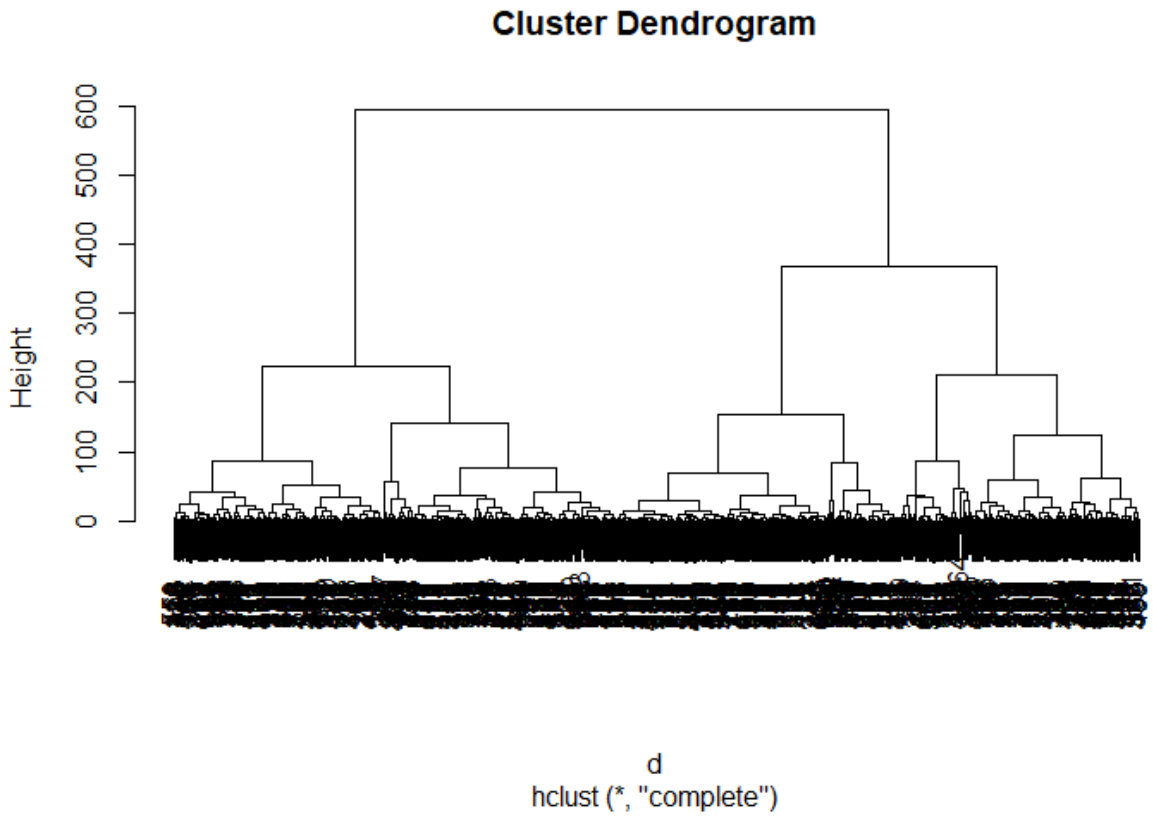
Project 4 : Clustering

Hierarchical Clustering :

It is a method of cluster analysis which seeks to find a hierarchy of clusters. It uses dendrogram to show the evolution of partions used to build the hierarchy.

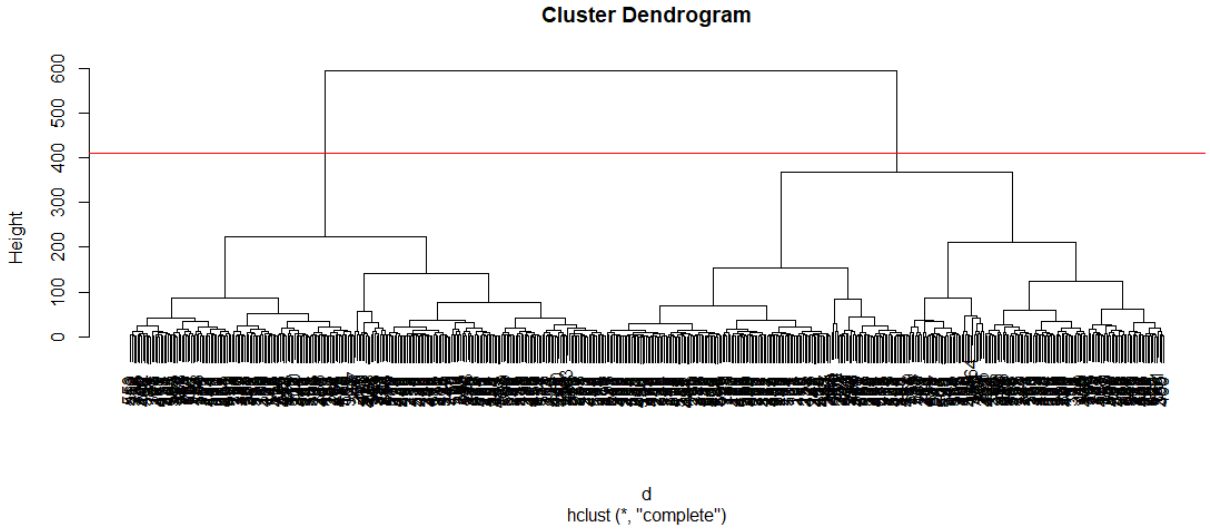
For the given data, when we do hierarchical cluster analysis, we find the following dendrogram,



