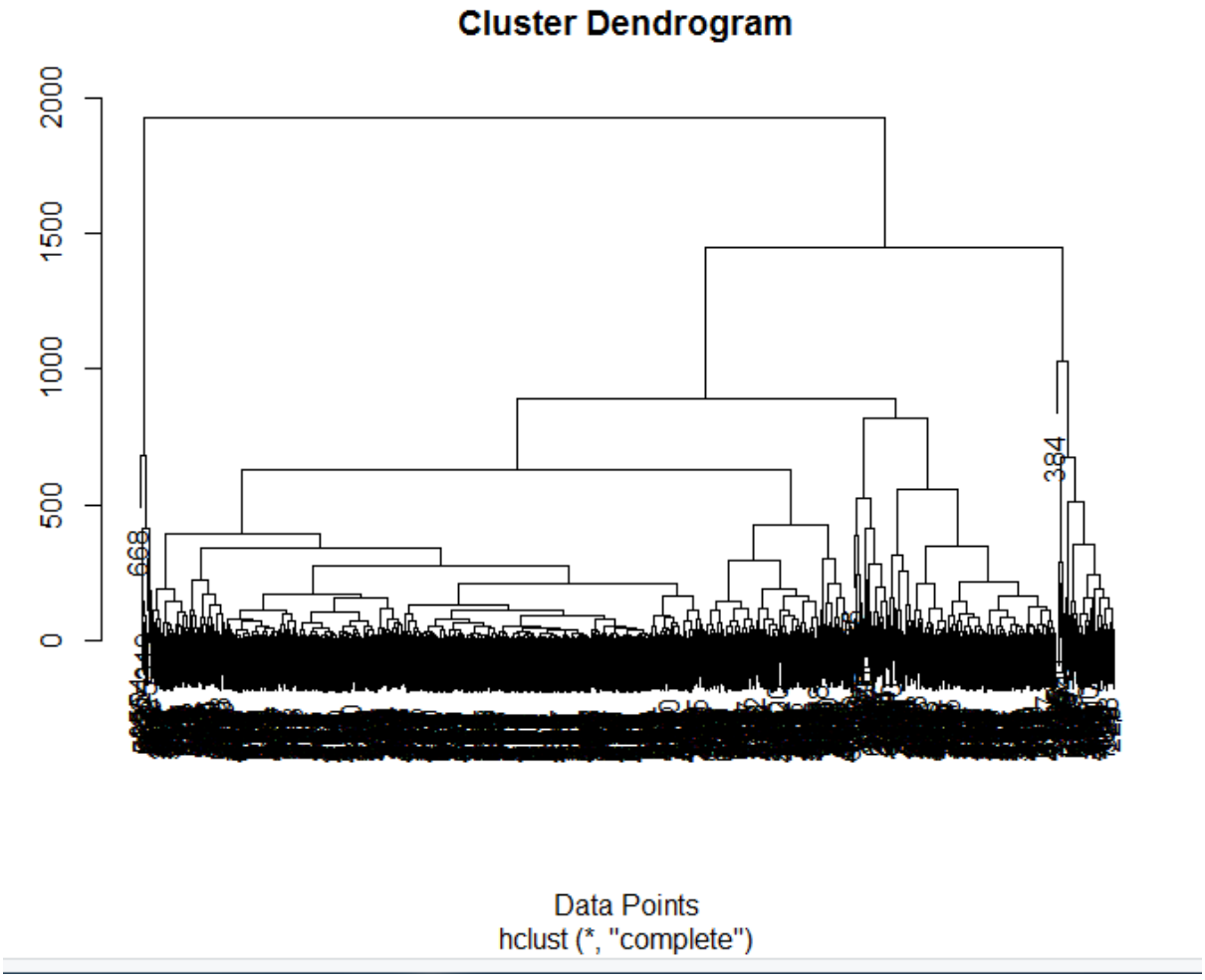
CSC – 591 Internet of Things Analytics (Project 4: Clustering)

