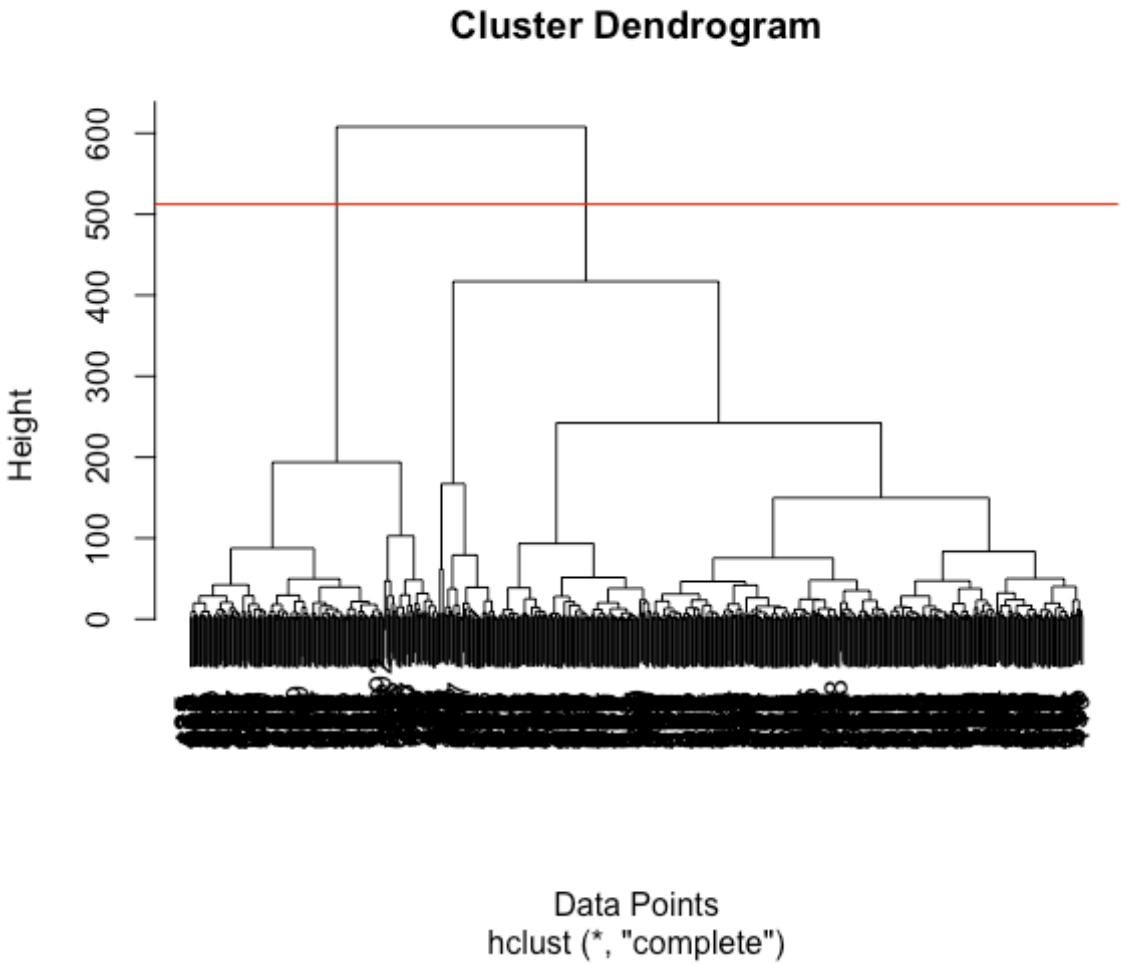
