# Assignment 3
# Report

## An introduction into clustering.

Student:

Amaan Makhani

[redacted]

Teacher:

Nishant Mehta

Course:

Seng 474 - Data Mining

# 1. Introduction

In this report, we will be using a 2D and a 3D dataset to explore clustering. Both datasets are plotted in **Figures 1 and 2** below. Task one and two use an implementation of Lloyd's algorithm. Task one uses random initialization, while task two uses k-means++ for initialization. Task three and four explore hierarchical agglomerative clustering using Euclidean distance. Task three uses a single linkage, while task four uses average linkage.

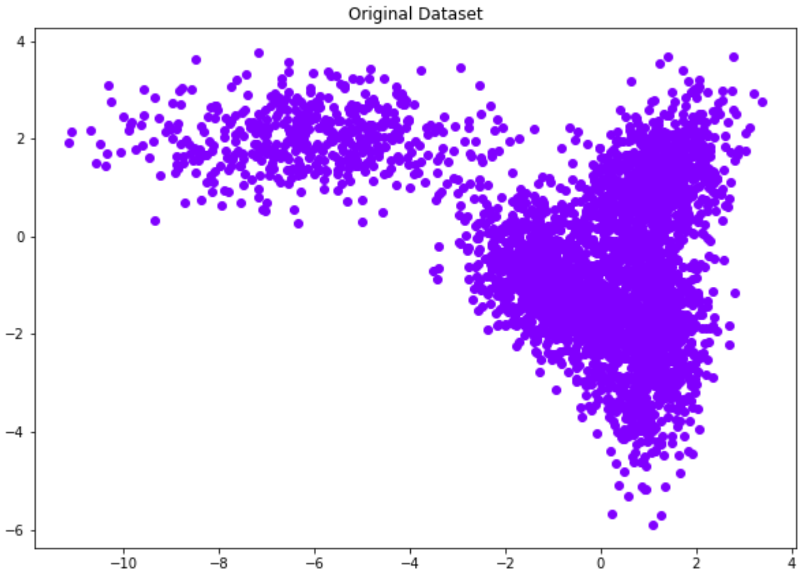


***Figure 1: Dataset 1 prior to clustering.***

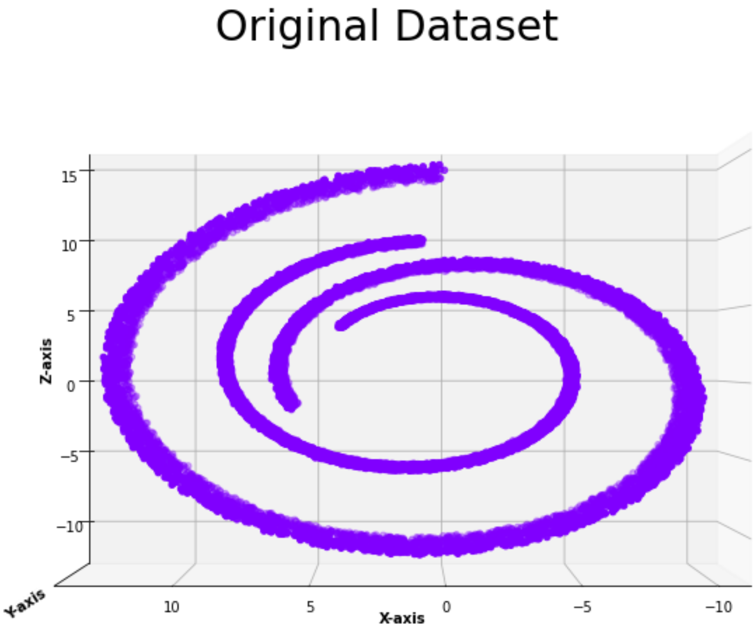
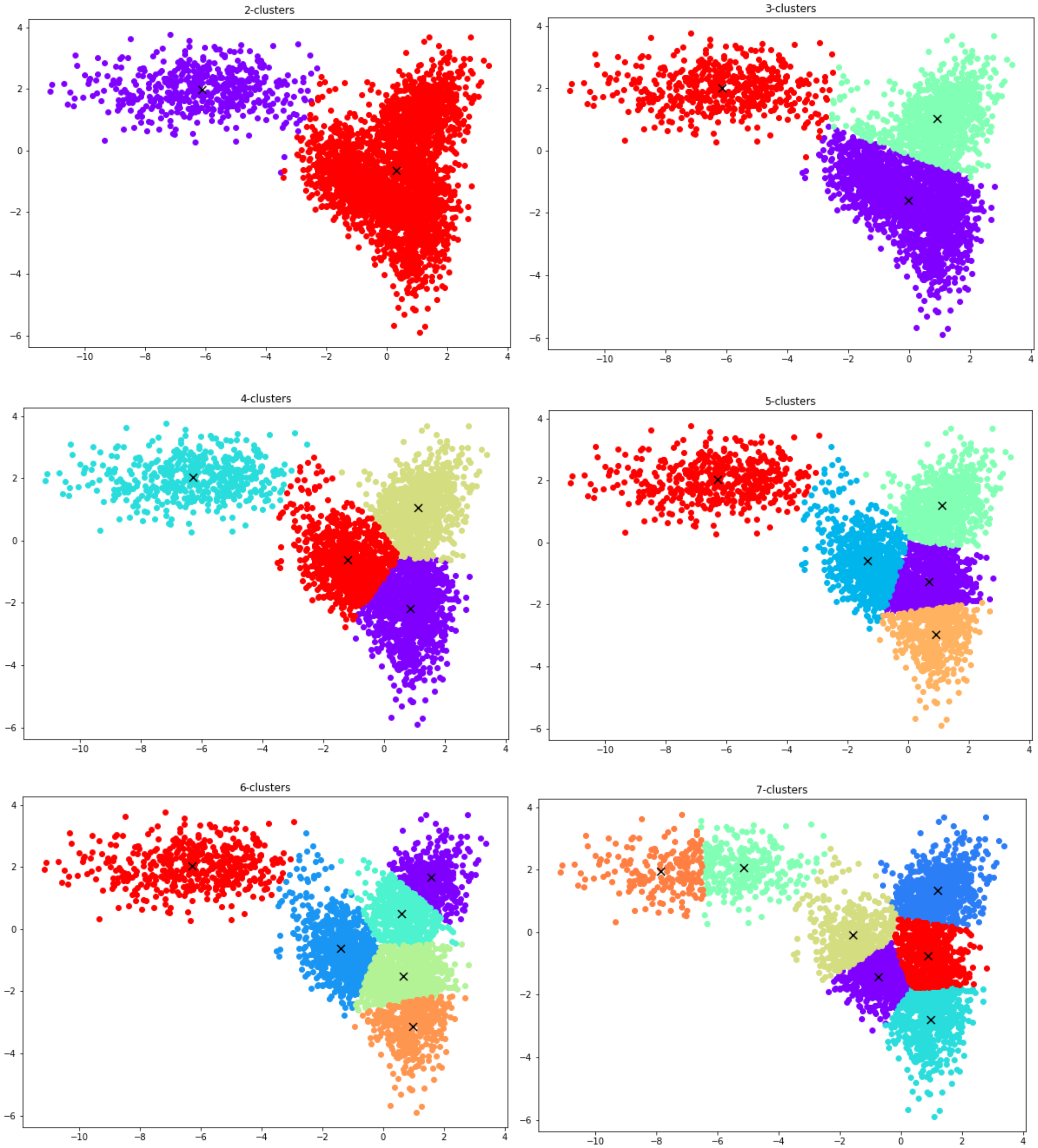


***Figure 2: Dataset 2 prior to clustering.***

## 2. Task One

In **figure 3** below, dataset one is shown with a varied number of clusters beginning from two and ending at ten clusters. To create the clustering, a uniform random initialization version of Lloyd's algorithm was used. These visuals are intended to allow for visualization of the clustering and allow for a better comparison of clustering using other parameters or methods.



