Hierarchical Clustering:

- It is a method of cluster analysis which seeks to build a hierarchy of clusters.
- Using function hclust() to perform Hierarchical Clustering.
- hclust() - Hierarchical cluster analysis on a set of dissimilarities and methods for analyzing it.

K means clustering:

- It aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
- Using function kmeans() to perform K means clustering.

Density-based Clustering:

- Discovers dense regions in the data space separated by regions of lower object density of arbitrary shape called clusters
- Using function hdbscan() to perform Density-based Clustering.
- HDBSCAN is a clustering algorithm developed by Campello, Moulavi, and Sander. It extends DBSCAN by converting it into a hierarchical clustering algorithm, and then using a technique to extract a flat clustering based in the stability of clusters.

Graph-based Clustering:

- a subset of nodes in a graph such that every two nodes in the subset are connected by an edge.
- Using method sNNclust() to perform Graph-based Clustering.
- sNNclust() – implements the shared nearest neighbor algorithm by Ertoz, Steinbach, Kumar.

Dataset 1:

Hierarchical Clustering without Normalization:

1	2	3	4	5	6	7	8
113	189	159	222	111	54	100	52

Hierarchical Clustering with Normalization:

1	2	3	4	5	6	7	8
91	117	256	131	81	96	127	101

Density based Clustering without Normalization:

HDBSCAN clustering for 1000 objects.

Parameters: minPts = 12

The clustering contains 8 cluster(s) and 340 noise points.

0	1	2	3	4	5	6	7	8
340	28	423	31	24	29	26	42	57

Density based Clustering with Normalization:

HDBSCAN clustering for 1000 objects.

Parameters: minPts = 13

The clustering contains 8 cluster(s) and 403 noise points.

0	1	2	3	4	5	6	7	8
403	23	418	17	20	23	53	25	18

Graph based Clustering without Normalization:

0	1	2	3	4	5	6	7	8
21	122	342	143	67	110	22	20	153

Graph based Clustering with Normalization:

0	1	2	3	4	5	6	7	8
881	16	19	14	13	13	13	13	14

k means Clustering without Normalization:

K-means clustering with 8 clusters of sizes: 102, 100, 152, 116, 120, 171, 117, 122

k means Clustering with Normalization:

K-means clustering with 8 clusters of sizes: 124, 112, 161, 112, 120, 101, 99, 171

Root Mean Square Deviation values:

1) K means Clustering Without Normalization vs Ground Truth labels: 3.538474

2) K means Clustering With Normalization vs Ground Truth labels: 3.587698

3) Graph-Based Clustering Without Normalization vs Ground Truth labels: 3.165789

4) Graph-Based Clustering With Normalization vs Ground Truth labels: 3.263707

5) Hierarchical Clustering Without Normalization vs Ground Truth labels: 3.013222

6) Hierarchical Clustering With Normalization vs Ground Truth labels: 3.325489

7) Density based Clustering Without Normalization vs Ground Truth labels: 3.025773

8) Density based Clustering Without Normalization vs Ground Truth labels: 2.898295

Accuracies:

1) K means Clustering Without Normalization vs Ground Truth labels: 0.133

2) K means Clustering With Normalization vs Ground Truth labels: 0.113

3) Graph-Based Clustering Without Normalization vs Ground Truth labels: 0.109

4) Graph-Based Clustering With Normalization vs Ground Truth labels: 0.125

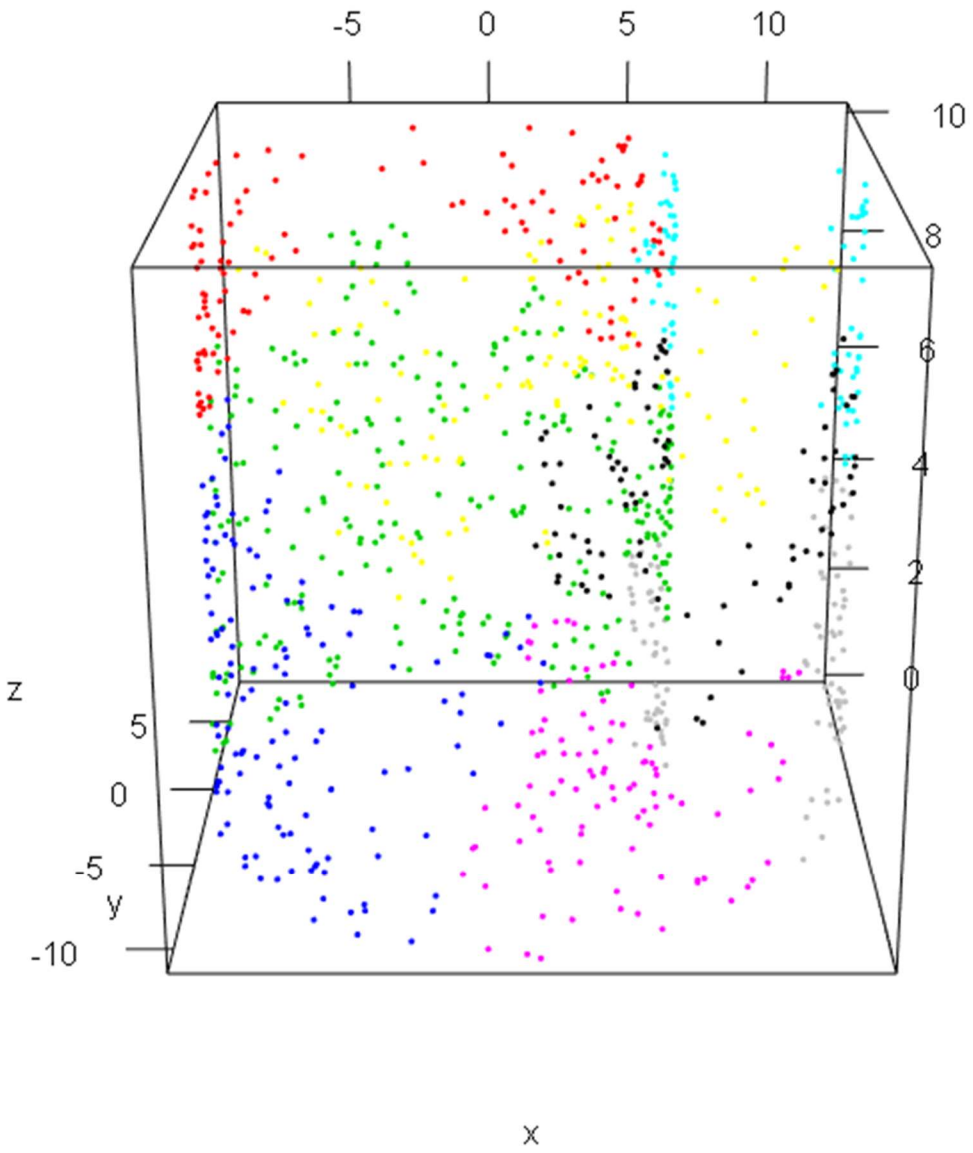
5) Hierarchical Clustering Without Normalization vs Ground Truth labels: 0.11

6) Hierarchical Clustering With Normalization vs Ground Truth labels: 0.141

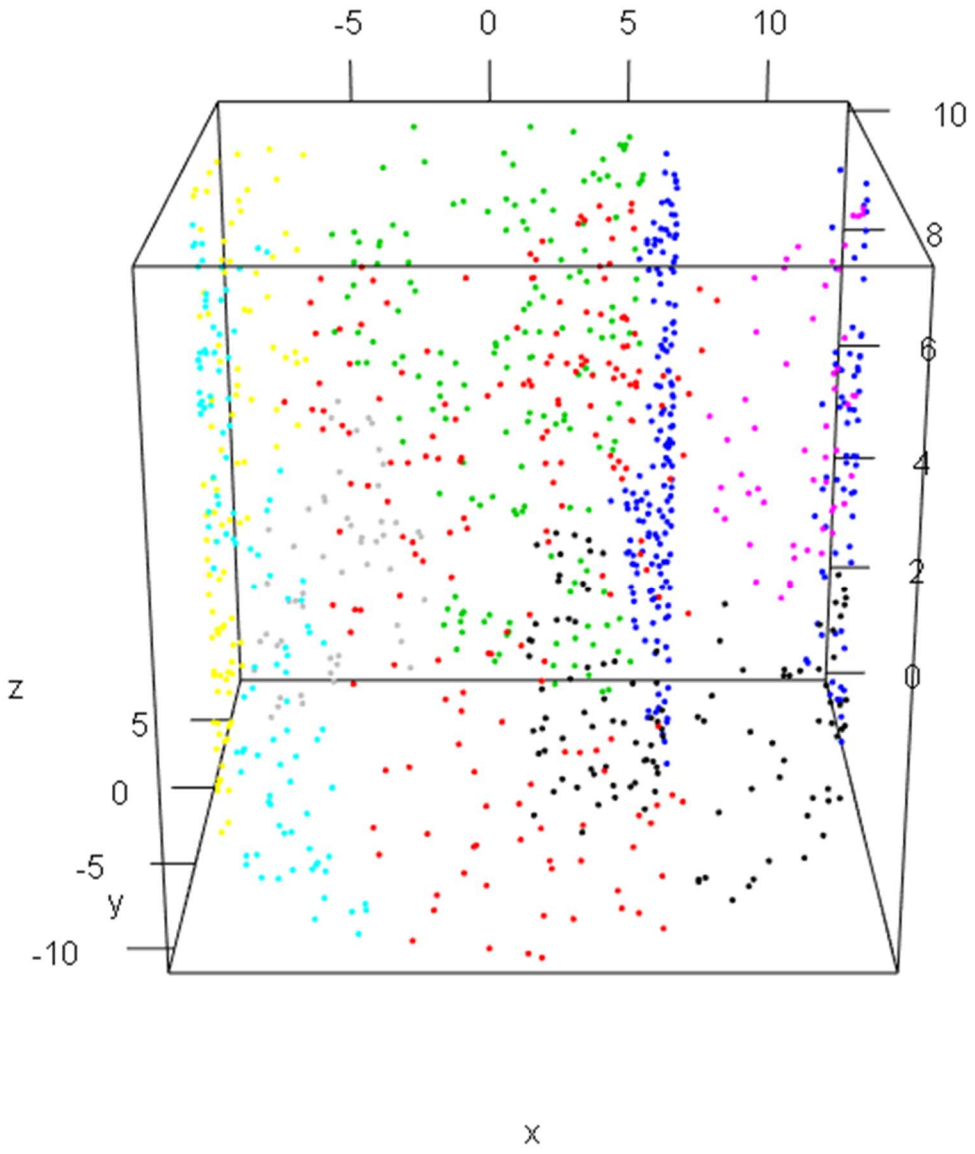
7) Density based Clustering Without Normalization vs Ground Truth labels: 0.113