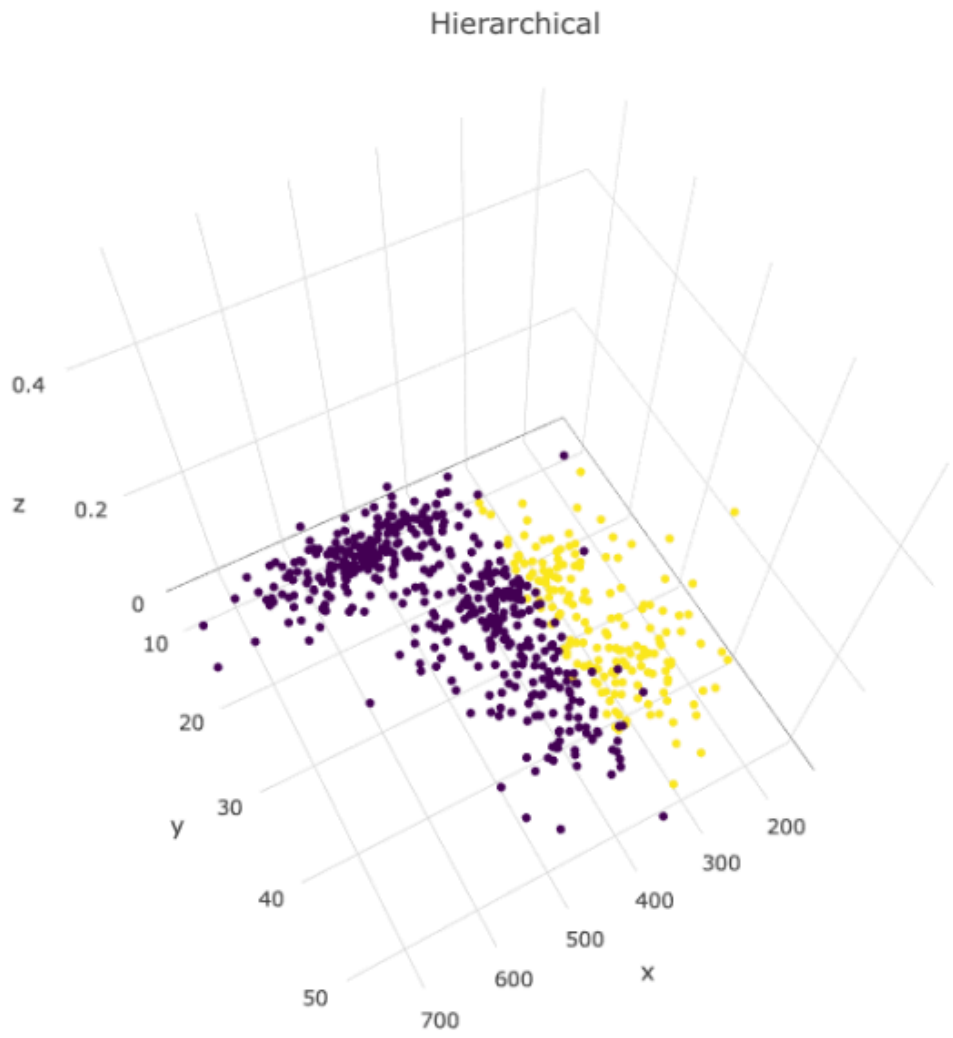
Clustering

