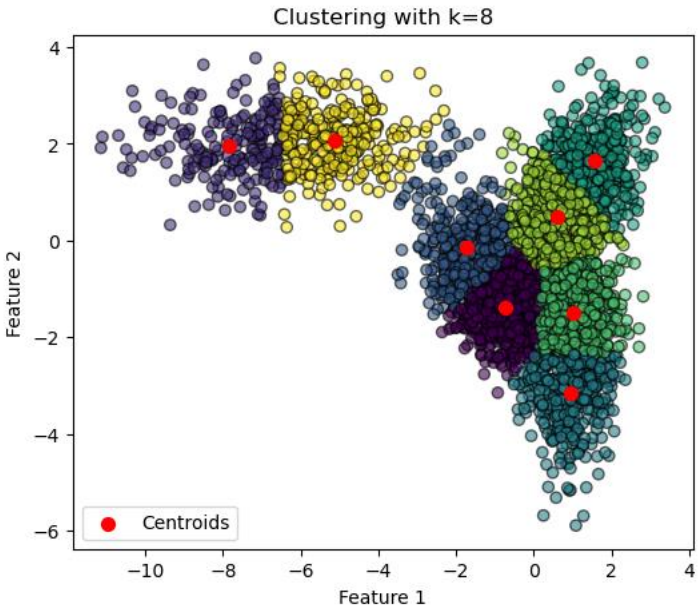

Figure 07. Clustering with eight centroids using ++Initialization.

Figures 06 and 07 shows the datapoints clustering with four and eight centroids respectively when Loyd's algorithm is used with Uniform Random Initialization.

To have a better understanding about the spreading of the datapoints relative to their respective clusters, the sum of the squared errors from each point to its cluster center (cost) was computed.

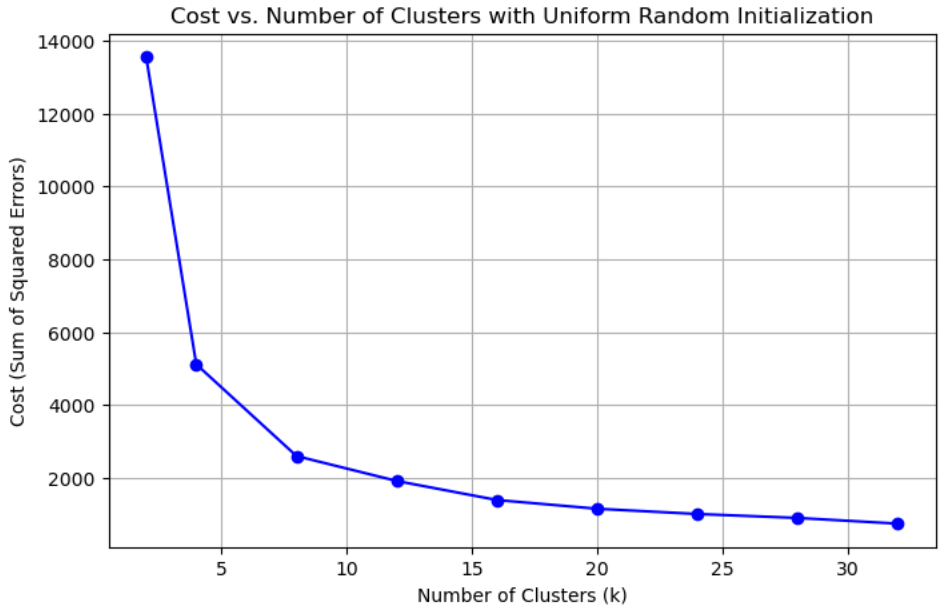

Figure 08. Datapoints cost against increasing number of centroids with Uniform Random Initialization.