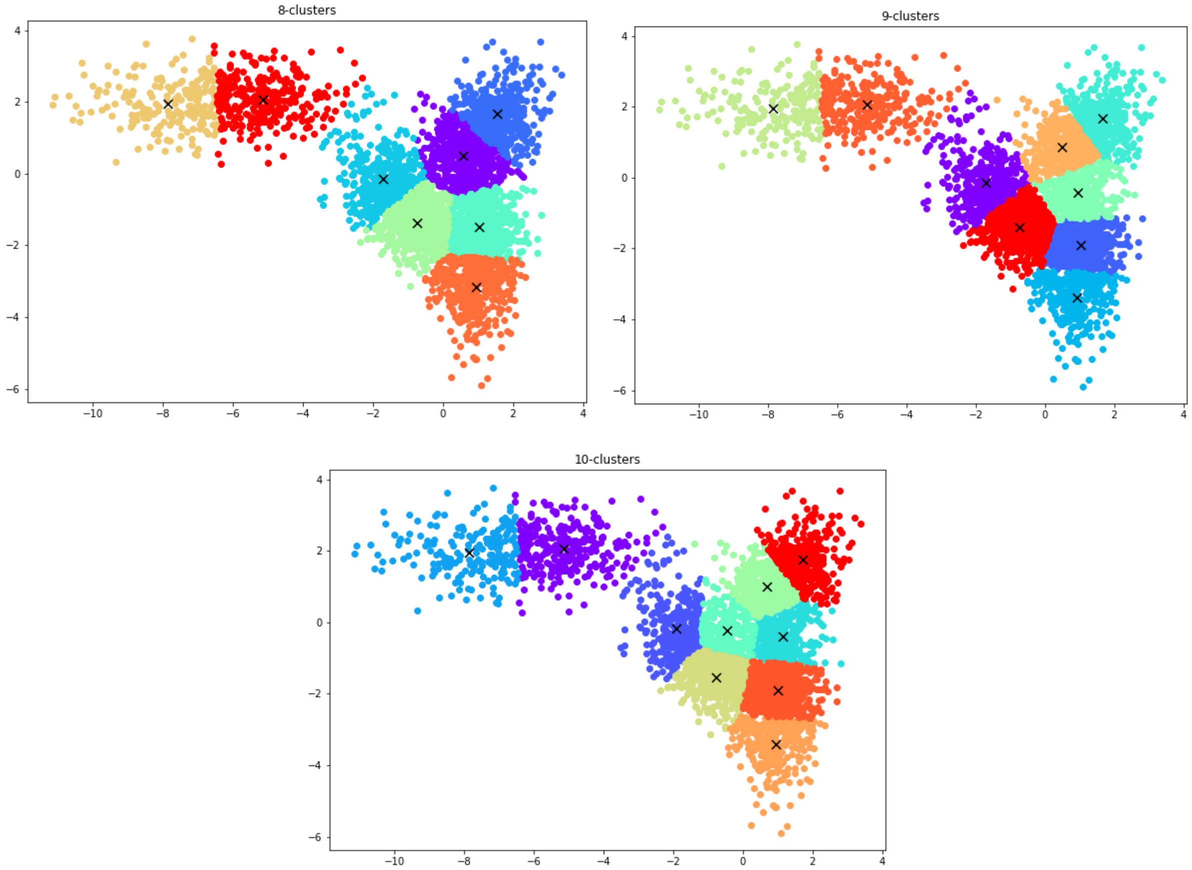


**Figure 3: Dataset 1 is shown with a varied number of clusters. The clusters were created using Lloyd's algorithm and uniform random initialization.**

In **figure 4** below, the cost of each clustering for dataset one is plotted against the number of clusters. Using this graph, I selected a clustering of 4 as this is the point where the graph reaches a kink in the curve. In other words, this is the point in which I feel like the graph stopped decreasing rapidly. To verify my choice, I also inspected the visuals of the clustering, 4 clusters seemed to have the best spacing of the cluster centers and appeared to have the most logical separation of datapoints. Dataset one divided into four clusters is shown below in **figure 5**.

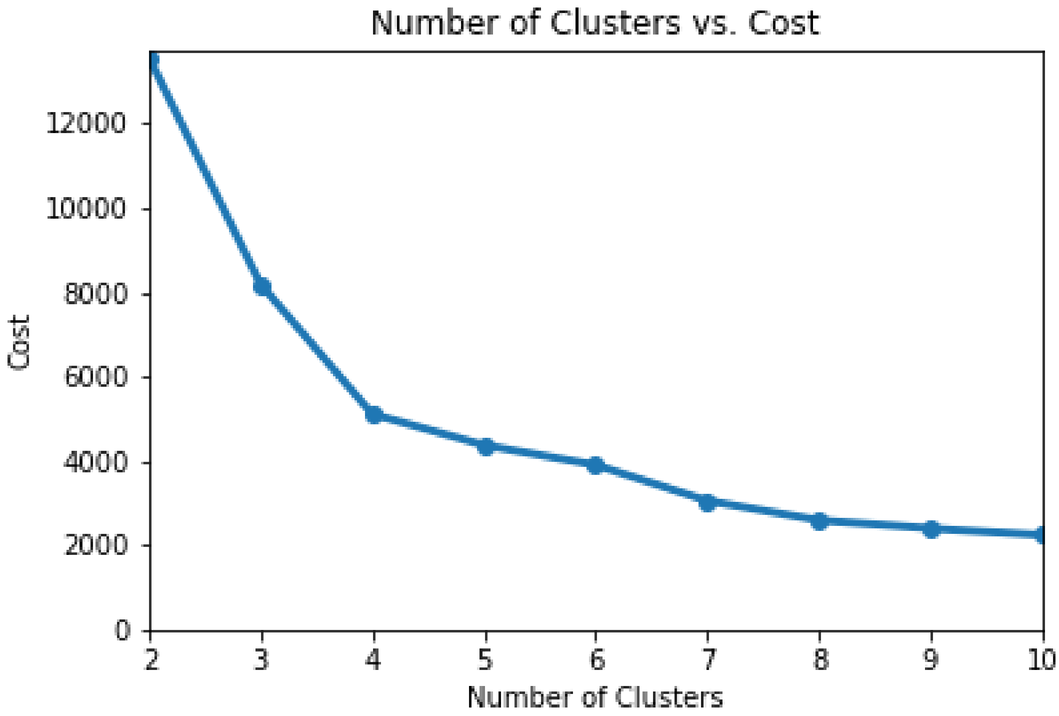


***Figure 4: Dataset 1 plotted with respect to cost vs. clustering.***

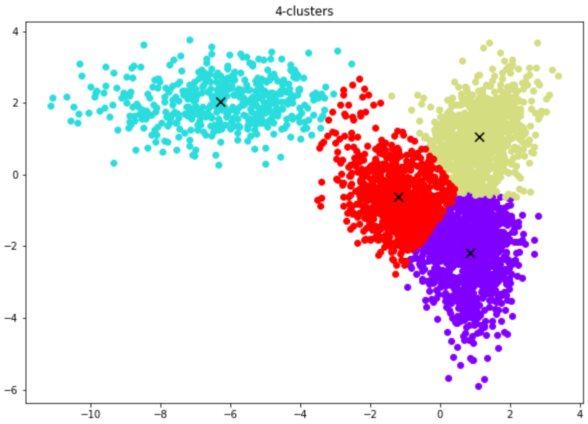
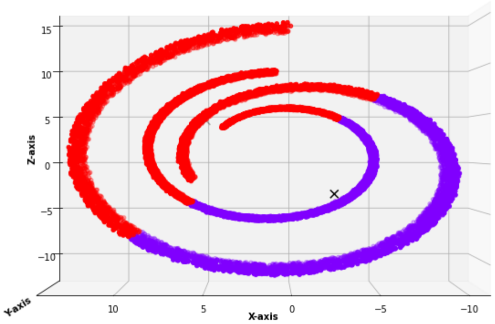


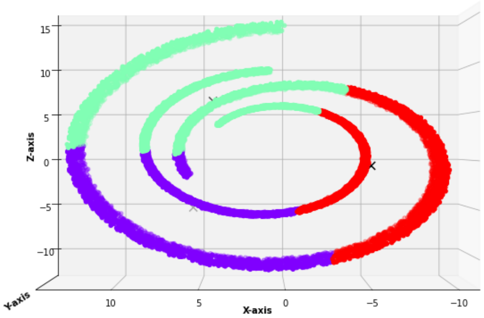
***Figure 5: Chosen clustering for task one and dataset one.***

In **figure 6** below, dataset two is shown with a varied number of clusters beginning from two and ending at ten clusters. To create the clustering, a uniform random initialization version of Lloyd's algorithm was used.

