IOT Project 4: Clustering Techniques

Ojas Barve - ovbarve@ncsu.edu

Task 1: Hierarchical Clustering

This task involved the clustering of data using the hierarchical clustering approach. This type of clustering comes under the unsupervised category of machine learning algorithms. Here we cluster data points based on a distance metric. We start of by considering that each data point is a cluster in itself. Then we use the bottom up approach to cluster these data points based on some linkage criterion.

For the project I experimented with different linkage criterion like minimum distance, maximum distance and weighted average distance. All the distances were calculated using standard euclidean distance metric. One could also use other popular techniques such as minkowski distance(p-norm), manhattan distance or the Mahalanobis distance. For this project I chose WARD linkage clustering which uses Ward variance minimization algorithm.

Figure 1. Below shows visualisation of dendrogram using the above linkage parameters

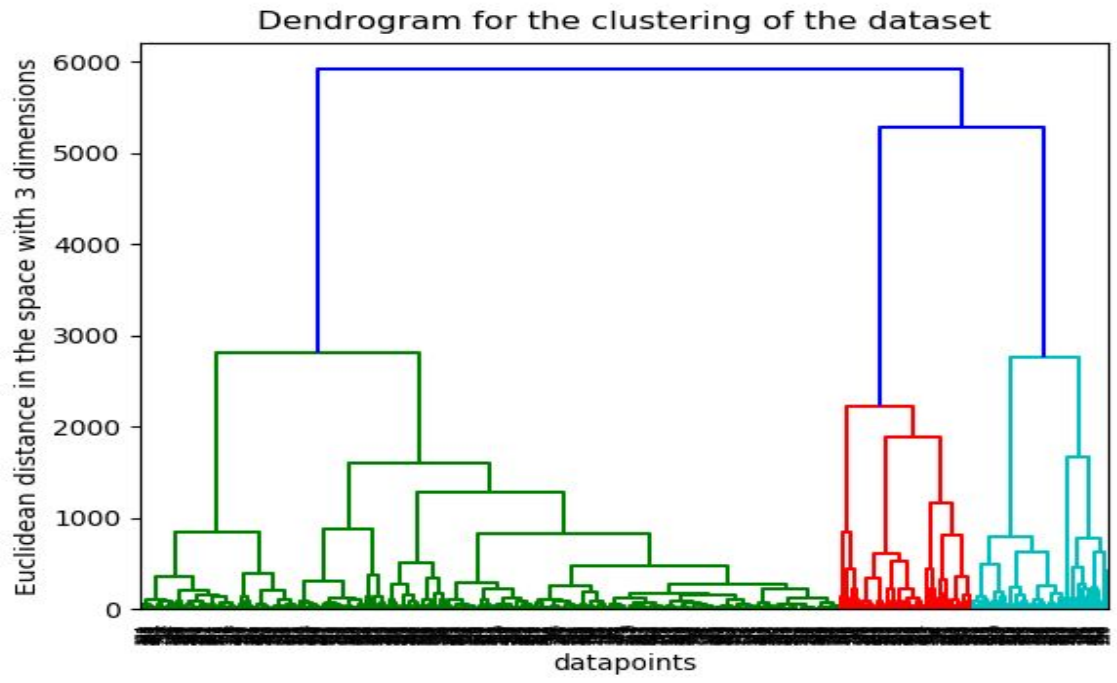


Figure 1. Dendrogram based on ward linkage.

Based on the Dendrogram we clearly observe that the maximum distance between cluster splits is obtained when the number of clusters is 3. The most stable clusters will be obtained for number of cluster equalling 3 for the linkage parameters used and the dataset given.

From hierarchical clustering approach I conclude that the dataset consists of 3 clusters. Below is a 3-Dimensional representation of the clusters.

