EAS 507: STATISTICAL DATA MINING II

Report for Homework #2

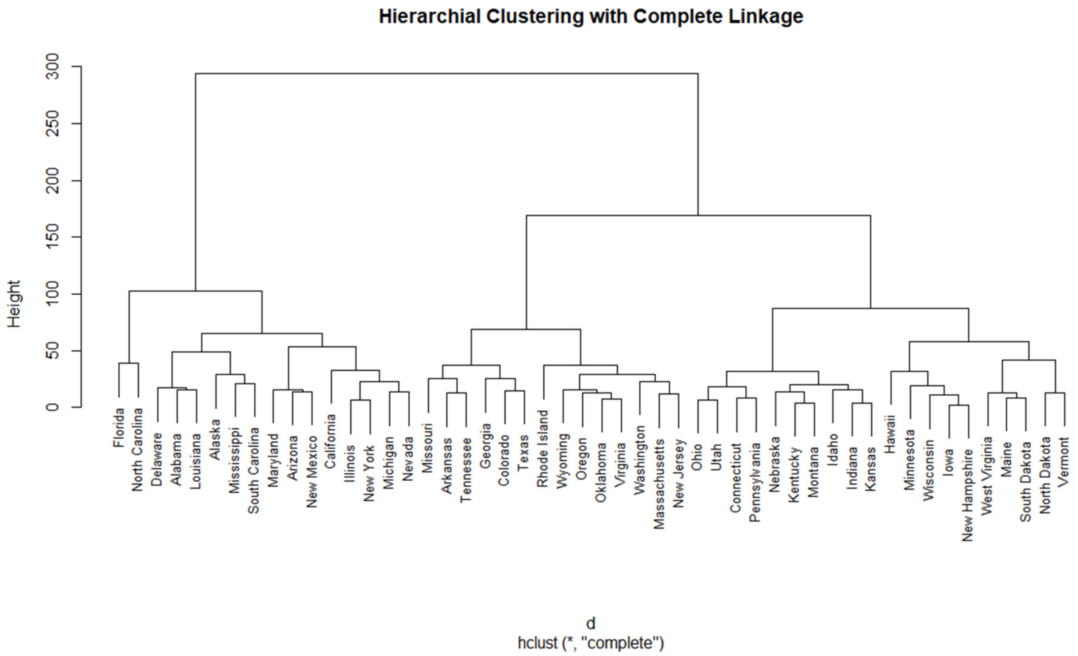
Kishore Ravisankar
UBIT: kravisan

1.

a)

The given dataset USArrests contains the statistics of arrests on counts of assault, murder and rape, per 100,000 residents in USA, for the year 1973. It also contains the population of each state. There are 50 observations, one pertaining to each state, and four predictors.

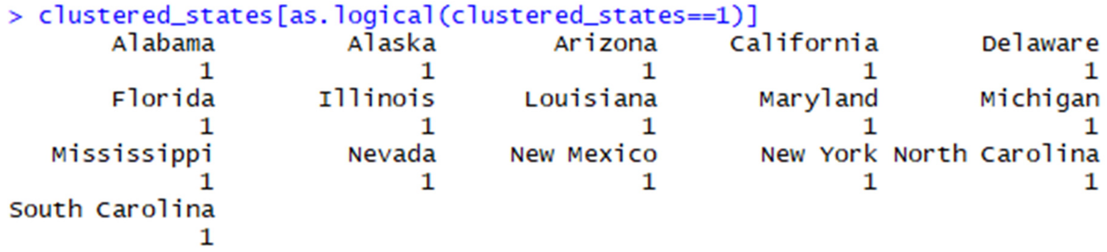
The distance matrix was calculated using Euclidean distance, and hierarchical clustering with complete linkage was performed. The states are clustered as follows:



b)

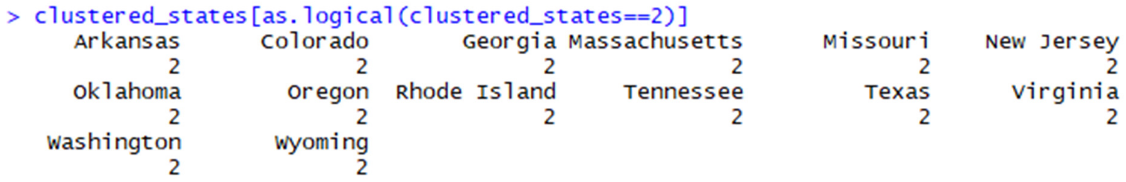
We see that at a height of approximately 150, we can obtain three clusters of the states.

The states that belong to cluster 1 are:



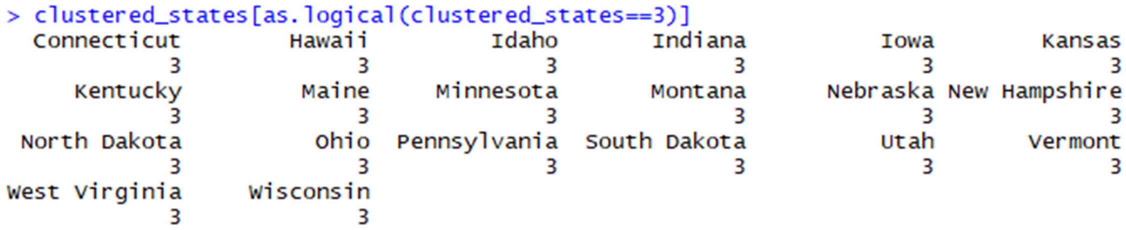
```
> clustered_states[as.logical(clustered_states==1)]
       Alabama         Alaska        Arizona     California       Delaware 
             1              1              1              1              1 
       Florida       Illinois      Louisiana       Maryland       Michigan 
             1              1              1              1              1 
   Mississippi         Nevada     New Mexico       New York North Carolina 
             1              1              1              1              1 
South Carolina 
             1 
```