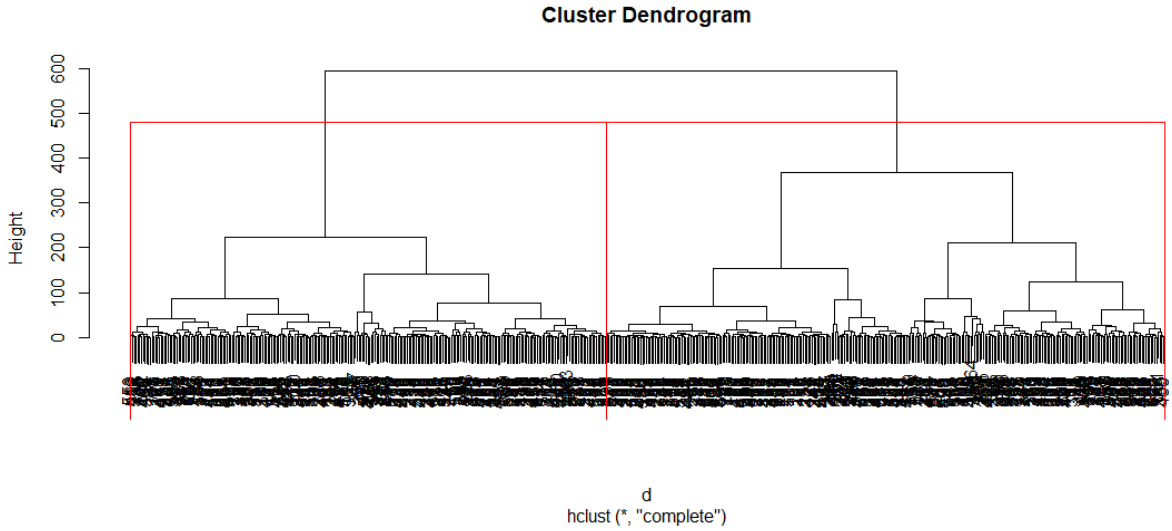
We can clearly see the hierarchy formed to show the various partitions. To get the minimum number of clusters we need to cut the dendrogram at d* .

We choose d* such that a slight change in d, should not lead to a completely different partition. Looking at the largest difference in our observed dendrogram, d* = (220 + 600)/2 = 410

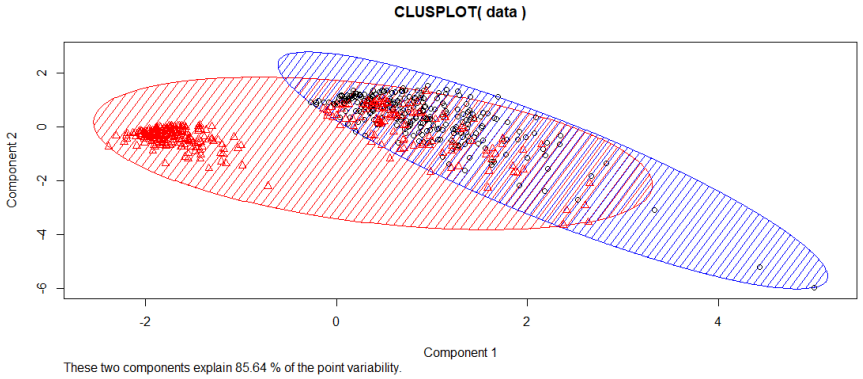
Now we draw a line to see our cluster cut at 410