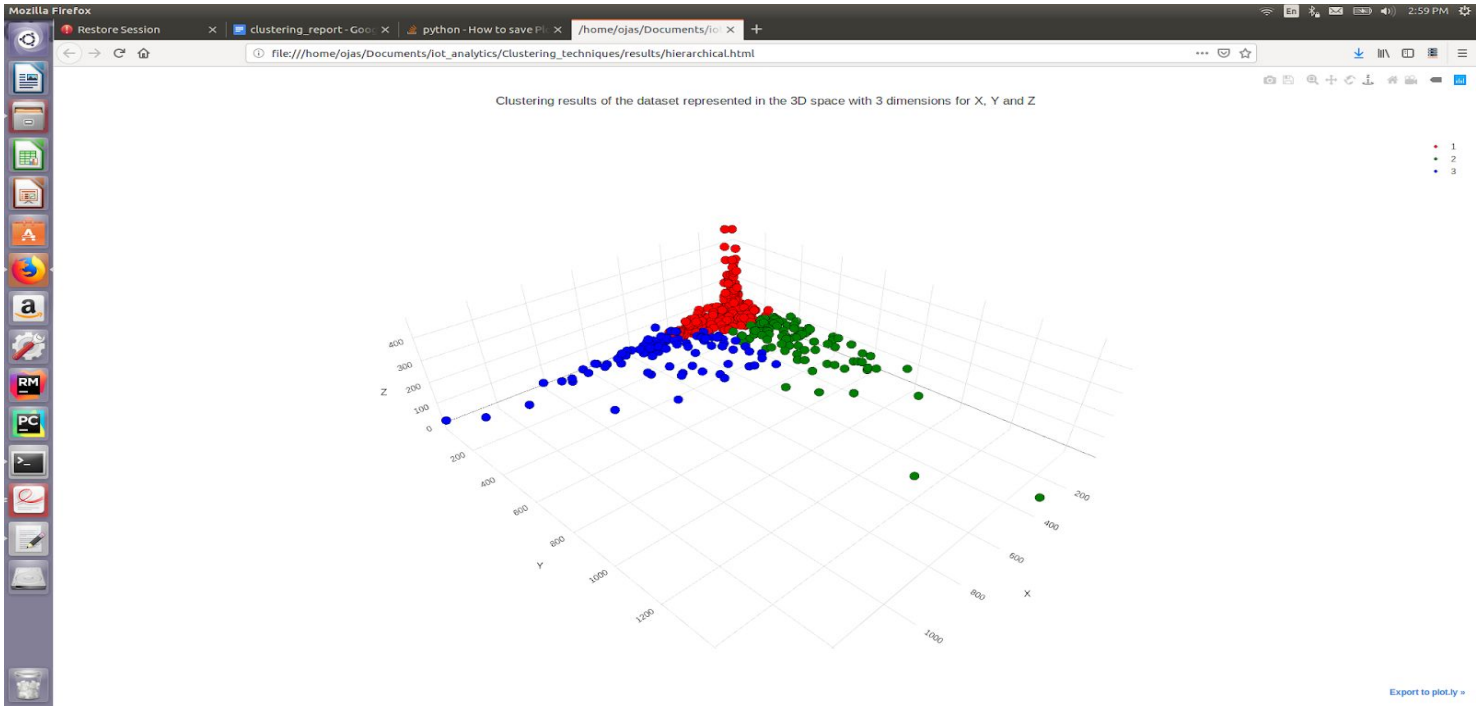


Figure 2. 3D Hierarchical clustering with number of clusters equal to 3

Task 2: K- Means Clustering

In this case we are interested in clustering the data using K means clustering approach. In this technique we decide the number of clusters before hand and then cluster data based on some distance criterion. Here we use the centroid method to carry out the k means clustering approach. Also we calculate the value of K using the elbow method. The elbow method is basically the plot of Sum of squared errors (SSE) v/s the value K.

