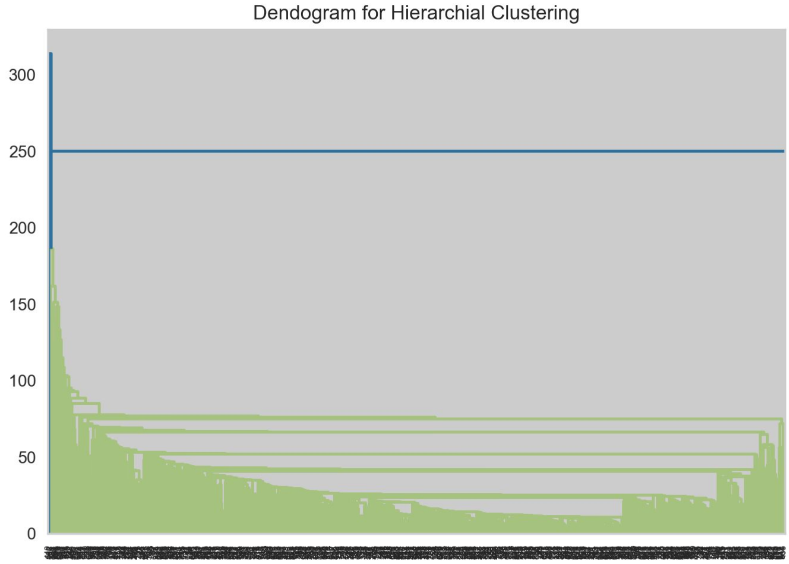
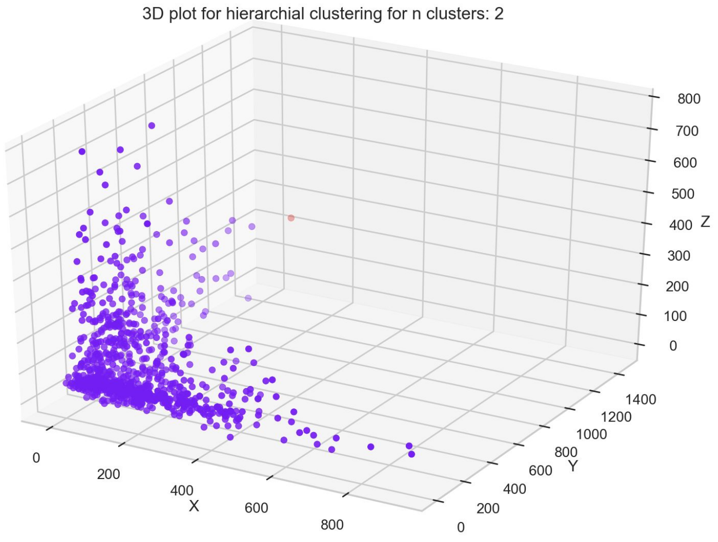
CSC 591 IOT Analytics
Project 4 Clustering
Name: Vidhisha Jaswani
Unity ID: vjaswan