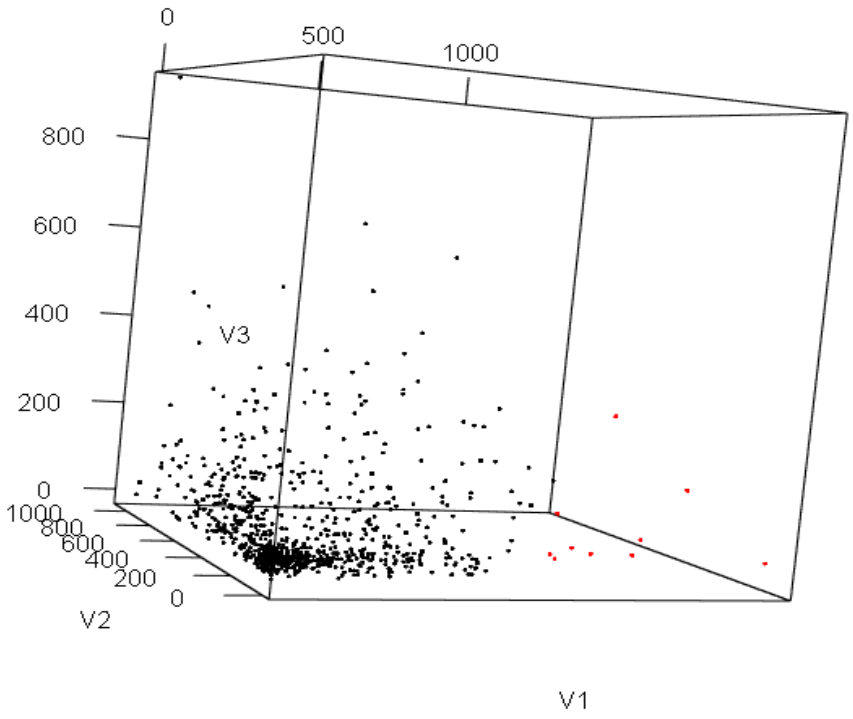
Name :- Harsh Jatinbhai Patel
Student Id :- 200258486
Unity Id :- hpatel8

1. Hierarchical clustering

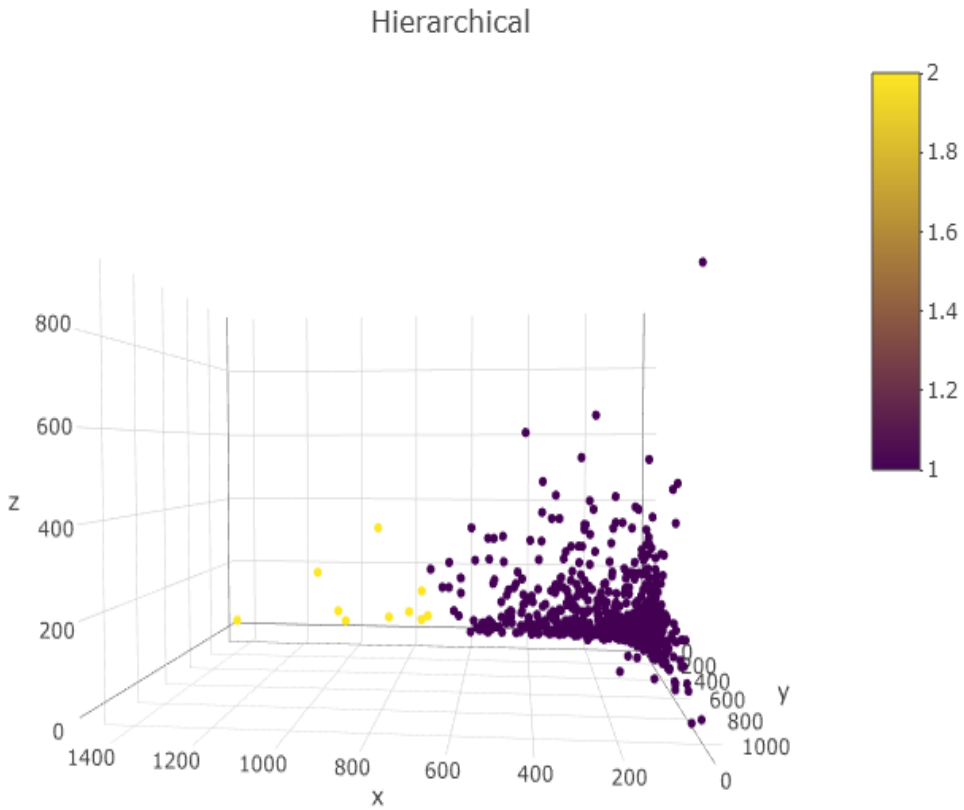
Here we apply the hierarchical algorithm to our data set and plot the dendogram as shown below.