Below is our elbow graph

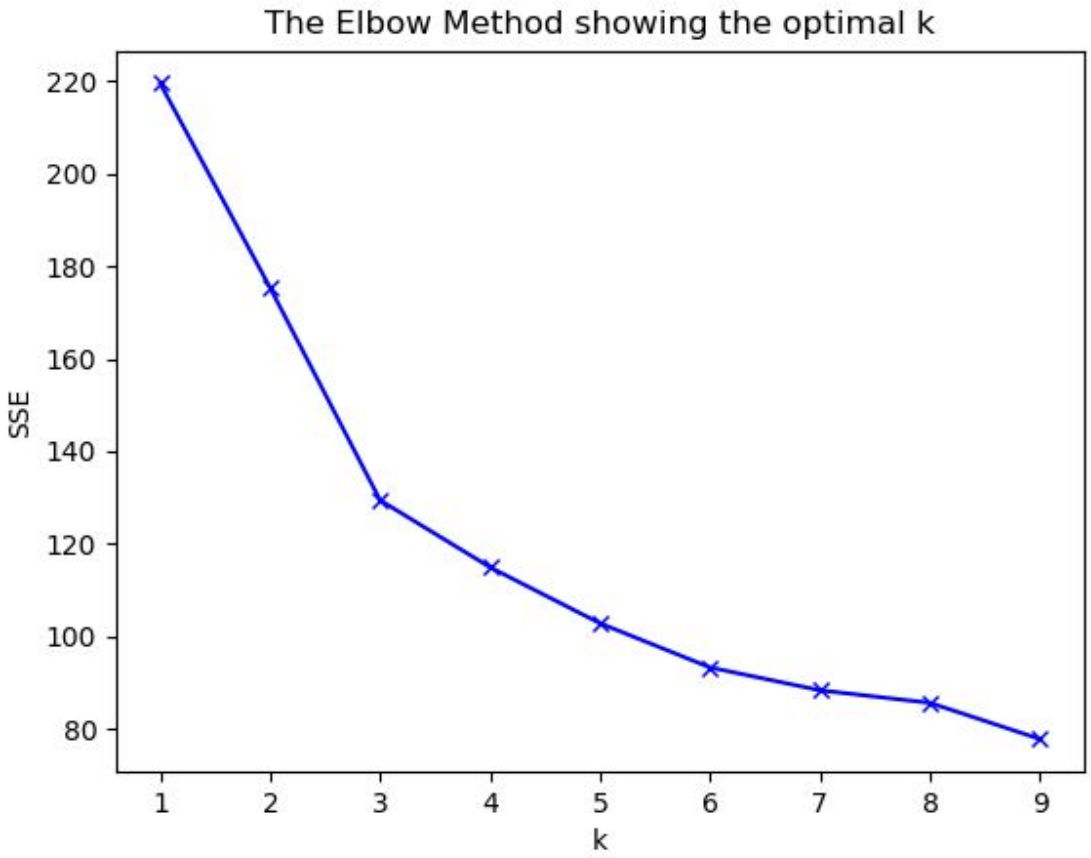


Figure 3: SSE v/s number of clusters K (Elbow Graph)

We see from the plot that the minimum value of RMSE occurs at a window size of 3.

Hence we use number of clusters equal to 3 for clustering our data. Furthermore we can also obtain the silhouette scores for each data point by varying K over a range of values and try and obtain the best value of k. This method of silhouette scores is mostly used when we don't have a well defined elbow. But in my case the elbow is clearly defined so we choose the value of k from the elbow itself.