Hierarchical Clustering

Here I compute Euclidian distance between every point and call hierarchical clustering model which outputs dendrogram as shown below.



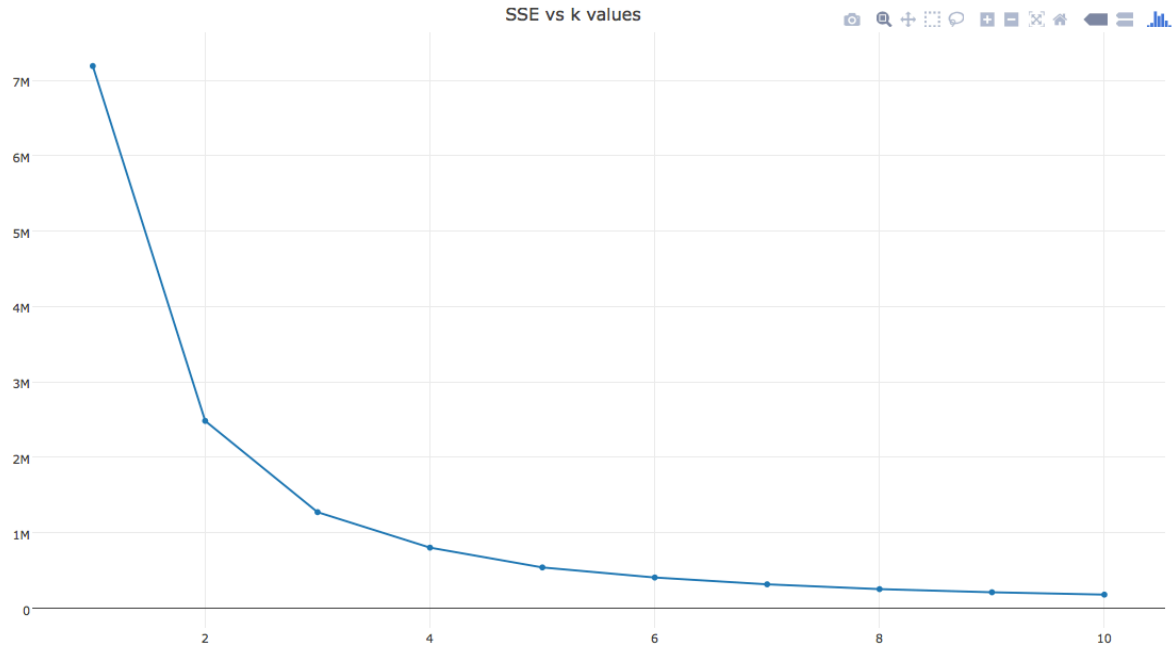
Then I compute difference between consecutive “Deltas” to find the delta to cut the dendrogram. Here the cut is shown with the red line. And then clustering is done based on that value which outputs as below.



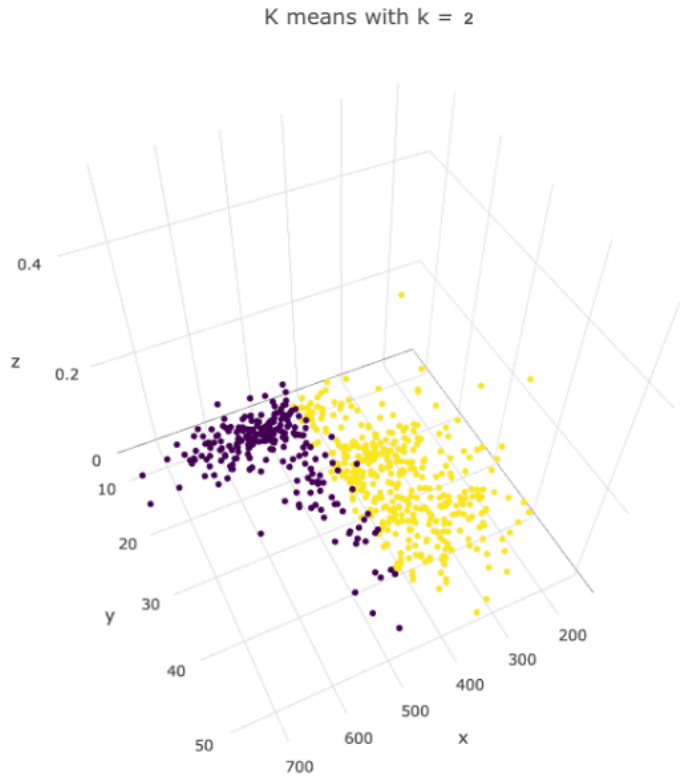
As we can see it identifies two clusters. The reason why clustering looks like this might be because data has 3 dimensions which vary in scale i.e. as we can see Z has most of the values below 0.2 and X and Y have values in higher ranges. So distance is dominated by X and Y. Apart from it single linking can be the reason of this kind of output. Because noise may combine clusters.

