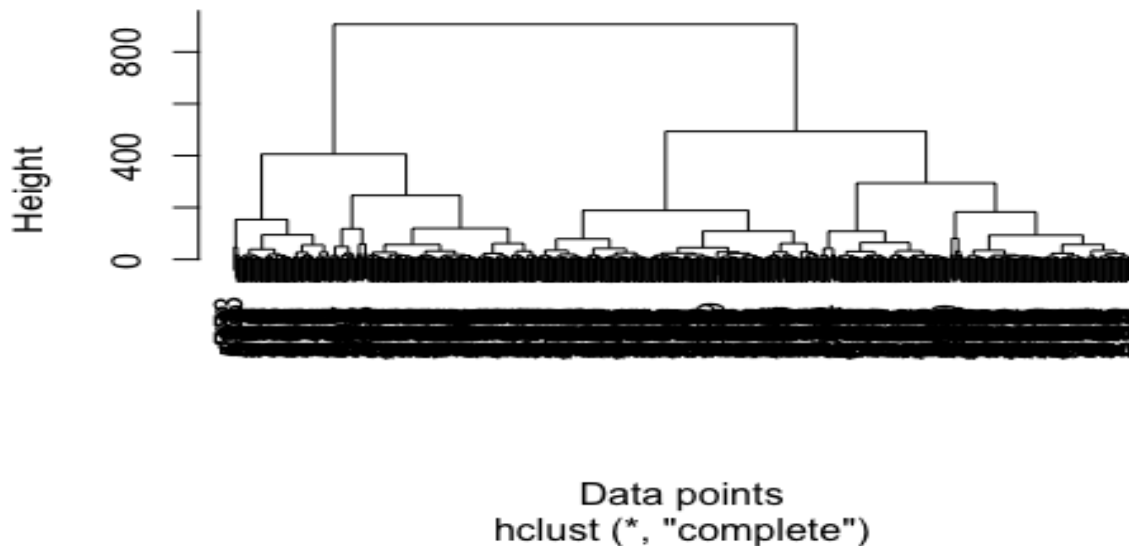


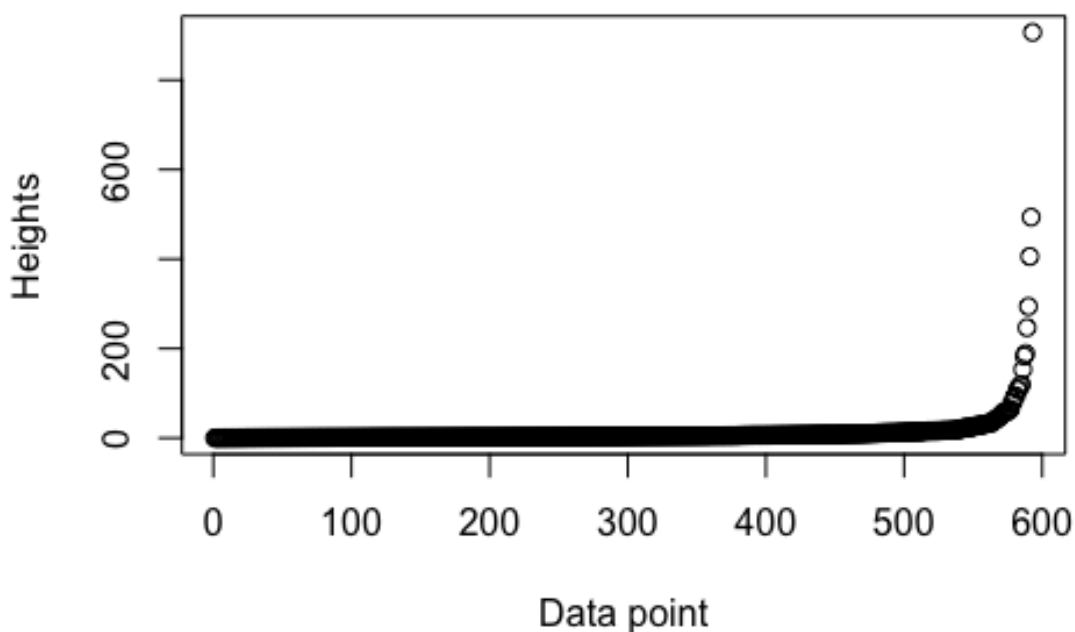
Hierarchical Clustering

1.1

The plot below is based on the Complete Linkage method of Hierarchical Clustering, which is better in terms of cluster creation over Single Linkage method.