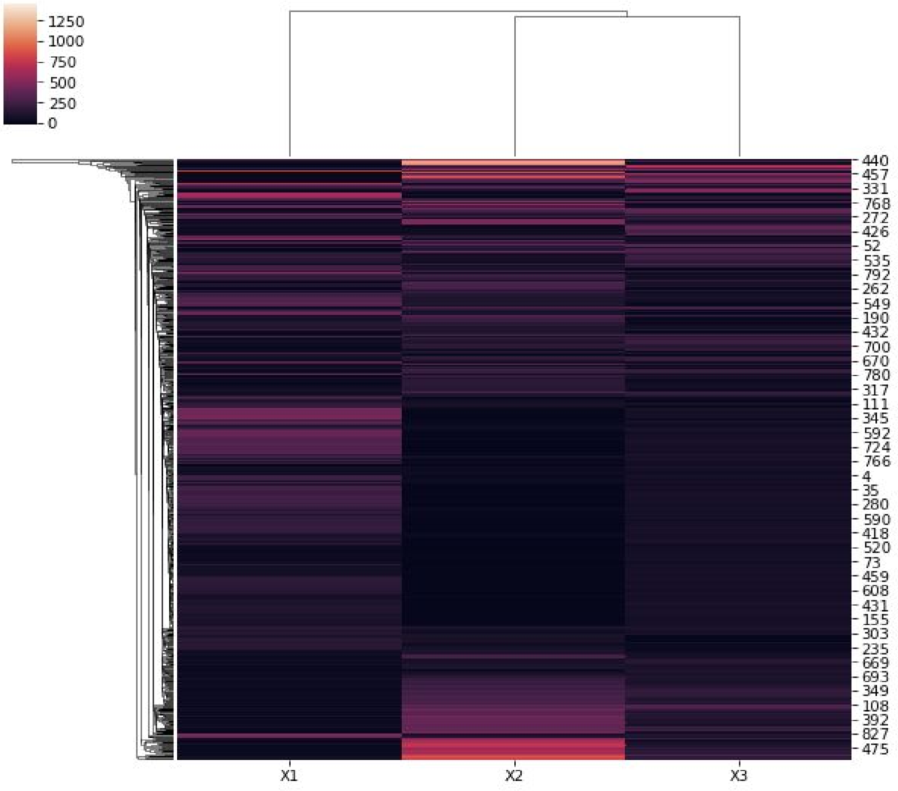
Task 1. Hierarchical Clustering

In hierarchical clustering, the following results are obtained for single linkage. After zooming in this plot, it was observed that only 2 clusters are obtained since *delta*=250.23.* For this clustering, there was just one point in one cluster and rest of the points in the second cluster. It seems that the individual point could be noise. This is also clear from the 3-D Scatter plot.

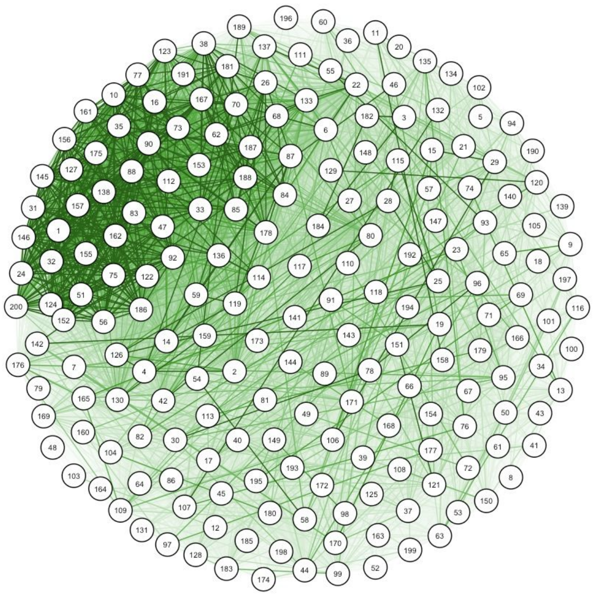




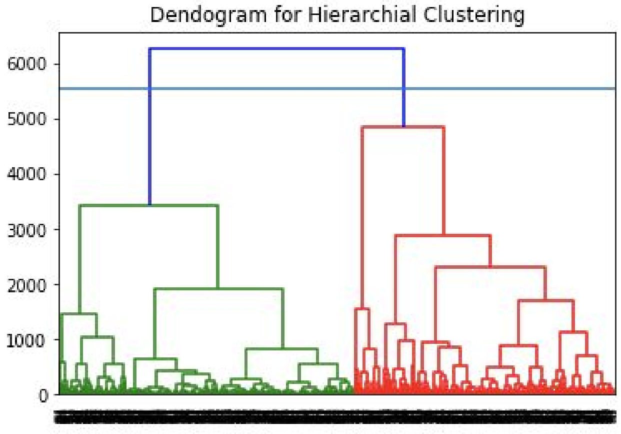
The distance graphs are also obtained as below

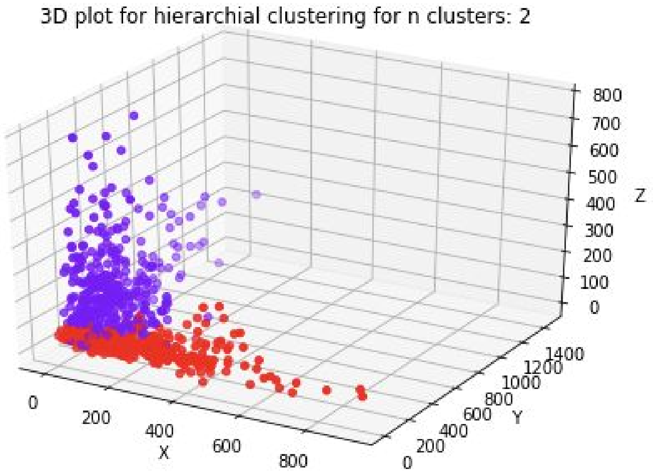


The below distance graph is obtained for only a subset of the data since it was really unclear for the whole data set.