The states that belong to cluster 2 are:



```
> clustered_states[as.logical(clustered_states==2)]
     Arkansas      Colorado       Georgia Massachusetts      Missouri    New Jersey 
            2             2             2             2             2             2 
     Oklahoma        Oregon  Rhode Island     Tennessee         Texas      Virginia 
            2             2             2             2             2             2 
   Washington       Wyoming 
            2             2 
```

The states that belong to cluster 3 are:

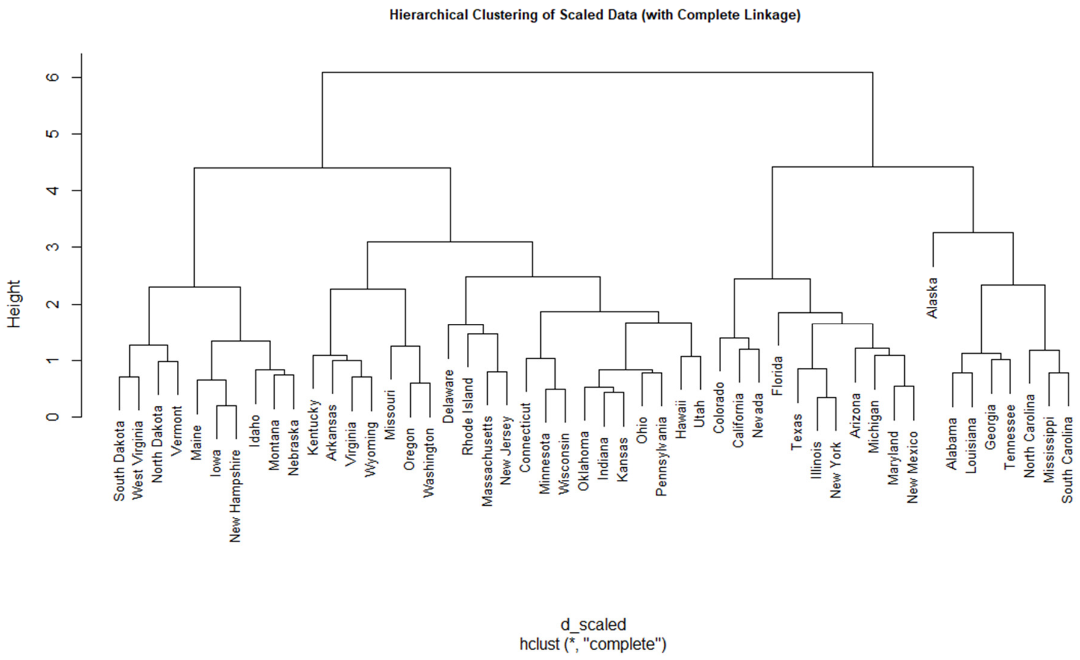


```
> clustered_states[as.logical(clustered_states==3)]
  Connecticut        Hawaii         Idaho       Indiana          Iowa        Kansas 
            3             3             3             3             3             3 
     Kentucky         Maine     Minnesota       Montana      Nebraska New Hampshire 
            3             3             3             3             3             3 
 North Dakota          Ohio  Pennsylvania  South Dakota          Utah       Vermont 
            3             3             3             3             3             3 
West Virginia     Wisconsin 
            3             3 
```

c)

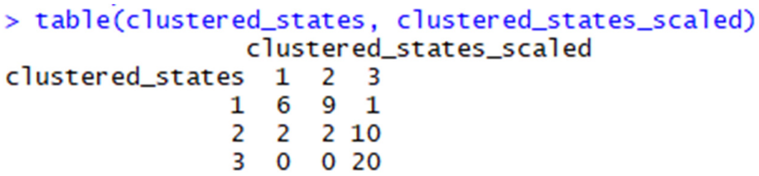
The data was scaled to have a mean of zero across all columns, and have a standard deviation of 1.

The dendrogram obtained is as follows:



d)

The clusters that the states belong to, before and after scaling the data, can be figured out by plotting a confusion matrix.



```
> table(clustered_states, clustered_states_scaled)
                 clustered_states_scaled
clustered_states  1  2  3
               1  6  9  1
               2  2  2 10
               3  0  0 20
```

We can see that only 56% of the data is clustered in the same groups before and after scaling the data. However, we can observe that the dendrograms obtained are very similar to each other.

The data should be scaled before computing the distance matrix, because it helps in obtaining the same scale of values across different columns. This helps in obtaining a distance matrix with equal contribution from various predictors. By not scaling the data, the data values which are high in magnitude will dominate the distance matrix.