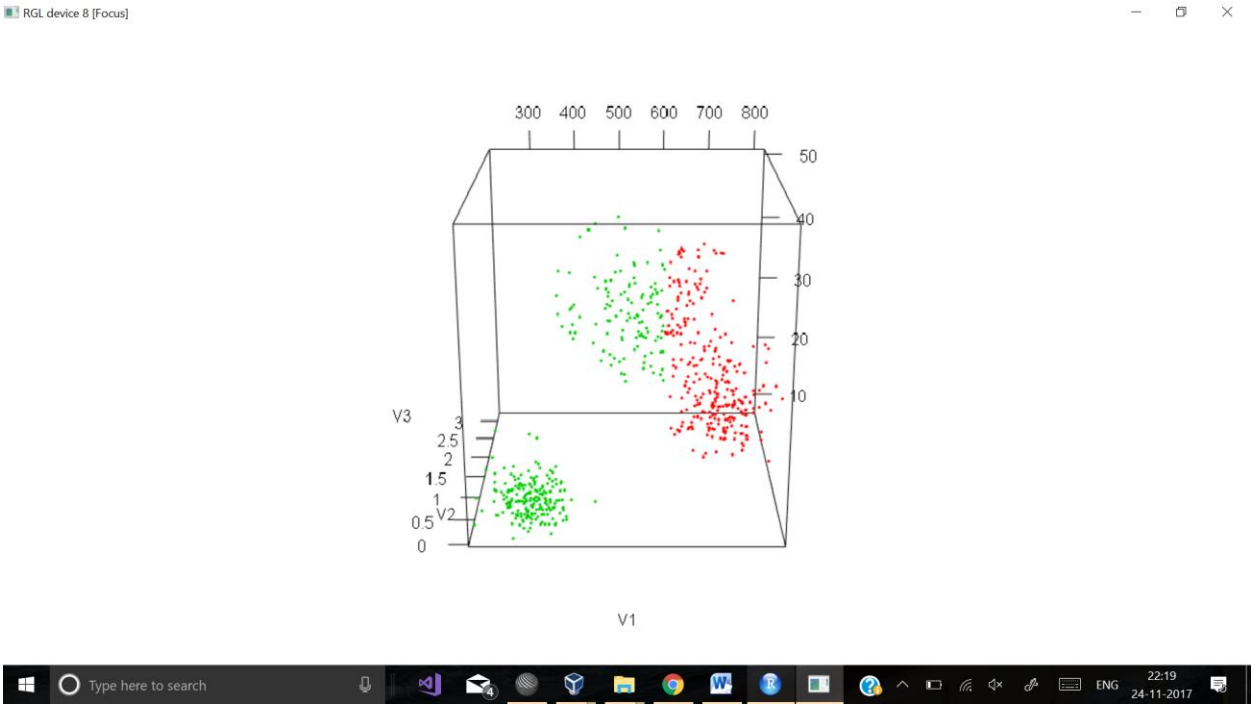
This gives us 2 cluster that we can see below :

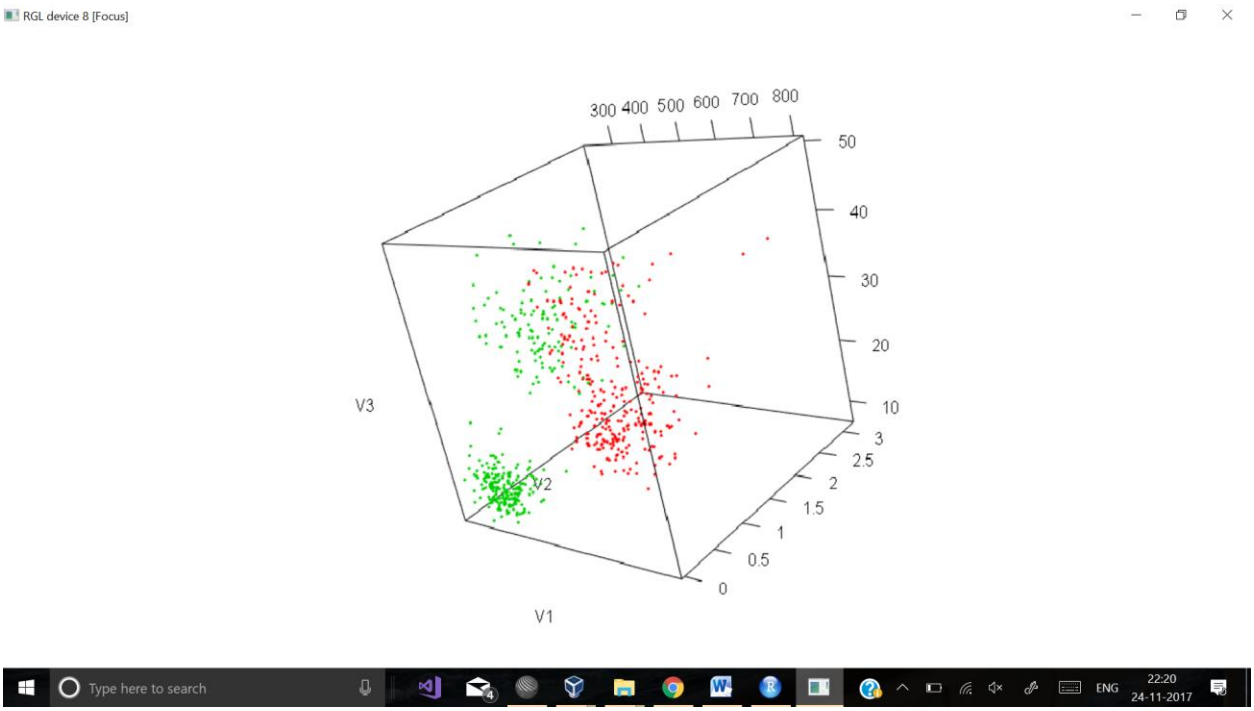


We can get a clear view from the coloured plot and scatter plot :



3d- plot shows the 2 clusters properly :





The RMSE value for this fit of hierarchical clustering is calculated to be 1.205577

K-Means Clustering

K Means Clustering tries to cluster data based on their similarity. We have to specify the number of clusters we want the data to be grouped into and then find the centroid of each cluster. Then, the algorithm iterates through two steps:

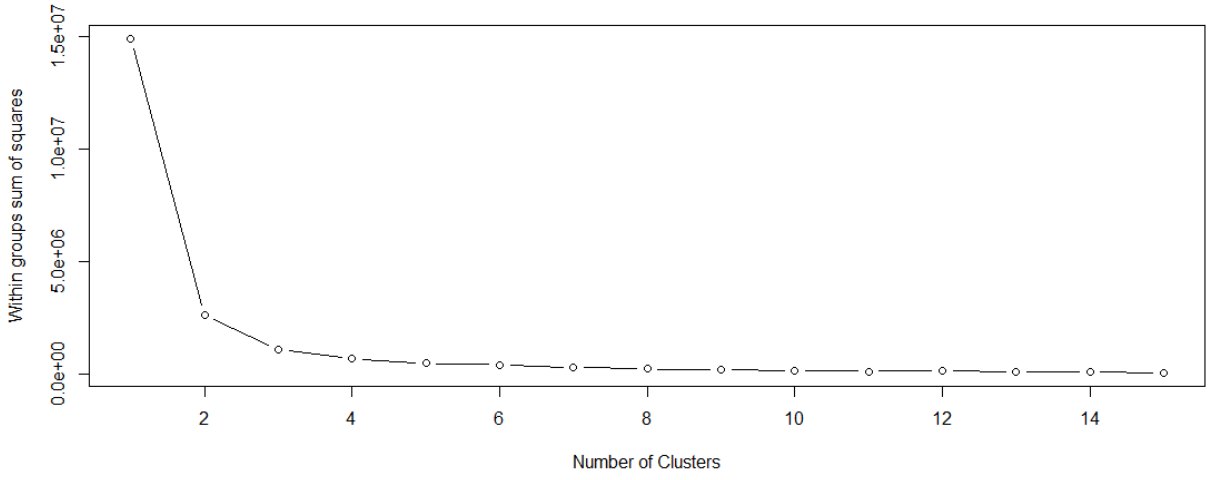
- Reassign data points to the cluster whose centroid is closest.
- Calculate new centroid of each cluster.

These two steps are repeated till the within cluster variation cannot be reduced any further. The within cluster variation is calculated as the sum of the euclidean distance between the data points and their respective cluster centroids.

Determining the best value of k :

The best value of k lies between 1 to n, where n is the total number of observations. We use the elbow method to determine the best value of k.

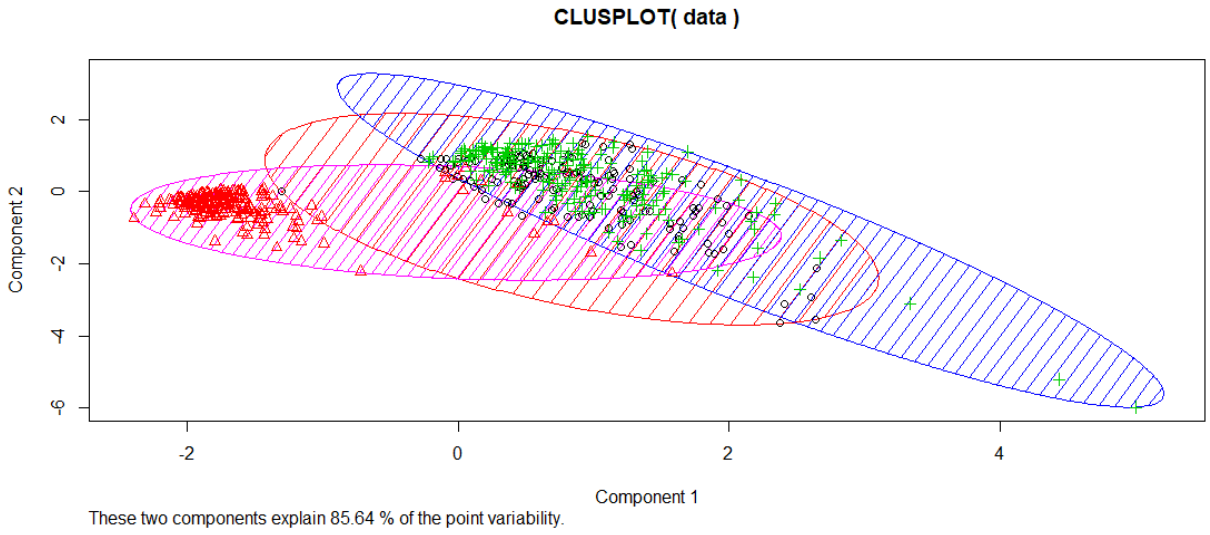
We calculate SSE (sum of squared errors) for various values of k(2 to 15. Then, we plot k vs SSE and elbow shaped graph is formed, such that with increase in the value of k, SSE decreases. We choose the value of k corresponding to the elbow.