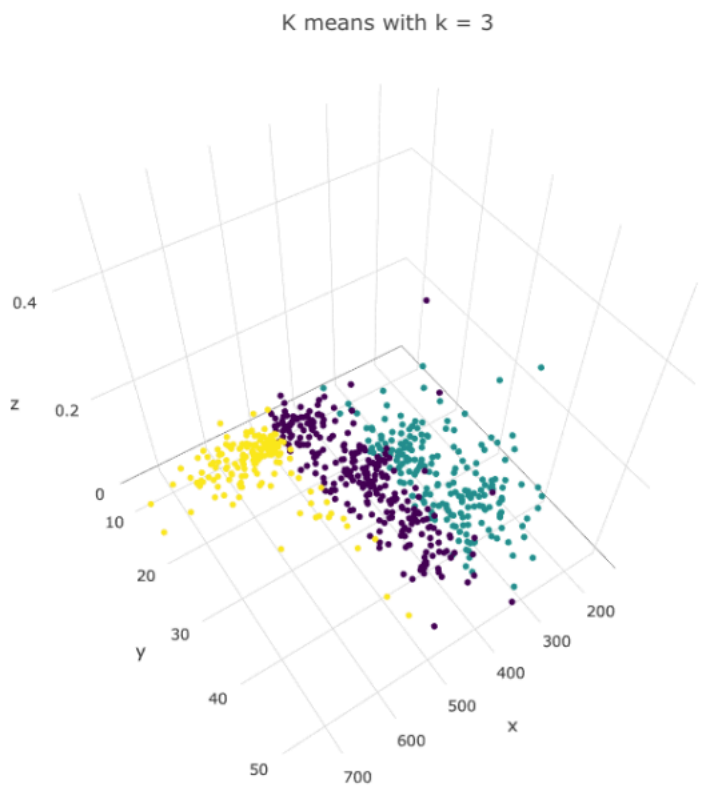
K-Means

First I try clustering with multiple k values i.e. 1,2,3…10. And plot it against Sum Squared Error which produces following graph with Y-axis representing SSE values and X-axis representing k-values.

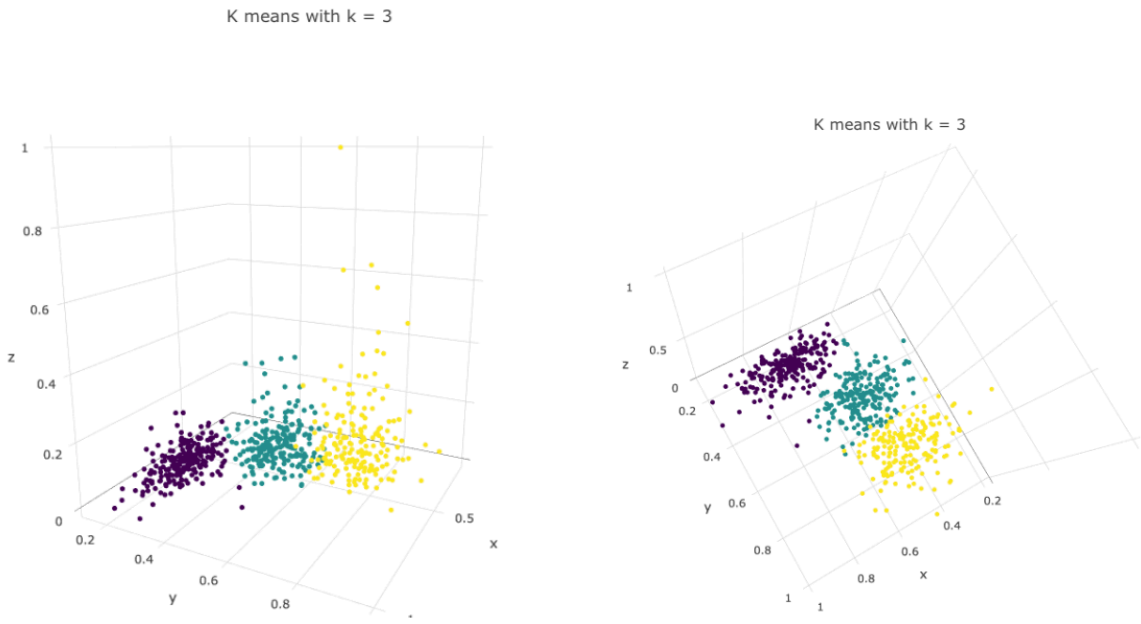


Based on above graph we can see that elbow or knee is at k value = 2 or 3. So we try k =2 and k = 3 and results are shown in below graphs.