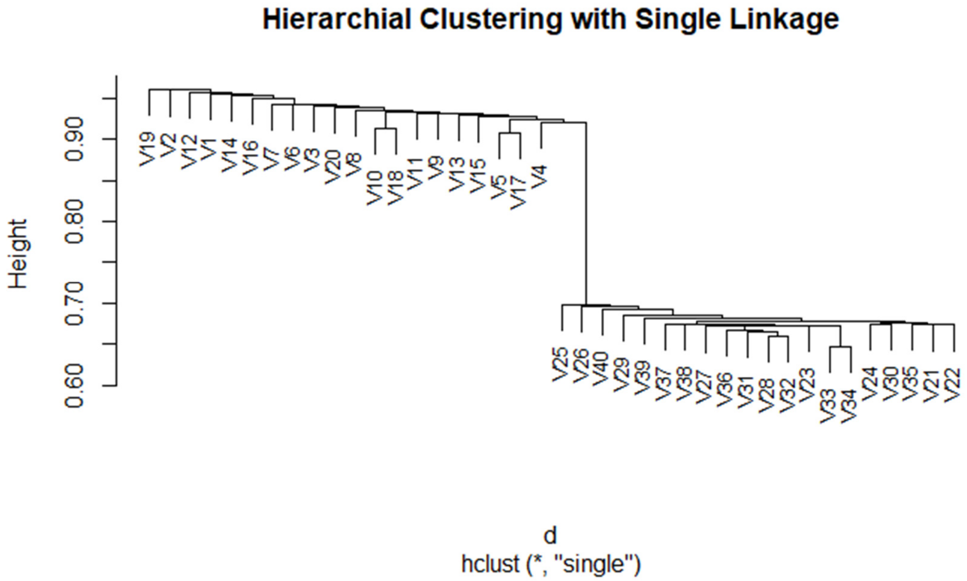
2.

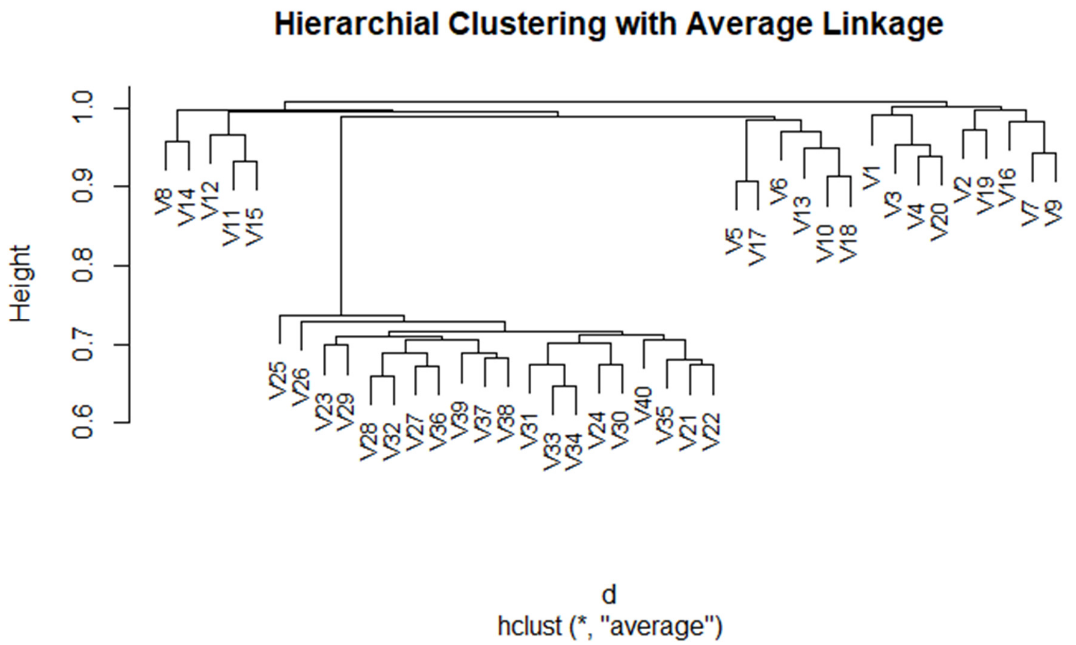
a)

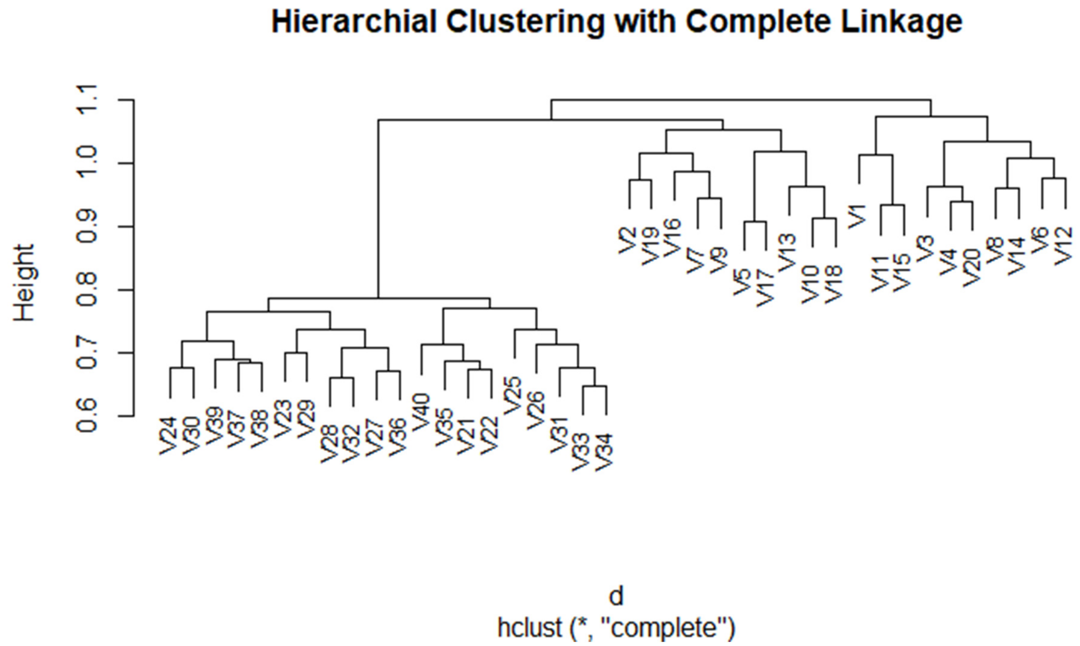
The gene expression data containing 40 tissue samples with measurements on 1000 genes, is loaded using the read.csv() function.