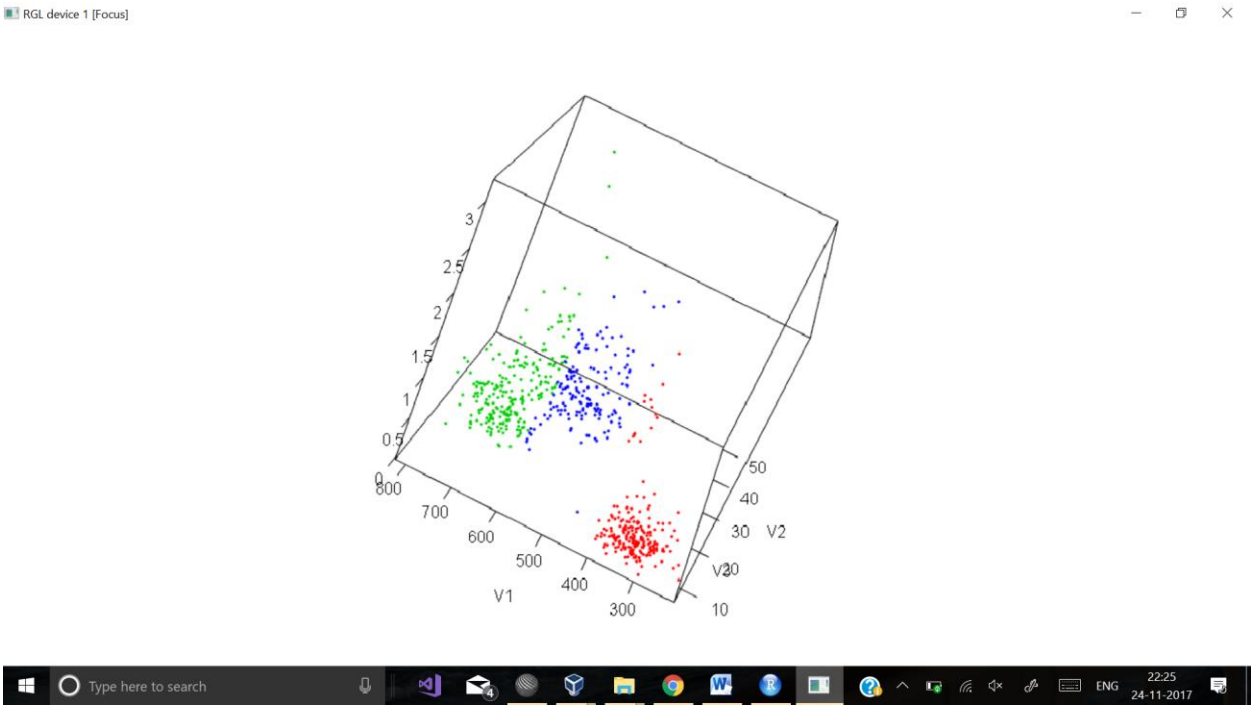
We find an elbow shaped graph and choose k = 3 as the elbow is corresponding to k=3.

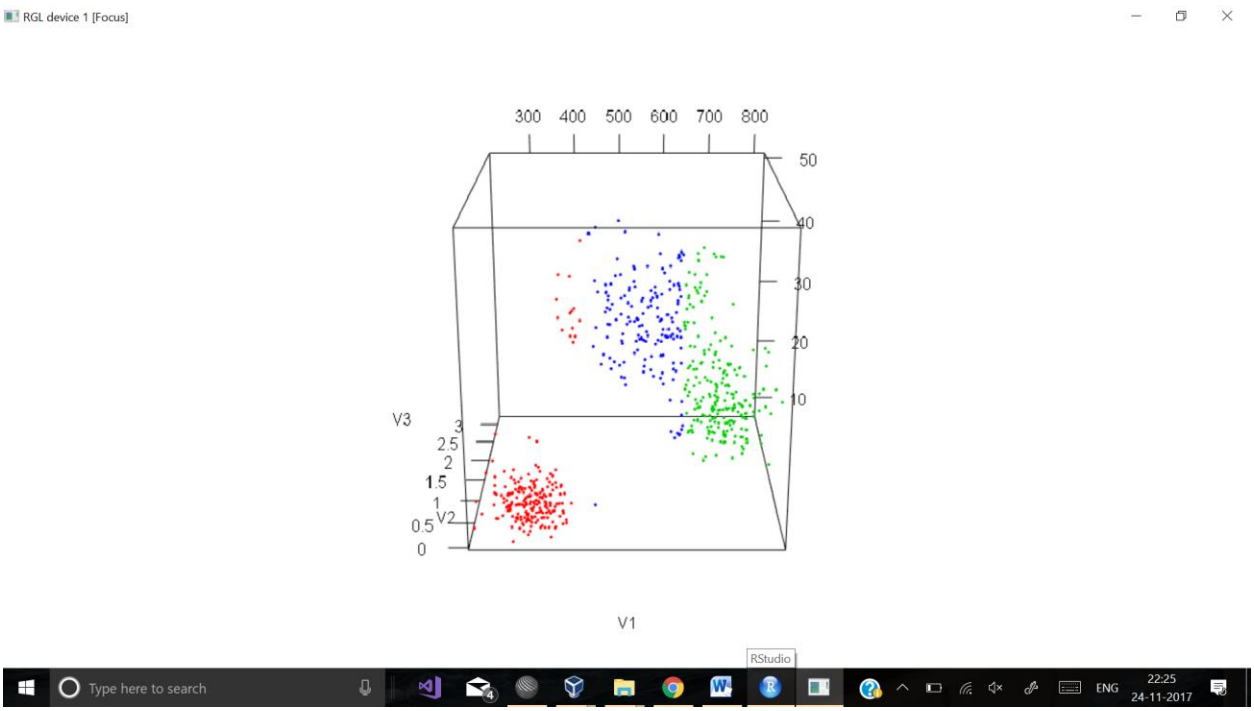
Ususally we don't choose k= 1, as it is trivial solution that there is only 1 cluster. And we don't choose k = 2 as this will again be a dendrogram.

