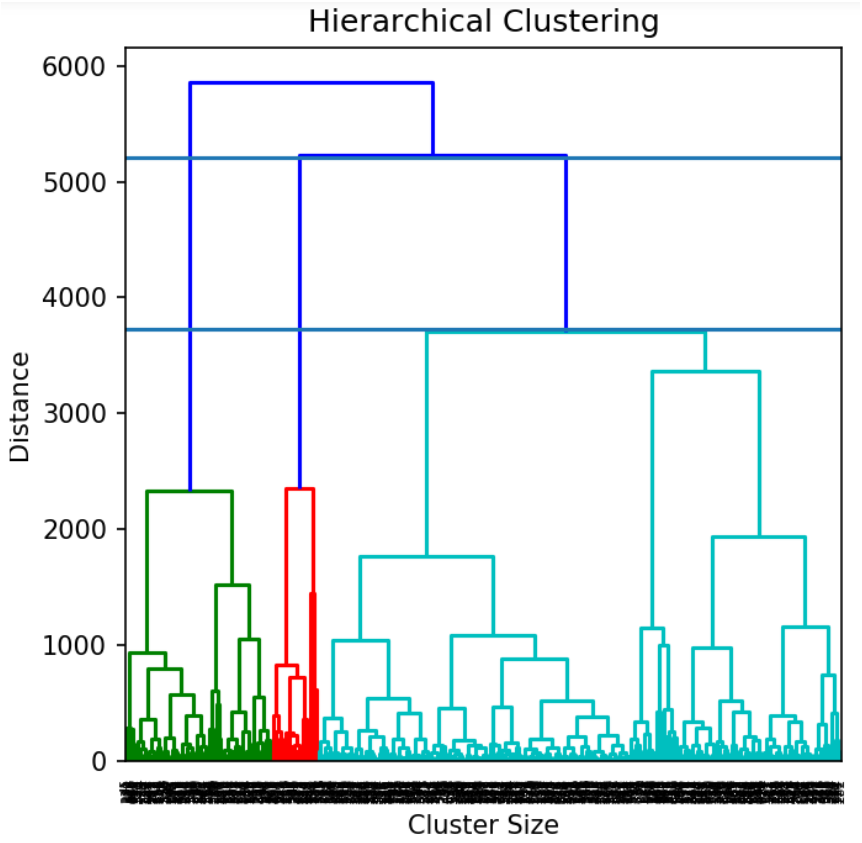
Clustering Assignment

 # Task 1

1.1) Dendrogram Plot

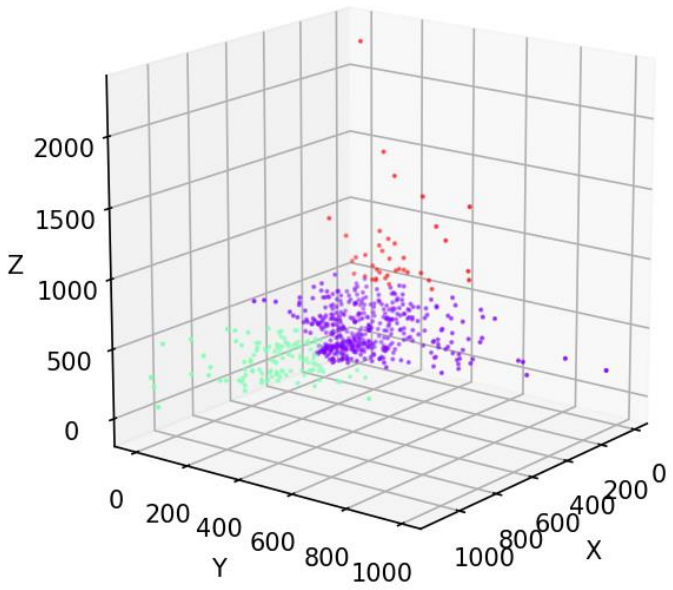


1.2) Determining the number of clusters from the Dendrogram

The 2 horizontal lines in the graph are the lines of maximum separation indicating the difference between the δ's is the maximum. Hence, if we put a horizontal line between these 2 horizontal lines, then the number of clusters is the number of vertical lines that this horizontal line intersects. In this case, it is 3 clusters.