b)

The inter-observation dissimilarities for the data is calculated based on correlation. Hierarchical clustering based on single, average and complete linkage is performed. The resulting dendrograms are as follows:





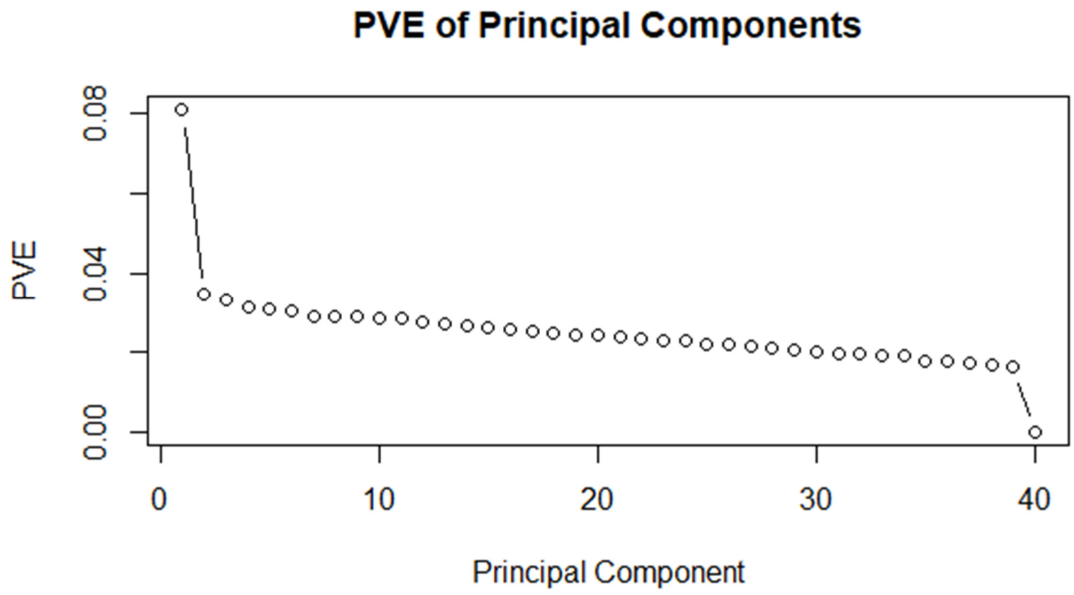


On observing the dendrograms obtained, we can clearly say that the clusters obtained depend on the method of linkage used. We can obtain two clusters for single linkage and complete linkage. We cannot obtain two clusters for average linkage. The minimum number of clusters that the data can be grouped into, in this case, is three.

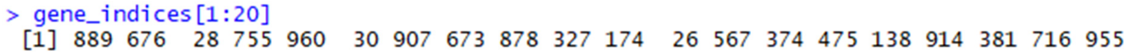
c)

To find the genes which differ the most across the two groups, we can use Principal Component Analysis (PCA). The rotation matrix obtained by performing PCA can be used to find the loadings, which indicates the weights of the genes. By arranging the absolute values of these loadings, we can find the genes which differ the most.

The proportion of variance explained by each principal component is follows:



The top 2% of genes which differ the most based on the calculated factor loadings are listed below, according to the sample number of the gene.



```
> gene_indices[1:20]
 [1] 889 676  28 755 960  30 907 673 878 327 174  26 567 374 475 138 914 381 716 955
```

3.

a)