We get three clear clusters shown below :



We can see more clearly in a 3d scatter plot :





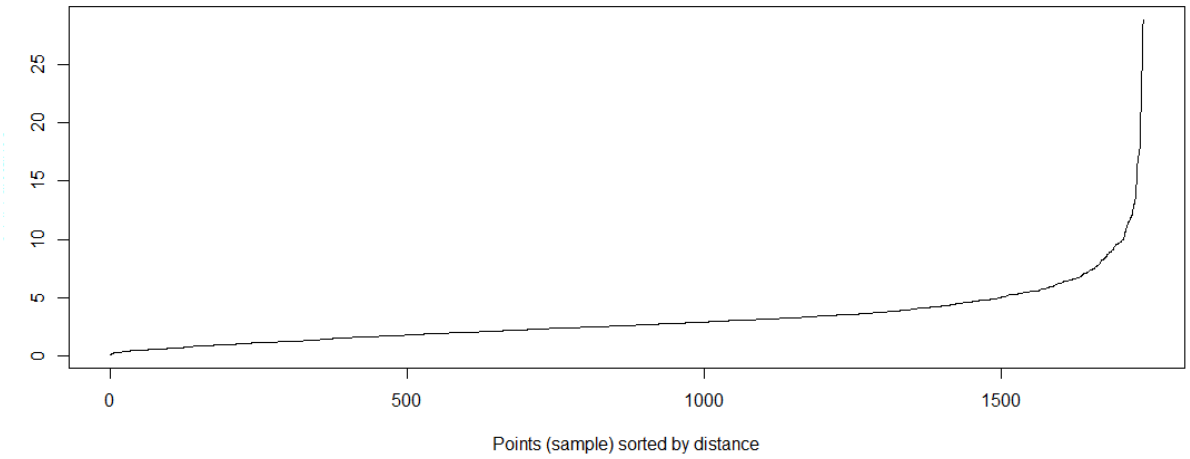
The RMSE value for this clustering is 1.405786

DBScan Clustering

Density-Based Spatial Clustering of Applications with Noise. Finds core samples of high density and expands clusters from them. Good for data which contains clusters of similar density.

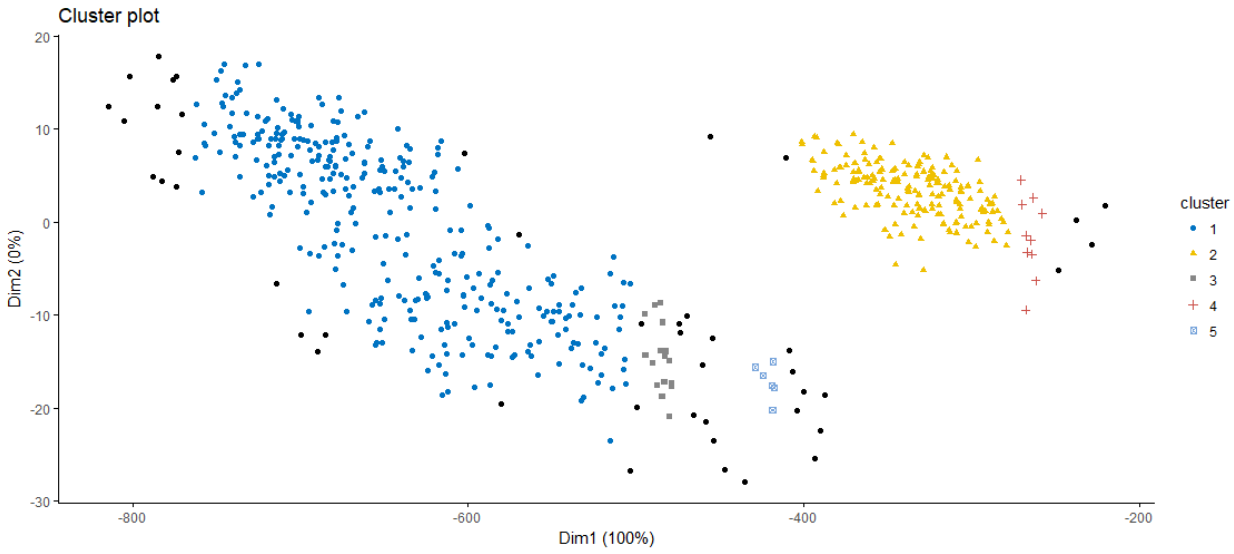
We choose minpoints and radius to find the best clustering.

For k = 3, we plot a graph corresponding to various radii based on the KNN distance.