1.3) 3D scatter diagram

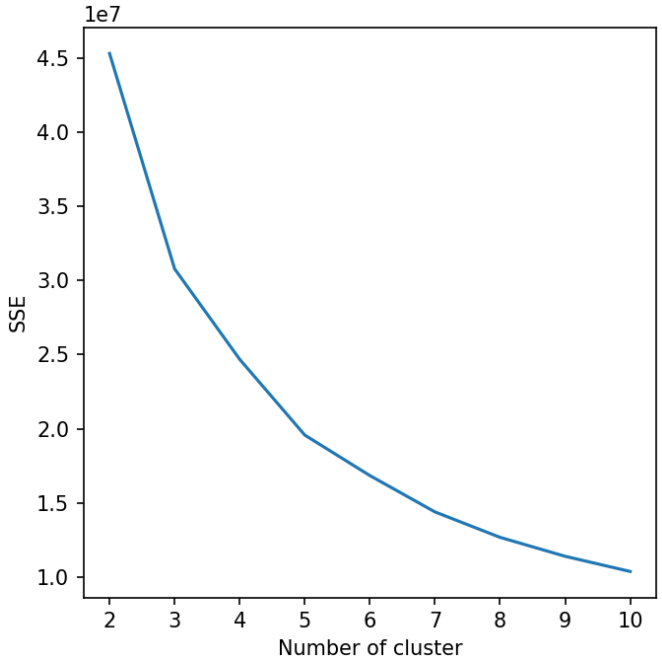
Below is the scatter plot for 3 clusters as obtained using the dendrogram. The dots of one color that are lighter than the other indicate that they are farther away from the other points of the same color from the point of user's view.



 # Task 2

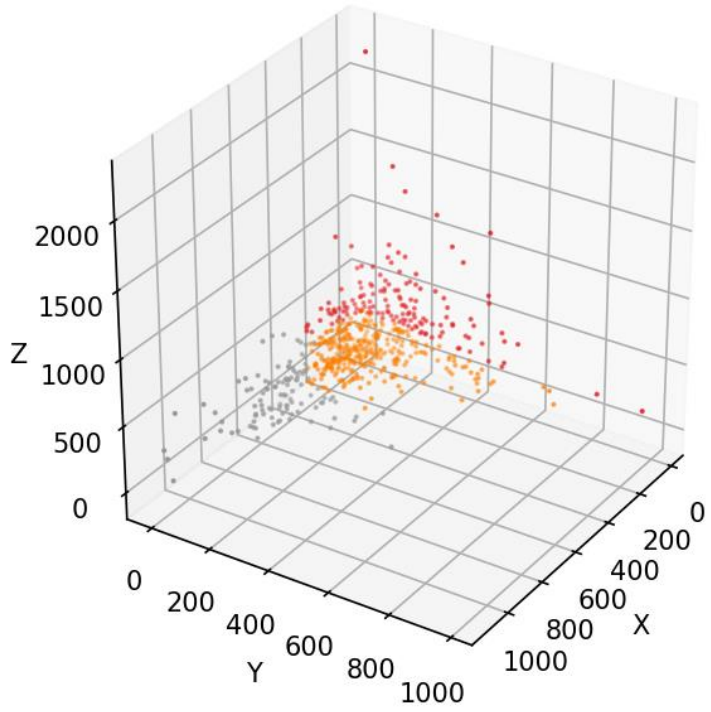
2.1) Done as a part of the Code

2.2) Elbow Method to determine the best value of k:



There is no clear 1 elbow. Since, there is a sudden change in slope for k=3, therefore choosing the number of clusters = 3.