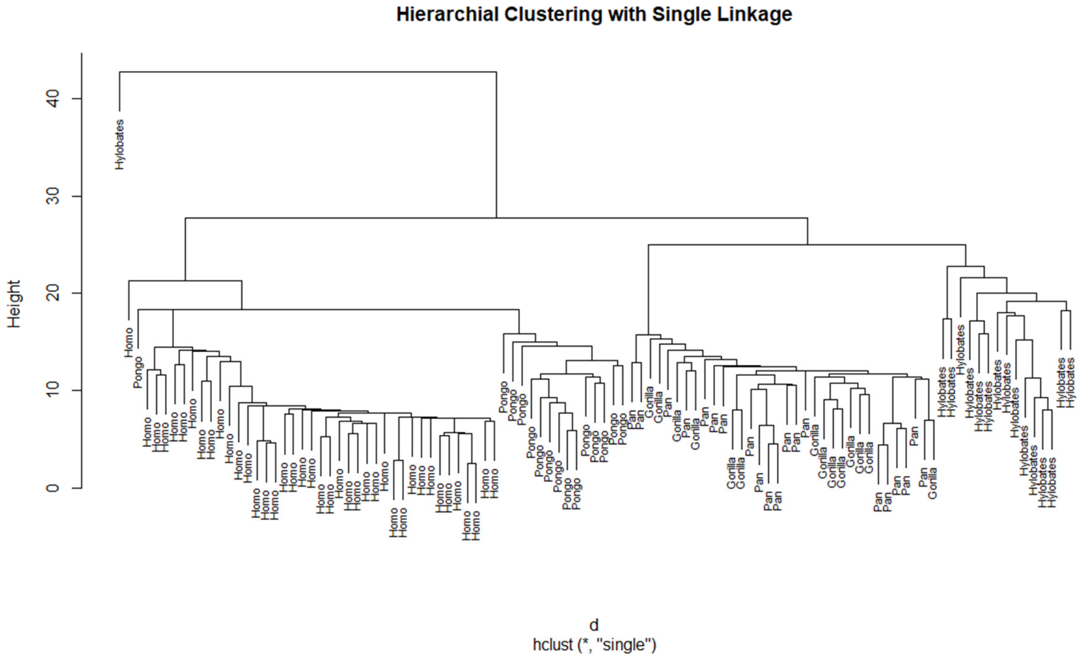
The dataset to be clustered in this exercise is the primate.scapulae dataset, which has 105 observations with 11 predictors.

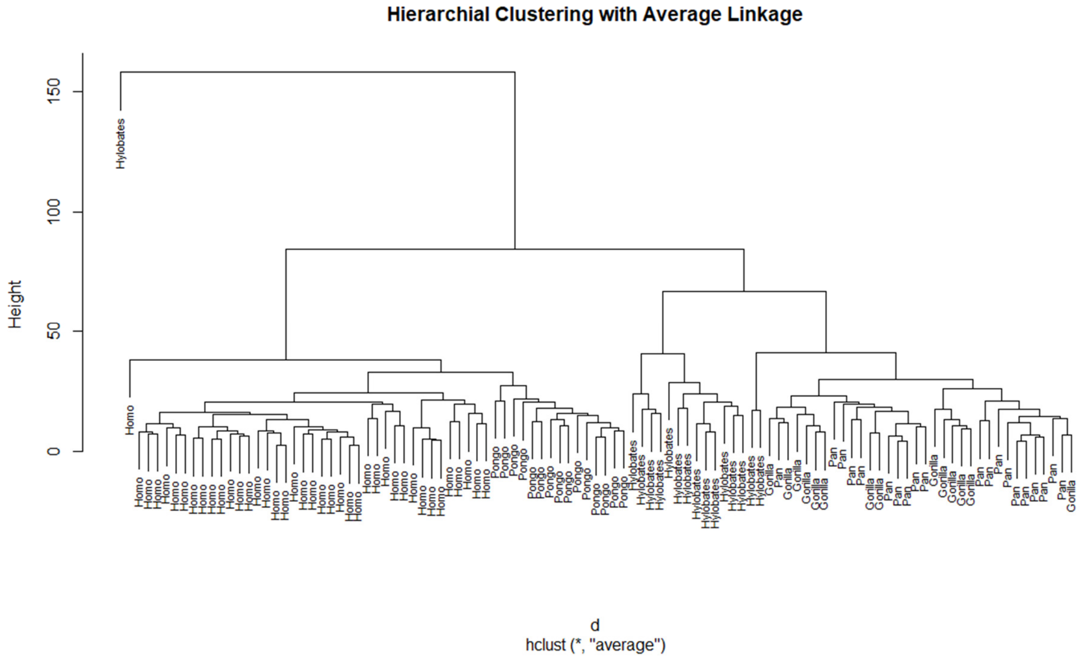
In the dataset, it can be observed that there are 40 NA values in the column named gamma. These values are replaced with the mean of the column. Removing this column and performing the analysis detailed below gave very similar results.

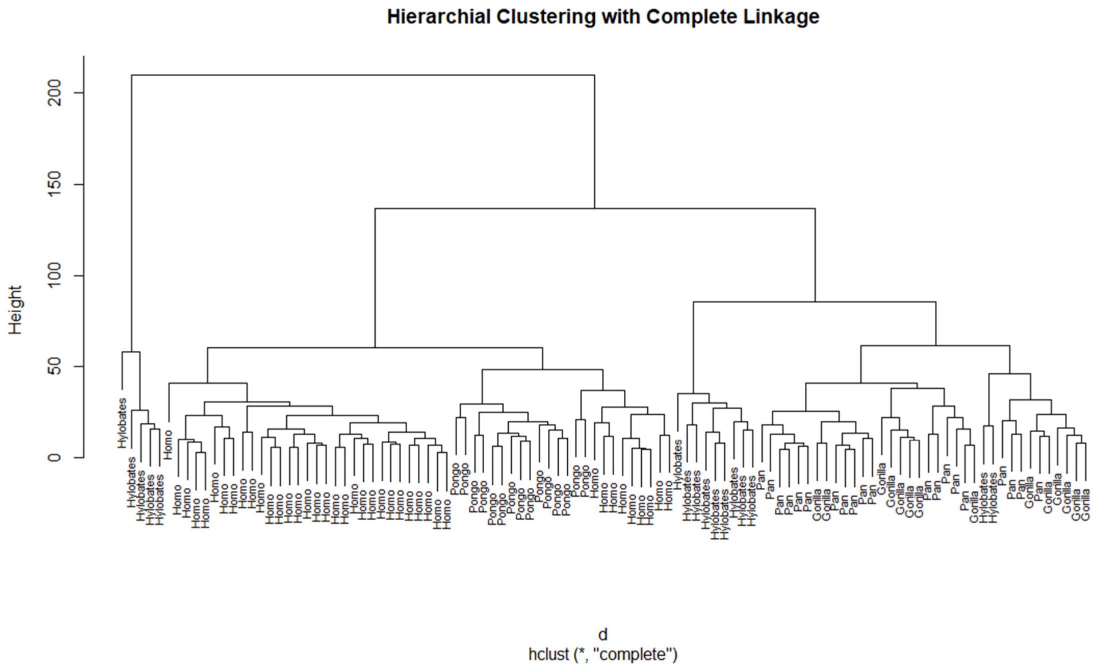
Also, the class column is dropped from the dataset, in view of the presence of characters. The presence of this column led to some issues while clustering the data using k-means. On the other hand, it is same as the column classdigit.