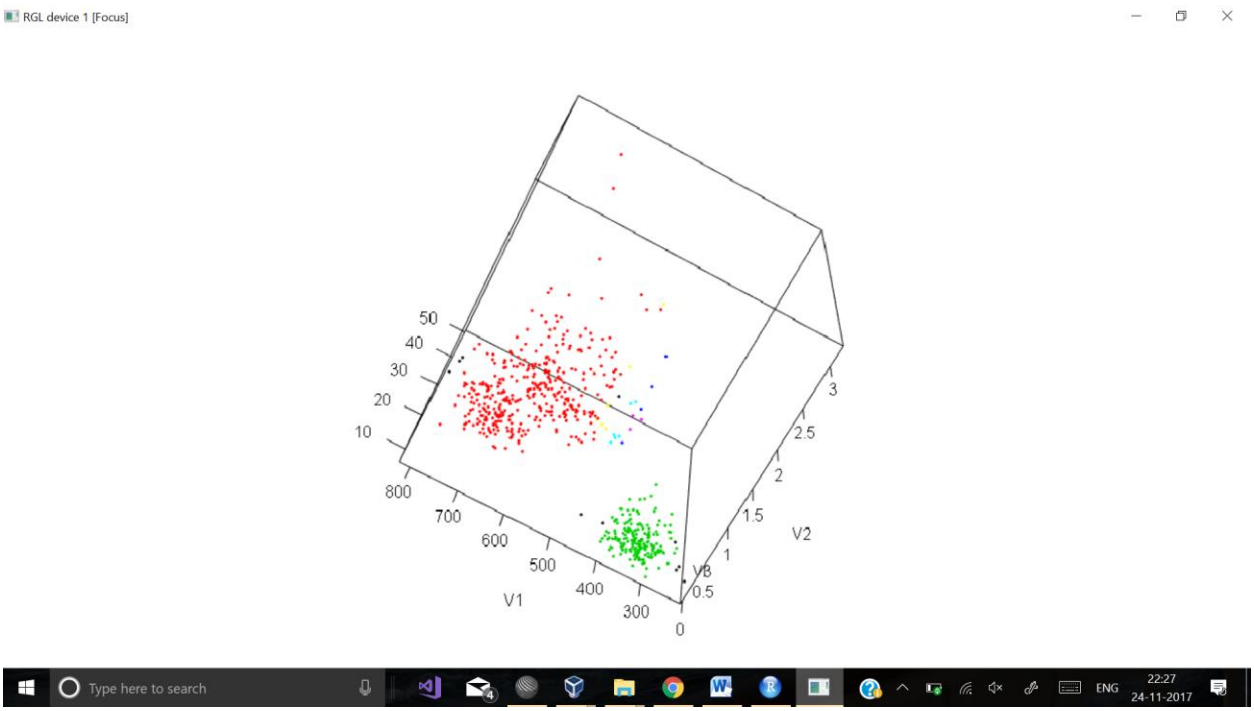


Here we find that value of radius is 7 that we take corresponding to knee.

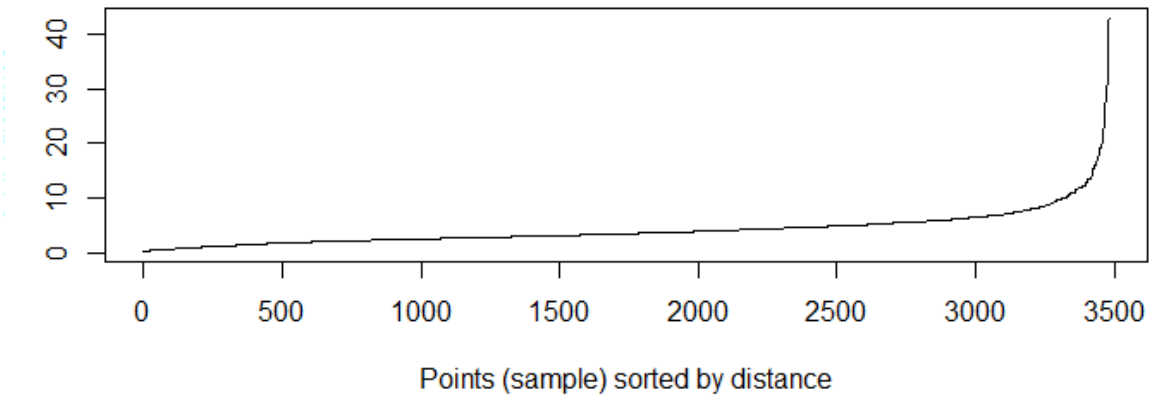
Clustering for k=3 is as follows :