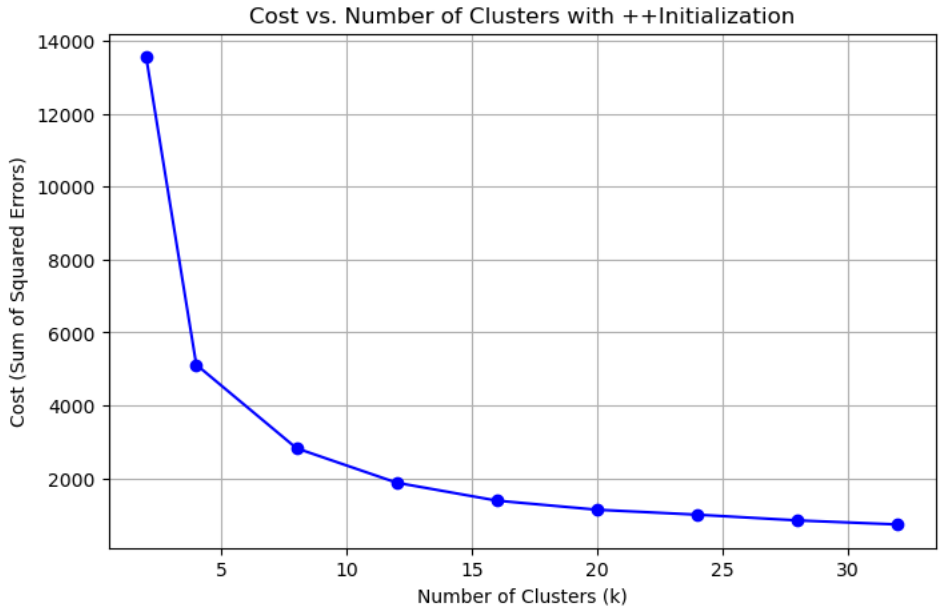

Figure 09. Datapoints cost against increasing number of centroids with ++Initialization.

At first glance the clustering with four and eight centroids seemed to be the most appropriate for this dataset. However, based on cost, clustering with twelve centroids is a middle point between cost lowering improvement and over clustering of the datapoints, for higher number of centroids the cost improvements are minimal and the risk of over clustering the data increases.

Hierarchical agglomerative clustering is an alternative to Loyd's algorithm. It merges or splits cluster successively to form nested clusters, for this experiment single and average linkages were used. Single linkage minimizes the distance between the closest observations of two clusters, in the other hand, average linkage minimizes the average of the distances between all observations of two clusters.