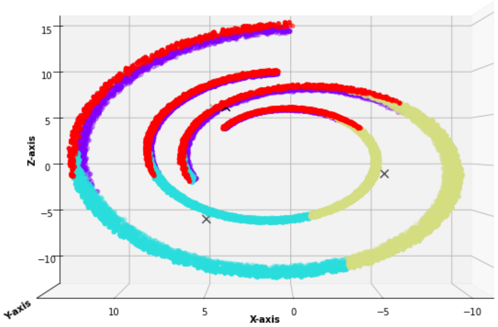
2-clusters



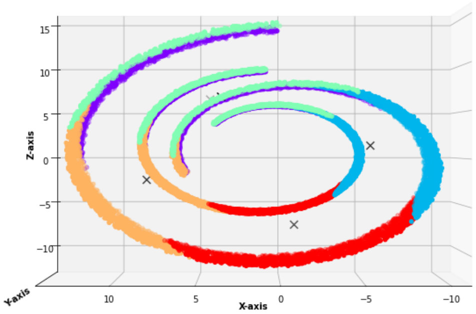
3-clusters



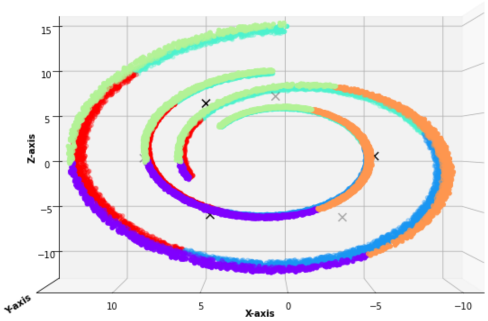
4-clusters



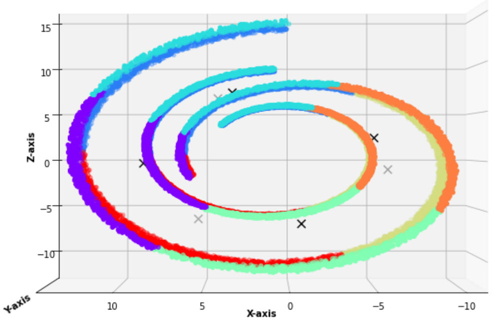
5-clusters



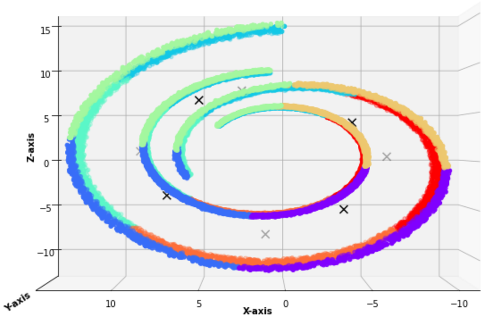
6-clusters



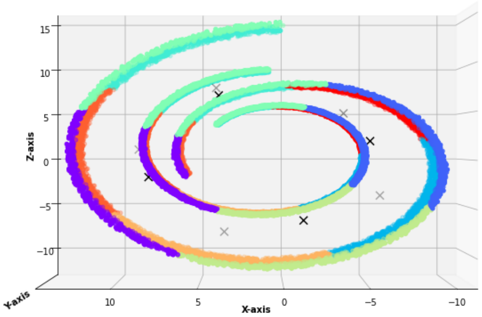
7-clusters



8-clusters



9-clusters



10-clusters

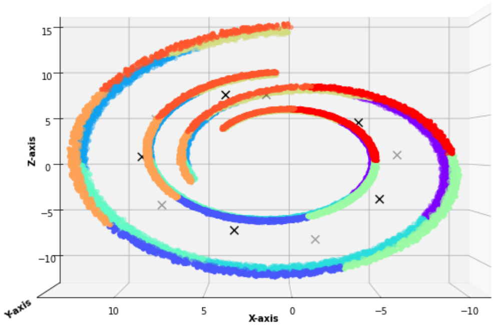


**Figure 6: Dataset 2 is shown with a varied number of clusters. The clusters were created using Lloyd's algorithm and uniform random initialization.**

In **figure 7** below, the cost of each clustering for dataset two is plotted against the number of clusters. Using this graph, I selected a clustering of 8, as this is the point where the graph reaches a kink in the curve. In other words, this is the point in which I feel like the graph stopped decreasing rapidly. To verify my choice, I also inspected the visuals of the clustering and found no clustering clustered the data efficiently. It seems like Lloyd's cannot cluster this dataset, which testifies to its inability to cluster non-globular data, as discussed in lectures and readings. The chose clustering is shown in **figure 8**.

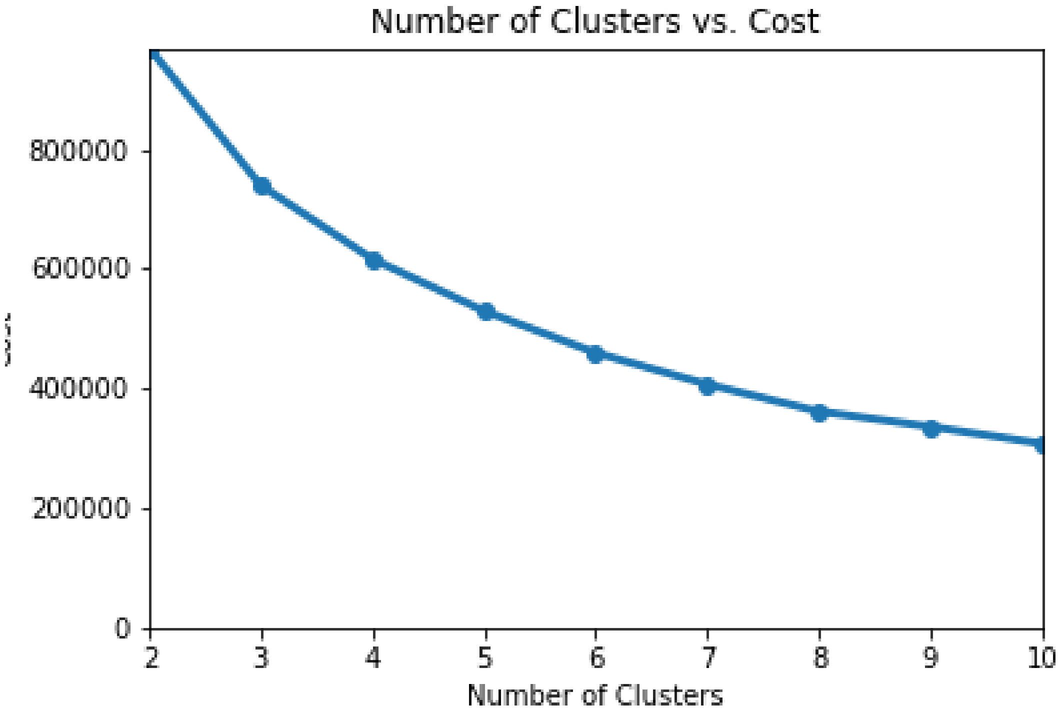


**Figure 7: Dataset 2 plotted with respect to cost vs. clustering.**

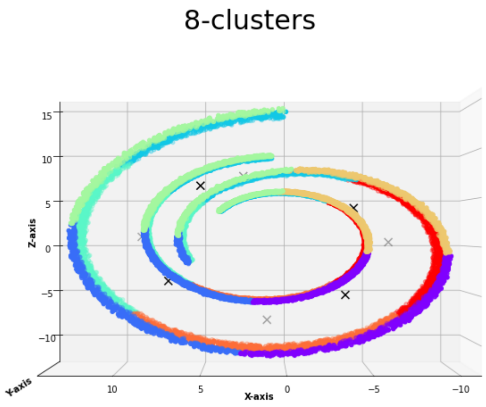
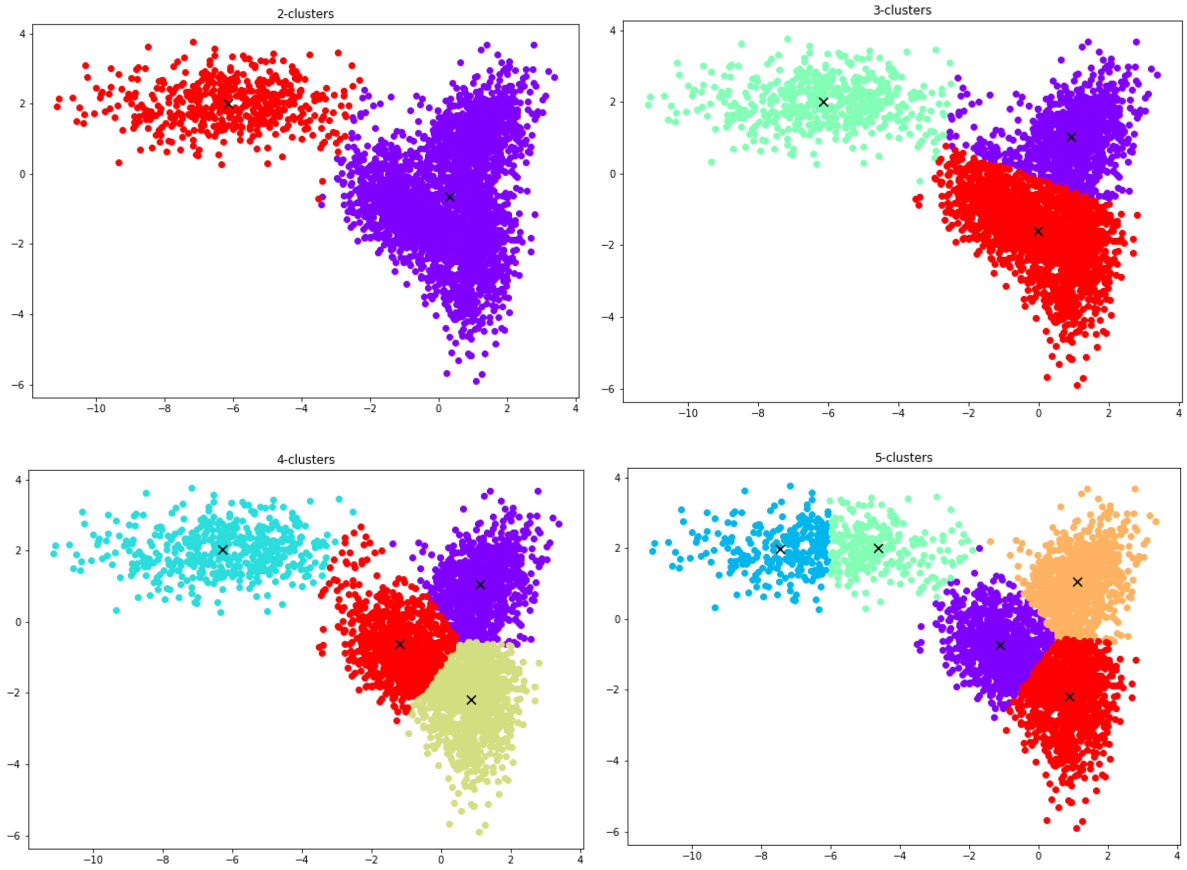