Below is a 3 Dimensional view of the clusters formed using the k-means approach

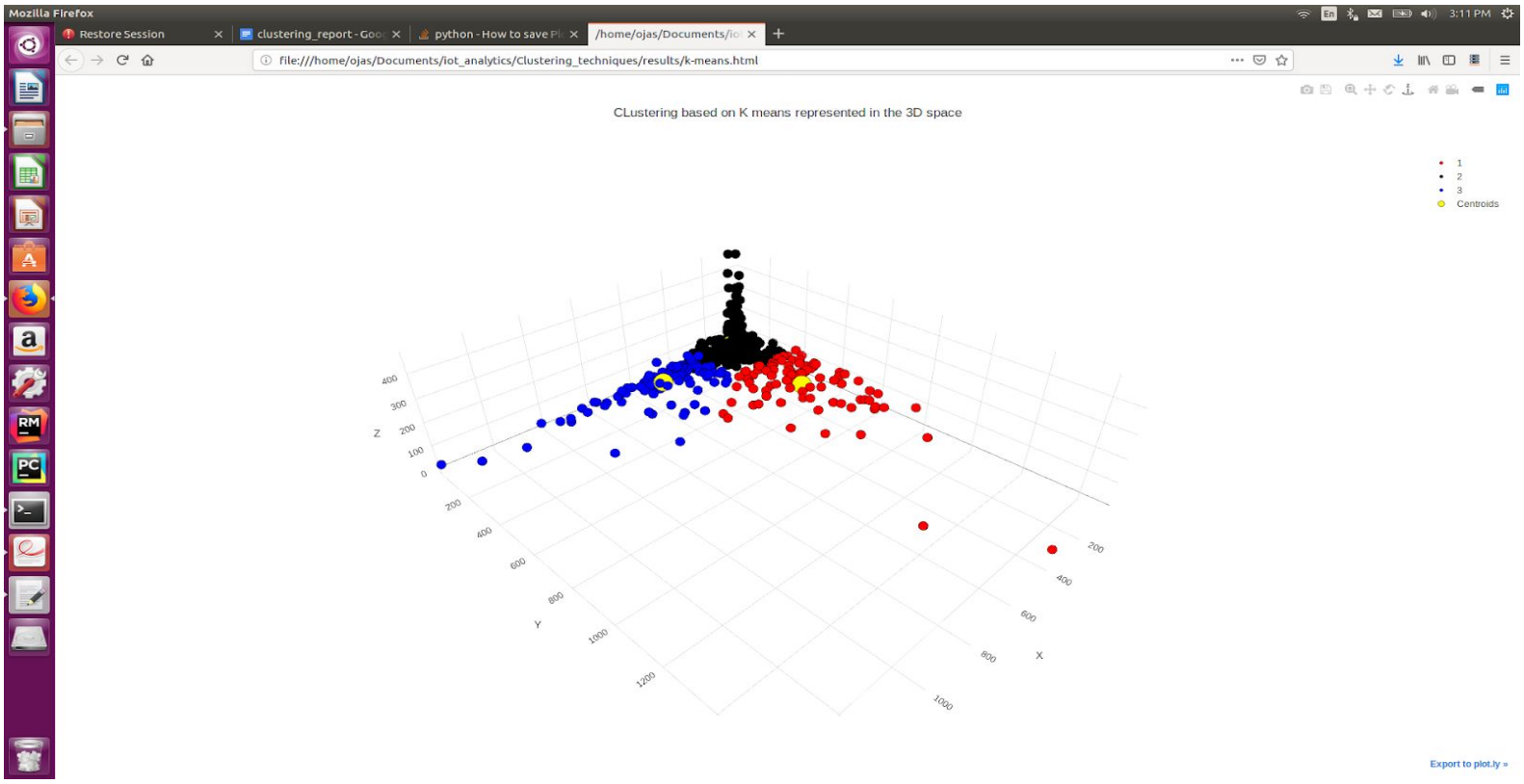


Figure 4: K-means clustering with 3 clusters.

Task 3: Density based Spatial Clustering (DBSCAN):

In this case we are interested in using the dbscan technique to cluster the provided data. The DBSCAN algorithm is useful in the case where there is noise between the clusters and normal clustering techniques like hierarchical, k means fail to cluster the data efficiently. Due to noise merging of clusters is observed using k means and hierarchical clustering techniques. This is overcome using DBSCAN using density based approach. We define points to be core points if and only if they have a certain minimum points within eps radius.

The value of Min Points is usually taken as the dimension of data plus one. The value of eps is obtained by elbow method. We plot the minimum value of eps for each data point to be a core point and reverse sort it. The elbow thus obtained is taken as the eps value.

For our project I the Minpoints is varied over the range [3, 4, 5, 6]

Below is shown one of the elbow graphs thus obtained.

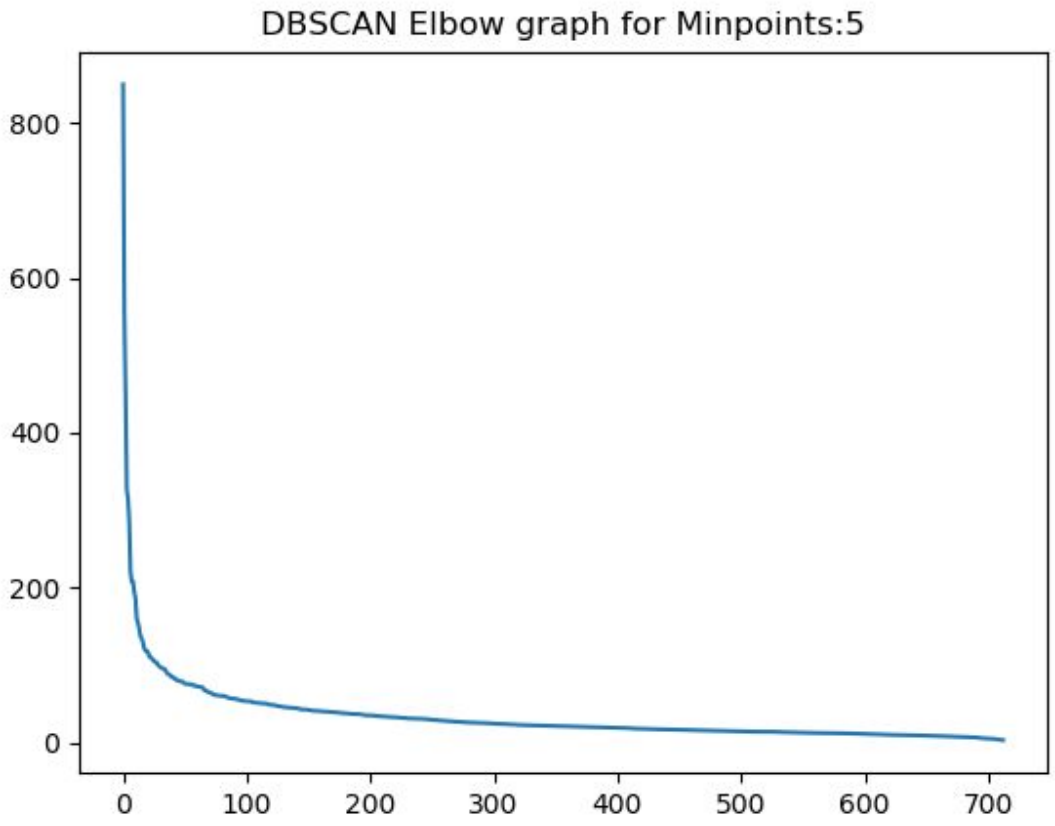


Figure 5: Minpoints equal to 5.

We approximate the elbow point at a value of 18.

Below we show the cluster obtained from taking the eps value as 22 and min points 3.