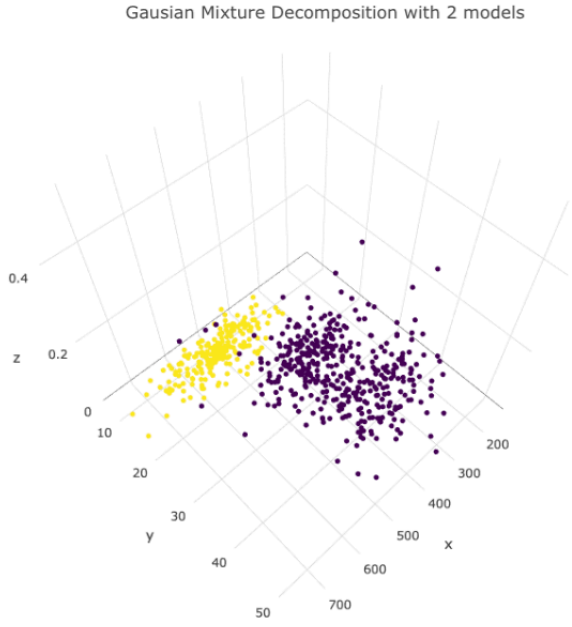


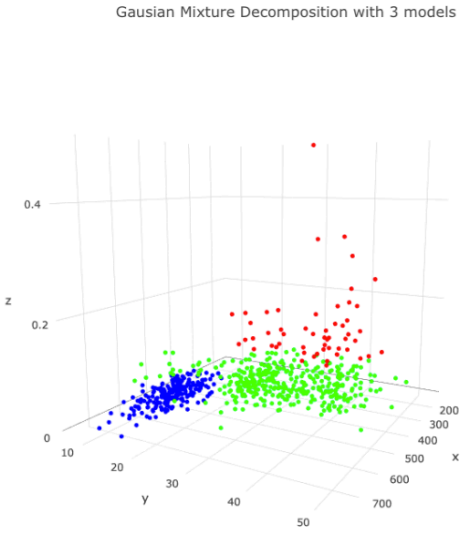
Here also results seem unintuitive. The reason might be the different ranges of values in x,y and z direction. So I did k-means after scaling the data to the range 0 to 1. And results are as below :



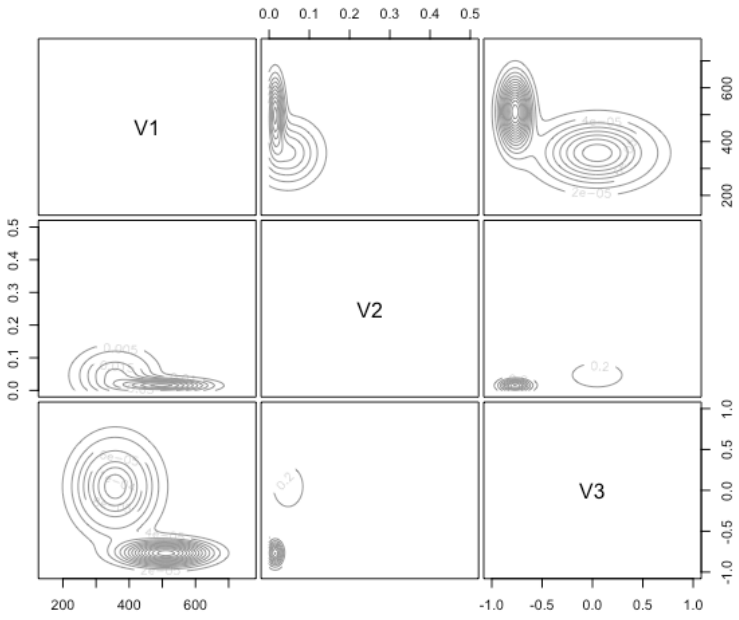
Which seems pretty good. Also we can observe that data points might be from some normal distribution. Like violate points coming from one distribution and green and yellow coming from two other normal distributions. Though data points does not look exactly like normal distribution curve. I tried Gaussian Mixture Decomposition with 2 numbers of normal

distribution model and 3 numbers of normal distribution models. Which produce following results.