*Figure 8: Chosen clustering for task one and dataset two.*

## 3. Task Two

In **figure 9** below, dataset one is shown with a varied number of clusters beginning from two and ending at ten clusters. To create the clustering, a k-means++ version of Lloyd's algorithm was used.



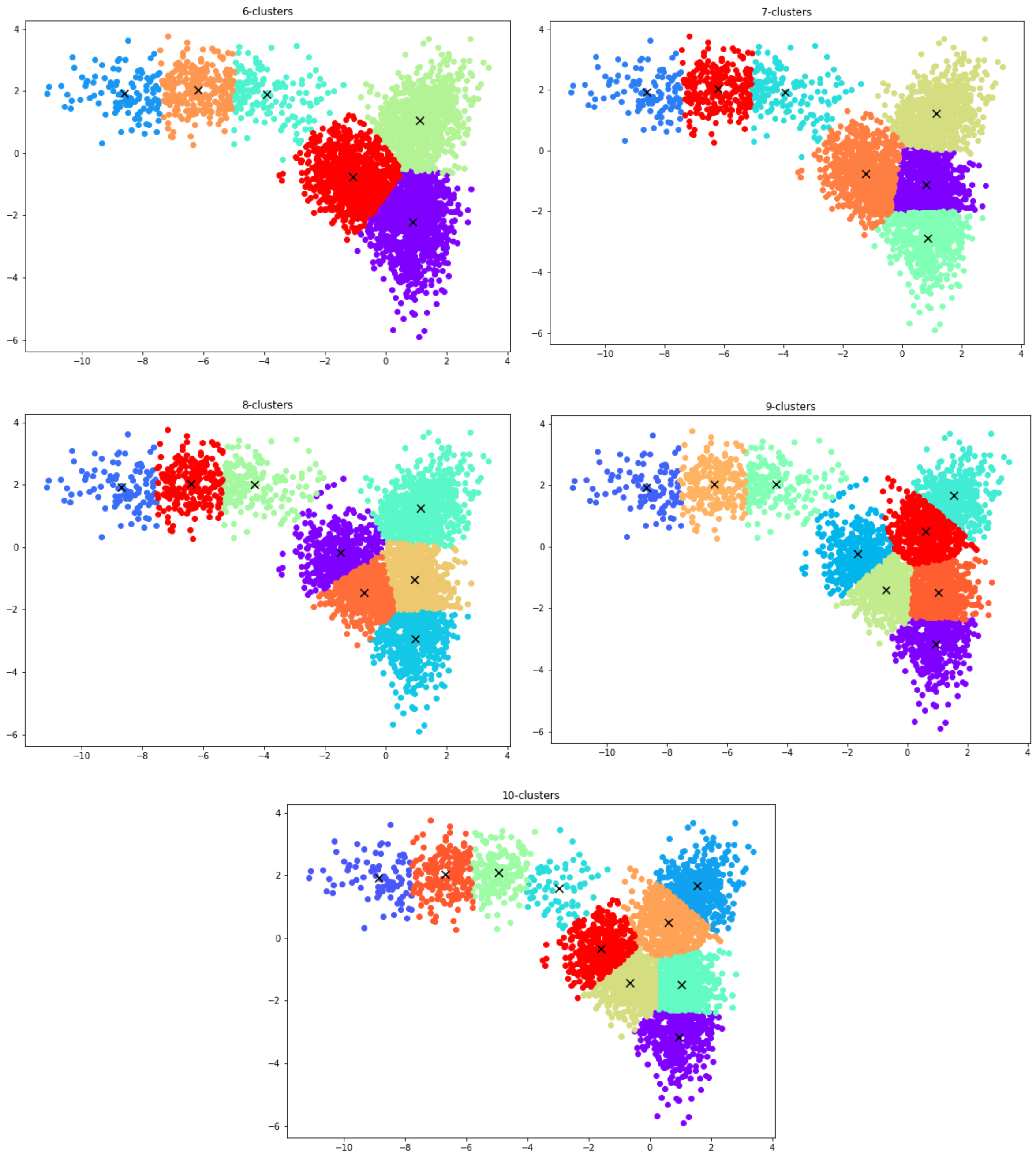


**Figure 9: Dataset 1 is shown with a varied number of clusters. The clusters were created using Lloyd's algorithm and k-means++ initialization.**

In **figure 10** below, the cost of each clustering for dataset one is plotted against the number of clusters. Using this graph, I would select a clustering of 5 as this is the point where the graph reaches a kink in the curve. In other words, this is the point in which I feel like the graph stopped decreasing rapidly. This is different than the results I got when using Lloyd's algorithm and uniform random initialization. Furthermore, the clustering with 5 clusters is also visually different. In the previous clustering of 5, the rightmost cluster had an additional cluster, whereas in **figure 11,** using k-means++, the leftmost cluster has an additional cluster.

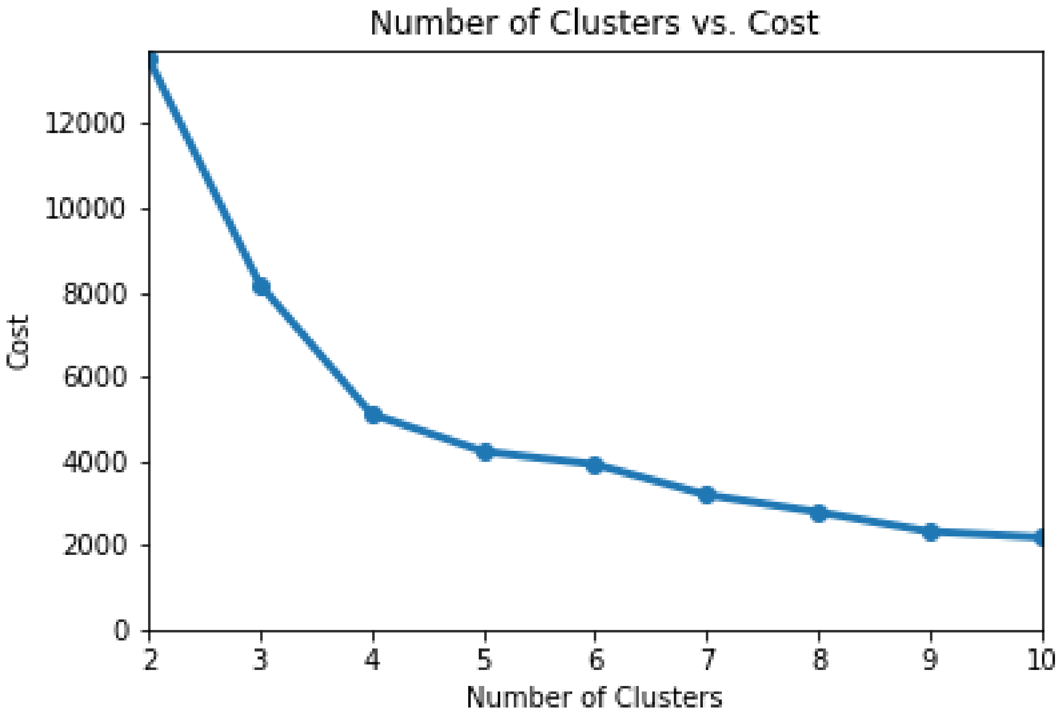


**Figure 10: Dataset 1 plotted with respect to cost vs. clustering.**

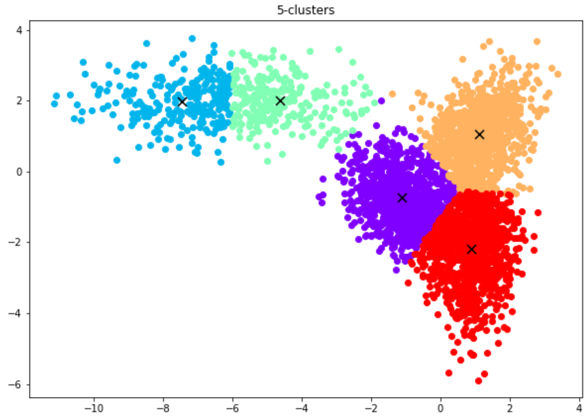
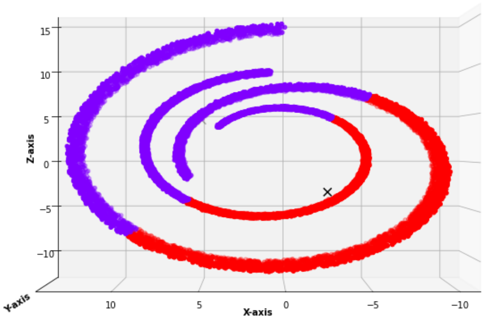