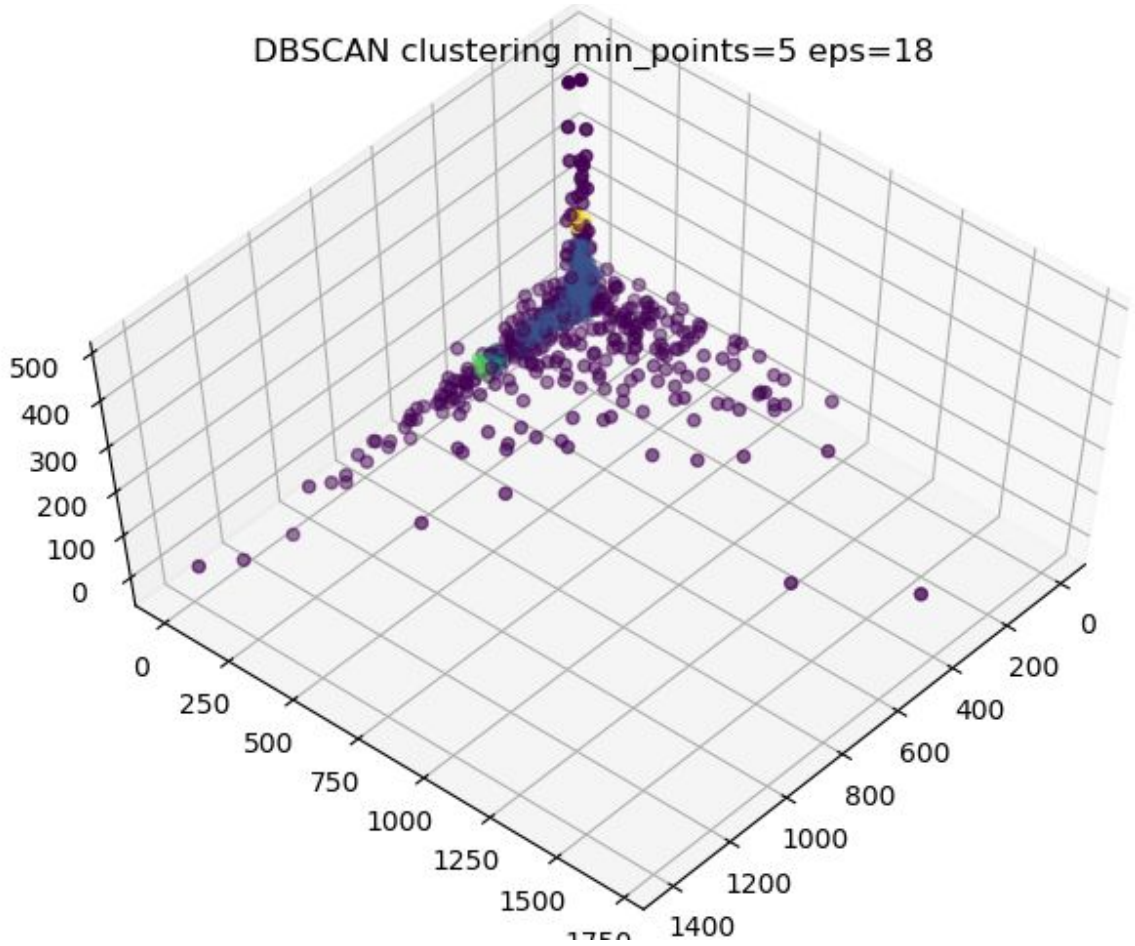


Figure 6: DBSCAN clustering eps=18, min points =5

For the above values of Min points and eps I obtain a total of 3 clusters for the given data. We see that most of the data is classified as noise which is depicted by violet color.

From the above observation I conclude that the DBSCAN algorithm is not suited for the dataset provided to me. The dbscan algorithm is better suited for the data where the density of the clusters and the density of the noise is different. The noise being more sparse than the actual data points. But in our case the data has a non uniform density throughout making it very difficult to cluster using the DBSCAN algorithm.

Results for other values of minpoints and eps values can be found in results folder.

Task 4: Gaussian Mixture Models

In this case we are interested in using Gaussian Mixture Models to cluster our data. In this approach we use the Expectation Maximisation algorithm for the optimization of the parameters of the k gaussian distributions. The value of optimal k can be found by plotting the maximum likelihood against k (number of clusters). This can be achieved by using the Akaike Inference Criterion or the Bayesian Inference criterion.

Below we plot the AIC and BIC v/s the k clusters.