8) Density based Clustering Without Normalization vs Ground Truth labels: 0.127

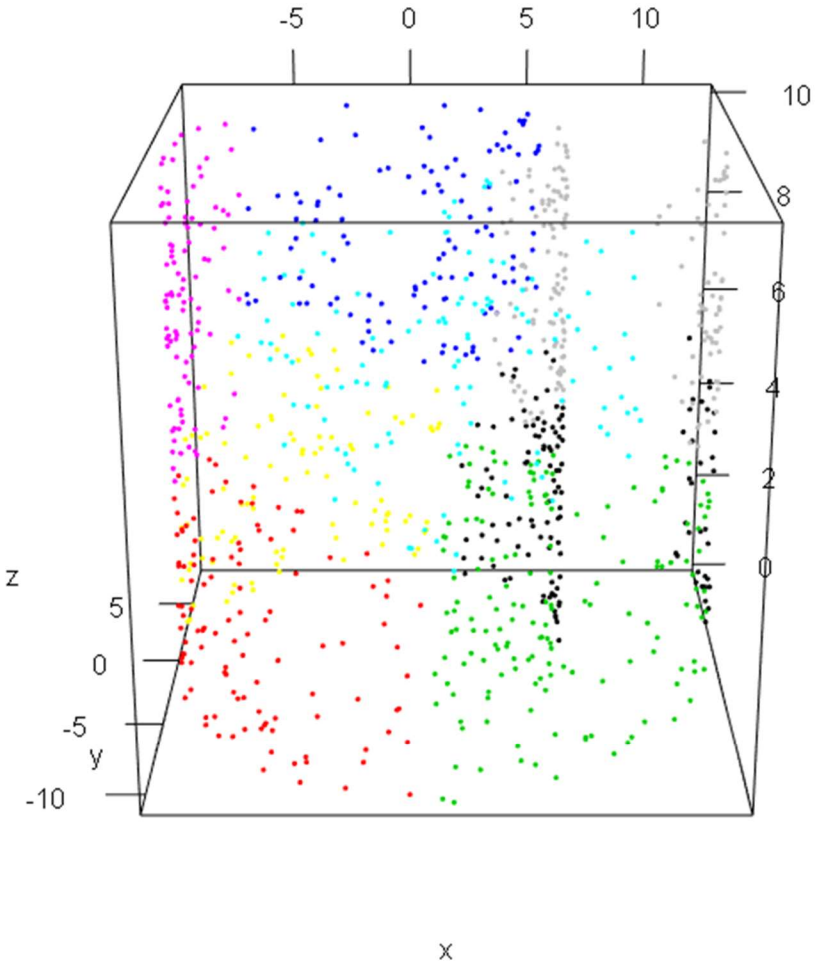
hierarchical clustering with normalization dataset1:



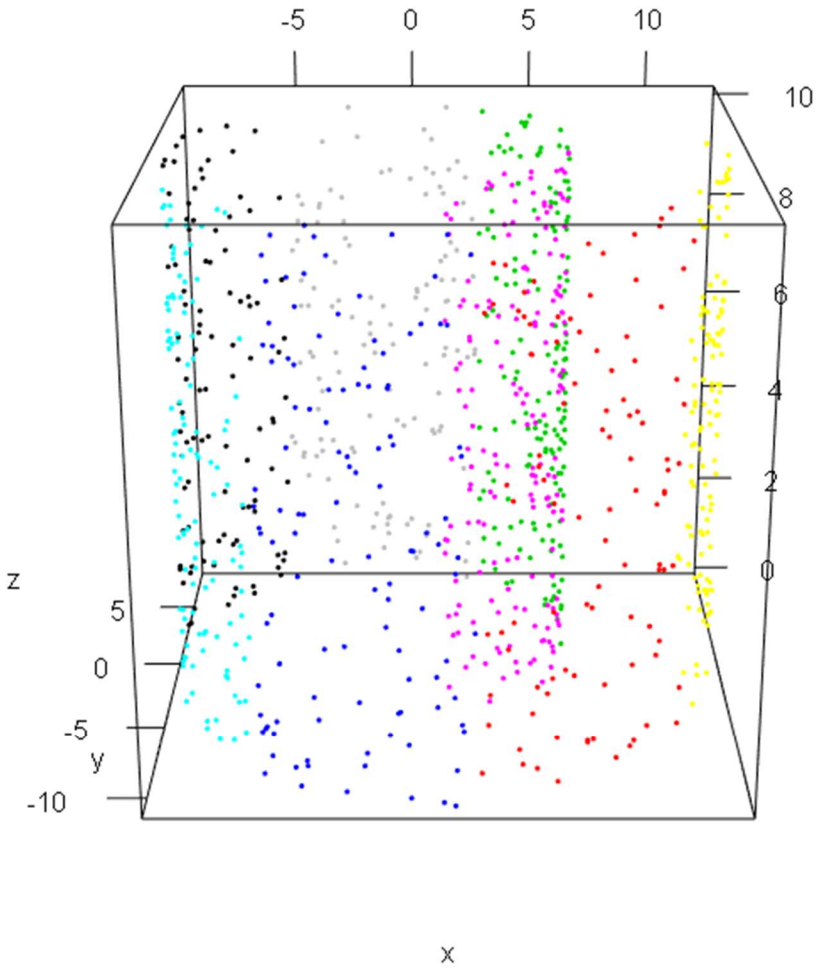
hierarchical clustering without normalization dataset1:



kmeans clustering with normalization dataset1:

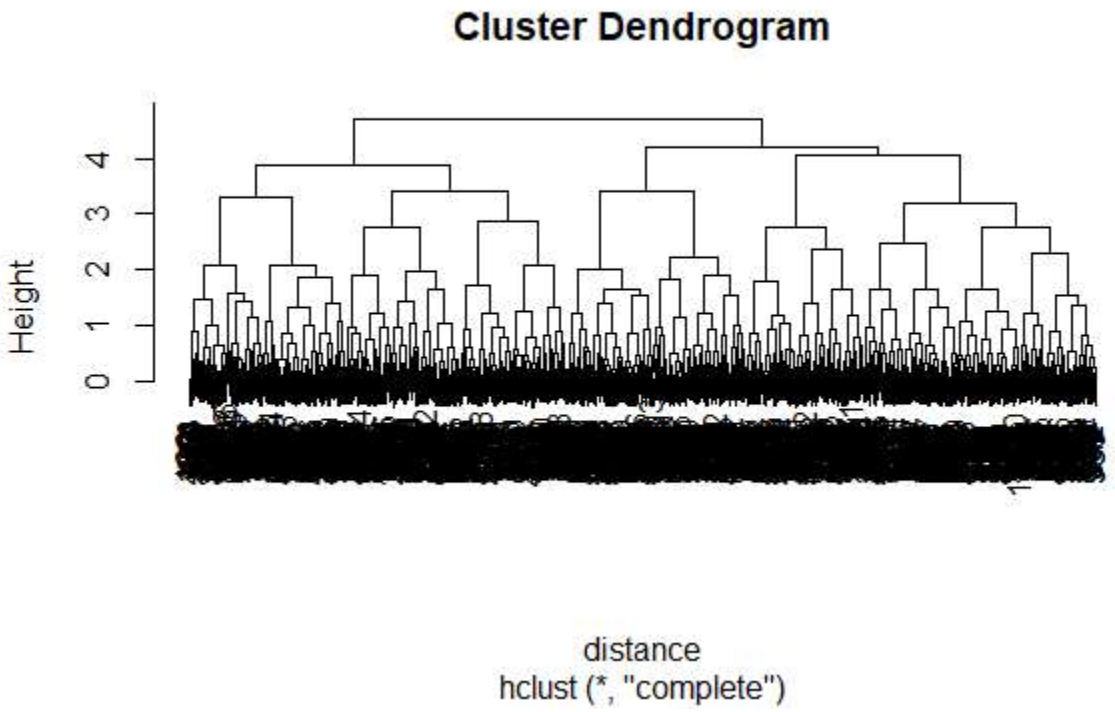


kmeans clustering without normalization dataset1:

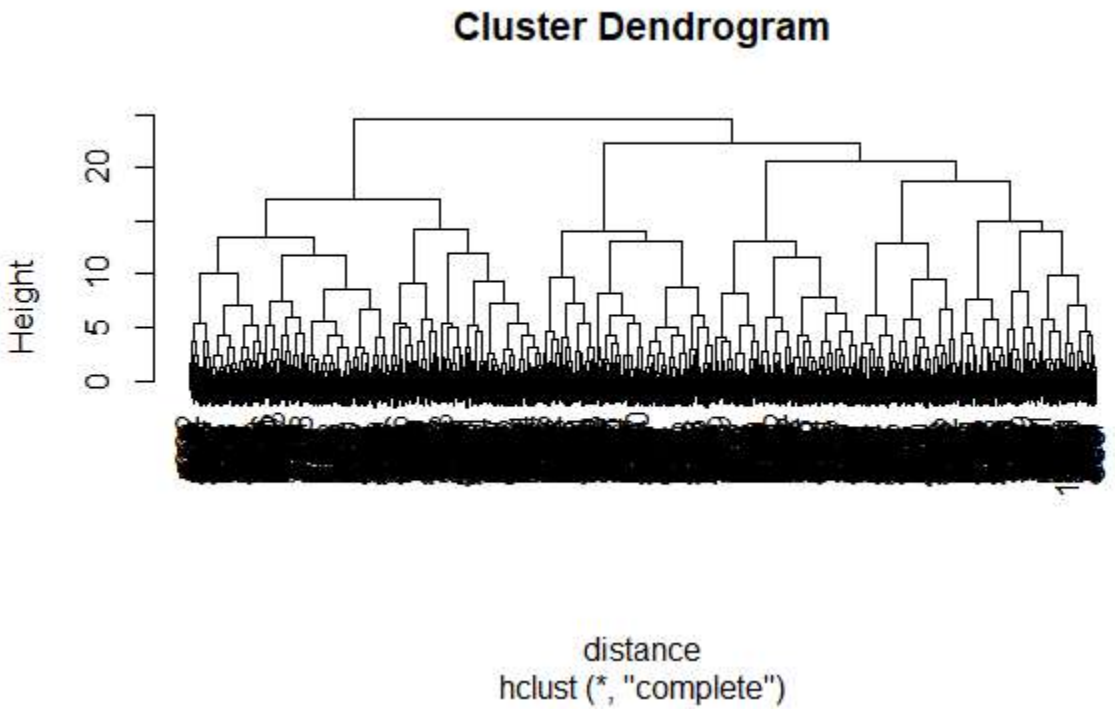


Note: Cannot print Graph Based and Density based Clustering 3d plots because of noise points(Cluster 0).

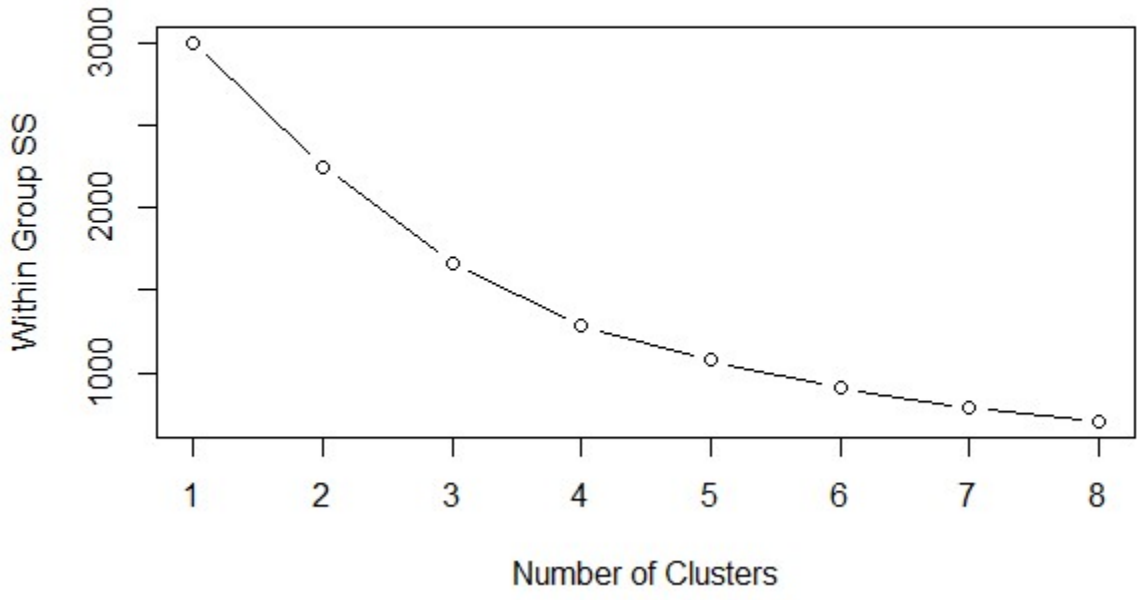
hierarchical clustering with normalization dataset1:



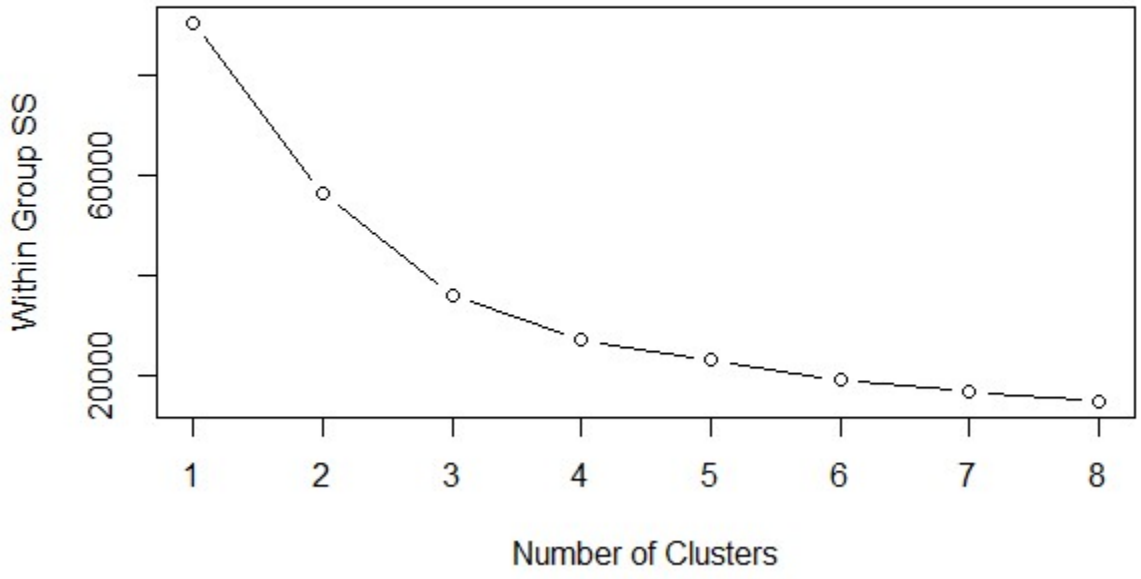
hierarchical clustering without normalization dataset1:



WSS with Normalization:



WSS without Normalization:



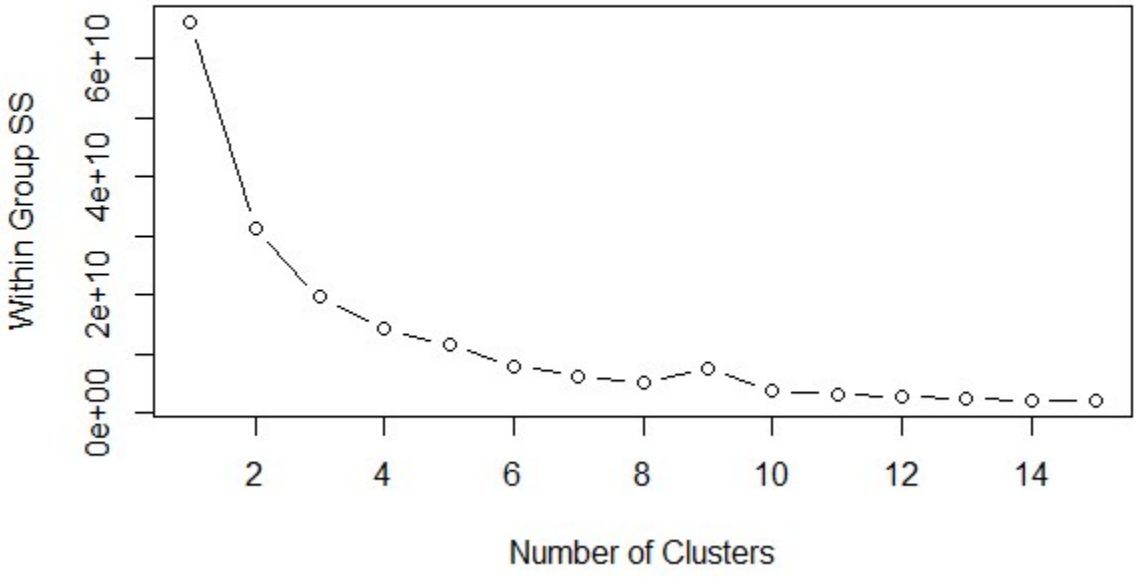
Dataset 2:

How many clusters you think is the best to seek?