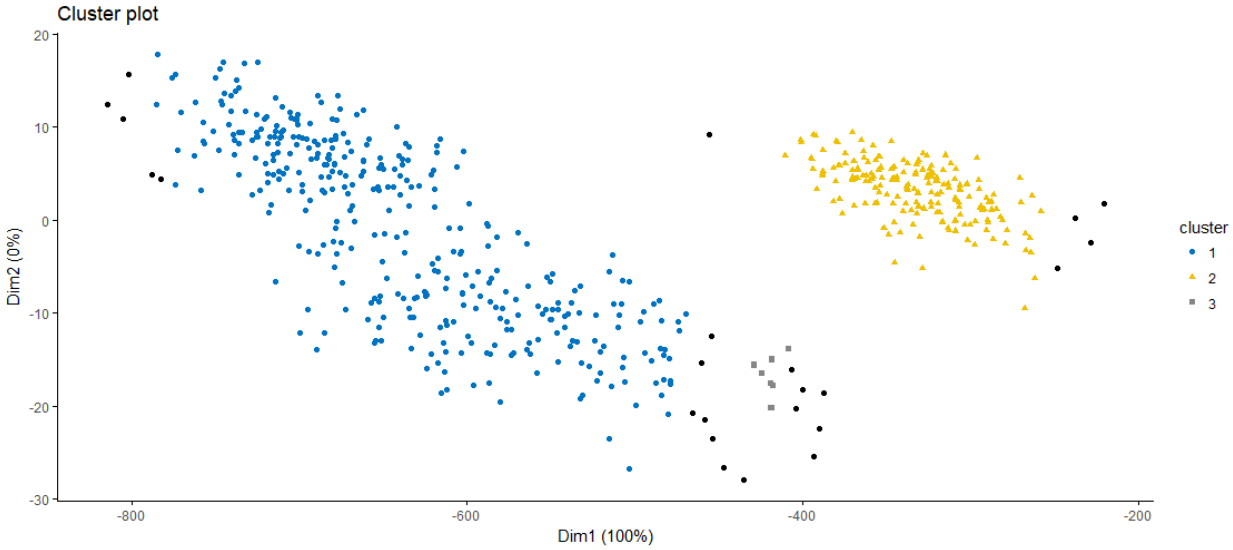
And the 3d plot looks like below



The best fit in case of DBSCAN clustering is found to be for k=6. The radius=10 which is found from the following graph



Three clusters are recognized and found to be as follows :



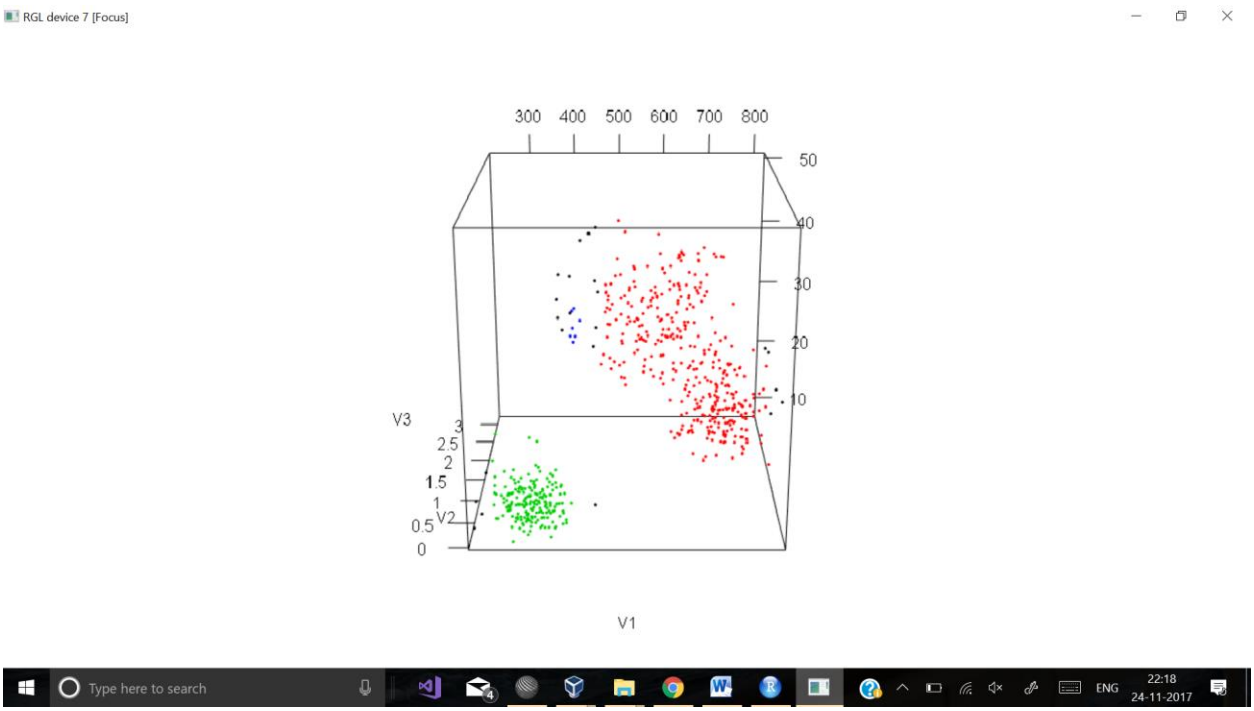
This fit is chosen taking the minimum error into consideration

Error = 1.482094, for k= 3

Error = 1.397276, for k= 4

Error = 1.41881, for k= 5

Error = 1.313662, for k= 6



For the data set given the best clustering is found to be hierarchical clustering as its error is the minimum and considering the the classification and 3D plots, we choose hierarchical clustering as the best methd to classify our data set.

Hierarchical

No. of clusters = 2

Error = 1.25

K-Means

No. of clusters = 3

Error = 1.405786

DBScan

No. of clusters = 3

Error = 1.313662