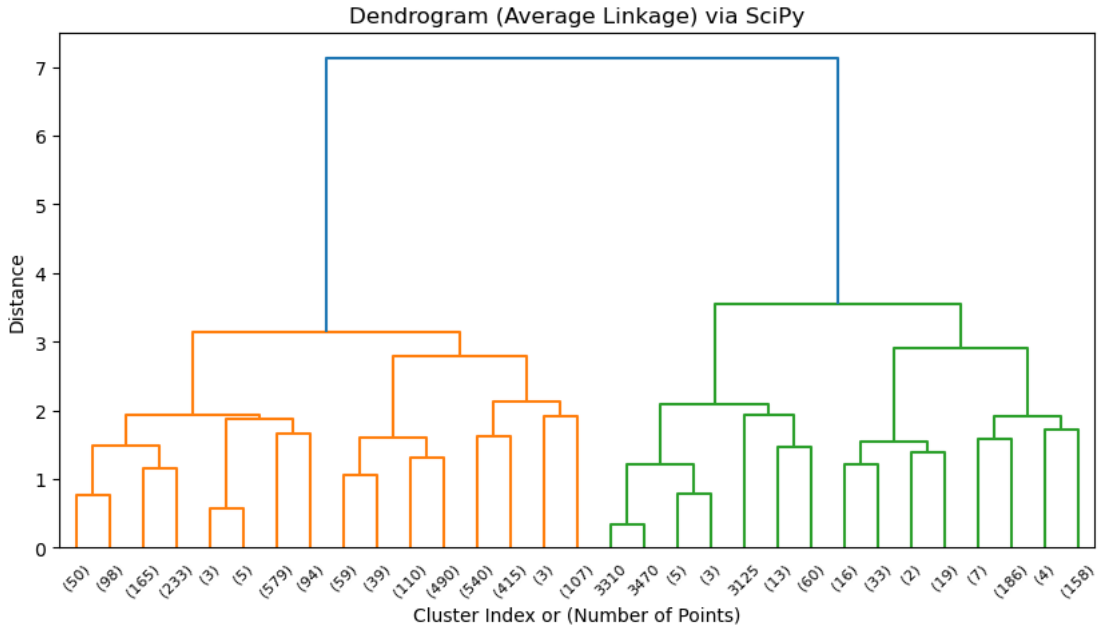

Figure 10. Dendrogram of Hierarchical Agglomerative Clustering with Single Linkage in the First Dataset.

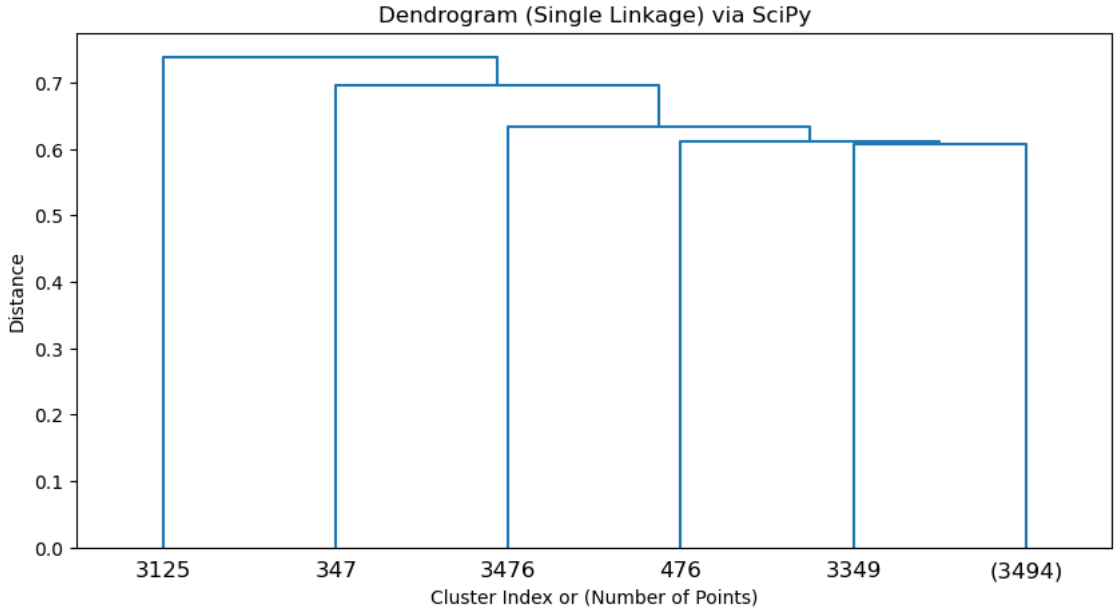

Figure 11. Dendrogram of Hierarchical Agglomerative Clustering with Average Linkage in the First Dataset.

The previous figures show the clusters distributions with the number of examples that each cluster contains as the tree leaves. Each dendrogram only shows four splits, where the most noticeable difference is the number of leaves that each tree has at this level, average linkage produced less leaves and therefore contains more examples in each of them, while single linkage generated a broader tree with less examples per leaf.

The next experiment was ran using Loyd's algorithm in the 3D dataset and these were the results:

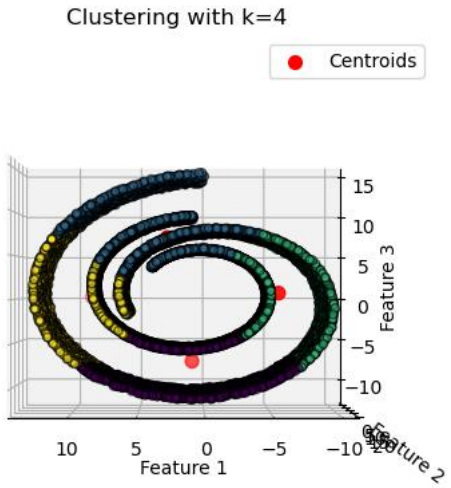

Figure 12. Clustering with four centroids using Uniform Random Initialization.