The rows containing the NA values were not removed, since it formed a significant portion of the data.

The distance matrix was calculated for this dataset, and hierarchical clustering was performed using single linkage, average linkage and complete linkage. The dendrograms obtained are as follows:







To cluster the data into respective clusters, we would require a minimum of five clusters, as there are five classes in the data. Moreover, on looking at the dendrograms for each linkage method, five is an optimal choice; though the clusters obtained with single linkage and average linkage are note very satisfactory.

The confusion matrix for each linkage method was computed, and the classification rates obtained are as follows:

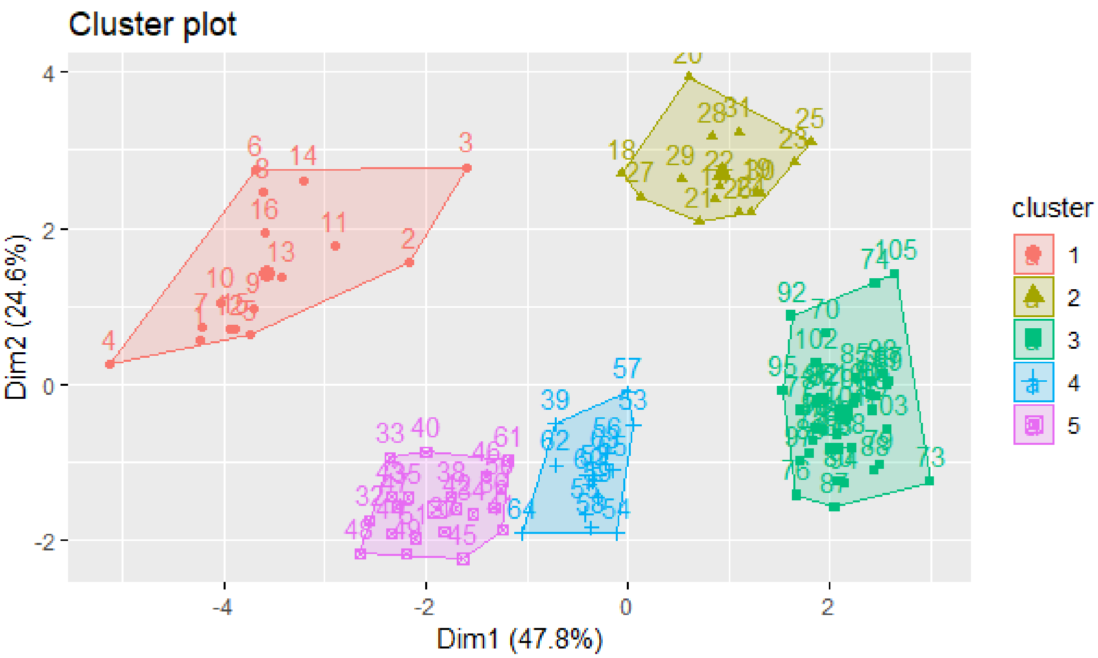
Linkage Method	**Accuracy**
Single Linkage	12.38%
Average Linkage	12.38%
Complete Linkage	4.76%

Complete linkage has the worst classification rate, whereas single and average linkage performed the best in this case. I expected complete linkage to give the best results, compared to the other two, but based on the table; it is just the other way round.