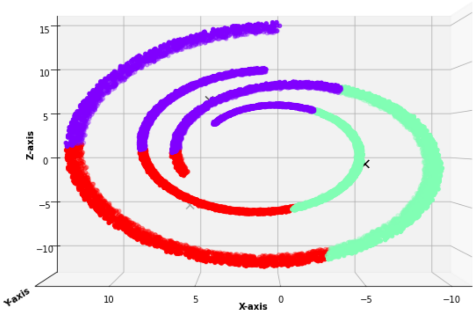
***Figure 11: Chosen clustering for task two and dataset one.***

In **figure 12** below, dataset two is shown with a varied number of clusters beginning from two and ending at ten clusters. To create the clustering, a k-means++ version of Lloyd’s algorithm was used.

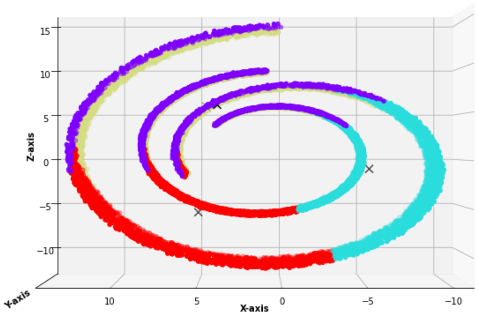
2-clusters



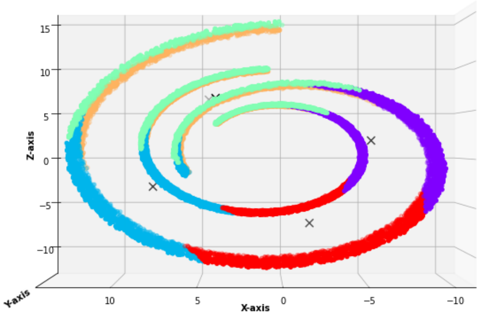
3-clusters



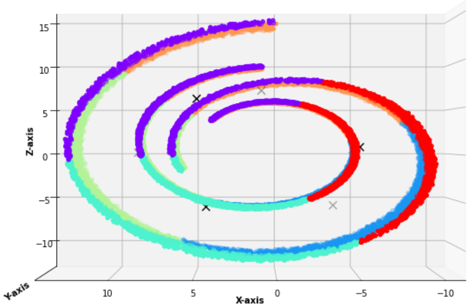
4-clusters



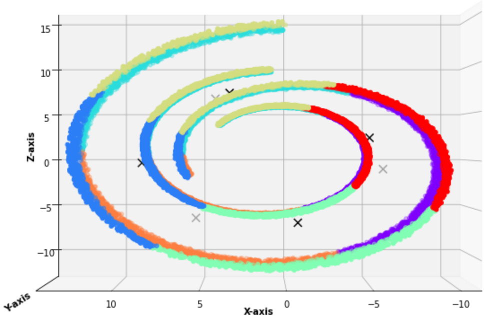
5-clusters



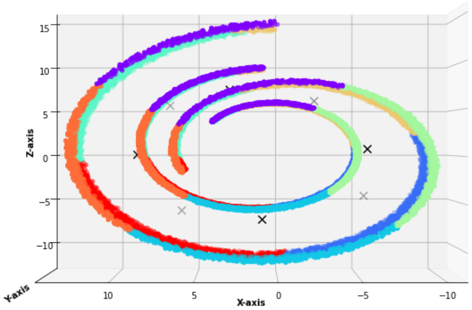
6-clusters



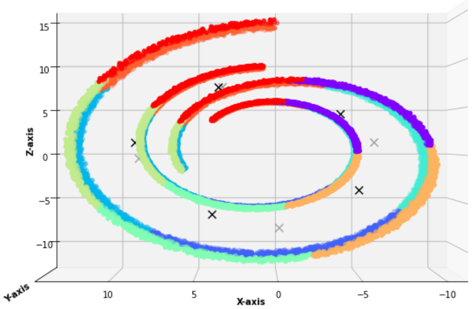
7-clusters



8-clusters



9-clusters



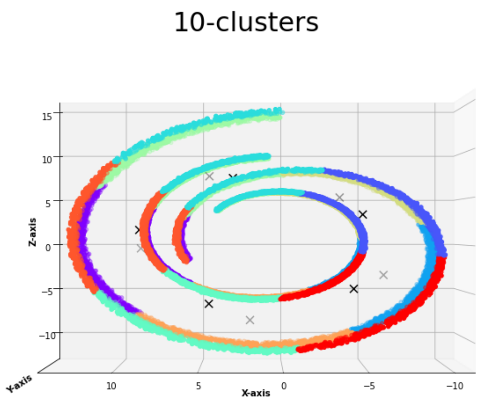


**Figure 12: Dataset 2 is shown with a varied number of clusters. The clusters were created using Lloyd's algorithm and k-means++ initialization.**

In **figure 13** below, the cost of each clustering for dataset two is plotted against the number of clusters. Using this graph, I would select a clustering of 8 as this is the point where the graph reaches a kink in the curve. In other words, this is the point in which I feel like the graph stopped decreasing rapidly. This is the same choice that I chose using Lloyd's with uniform random initialization. When comparing the two figures, they use different colours but have the same grouping. To reiterate from the previous observation, Lloyd's cannot cluster this dataset, which testifies to its inability to cluster non-globular data. The chosen clustering is shown in **figure 14**.