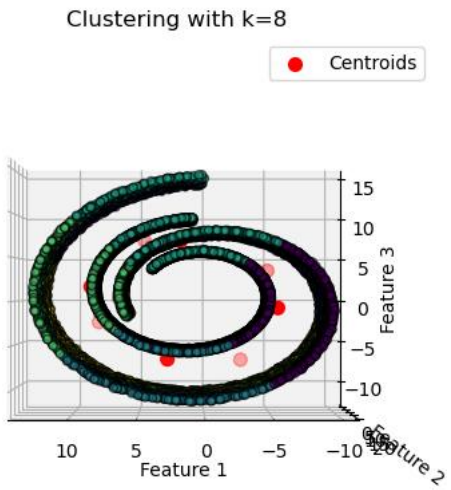

Figure 13. Clustering with eight centroids using Uniform Random Initialization.

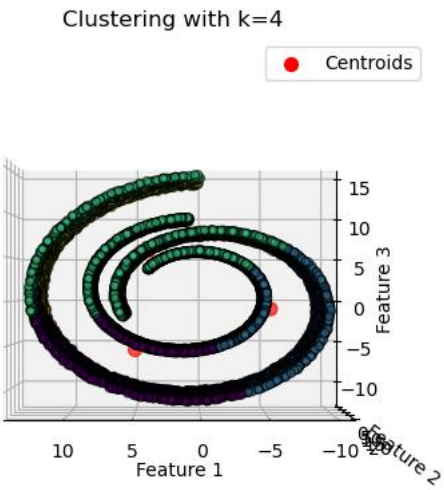

Figure 14. Clustering with four centroids using ++Initialization.

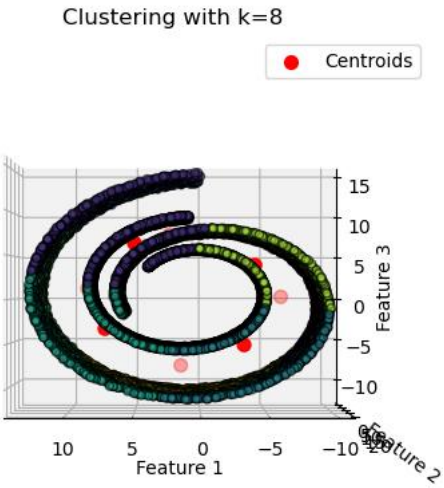

Figure 15. Clustering with four centroids using ++Initialization.

From this perspective the in both cases, Uniform Random Initialization and ++Initialization it is only possible to see a maximum of four clusters. This might imply that the correct number of clusters is four.