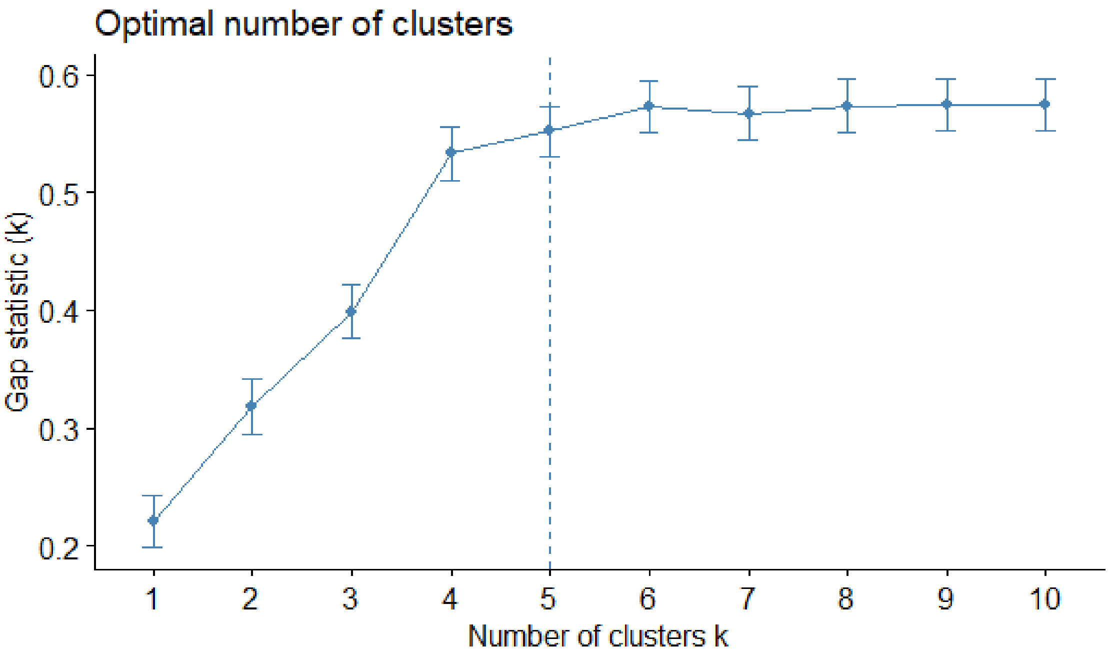
b)

The data is now clustered using k-means, by taking five centres to cluster the five classes in the data.

The accuracy of this clustering method obtained was 41.9%. Clearly, the results obtained using k-means clustering are much better compared to agglomerative clustering.



However, we must find the optimal value of k to cluster the data instead of choosing a value to our liking. We use the gap statistic method to find the optimal value of k. The graph obtained from calculating the gap statistic is as follows:



It can be seen that the optimal value of k is 5, which is line with the value of k that was chosen earlier.

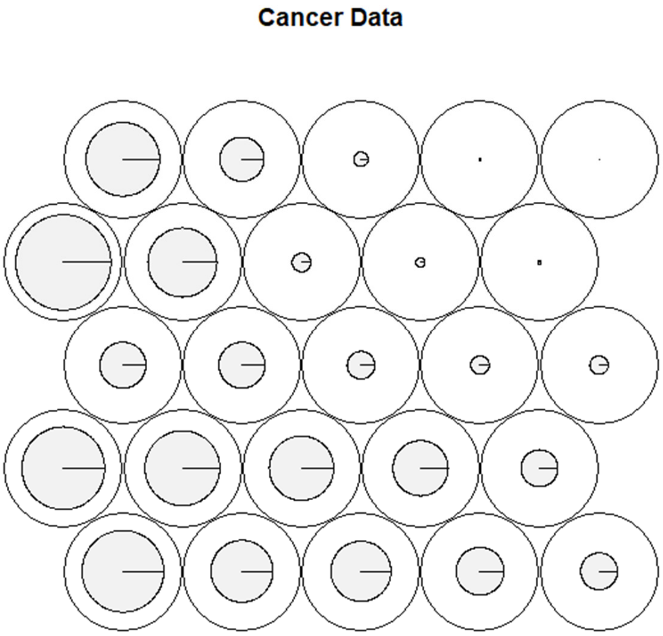
Despite the poor accuracy, I would consider agglomerative hierarchical clustering, as it is very flexible. Based on the dendrogram obtained from hierarchical clustering, we can cluster the data into any number of groups, instead of pre-determining a value of k for k-means. However, I would prefer k-means, which clusters data much faster than agglomerative clustering does.

4.

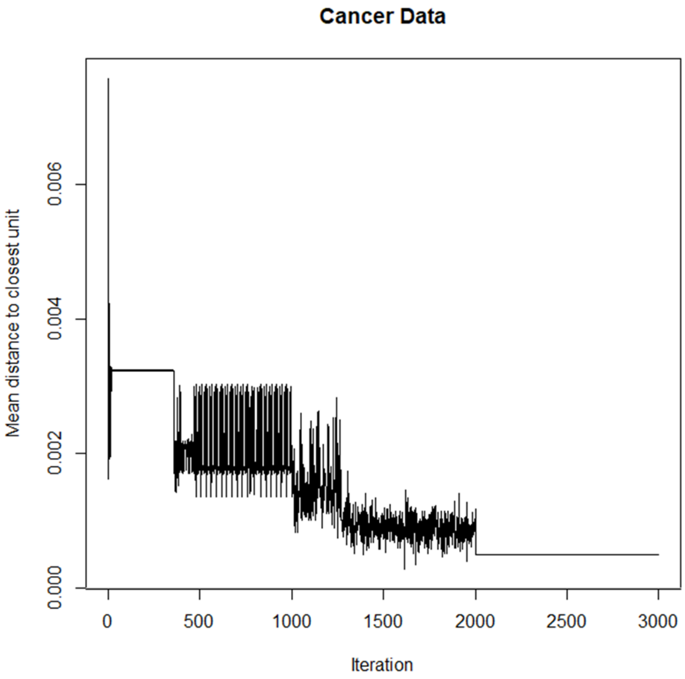
The dataset to be used for this exercise is the Wisconsin Breast Cancer data, which has 699 observations of 11 measurements of cells in suspicious lumps. The samples are either classified as benign or malignant.