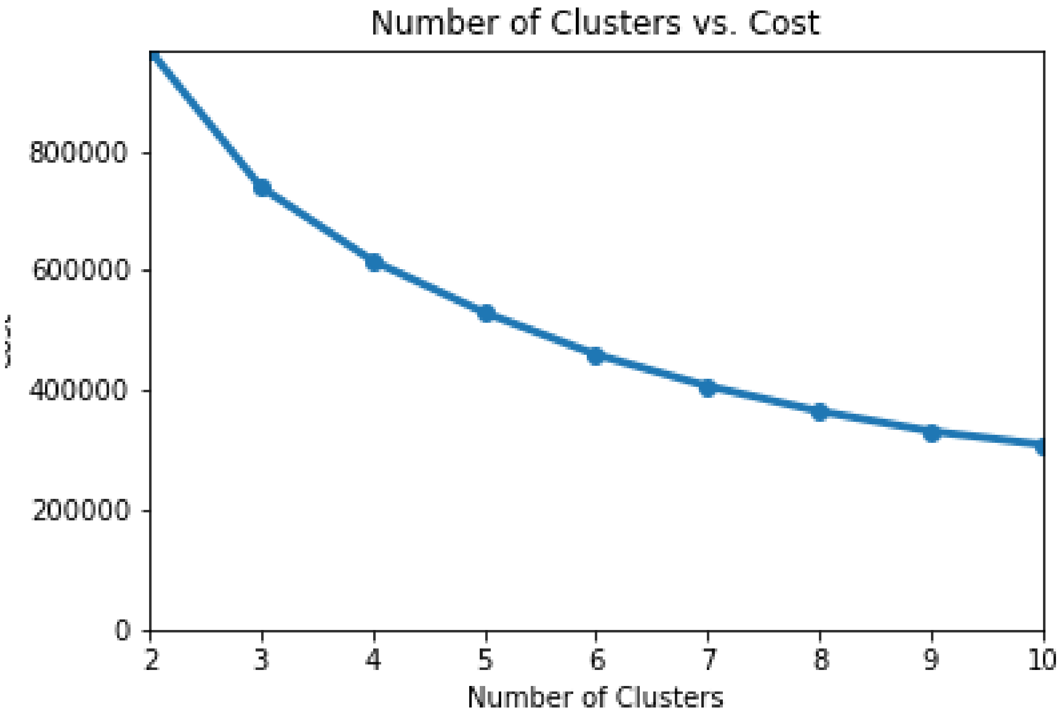


**Figure 13: Dataset 2 plotted with respect to cost vs. clustering.**

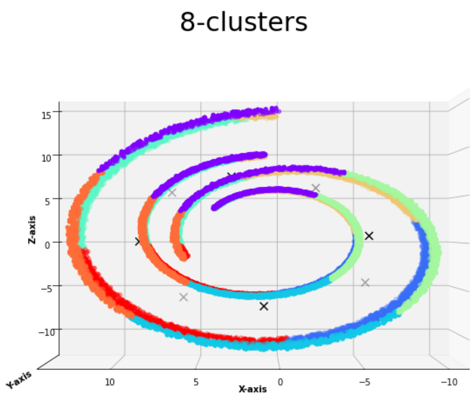


*Figure 14: Chosen clustering for task two and dataset two.*

## 4. Task Three

The dendrogram for dataset 1 using a single linkage is shown below in **figure 15**. The cut is shown in **figure 16**. The cut resulted in eight clusters. I chose to make the cut here for a few reasons. The first reason was, just above the start of the cut on the left-hand side, we can see an increase in the distance or dissimilarity between clusters. This increase in dissimilarity is pointing towards separate clusters being connected. However, as you can see in **figure 17,** five clusters are comprised of single datapoints. This is an inefficient clustering, and when trying other values, I saw this trend of clusters consisting of single points continue. This leads me to believe that a single linkage is ineffective for globular clusters.

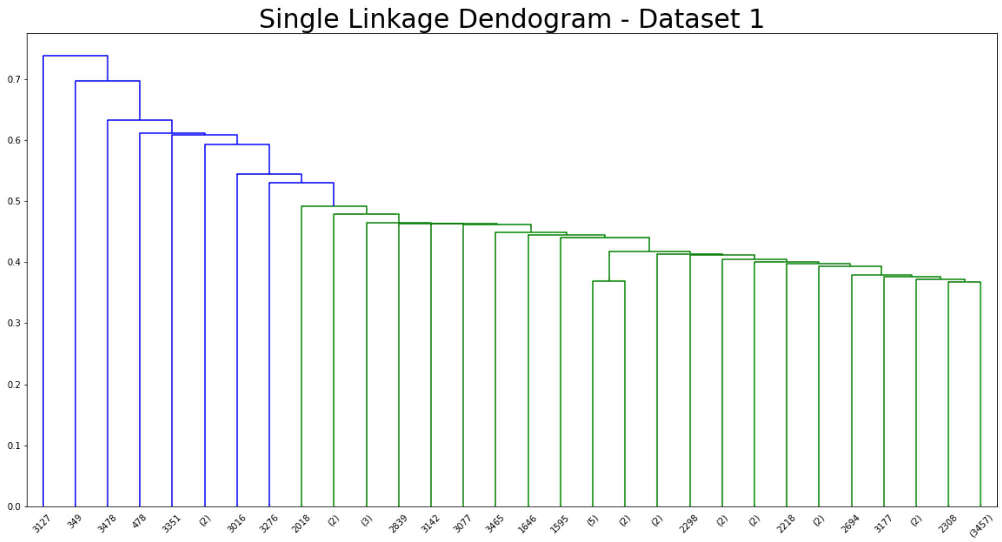


*Figure 15: Dendrogram for HAC using a single linkage for dataset 1.*

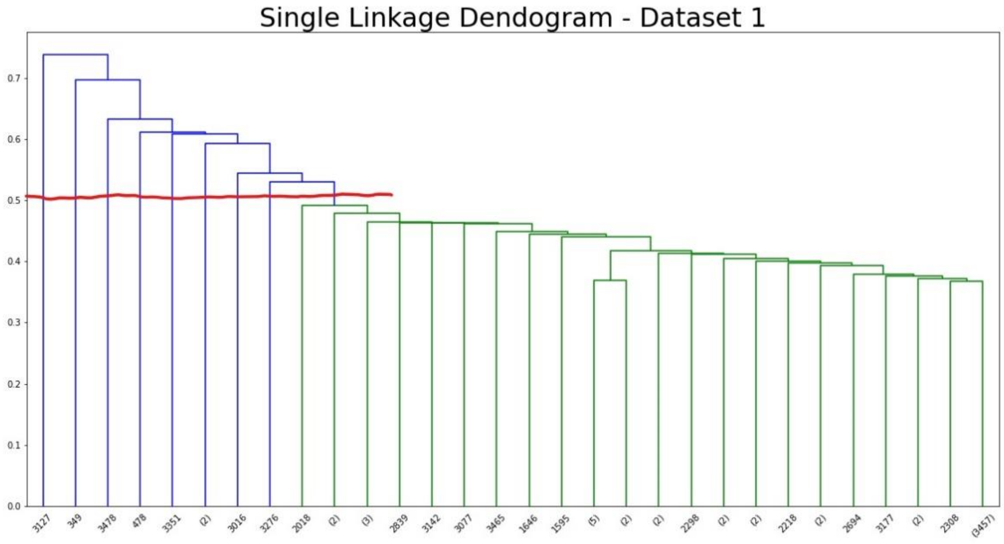


***Figure 16: Dendrogram and the chosen cut for HAC using a single linkage for dataset 1.***

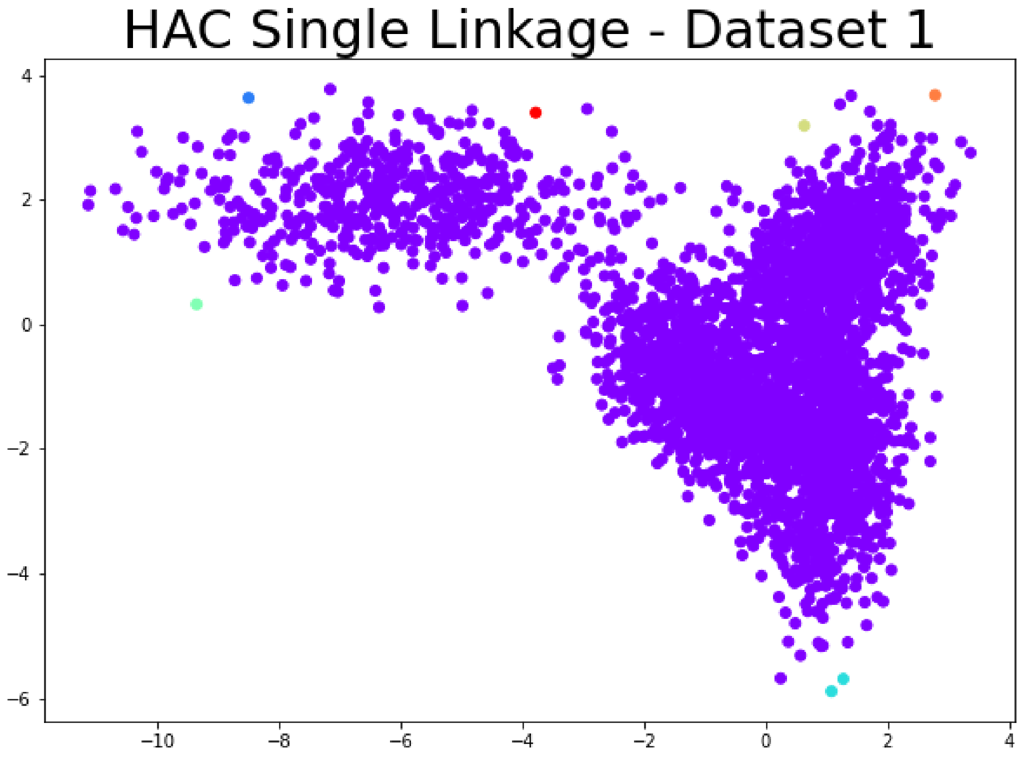


***Figure 17: Resulting clustering for HAC using a single linkage for dataset 1.***

The dendrogram for dataset 2 using a single linkage is shown below in **figure 18.** The cut is shown in **figure 19**. I chose to make the cut here due to the final cluster connection spanning the most distance. This increase in dissimilarity was around two which is relatively high, considering the range of the dissimilarity distance. I also chose this cut through visual analysis. This current cut produced two clusters. As shown in **figure 20**, two clusters capture the clustering logically correct based on the pattern we see. I explored putting the cut lower and found with 5 clusters, 3 of those clusters consisted of a single point. Therefore two clusters were the only clustering with relevant information. This is similar to my previous observation that single linkage is ineffective for globular clusters but can capture the line like structure of functions. This seems intuitive because the points on a line are closely spaced and should be connected to one another.

