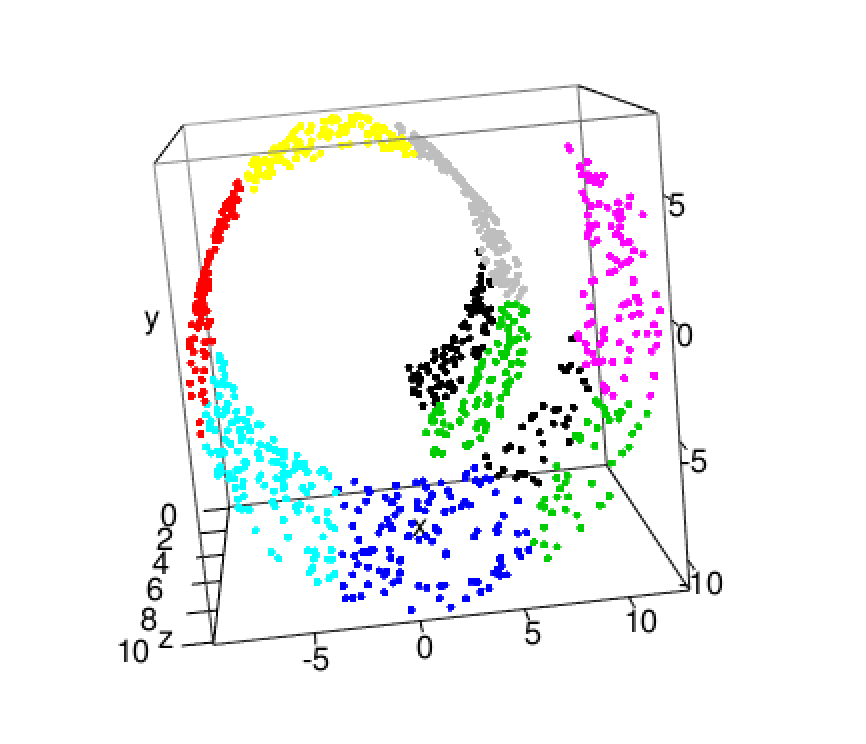
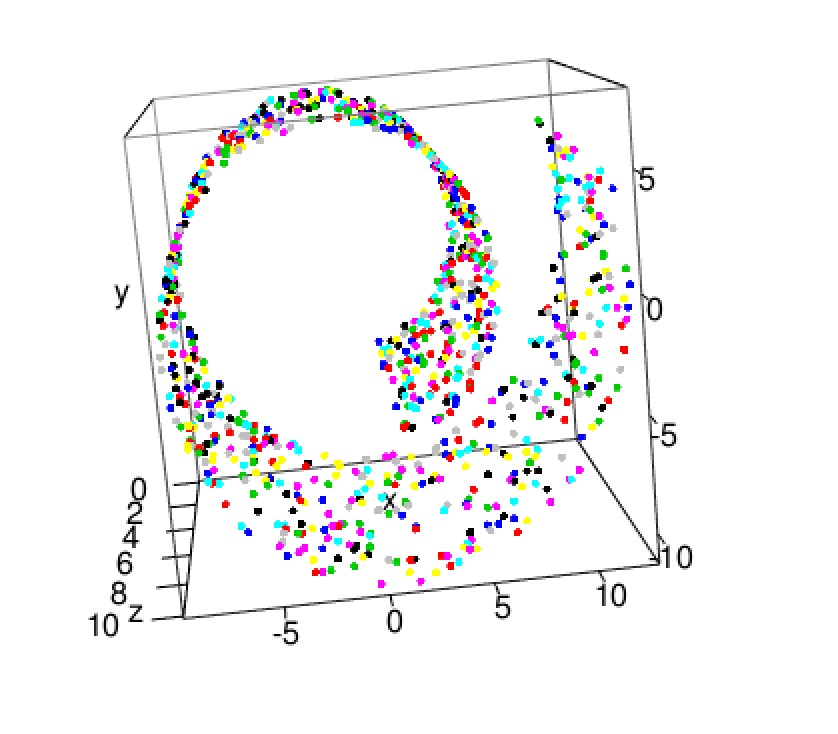
**Part 1:**

Distance based:



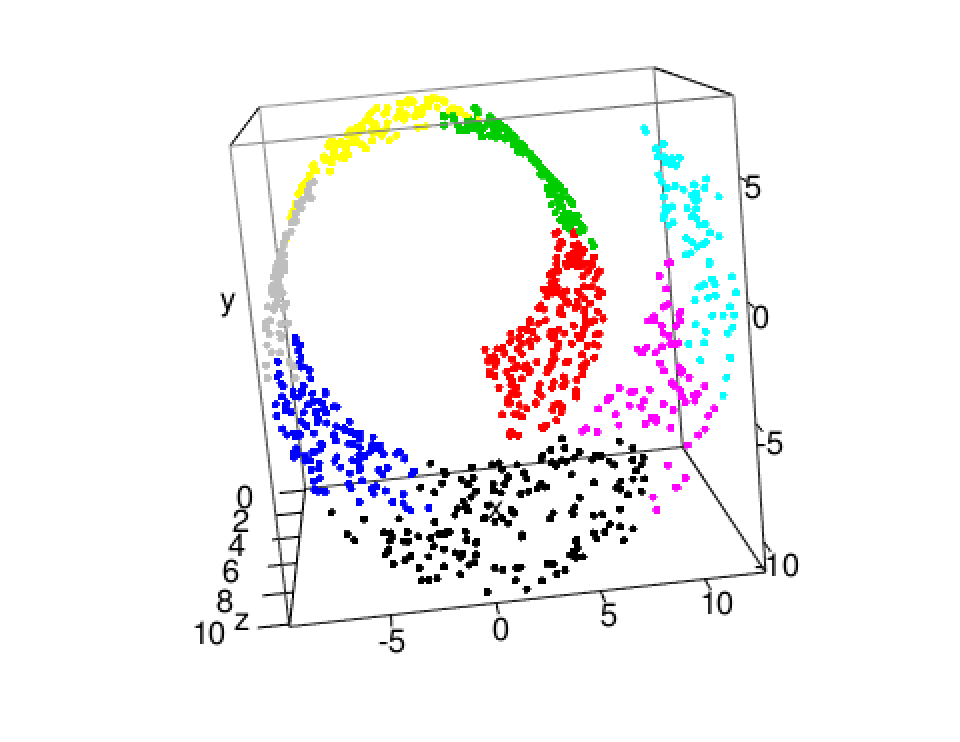
This makes sense because there are clear centers where the clusters are centered around.

Density based:



The density based graph is pretty bad. We expect to see this though because there aren’t any clear outliers in the graph and dbscan clusters based on the density in areas. Since the density is pretty uniform, the clusters are all over and don’t have clear centers.

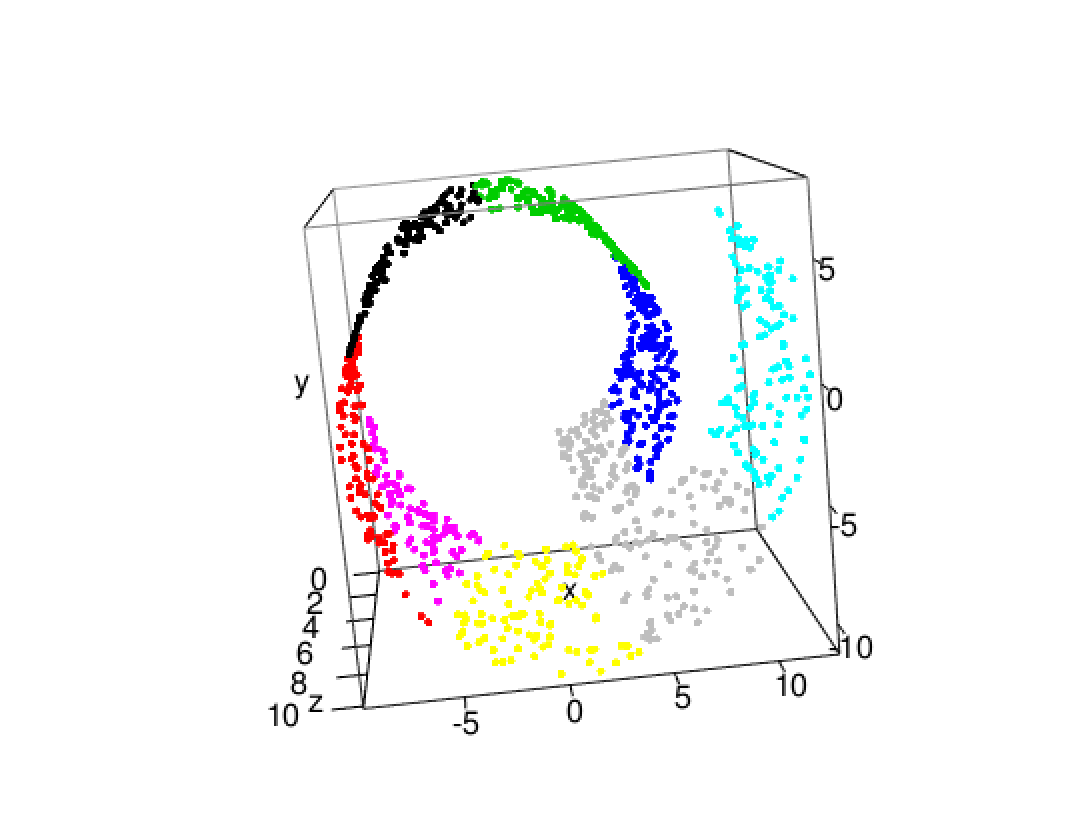
Graph based:



The graph based is what we expect since the algorithm uses connectivity to create the clusters. These clusters are created by connecting clusters on a bottom up scheme.

**Part 2:**

This method improved the outcome because x is weighted more than y which is weighted more than z. The z axis is basically looked at the least since that’s where there is the least variation, and it’s least important. The x axis has the most variation so it is looked to the most by weighting it more. This is why this method works better.



**Part 3:**

I used a fuzzy c mean. Fuzzy c mean is good for large datasets because the time complexity is low. In determining how many clusters there should be, I used the elbow test. The elbow test determines when the biggest jump in predicted variance occurs. The number at which the biggest jump occurs is then the optimal number of clusters.

Before I did anything on the real set I did a few tests with a sample set of the data.

I would have used k-means for this since the dataset is large. However, since this can’t be plotted there’s no way to know if the clusters are globular.