Hierarchical Clustering Dendrogram

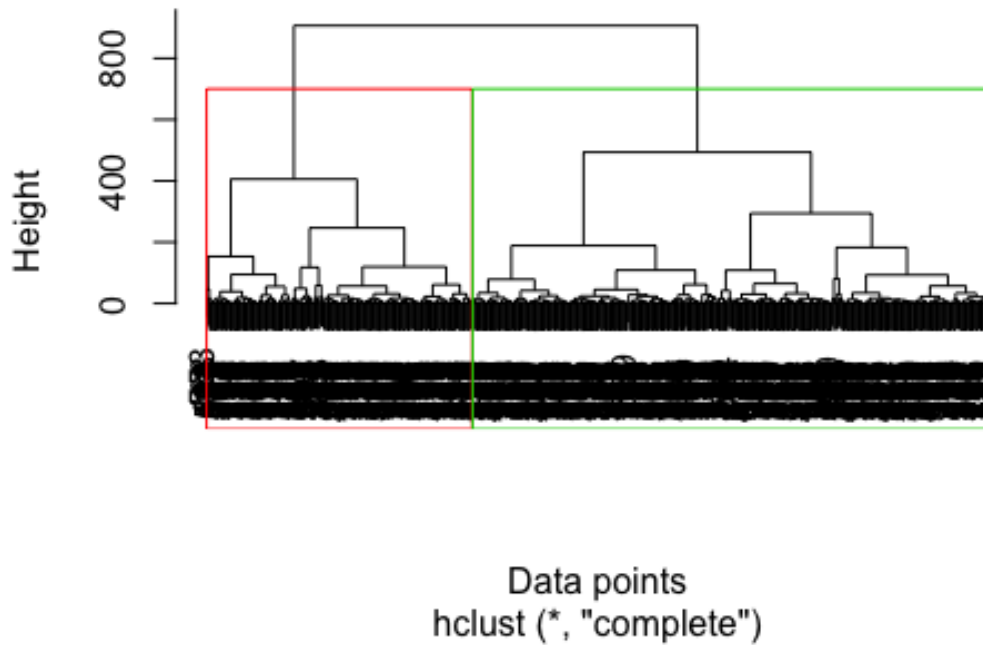
The plot below is for Distance plot between the points after performing the clustering. As per the given distance plot, the maximum distance between heights 493.6553736 and 906.6742595.

Hierarchical Clustering Distance plot

1.2

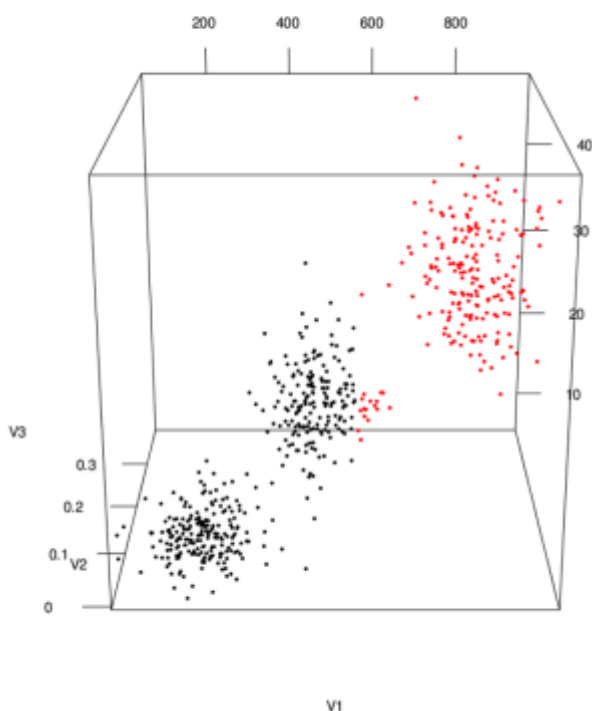
The average between the maximum visible distances: 910(approx.) and 500(approx.) = 705(approx.). Cutting the cluster from there, we get two clusters like below: **Hence, K = 2 clusters.**