- From the Dendogram we can see that the number of clusters formed by applying the hierarchical algorithm is 2.
- Now we plot a 3D Scatter plot to visualize the clusters and analyse how did the algorithm perform. Here (V1 = X-Axis, V2 = Y-Axis , V3 = Z-Axis)

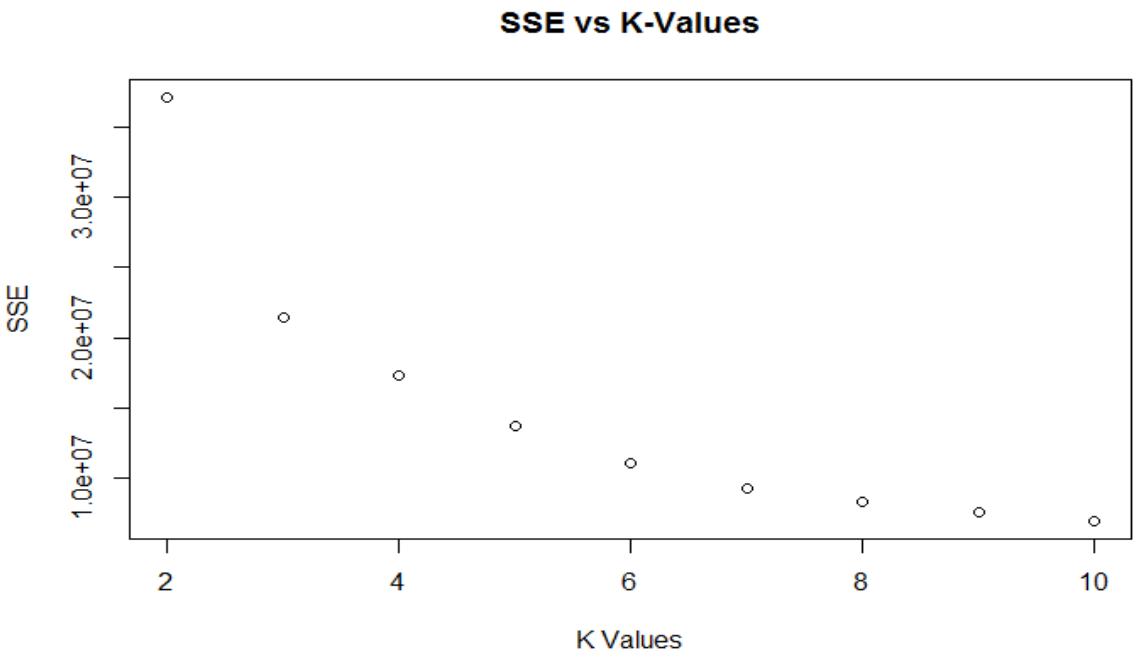


❖ 3D Scatter Plot with more measurements and clarity using the Plotly Library in the R Programming Language.