2.3) 3D Scatter Plot

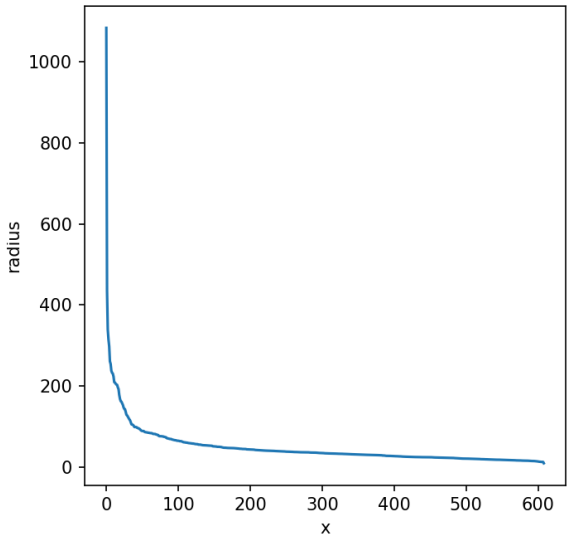


 # Task 3

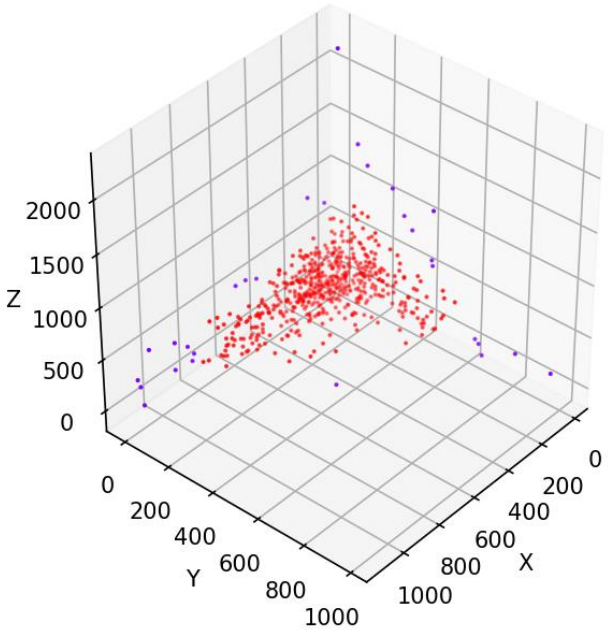
3.1) Done as a part of the code

3.2) Done as a part of the code

3.3) Plot of data points vs radius for minpts = 4



Taking y = 100 (for x = 38.05) from the graph since there is a large change of slope for this value of x. Hence, ε = 100. Applying the DBScan Algorithm for minpts = 4 and ε = 100.



This scatter plot is very much expected when applying DBScan algorithm. Even on varying the values of ε and minpts, the number of clusters obtained is only 1 (depicted here by red dots). The noises (outliers) are shown as purple dots. Being a density-based scanning algorithm, it puts all the points in a region that are packed together having similar densities in one cluster. Visually seeing the 3D plot also confirms this point that there is only one cluster of points.

3.4) Comparison of all the 3 types of clustering algorithms

Of all the 3 clustering techniques, based on visual inspection and then comparing it with output from the algorithms, DBScan gives the best result in the sense that it correctly predicts the existence of only 1 cluster with noise around the data. This is because the points cluttered together having similar density are all considered as a part of one group by DBScan algorithm. As the density of the points decreases, it then classifies those points as noise. This is the behavior obtained for the given dataset and the expected behavior.