Hierarchical Clustering Dendrogram



1.3

The 3D plot of clusters above is created after the cluster is



KMeans Clustering

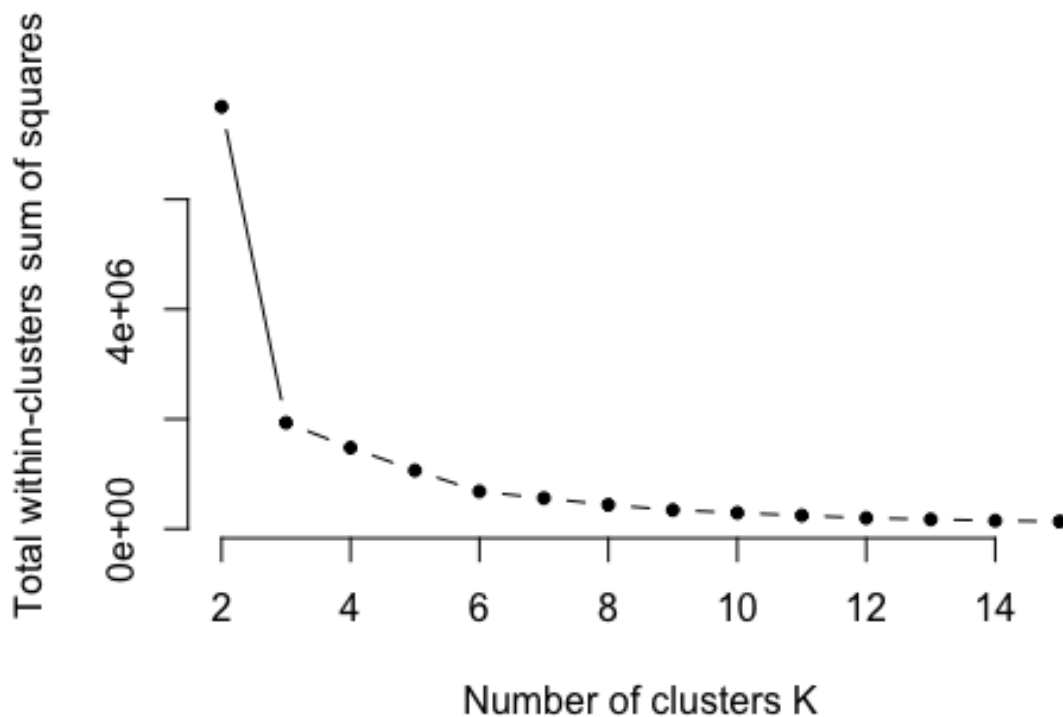
2.1

Table below is the table for representing k vs SSE values:

K	SSE
2	7684590.89335842
3	1933681.13238963
4	1479237.80028585
5	1065488.98471199
6	683164.589616959
7	559427.578834073
8	439073.186403395
9	345079.326210871
10	293506.648669897
11	245235.349948277
12	199234.36877875
13	173454.16380841
14	152431.80293779
15	135171.59087471