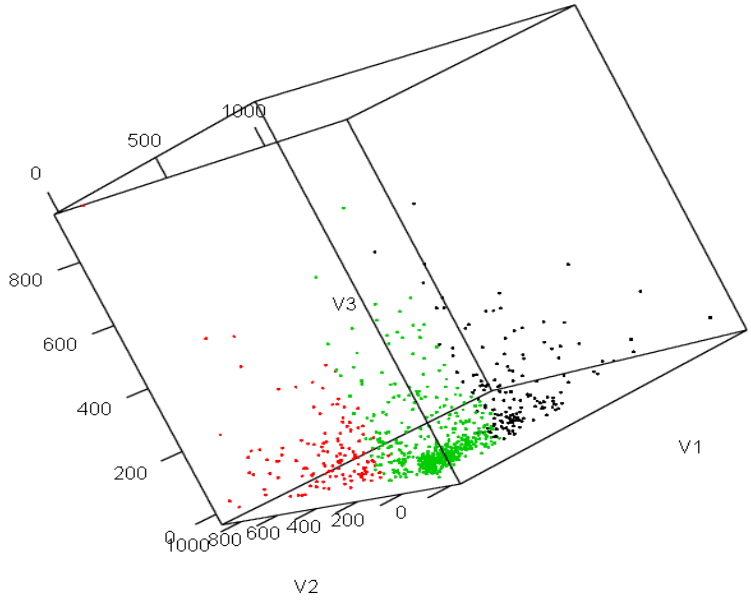
2. K-Means Clustering

Here we apply the algorithm for several values of k starting from k = 2. Here I continue the iterations till k = 10 and plot the elbow plot as shown below.

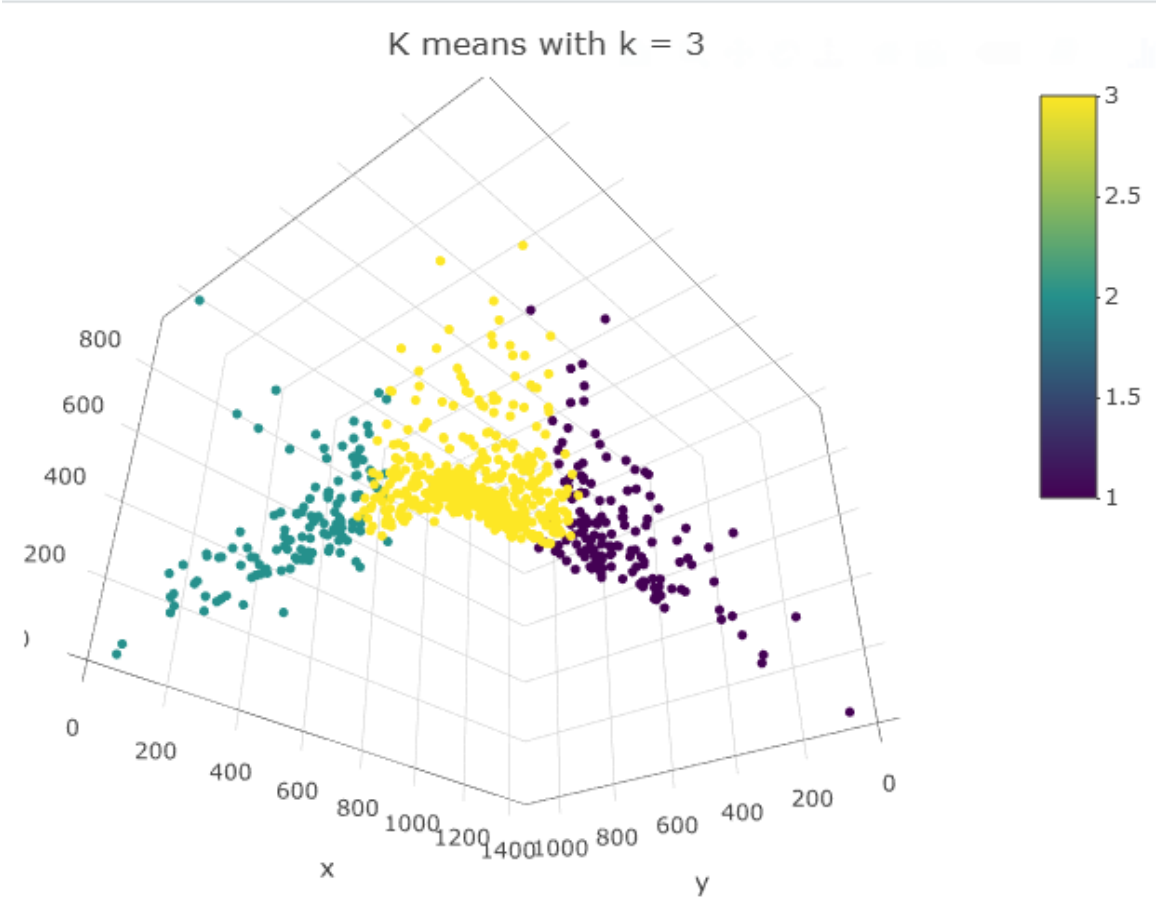


- Here from the above plot we can see that the elbow point is at **K = 3** and so that is our best value of k which is the number of clusters we shall use to plot the 3D Scatter plot as shown below.