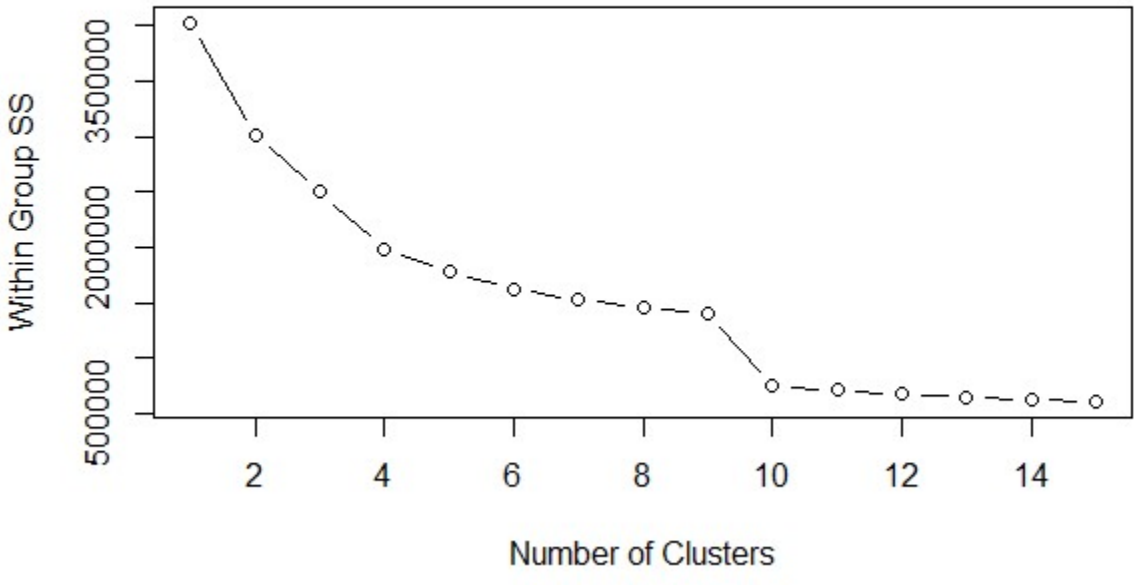
In order to find the no of clusters to use, I have used the **Elbow method** which is often used to identify the optimal number of clusters and it involves observing a set of possible numbers of clusters relative to how they minimize the within-cluster sum of squares. In other words, the Elbow method examines the **within-cluster dissimilarity** as a function of the number of clusters.

In order to implement it I have tried cluster sizes from 2 to 15 and plotted the values on the graph. In the process I used **nstart option set to 3** so that the starting position is different each time and best value among them is taken. Below was the graph for nstart = 3. **(PS – I have changed the nstart option to 1 again now in the script as it takes time to get results using higher values of nstart)**

WSS Without normalization:



WSS With Normalization:

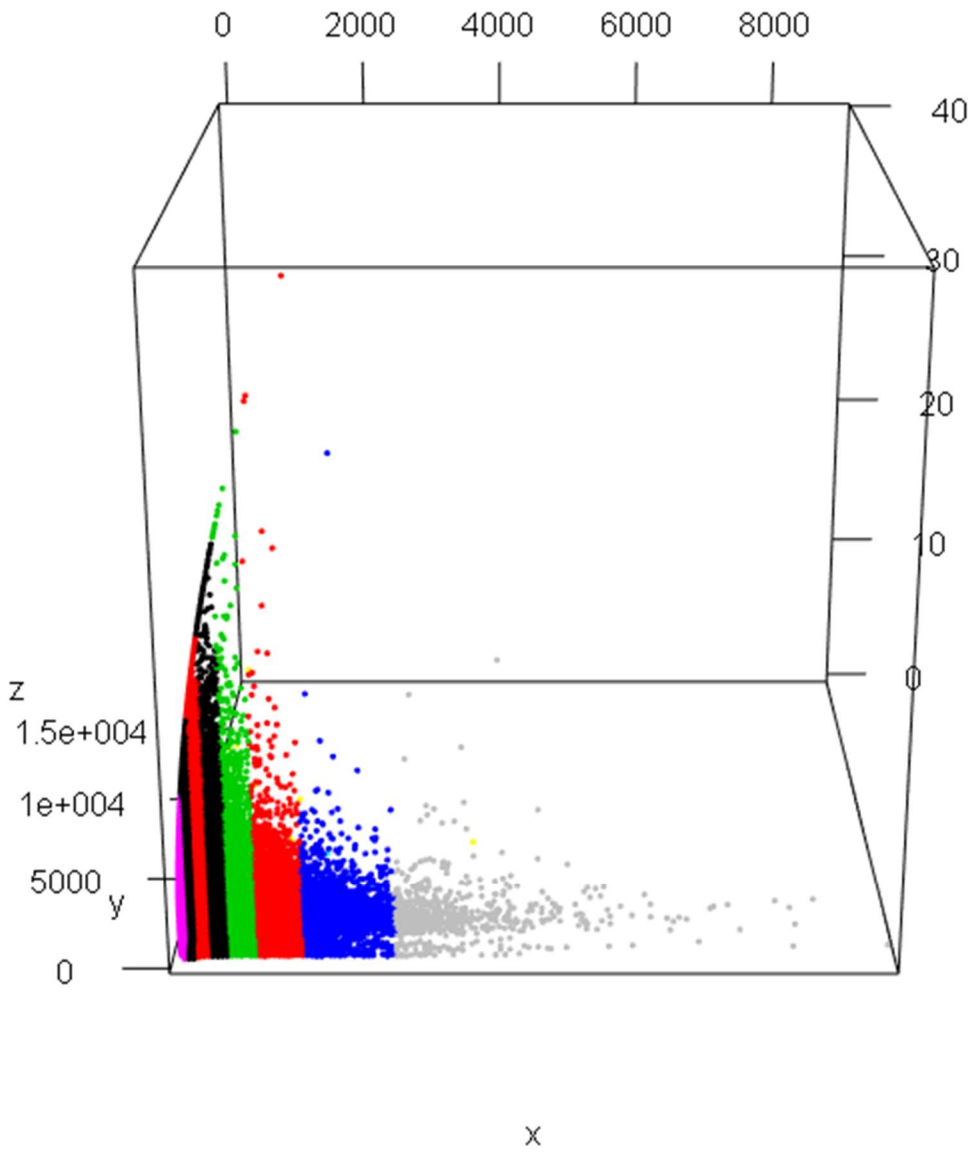


From both the graphs we can say that after 10 clusters the observed difference in the within-cluster dissimilarity is not substantial. Consequently, we can say with some reasonable confidence that the optimal number of clusters to be used is 10. Hence, I have used cluster value of 10 in the final r script submitted.

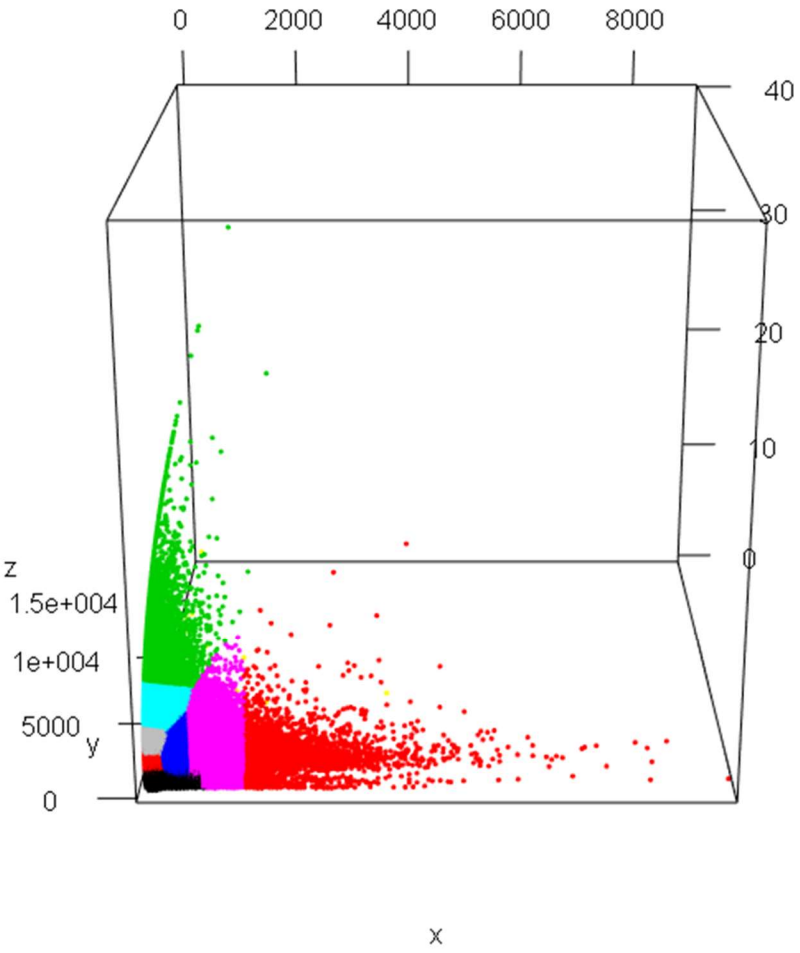
Type of clustering used:

I have only applied the k means clustering technique for want of memory for the other clustering methods(**Error: cannot allocate vector of size 3771.5 Gb**).

kmeans clustering without normalization dataset2:



kmeans clustering with normalization dataset2:



Conclusion:

1) For Dataset 1, almost all the clustering techniques have pretty much similar accuracies between 0.1 and 0.15 with Hierarchical Clustering With Normalization having the highest accuracy(0.141) and Graph-Based Clustering Without Normalization having the lowest accuracy(0.109)
2) We found that there are limitations on which clustering techniques can be used depending on the size of data.
3) We also found out that the clustering results are neither very accurate nor very consistent.

References:

1) http://planspace.org/2013/02/03/pca-3d-visualization-and-clustering-in-r/
2) https://www.youtube.com/watch?v=5eDqRysaico
3) https://cran.r-project.org/web/packages/dbscan/README.html