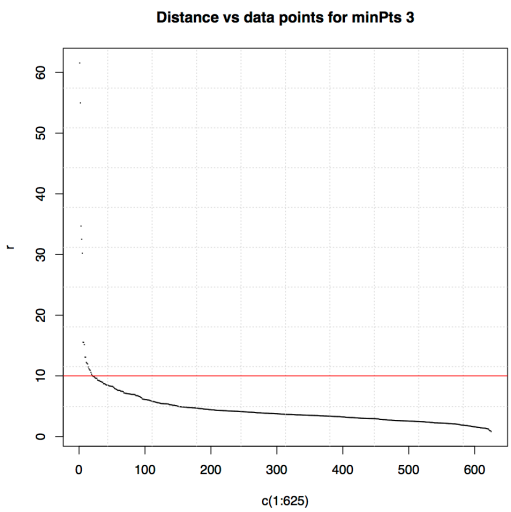
With two Gaussian model produces very good results. We can also see that from contour plot.

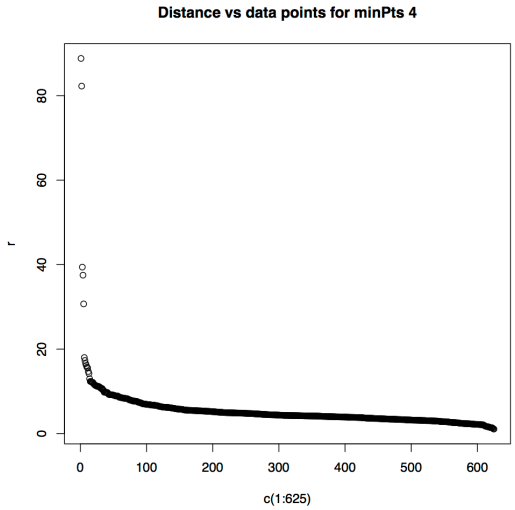


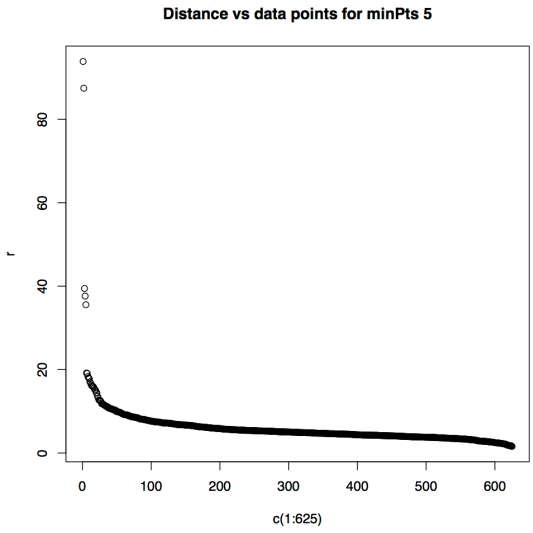
We get similar results for scaled data too.

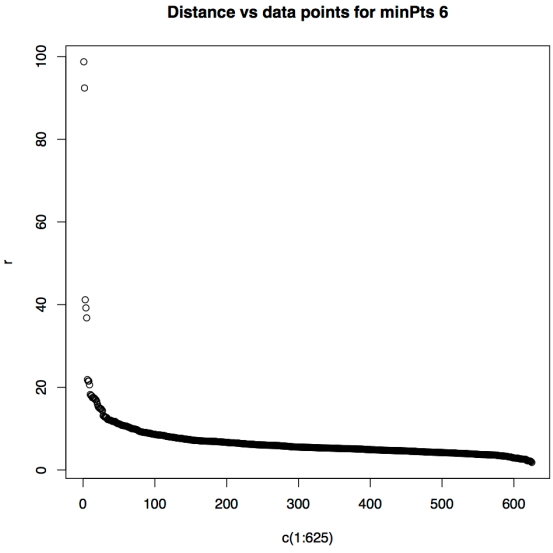
DBSCAN

For the DBSCAN first we need to decide ‘Epsilon’ for given minPts. So starting with minPts = 3 we plot Radius for each point which would include 3 number of points and we plot those readius against data point for minPts = 3.