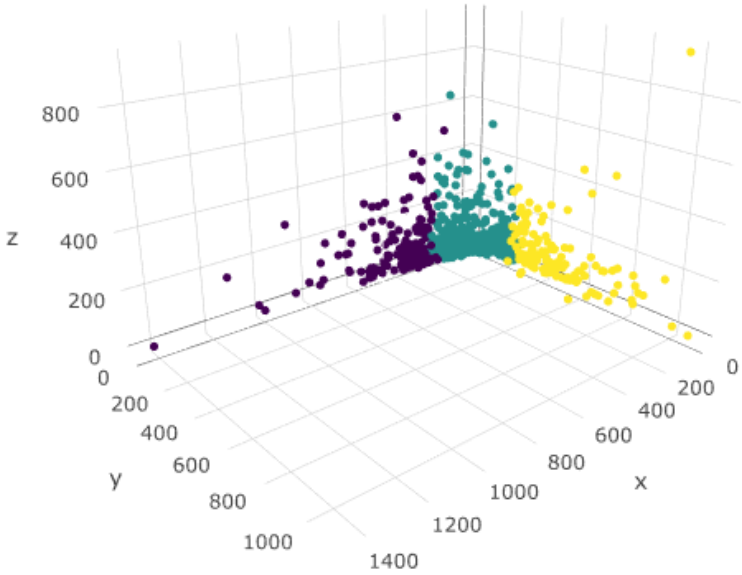
3D Scatter Plot using Boxgl



❖ **Rotated 3D Scatter Plot with more visual clarity. (Using the Plotly Library)**



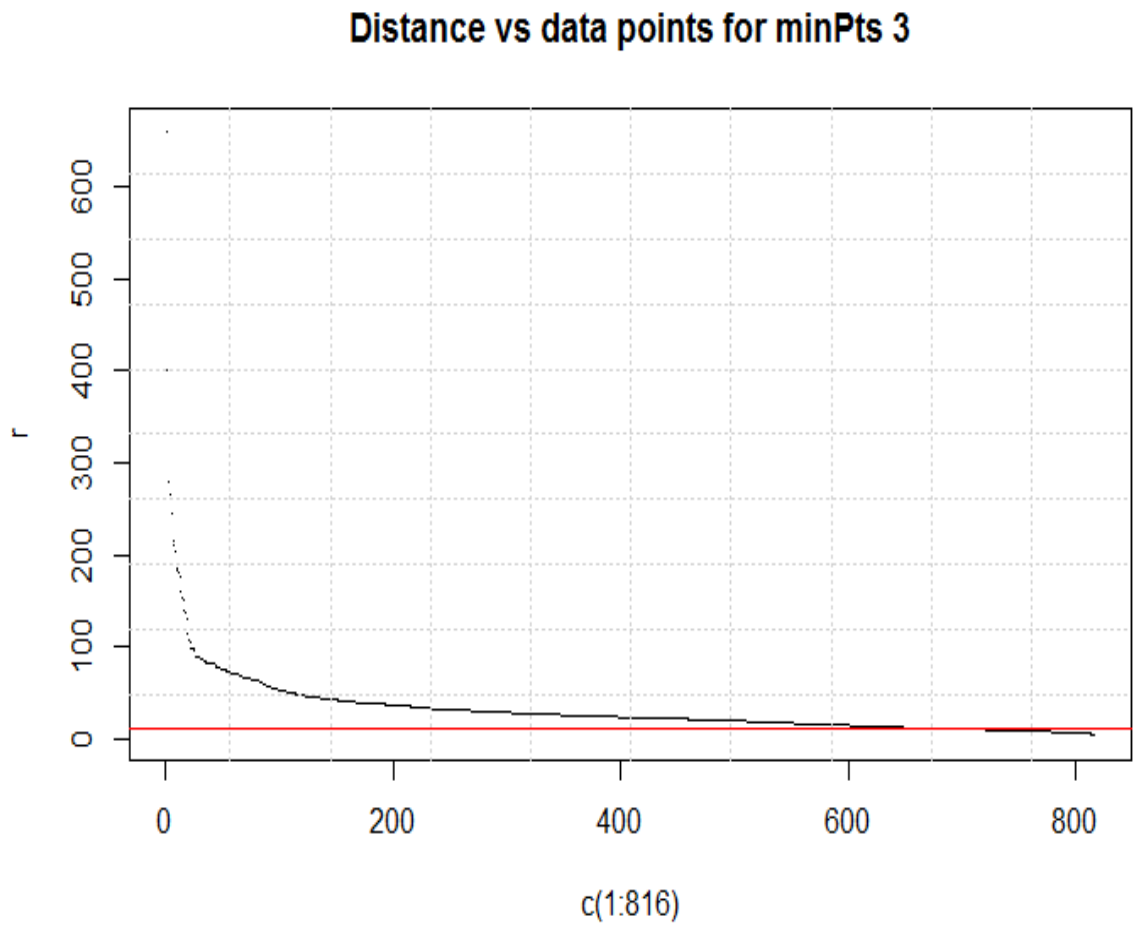
❖ **<u>Zoomed in 3D Scatter Plot to get measure of the plotted points.</u>**