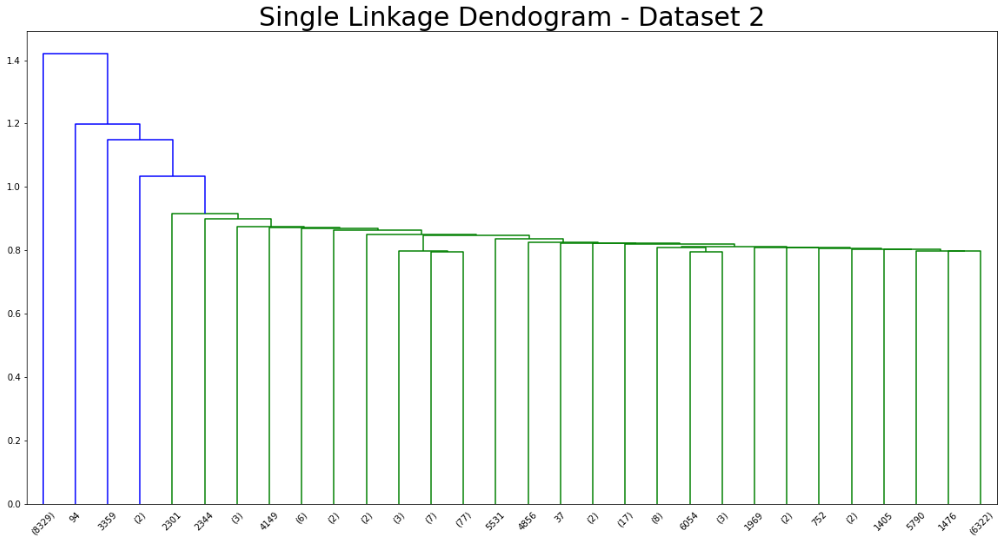


*Figure 18: Dendrogram for HAC using a single linkage for dataset 2.*

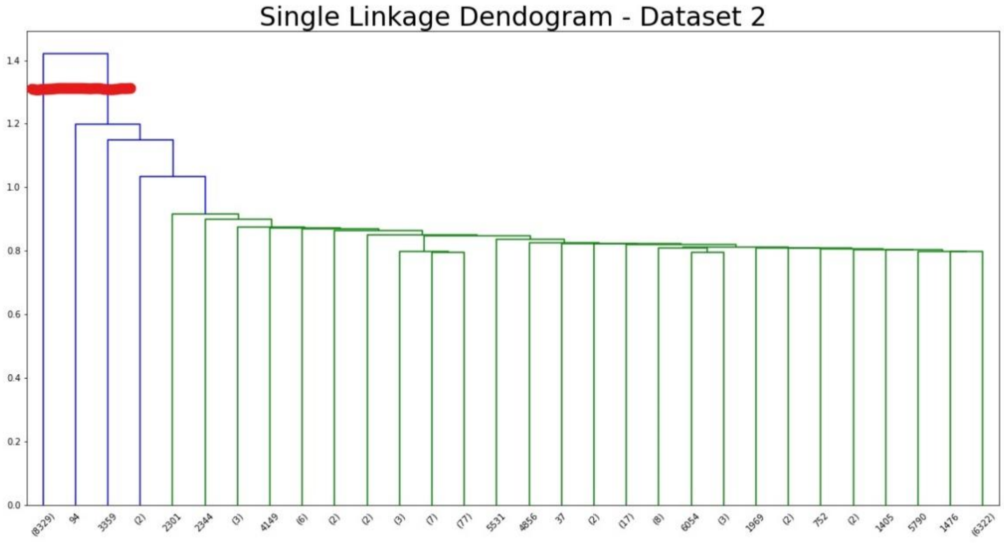


*Figure 19: Dendrogram and the chosen cut for HAC using a single linkage for dataset 2.*

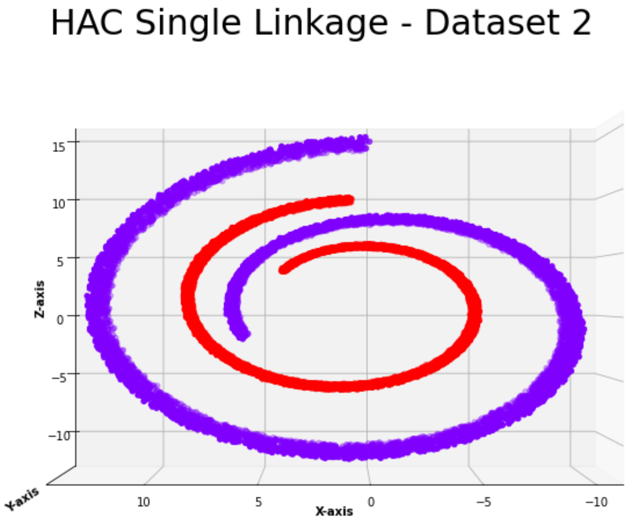


*Figure 20: Resulting clustering for HAC using a single linkage for dataset 2.*

## 5. Task Four

The dendrogram for dataset one using average linkage is shown below in **figure 21**. The cut is shown in **figure 22**. I chose to make the cut here for a few reasons. The first reason was that the dissimilarity measure of four was significant and, by far, stood out as the most dissimilar clustering. This cut created a final clustering of two. The other connections had a rather small dissimilarity value compared to this last connection of clusters. A shown below in **figure 23**, I was able to verify that this clustering worked correctly and captured a logical clustering. The visualization of this clustering seems like an efficient linkage technique and works well globular clusters.