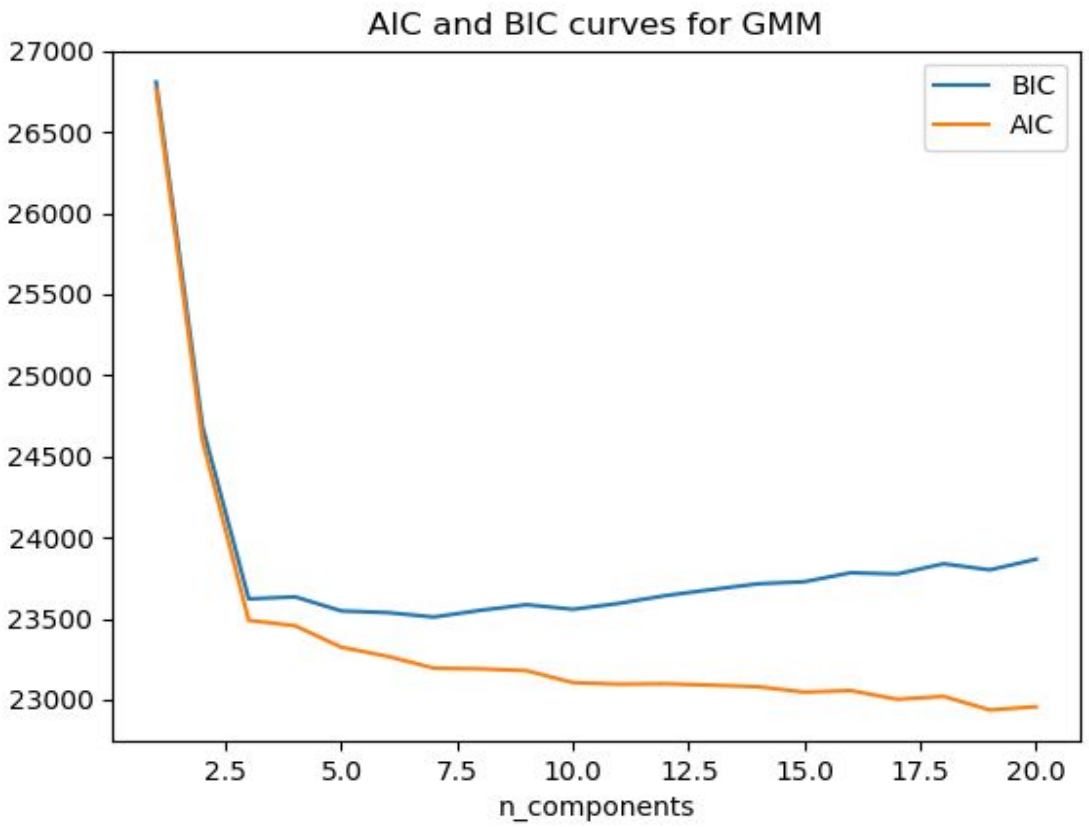
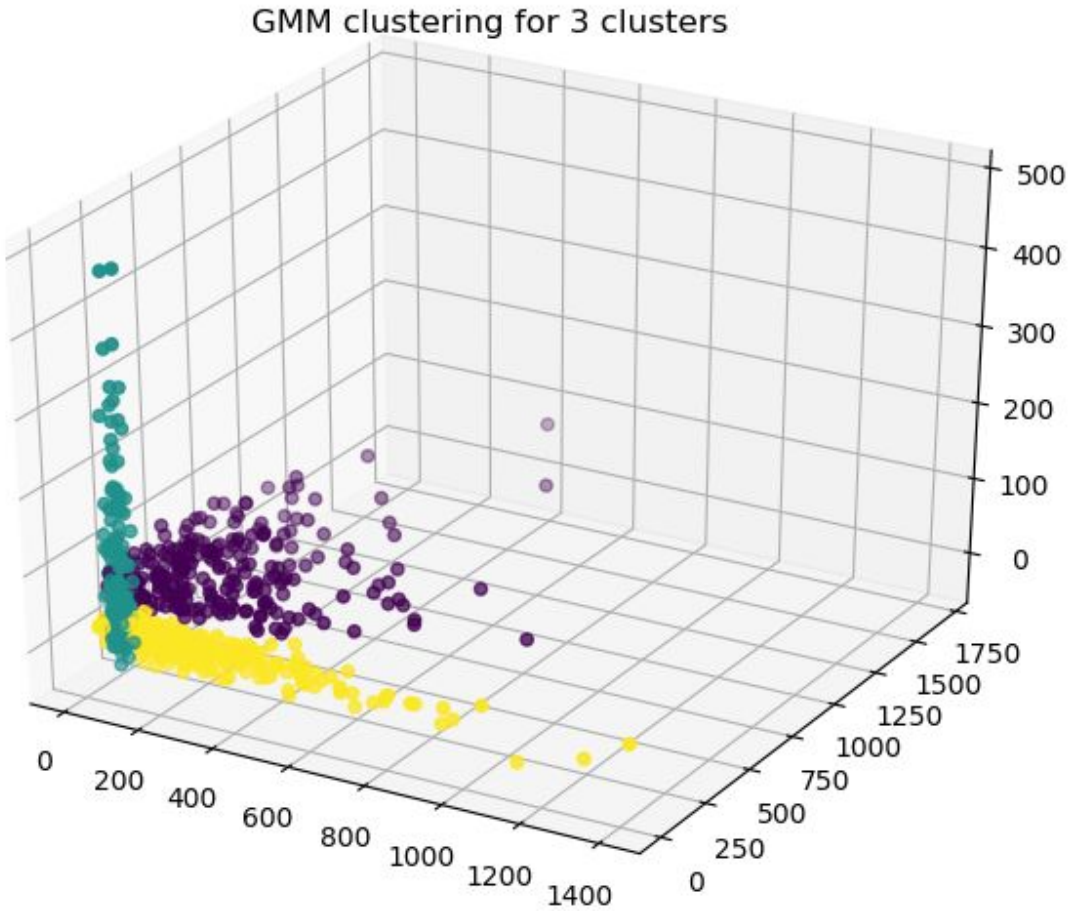
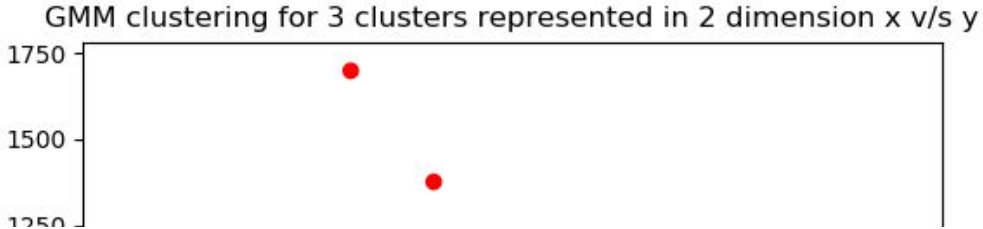