3. DBSCAN Clustering

- DBSCAN for Minpts = 3 and plotting the elbow graph for determining the best value of epsilon.

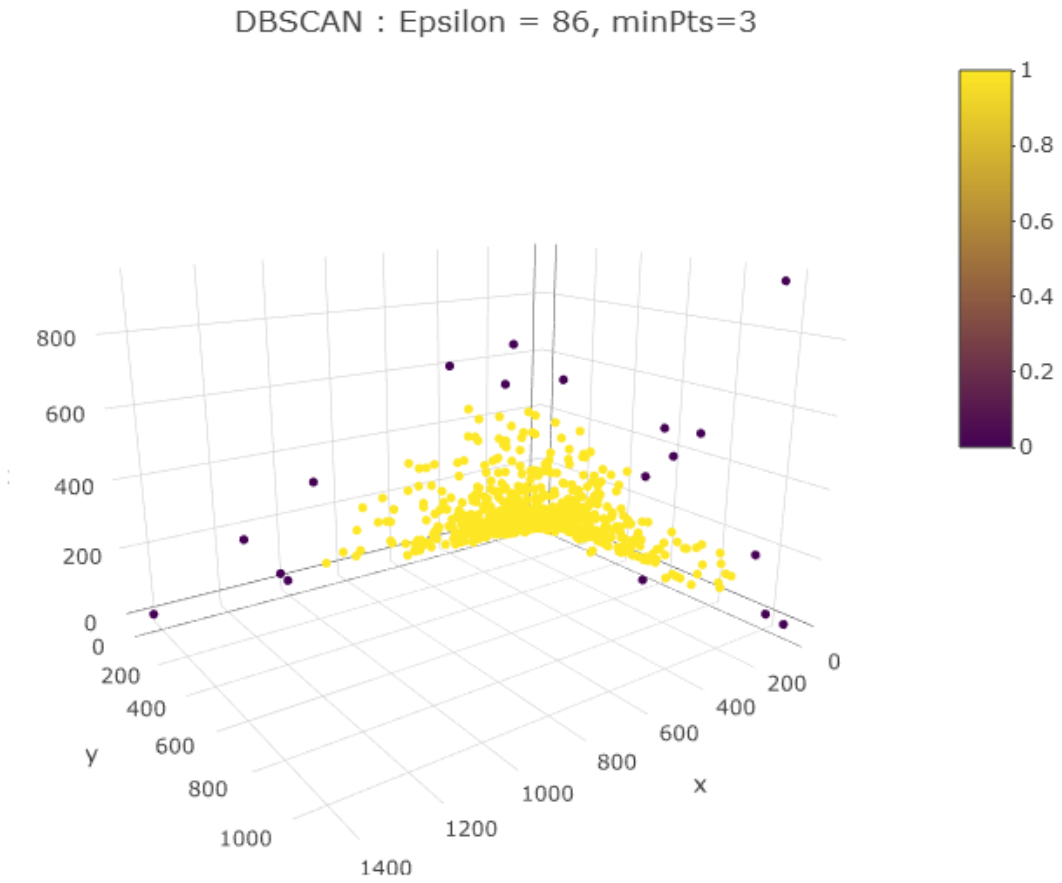
❖ Below is the elbow plot for our DBSCAN algorithm to select the best epsilon.