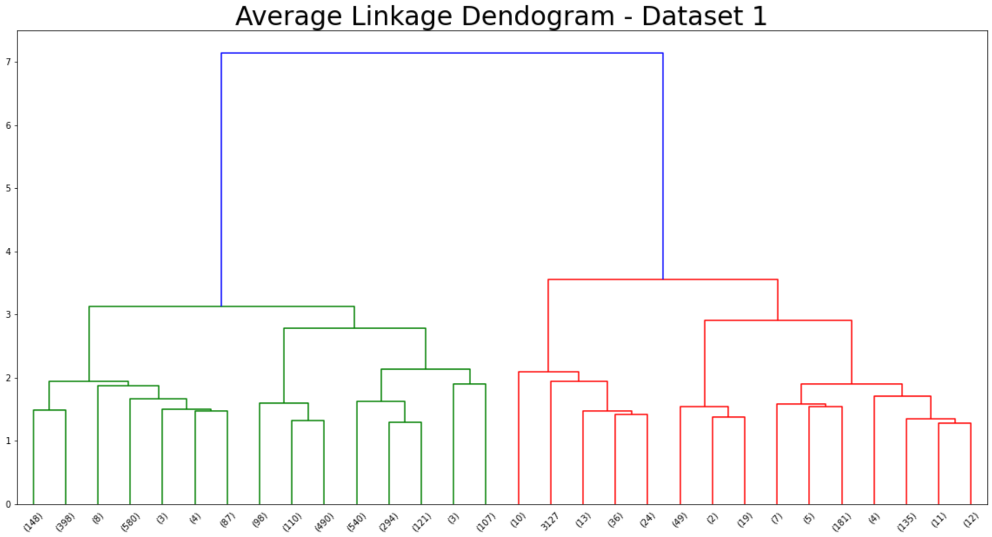


*Figure 21: Dendrogram for HAC using average linkage for dataset 1.*

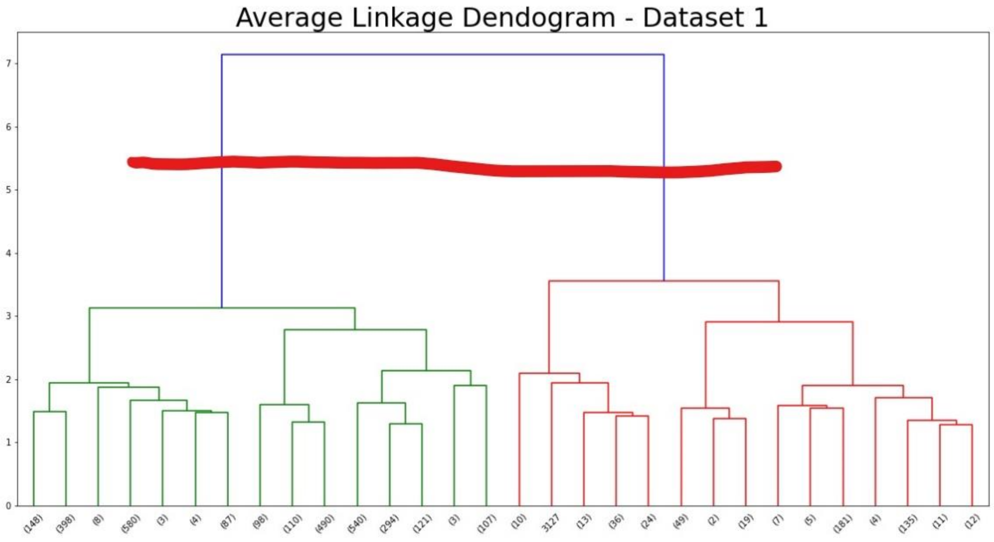


***Figure 22: Dendrogram and the chosen cut for HAC using average linkage for dataset 1.***

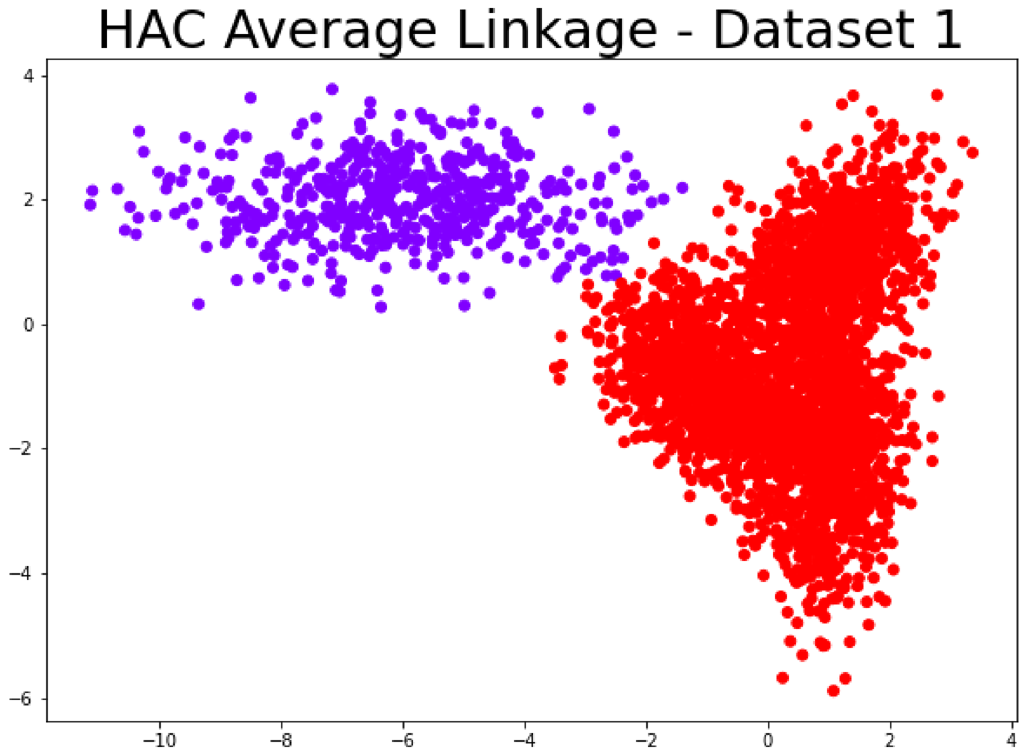


***Figure 23: Resulting clustering for HAC using average linkage for dataset 1.***

The dendrogram for dataset 2 using average linkage is shown below in **figure 24**. The cut is shown in **figure 25**. I chose to make the cut here for a few reasons. The clusters connected below the cut were connected to dissimilar clusters, which were then connected above the cut. This connection of two dissimilar clusters is why I chose to place the cut where I did. Many cuts in this dendrogram were more similar, but the connections above the cut had a relatively high dissimilarity distance. This increase in dissimilarity was relatively high, considering the range of the dissimilarity values. When using visual analysis, I saw decreasing or increasing the numbers of clusters had little effect and failed to capture the perceived logical clustering. As shown in **figure 26,** the cut resulted in eight clusters, which failed to capture the real clustering. This builds on my previous observation that average linkage is ineffective for graph-like data but is able to capture the clustering of globular data.

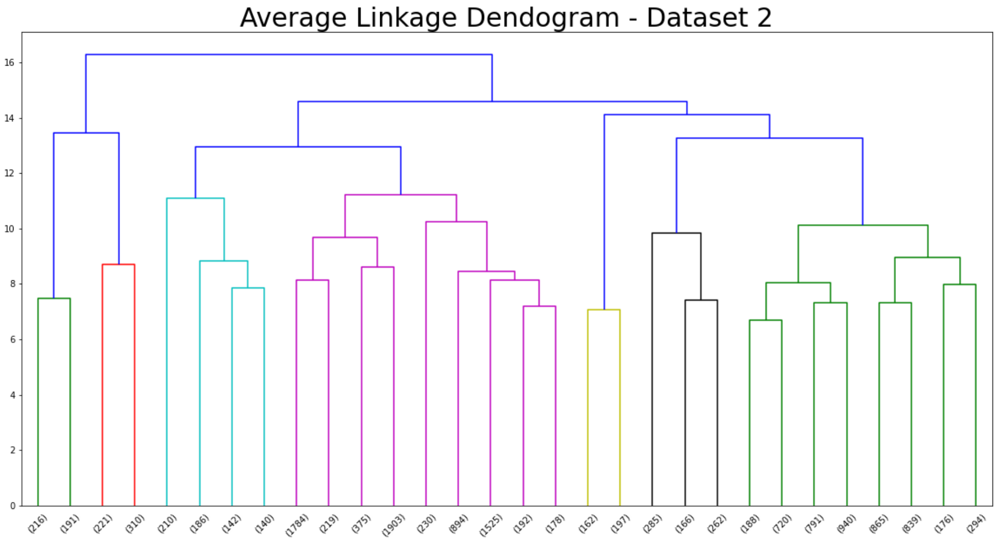


***Figure 24: Dendrogram for HAC using average linkage for dataset 2.***

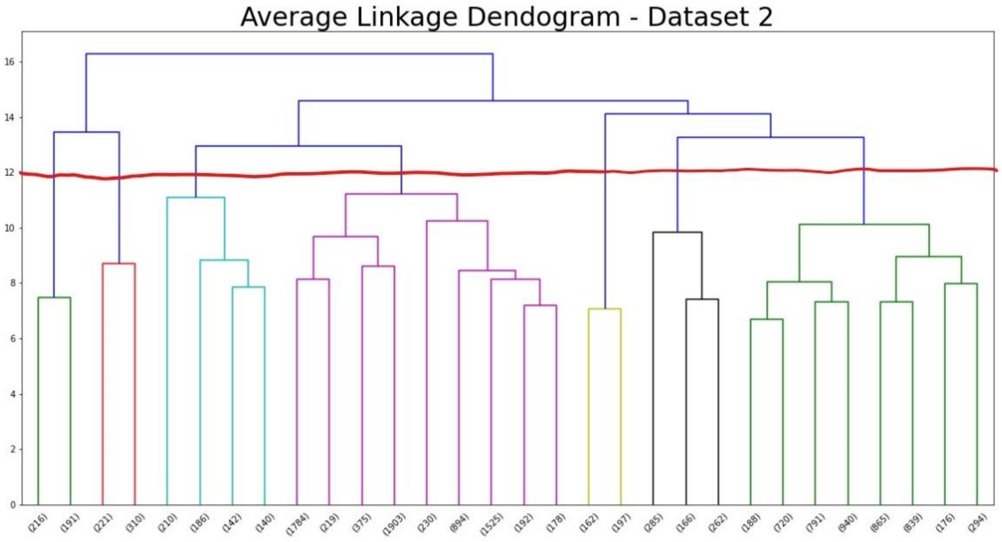


***Figure 25: Dendrogram and the chosen cut for HAC using average linkage for dataset 2.***

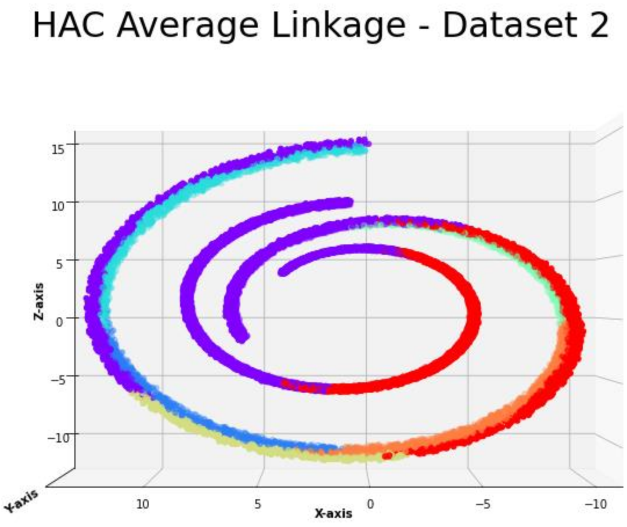


***Figure 26: Resulting clustering for HAC using average linkage for dataset 2.***

## 6. Conclusion

In conclusion, it can be seen that Lloyd’s algorithm was ineffective in clustering dataset 2. This supports our discussion that a disadvantage of Lloyd’s algorithm is the inability to cluster non-globular data. Both initialization techniques used in Lloyd’s produced similar results with a few variations on the location of clusters but overall captures the shape of globular data well. It was observed while running Lloyd’s using k-means++ convergence occurred at a much faster rate. I also found that Hierarchical agglomerative clustering using a single linkage worked best for graph-like clusters. In contrast, average linkage worked best for globular data.