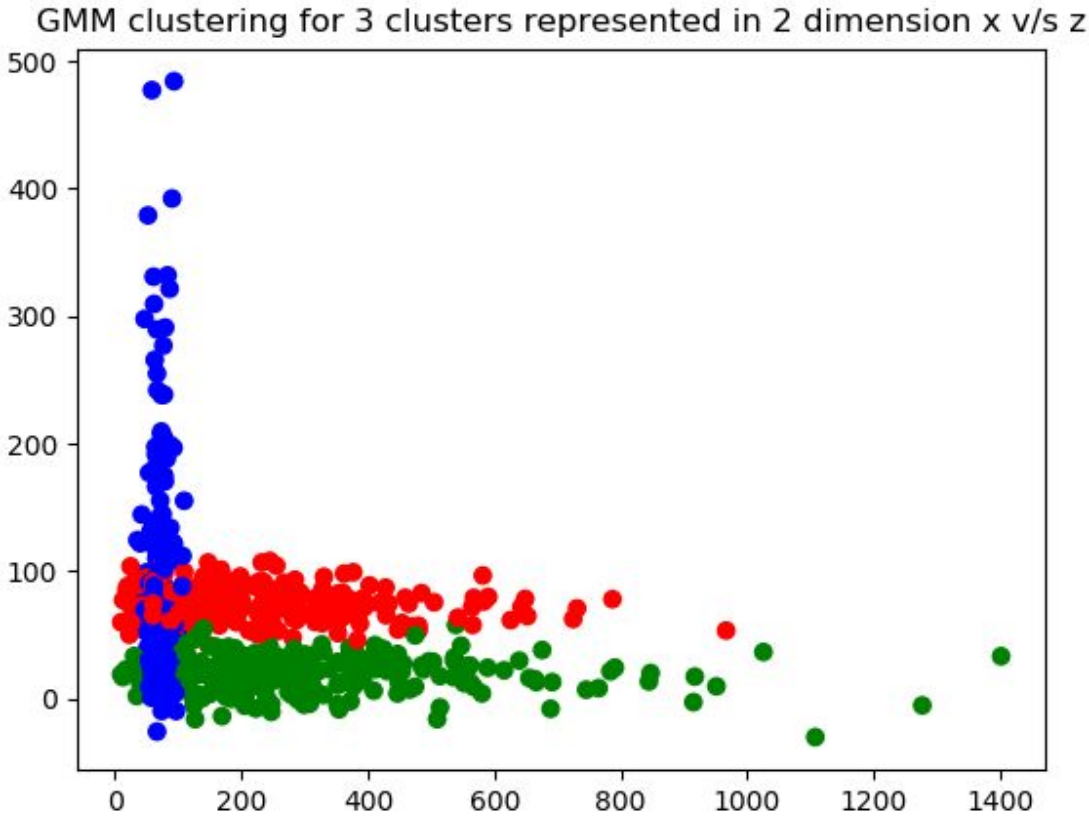
Figure 7: AIC and BIC for K -clusters.