For the k-means clustering, the number of clusters is provided based on the reduction observed in the value of SSE. Since, there is a significant decrease in the value of the SSE from k = 2 to k = 3, 3 partitions of the data are chosen. The data follows the shape of alphabet “L” in 3D. Hence, the points that are at the top of L compared to the points that

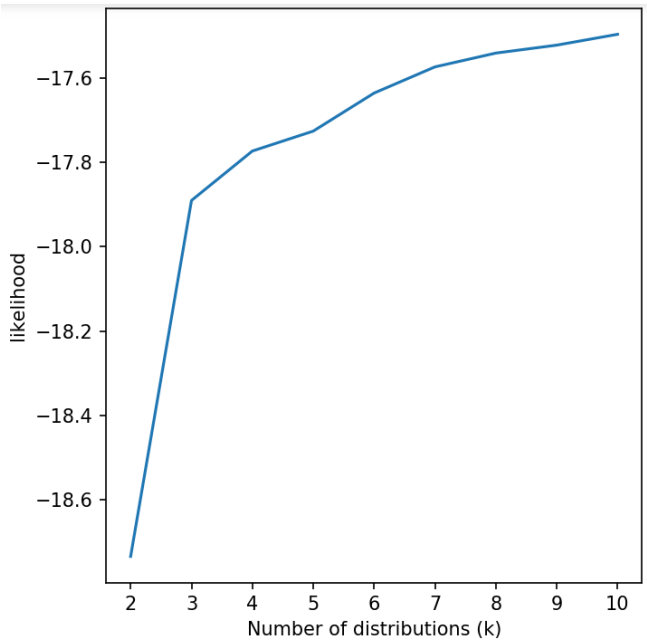
are at the bottom right corner have a large separation of distance between them. As a result of which a single cluster has a large SSE. Therefore, the entire dataset broken into 3 sets significantly reduces the SSE and hence this algorithm provides 3 clusters. Since there is no concept of separating outliers in the k-means clustering technique and they have to be a part of one of the clusters, these outliers tend to drag the cluster centroids and hence vary the shape of the cluster by moving certain points that should be a part of one cluster to be considered as a part of the other cluster. This is intuitive and also visible from the graph for the dataset provided. Assuming that we remove a few outliers, this algorithm would have given different (better) clusters but since it cannot segregate noise, it is not as good as the DBScan algorithm (especially for cases when there are outliers and the dataset is not clearly clustered).

Hierarchical Clustering is also not able to identify noises and hence is not a very good clustering algorithm especially for the cases when there are a few points (noises) that are bridging 2 or more clusters together. In the provided dataset, there is only one large cluster with noises around but hierarchical clustering splits this large cluster into three separate small clusters.

Hence, best clustering technique: DBSCAN

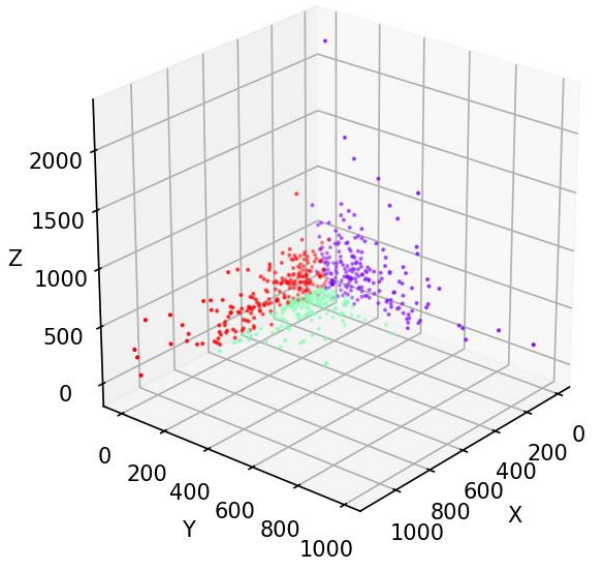
Extra Credit

1) Plot of likelihood vs k (number of distributions



The likelihood increases drastically from k=2 to k=3. Hence, taking the number of gaussian Mixture models to be 3.

2) 3D Plot for the distribution



3) The gaussian mixture model provides for 3 clusters. Points lying approximately in a plane are clustered together. This is different from the k-means and hierarchical clustering where points in different planes may be a part of the same cluster.