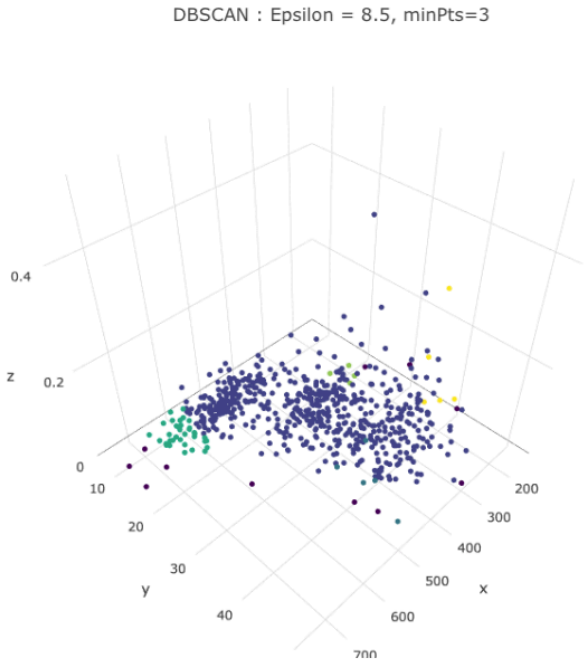




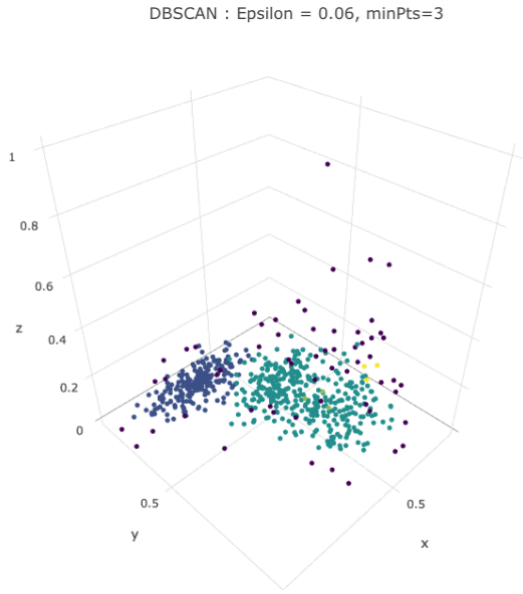


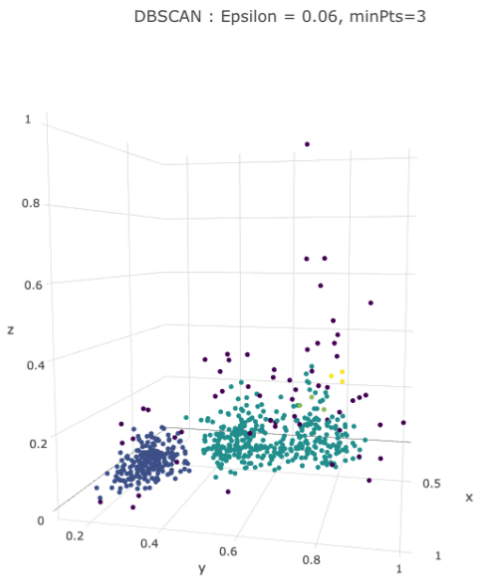


And then based on elbow method we decide epsilon and do DBSCAN. I tried all the combination and best DBSCAN clusters according me is shown below.



Here also results do not seem very good. So I tried scaling data first and results look nicer after scaling which is shown below.





Best Method

From the data it looks like that it has 3 clusters. And according to intuition and after trying all these methods with various parameters *K-Means on scaled data with k=3* provides best clustering.