Since single linkage did not give very good results, I also tried ward linkage which gave much better results. For this, delta* is obtained at 5550.25. The dendrogram is as below. This indicated two clusters as well but with much better distribution as compared to the one is to rest ratio obtained above.

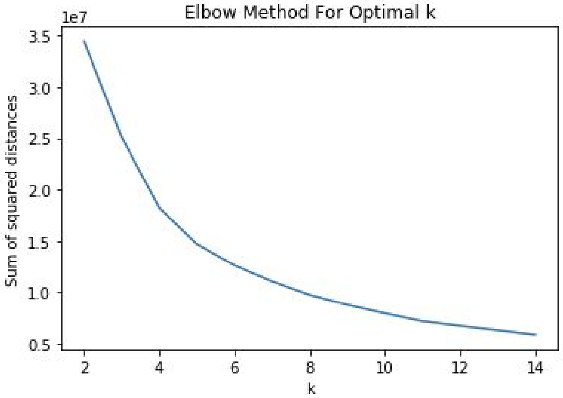




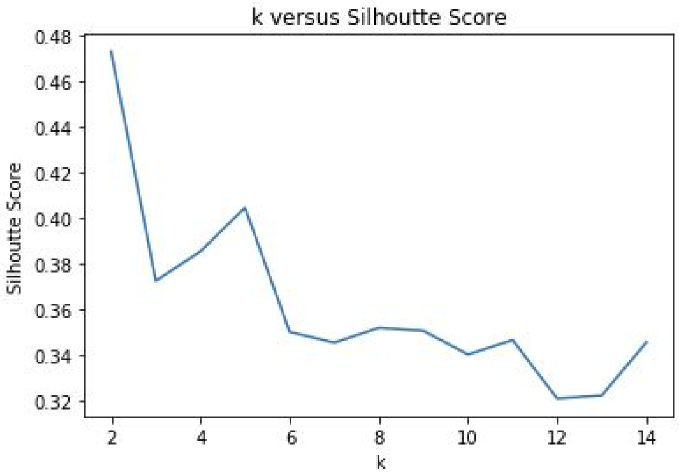
By both the methods the number of clusters obtained is 2.

Task 2. K-Means

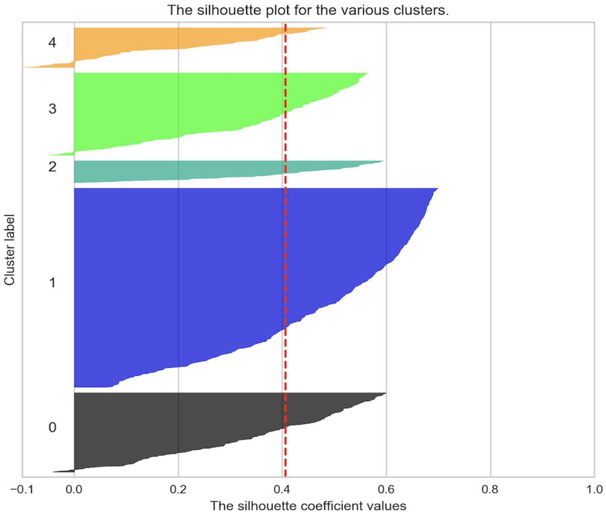
Below is a graph for plotting the SSE versus k for K-Means. Here k varies between 2 and 15.

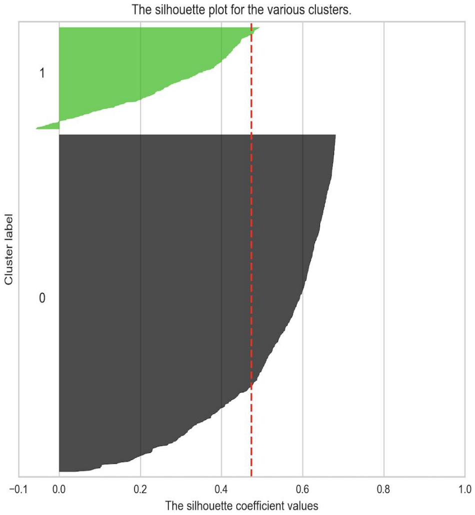


The elbow seems to be at 4 or 5 but since it is not a smooth curve, an alternate method is also tried. Here Silhouette score is plotted against k. The highest values for average silhouette score are observed at k=2 and k=5.