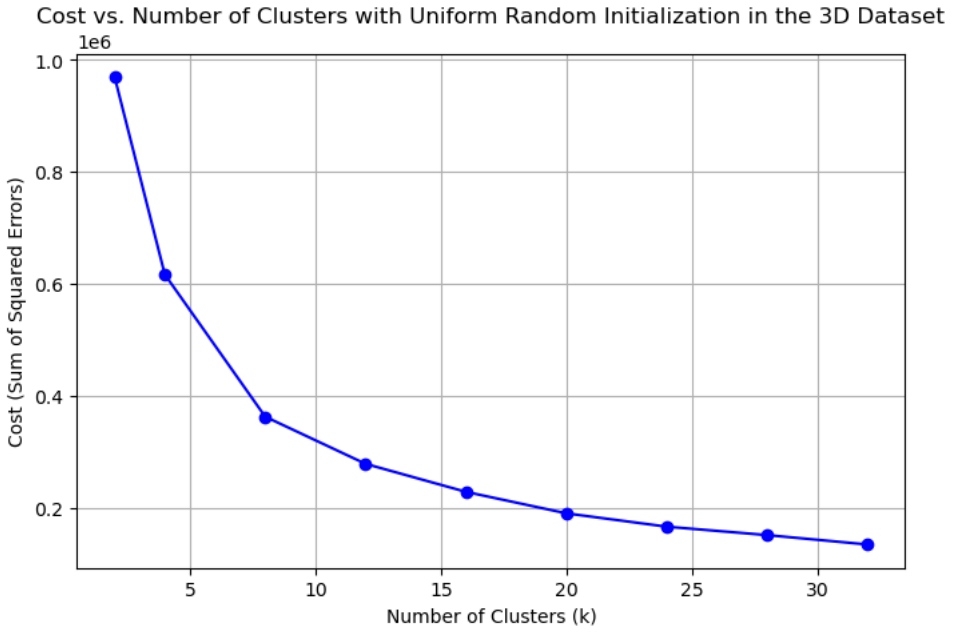

Figure 16. Datapoints cost against increasing number of centroids with Uniform Random Initialization in the second dataset.

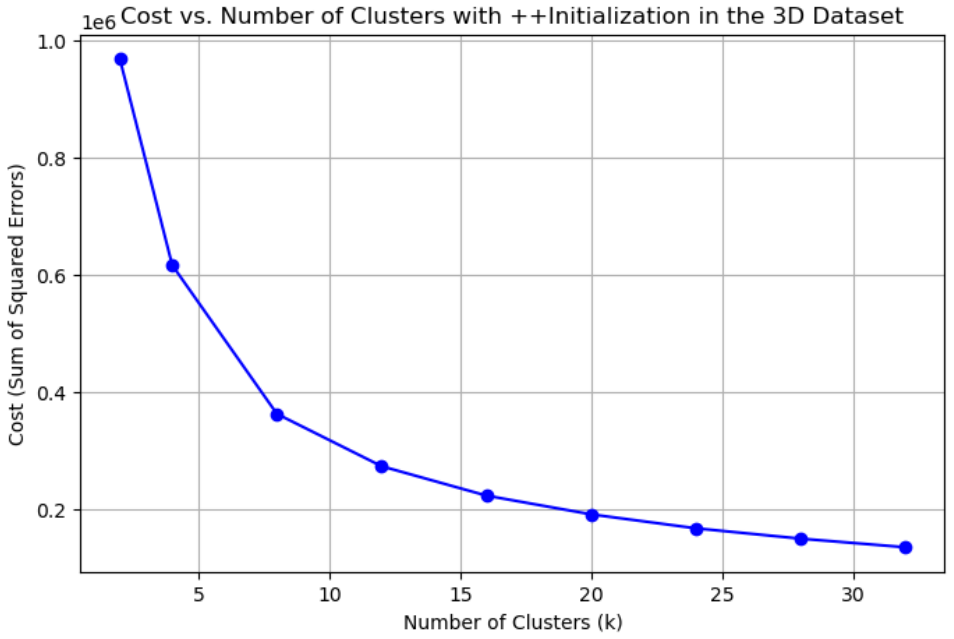

Figure 17. Datapoints cost against increasing number of centroids with ++Initialization in the second dataset.

Based on the cost, at simply glance the bests number of centroids are twelve and sixteen because from that point the cost reducing is almost negligible.