❖ Here from the above graph we can see that the elbow point in our graph is at point epsilon = 86 which we shall select as our epsilon to fit our DBSCAN algorithm.

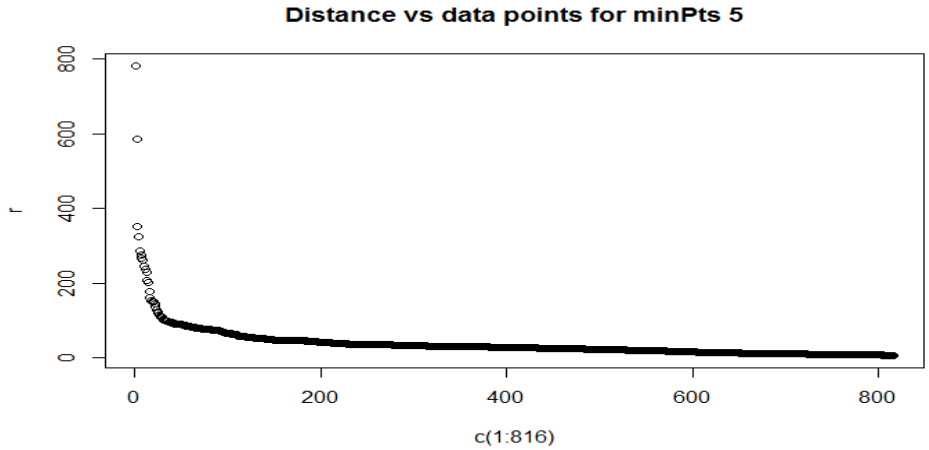
❖ Now using Minpts =3 and Epsilon = 86 we plot our 3D Scatter Diagram.