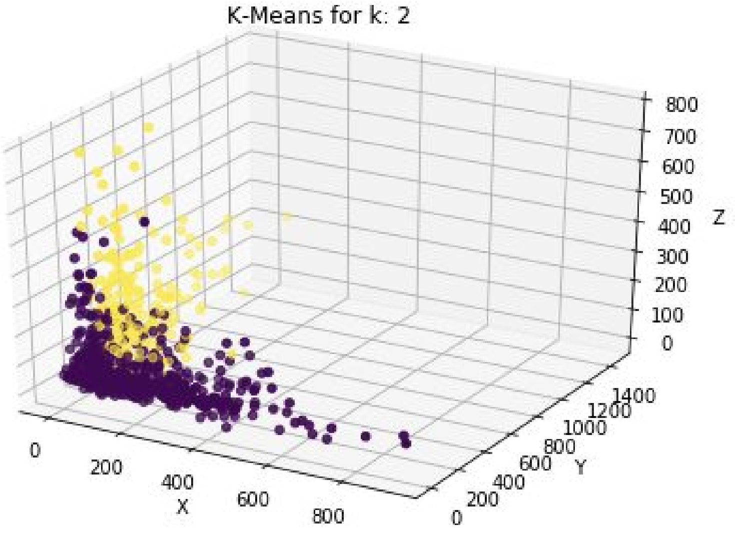


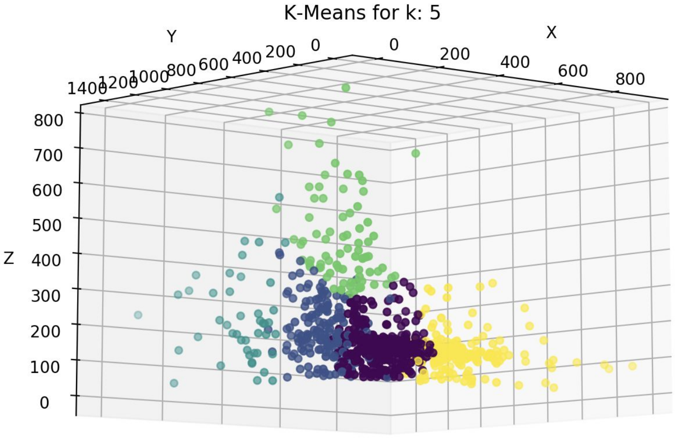
Another, method of deciding k is plotting silhouette score of individuals points in the cluster and observing how many points have a negative score. A negative score indicates bad clustering. Below plots are obtained for k=2 and k=5 as they gave the best results. For k=2, there are some negative values in Cluster 1 but the average is pretty high. For k=5, there are few negative values as well in 3 clusters. Although, the average for k=2 is lower than for k=5. The code for this has not been included in the submission as they have been referred from pythons scikit library.
(Reference:https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html)





So, I have plotted the 3D plots for k=2 and k=5.