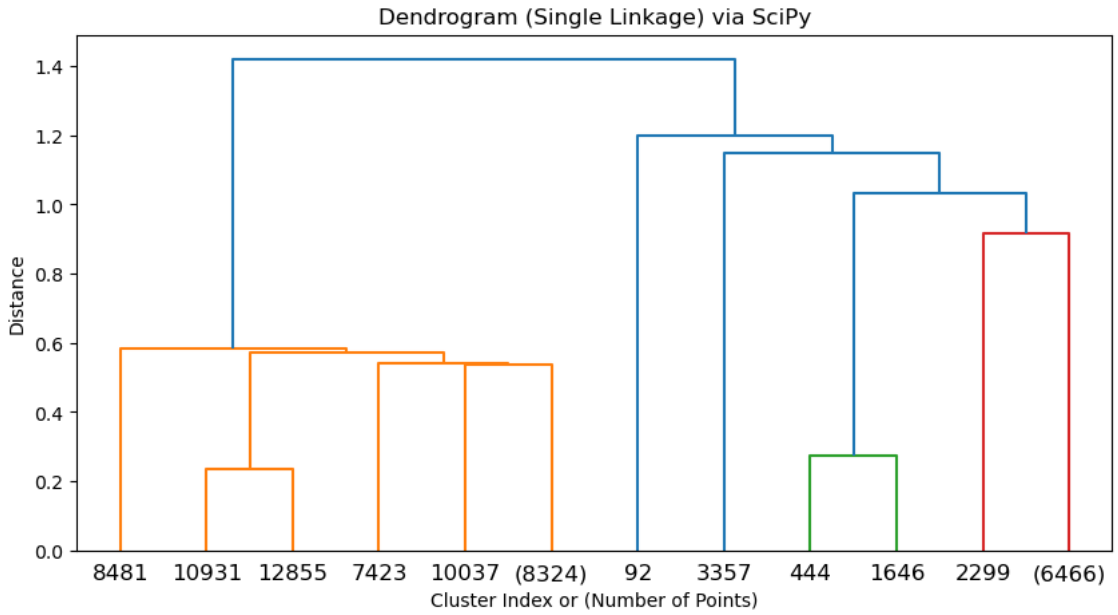

Figure 18. Dendrogram of Hierarchical Agglomerative Clustering with Single Linkage in the Second Dataset.

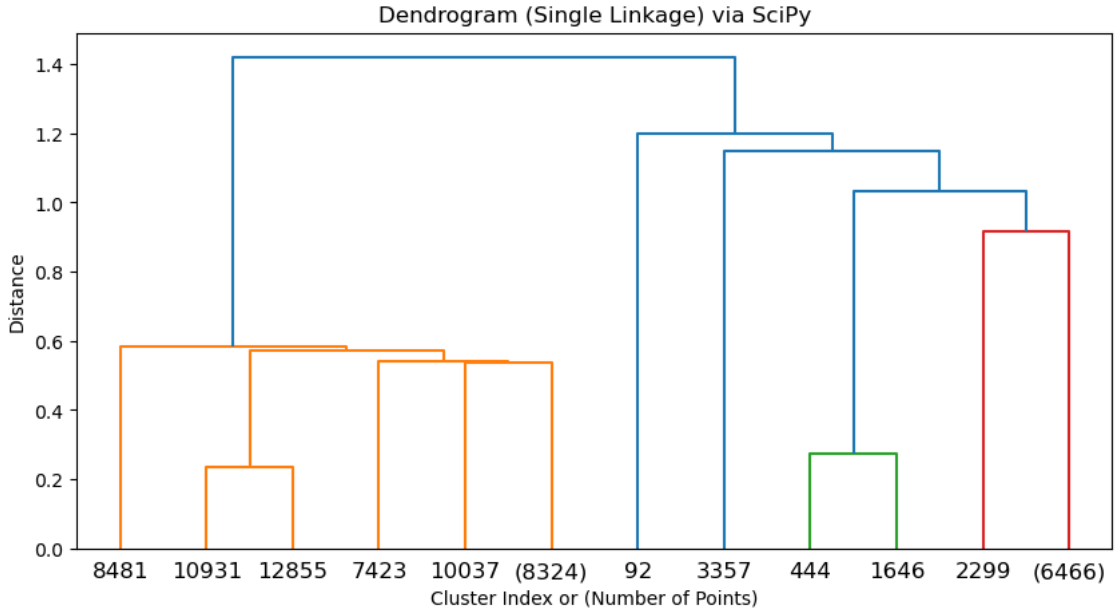

Figure 18. Dendrogram of Hierarchical Agglomerative Clustering with Average Linkage in the Second Dataset.

Finally, the previous figures show the different dendrograms, of the two types of Hierarchical Agglomerative Clustering in the second dataset, where agglomerative clustering seems to have a more even distribution of examples per leaf.