❖ **3D Scatter Diagram for MinPts = 3 and Epsilon = 86.**

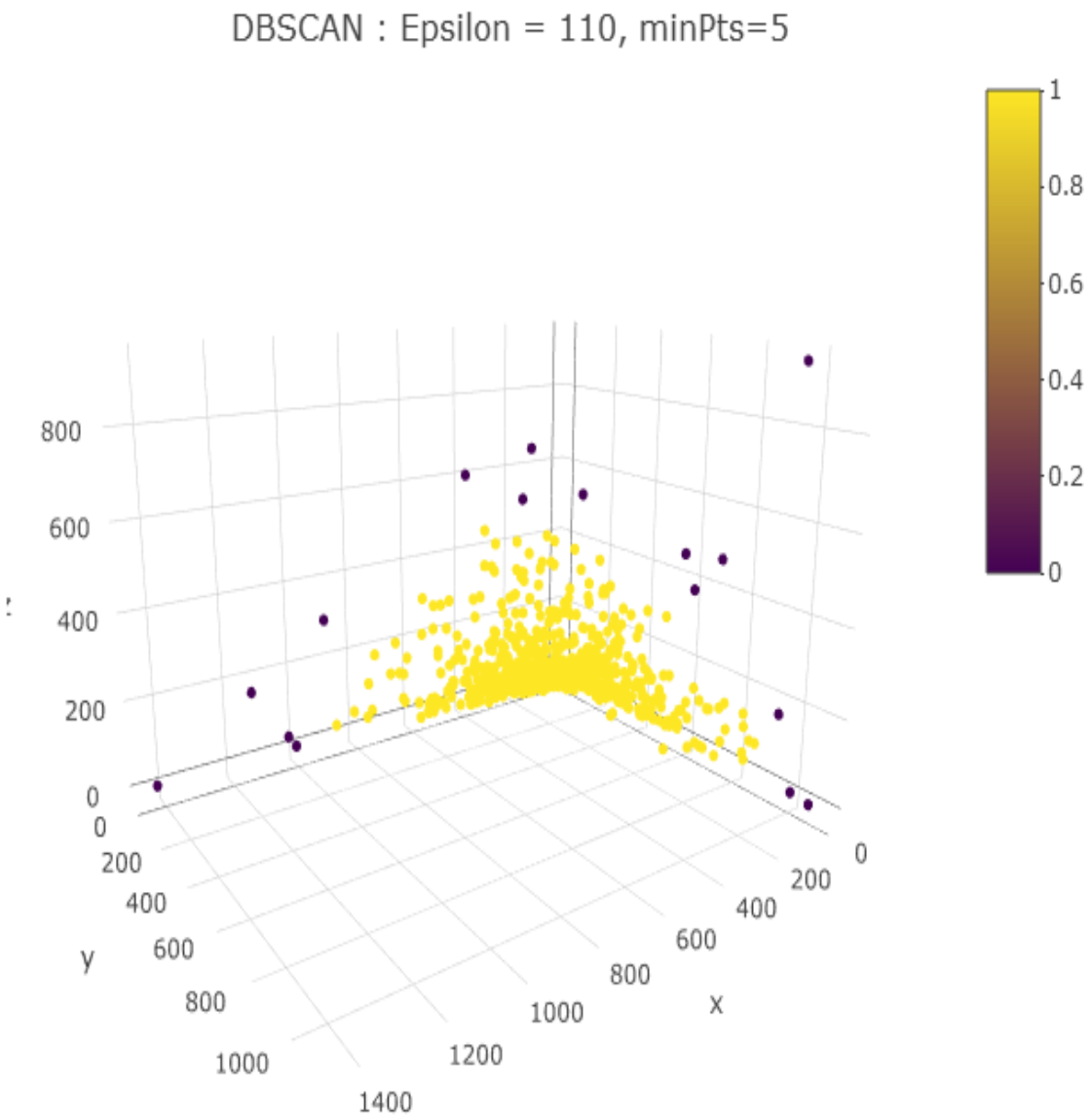


- Now we want to find the best clustering fit for DBSCAN and so we try it for Minpts = 4,5,6. On trying it for other Minpts I found that for Minpts = 5 we get a good clustering fit for our data set with the points mapped to a cluster better than other Minpts.

- **For Minpts = 5 we plot the Elbow plot to get the Epsilon Value**



- Here we get the Elbow point at Epsilon = 110 from the above graph and now we fit the DBSCAN Algorithm for and plot the 3D Scatter Plot.



❖ Here we can see two clusters, one in yellow and other in purple. Here compared to Minpts = 3 we get a clear demarcation between two clusters and in MinPts = 3 it mapped some outer points into the yellow cluster and so taking MinPts = 5 gives us two clusters with points mapped correctly in each cluster.

4. Compare and discuss results from all three methods. Identify the best clustering of the dataset.

- In the Hierarchical clustering method we get two clusters as and here we can see that some of the outer points are also merged into the single cluster and there is no clear demarcation between the points and the clusters as it is sensitive to the noisy data present in our data set. Hence this algorithm does not give a conclusive clustering when applied to our data set.

- In the K-Means clustering we apply the algorithm for K=3 and even as our the data points in our data set are clustered together at the origin K-Means method does a good job in differentiating them into three clusters and mapping them to their respective clusters. Here the algorithm takes into consideration the fact that even if the data points in our data set are near to each other it is still able to differential and find three clusters among them and we can see them in different colours in our scatter plot.

- In the DB-SCAN algorithm we can see that by trying for different epsilon and different MinPts we can see that as the points are clustered near 0 to 800 on X-axis, 0 to 800 on Y-axis and 0 to 400 on Z-axis so as we can see from the scatter plots as all the points are near to each other and the distance is small DBSCAN clusters all the points near the origin and in range of 0-600 on X and Y axes as yellow and the remaining points which tend to seem outliers or noise as shown in purple.

- Hence the K-Means clustering algorithm performs the best among all the three algorithms above and gives us the best result