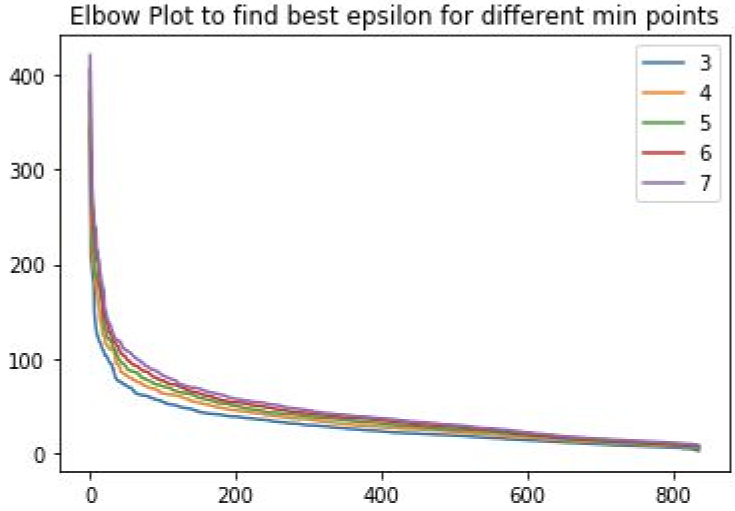




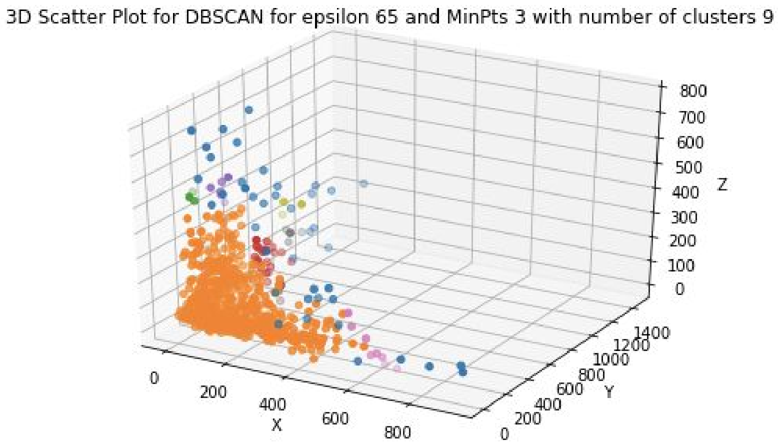
Both clusterings look good as there is some density around X=0 and then tends to deviate in all directions but to decide which clustering is best we do need some more knowledge about the data.

Task 3 DBSCAN

Below is the elbow plot to find out the best value of epsilon for different number of MinPts. I have varied MinPts from 3 to 7. Best epsilon is around 65-70 for MinPts 3.

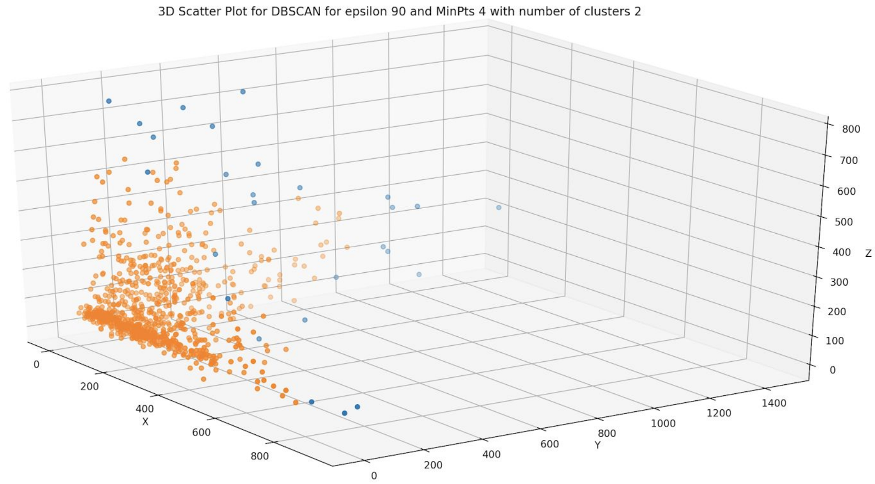


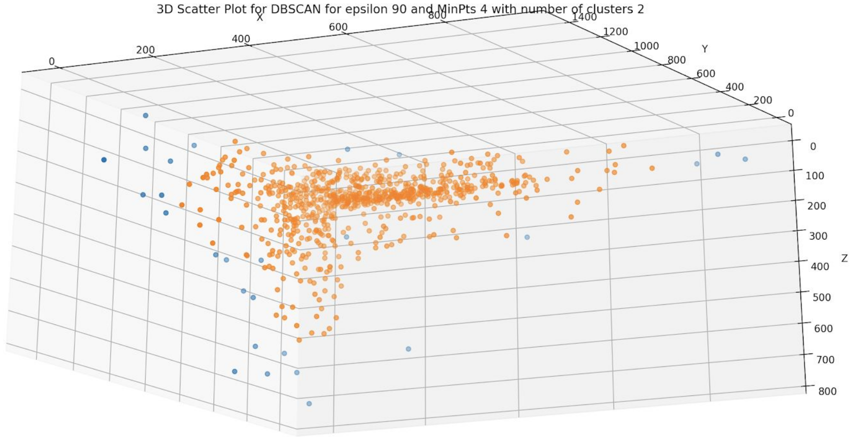
For epsilon 65 and min points 3, 9 clusters are obtained. This seems like bad clustering.