2.2

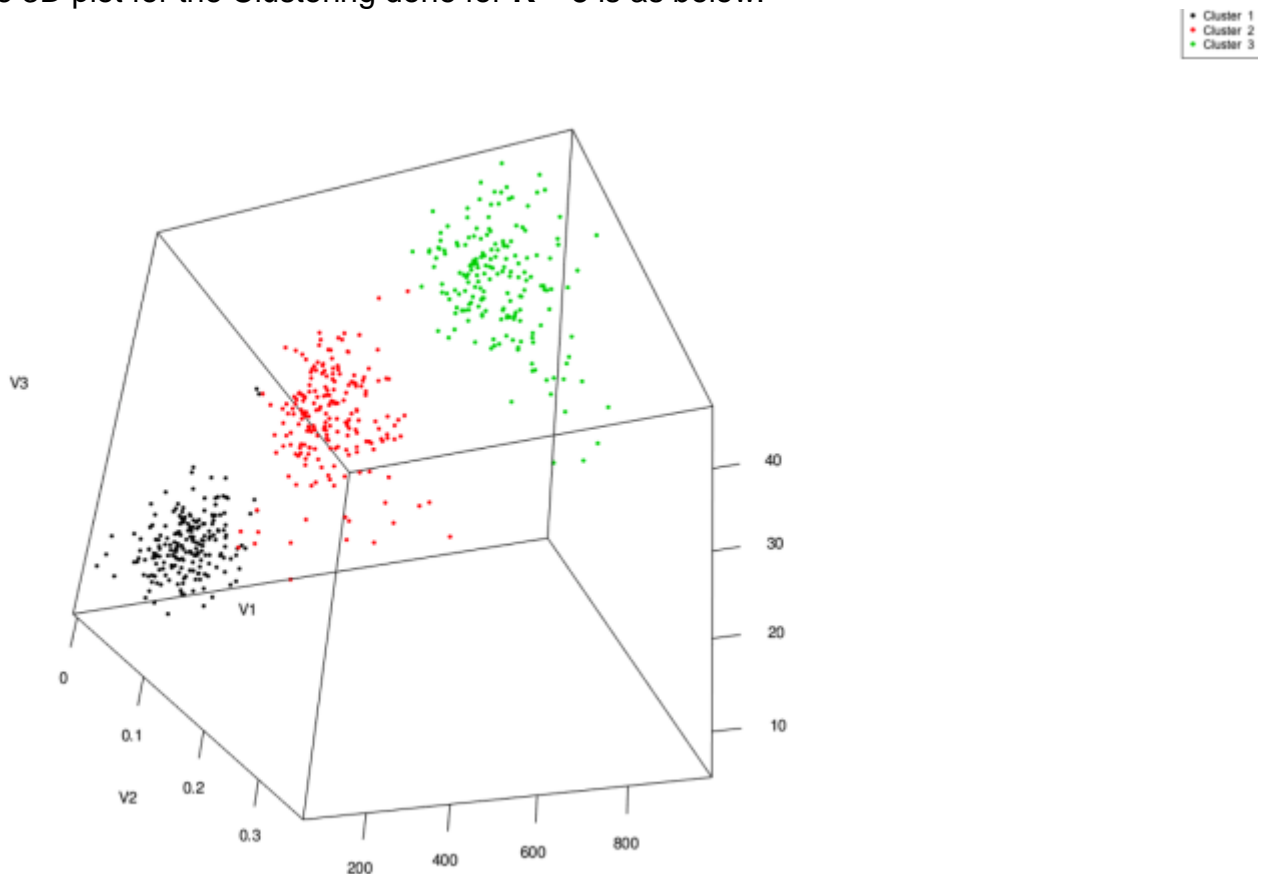
After performing K-Means on various k values, the elbow curve looks like below:



The best value of $k = 3$ after checking the elbow graph and hence, will be used to do clustering.

2.3

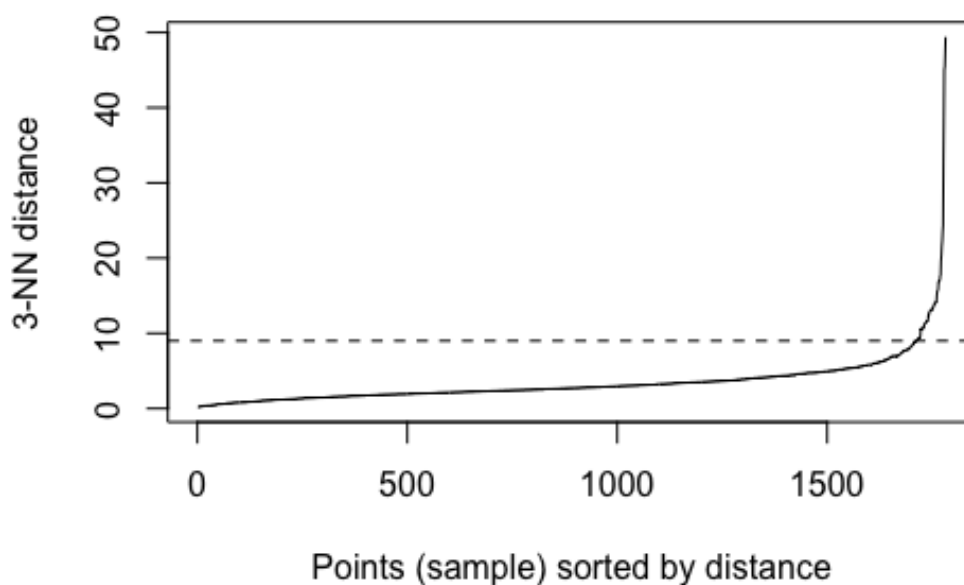
The 3D plot for the Clustering done for $K = 3$ is as below:



DBSCAN

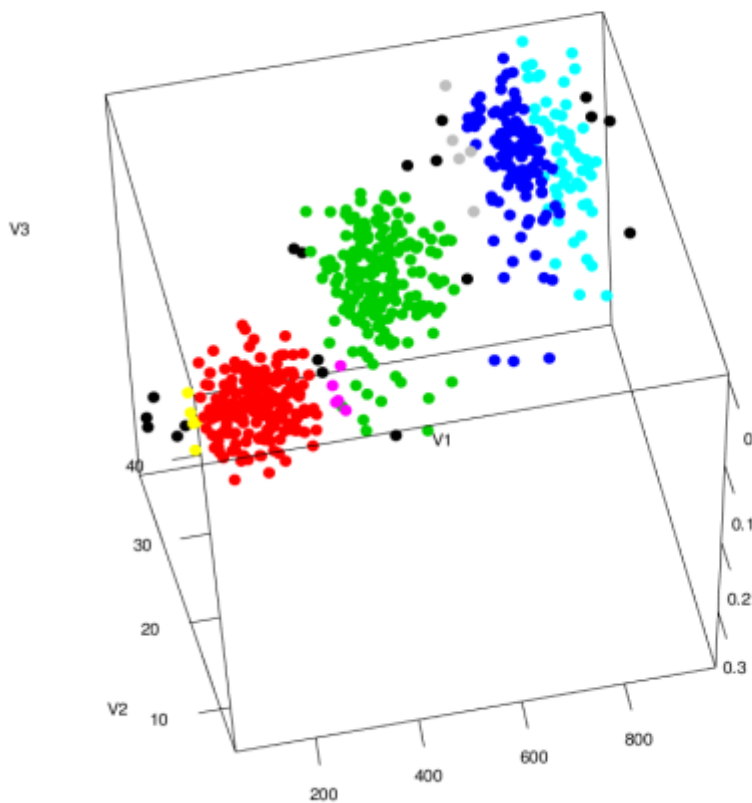
3.1

For $\text{minpts} = 3$, the kNN distance plot for ϵ .



The ϵ comes out to be 9 in this case after carefully observing the kNN plot.

- Cluster 0
- Cluster 1
- Cluster 2
- Cluster 3
- Cluster 4
- Cluster 5
- Cluster 6
- Cluster 7



In this plot, we can observe that there are 8 clusters where one cluster is reserved for outliers.

3.2

The results for the given values of minpts in the questions are in abhardw2_project4_3.2.pdf file.

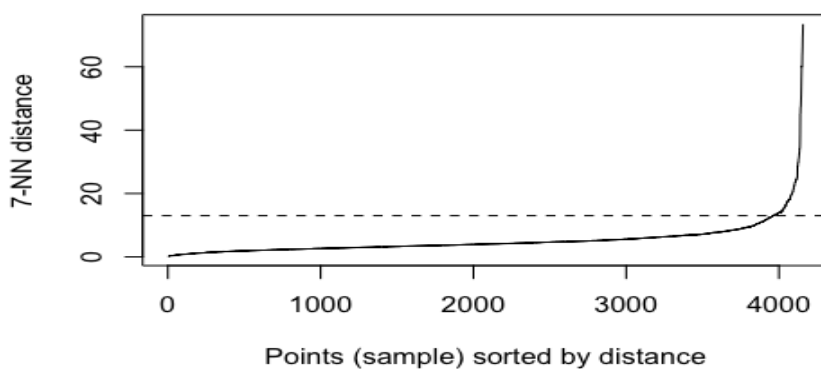
Min Points	ϵ value
4	11
5	10
6	12
7	13

The above values are results from the kNN distance plot done for each min point.