To perform a batch-SOM analysis, the data is first scaled to ensure uniformity across along all columns. Using the mode="batch" parameter in the supersom() function, a batch analysis is performed.

The SOM obtained is represented as follows.

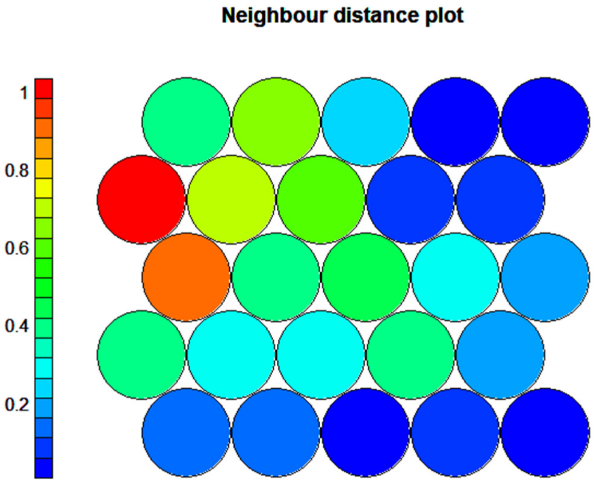


The graph below represents the changes in the codebook factors across the 3000 iterations performed on the dataset. We can see that convergence starts occurring from the 1500th iteration.

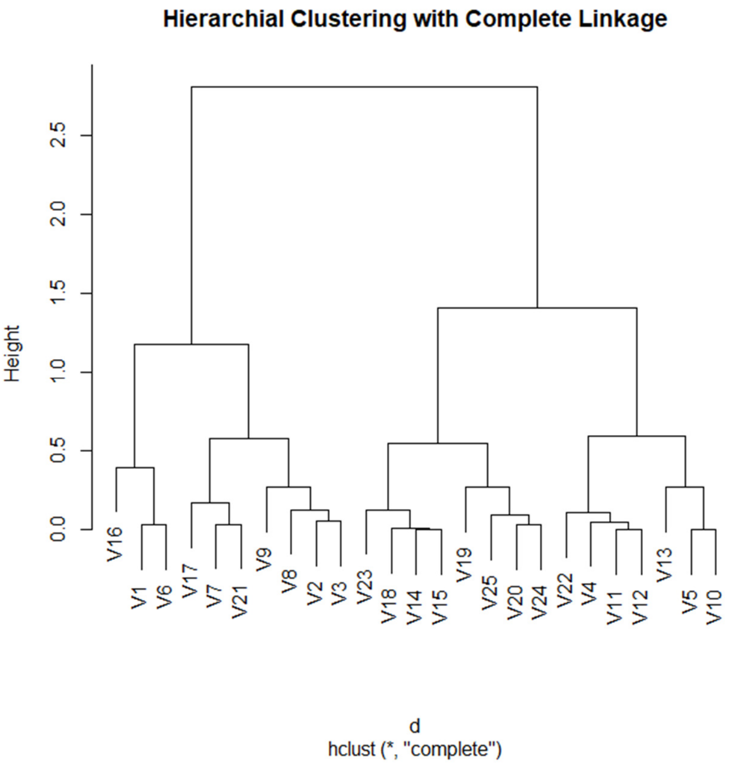


One other significant plot from this batch analysis is the neighbour distance plot, or the U-matrix. The U-matrix basically represents the Euclidean distance between the codebook vectors of neighbouring nodes. It can be used to identify clusters in the data.

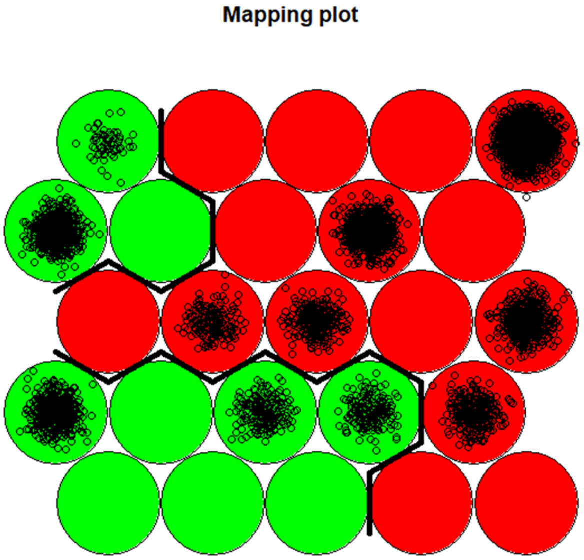
In the plot that we have obtained below, we can possibly have two or three clusters, by grouping similar colour shades.



Now, hierarchical clustering with complete linkage is performed on the dataset. On observing the plotted dendrogram, we can obtain two distinct clusters, by cutting the dendrogram at height 2.5.


Hierarchial Clustering with Complete Linkage

To see how well SOM clusters the data into benign and malignant, we use the clusters obtained above to obtain a mapping plot of the SOM.


Mapping plot

The mapping plot above fairly distinguishes and clusters the data well. We see that 15 nodes (60% of the data) are classified as benign, and 10 nodes (40% of the data) are classified as malignant. This is quite close in line with the given data, where 444 observations (65% of the data) are benign, and 239 observations (35% of the data) are malignant.