After that, for the combinations of
a) epsilon 90 and min points 4
b) epsilon 100 and min points 5
c) epsilon 120 and min points 6
d) epsilon 120 and min points 7

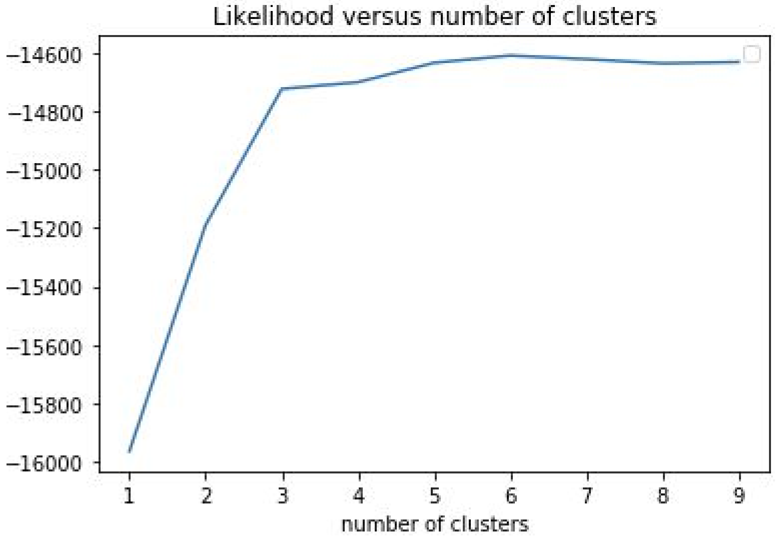
A same 3-D Scatter plot is obtained with number of clusters as 2 indicating that number of clusters are indeed 2. Best scatter plots are below. However, it does not seem like they are 2 clusters, it seems like one dense cluster with noise around it.