From the above graph we obtain the value of k for our Gaussian mixture model as 3. I used this value of K and obtained the 3d scatter plot of the clustered data.

Below you can see the 3d scatter plot as well as individual 2d projects of these gaussian clusters on 2 dimensional axes.






GMM clustering for 3 clusters represented in 2 dimension x v/s z
500
400
300
200
100
0
0
200
400
600
800
1000
1200
1400

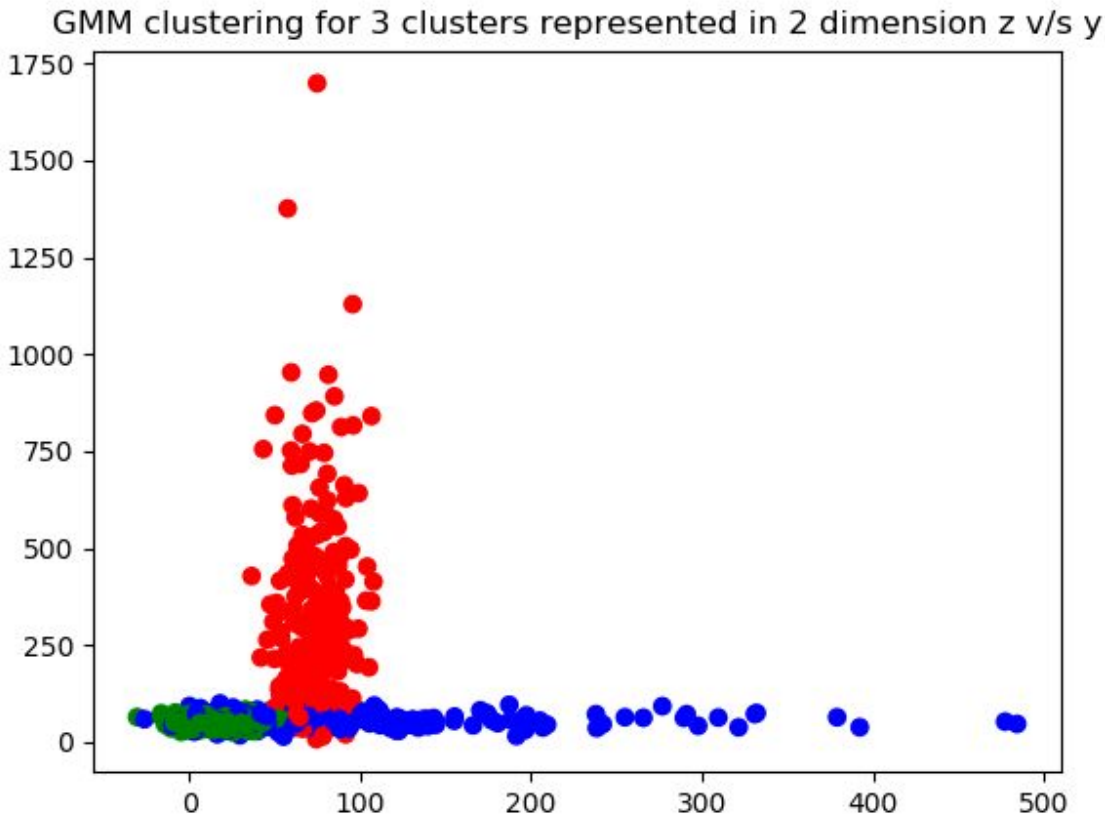


Figure 8. Gaussian Mixture Model K=3

We can see from the Figure 8 above GMM fits the data quite well. The gaussian mixture model performs well with respect to other clustering techniques.

Task 5 Final Comment - Comparison of All Techniques:

Based on all the techniques that I implemented for the given dataset I conclude that the number of clusters in the dataset is 3 as we get similar results from hierarchical, K means and GMM clustering approaches. When we plot the 3d scatter plot of the data we observe that the data has an uneven density throughout the data is more dense towards a center a is spread out in three directions. This helps us in understanding why the DBSCAN algorithm fails on the provided data. The dbscan algorithm will only work when the clusters have even density and the noise in between is sparse.

Coming to the other algorithms I found that the hierarchical technique works well when

I use the WARD linkage criterion. For other kinds of linkage the dendrogram is a little difficult to interpret. When we cluster using the agglomerative technoque we obtain 3 clear clusters.

For the K-means approach we obtain a clear elbow at k = 3 so we perform the clustering based on this value. The obtained cluster scatter plot only helps us in confirming our assumption that the number of cluster k obtained from hierarchical is correct.

Finally we use the Gaussian mixture models which confirms our assumptions further about the number of clusters being 3.