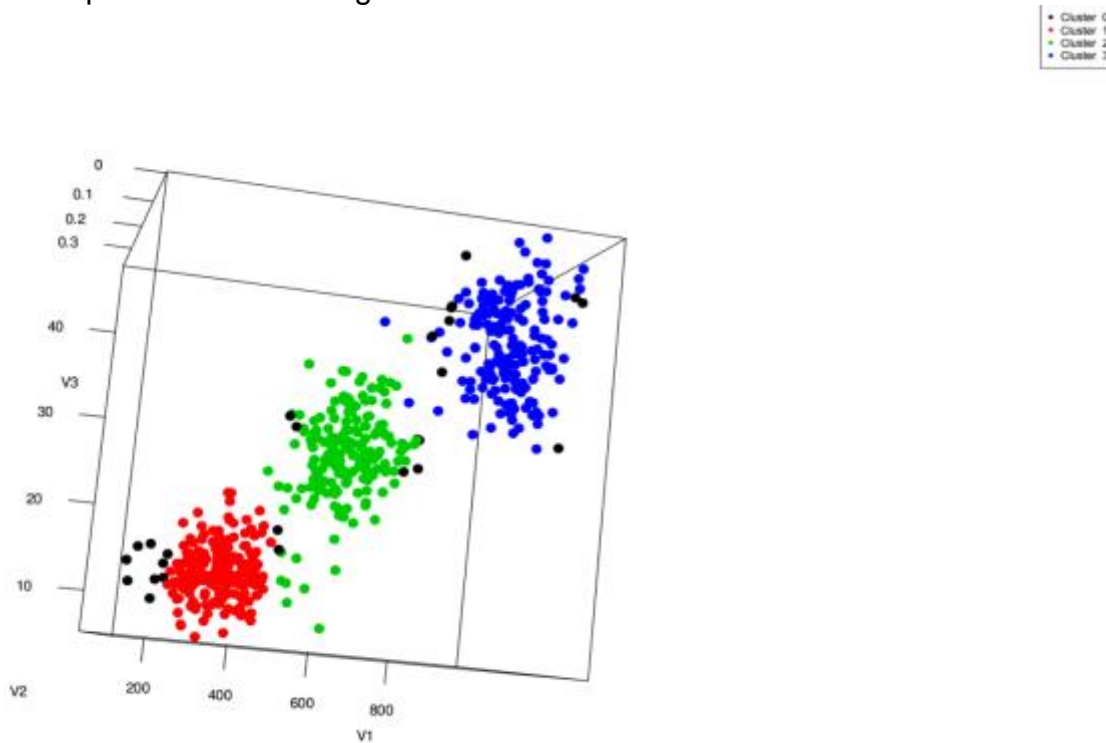
3.3

The best clustering results appears at min points = 7 and $\epsilon = 13$.

The kNN distance plots is as give below:



The 3D plot of the clustering is as below:



This plot contains 3 clusters which seems correct on observing the data plots and it also contains some outliers marked with black color. The outliers are little lesser than the 6NN plot. There are no major changes in the clusters upon increasing the minimum points and ϵ value. So, the values: min pts = 7 and $\epsilon = 13$ is sufficient for clustering in DBSCAN.

4. Compare the results

The comparison results on observing the plots are as follows:

1. The dataset plot clearly shows that there are 3 clusters in the data.
2. **Hierarchical Clustering** produces 2 clusters which are divided as per **complete linkage method** of clustering. The scatter plot on observation shows that there is need for another cluster in order to identify all the data points correctly.
3. **Kmeans Clustering** produces 3 clusters which can be easily visible in the **Elbow Method** plot of the Kmeans algorithm. The data points clustering looks correct and better than Hierarchical clustering where some data points were misclassified.
4. **DBSCAN Clustering** also produces 3 clusters at **MinPts = 7** and $\epsilon = 13$. Here, some points are classified as outliers which looks correct given the plots. This method has marked outliers which could be part of the clusters.

Conclusion:

In the conclusion, both DBSCAN and Kmeans algorithm can be used to create clusters. If we consider the outliers DBSCAN is a better choice over Kmeans, otherwise, to cluster all the data points in at least one cluster then Kmeans also gives better division of points among Clusters.