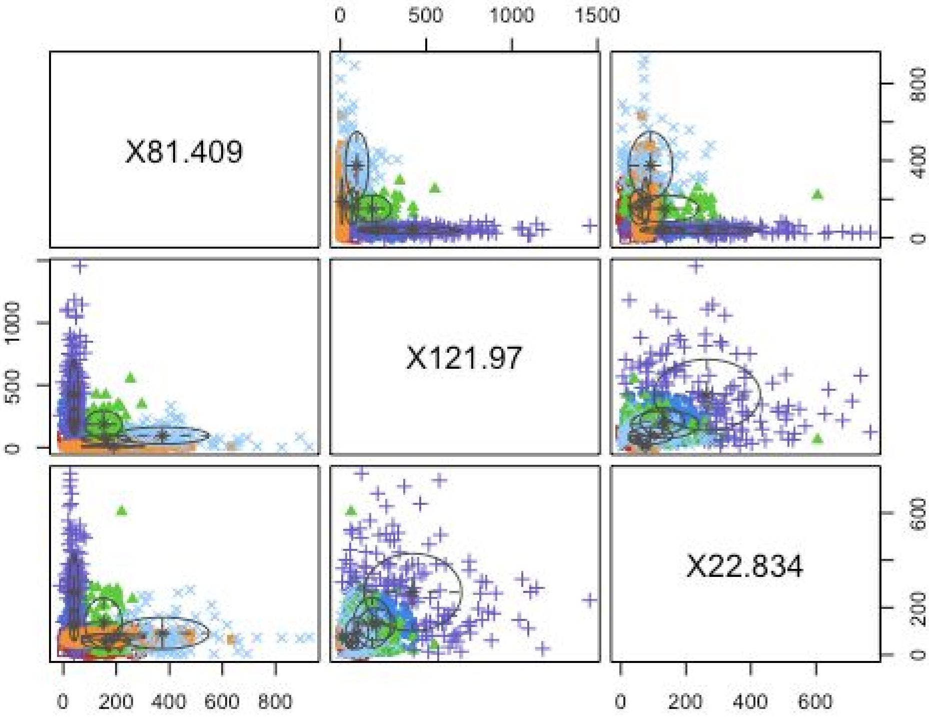
Extra Credit

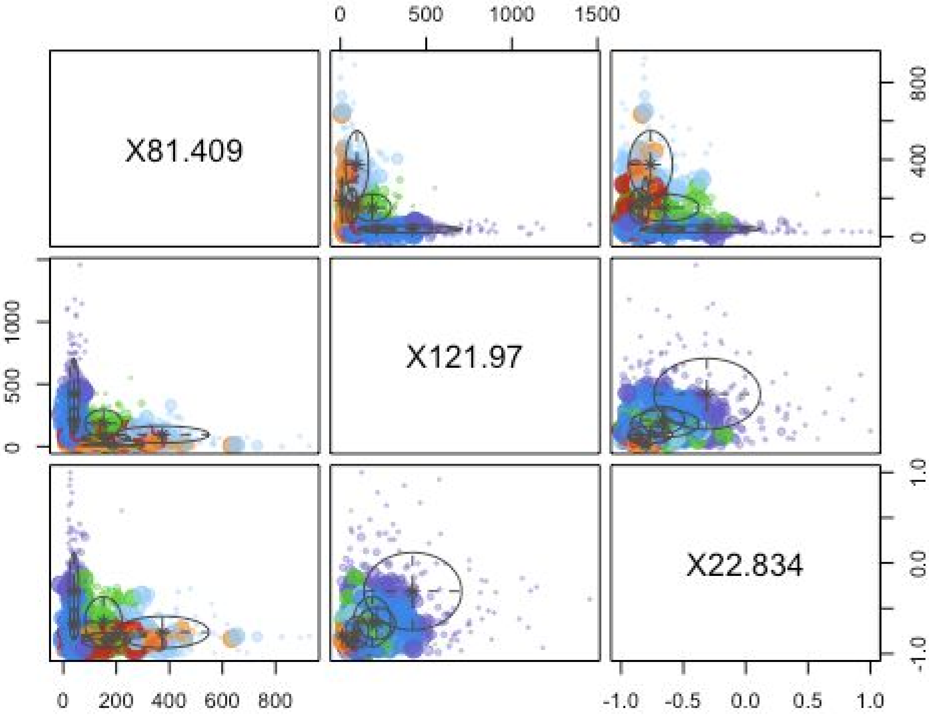
Below is the plot for Likelihood versus number of clusters. So we have to look for maximum value of Likelihood which comes at number of clusters equals 6.

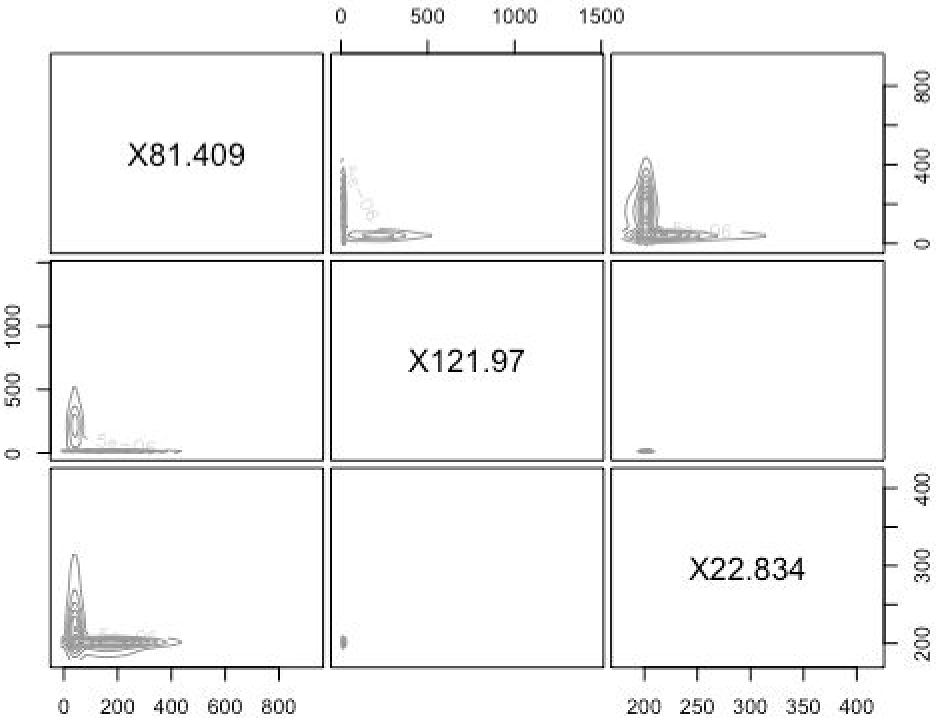


The minimum value is obtained for number of clusters equals 6. It plots XY,XZ in first row. YX and YZ in second row and ZX and ZY in the last row. The XY and YX plots look symmetrical to each other and somewhat also align with the scatter plot of 2 features taken at a time in terms of their densities and distributions. The code to obtain the below plots is in R. The results are as below:

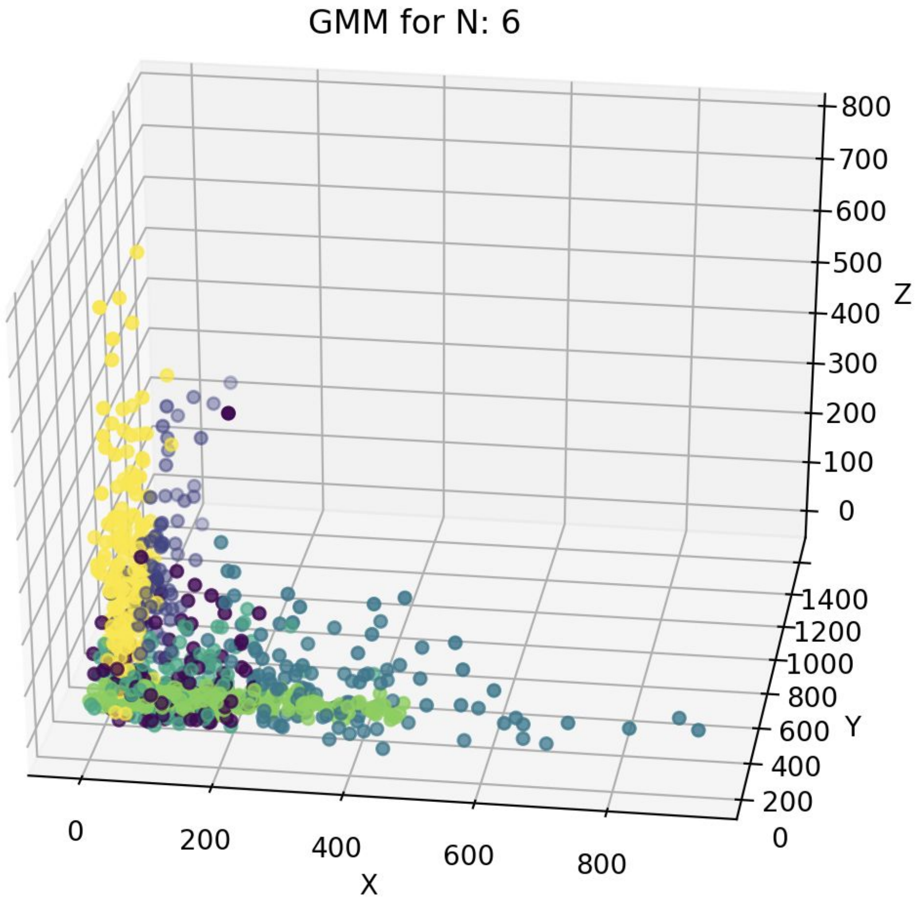
```
----------------------------------------------------------------------------------------------------------------------
Mclust VVI (diagonal, varying volume and shape) model with 6 components:

 log.likelihood   n df       BIC       ICL
      -14439.05 836 41 -29153.97 -29483.17

Clustering table:
  1   2   3   4   5   6 
176 151  63 127 240  79 
----------------------------------------------------------------------------------------------------------------------
```


X81.409
X121.97
X22.834
0
500
1000
1500
800
400
0
1000
500
0
600
400
200
0
0
200
400
600
800
0
200
400
600


X81.409
X121.97
X22.834
0
500
1000
1500
800
400
0
1000
500
0
1.0
0.0
-1.0
0
200
400
600
800
-1.0
-0.5
0.0
0.5
1.0



The scatter plot for gaussian mixture decomposition with n=6 is



Which looks like a really good plot. Points scattered in one direction are labelled as one cluster.

Comparison of Results

According to me, since our purpose is to do clustering, gaussian mixture with n=6 gives the best results. Apart from either hierarchical clustering with ward linkage with number of clusters is 2 or K-Means with K=2 work best as they cluster denser data together. DBSCAN is suitable had our purpose been to remove noise from the data.