STATISTICAL DATA MINING

Home Work - 2

Ravi Teja Sunkara

UBIT Number: 50292191

UBIT Name: rsunkara

Email: rsunkara@buffalo.edu

Question 1

Consider the USArrests data. We will now perform hierarchical clustering on the states.
(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.
(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?
(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.
(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

Solution:

a &b) <u>Hierarchical clustering with complete linkage and Euclidean distance without Scaling:</u>

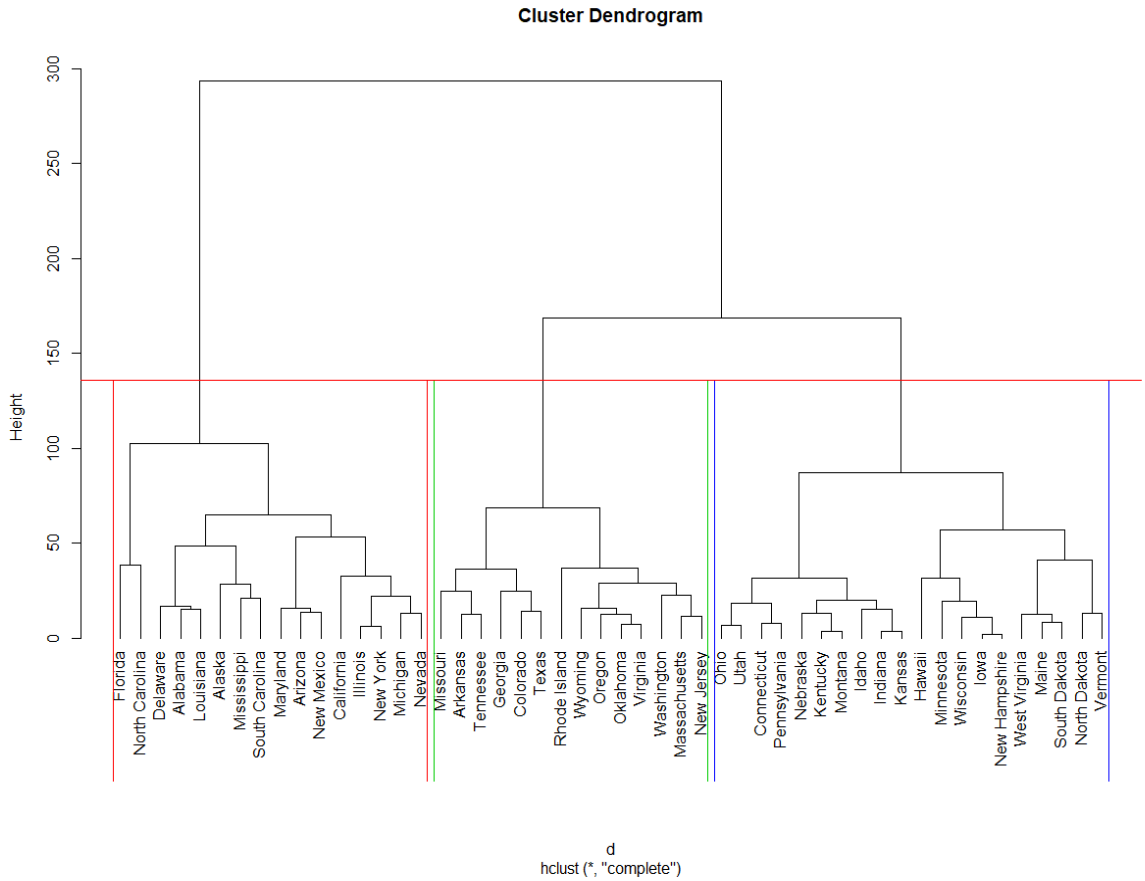


Fig: Dendrogram before Scaling

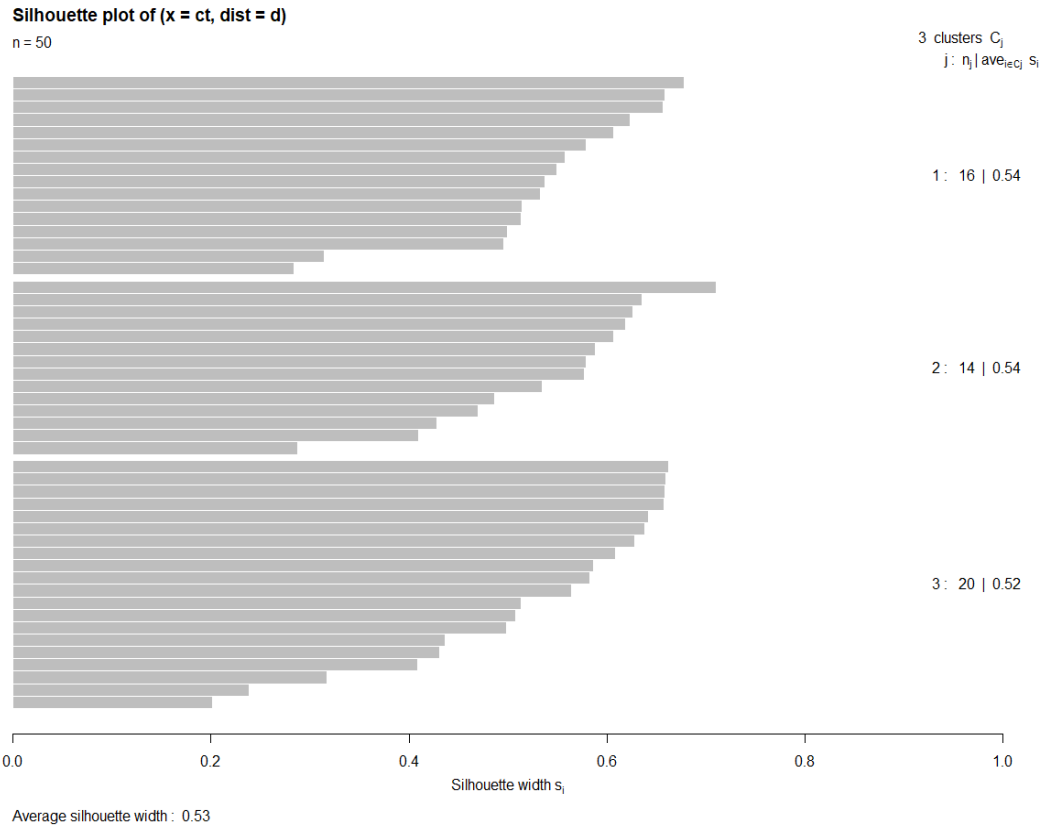


Fig: Silhouette before scaling

c) <u>Complete linkage with Euclidean distance After Scaling:</u>

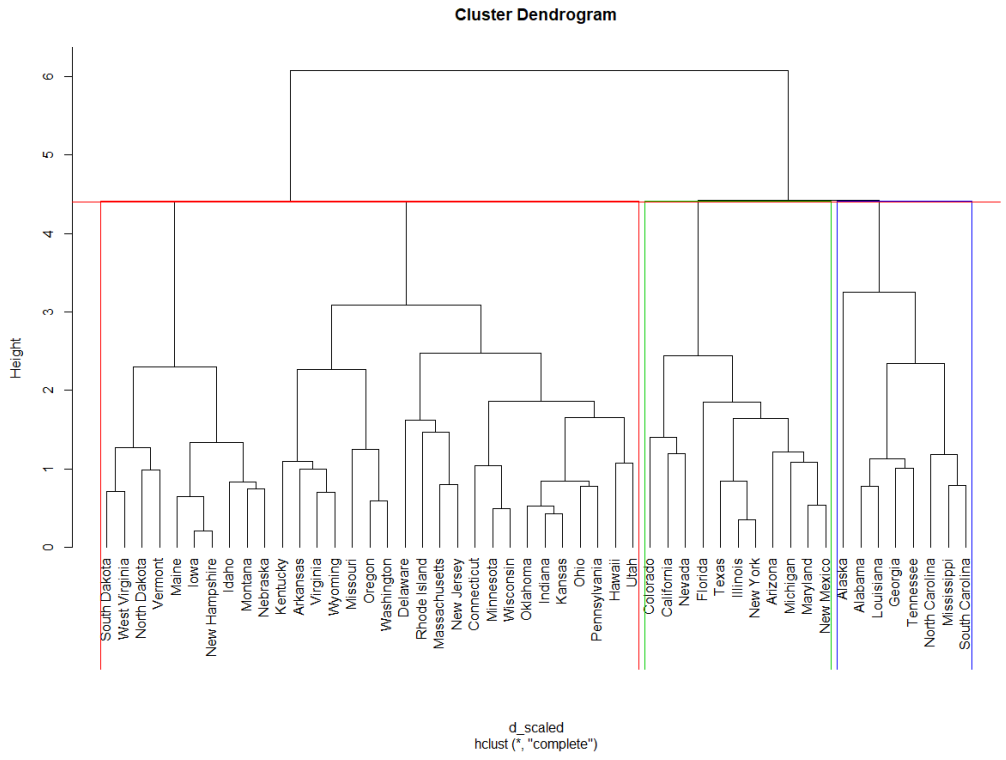


Fig: Dendrogram after scaling

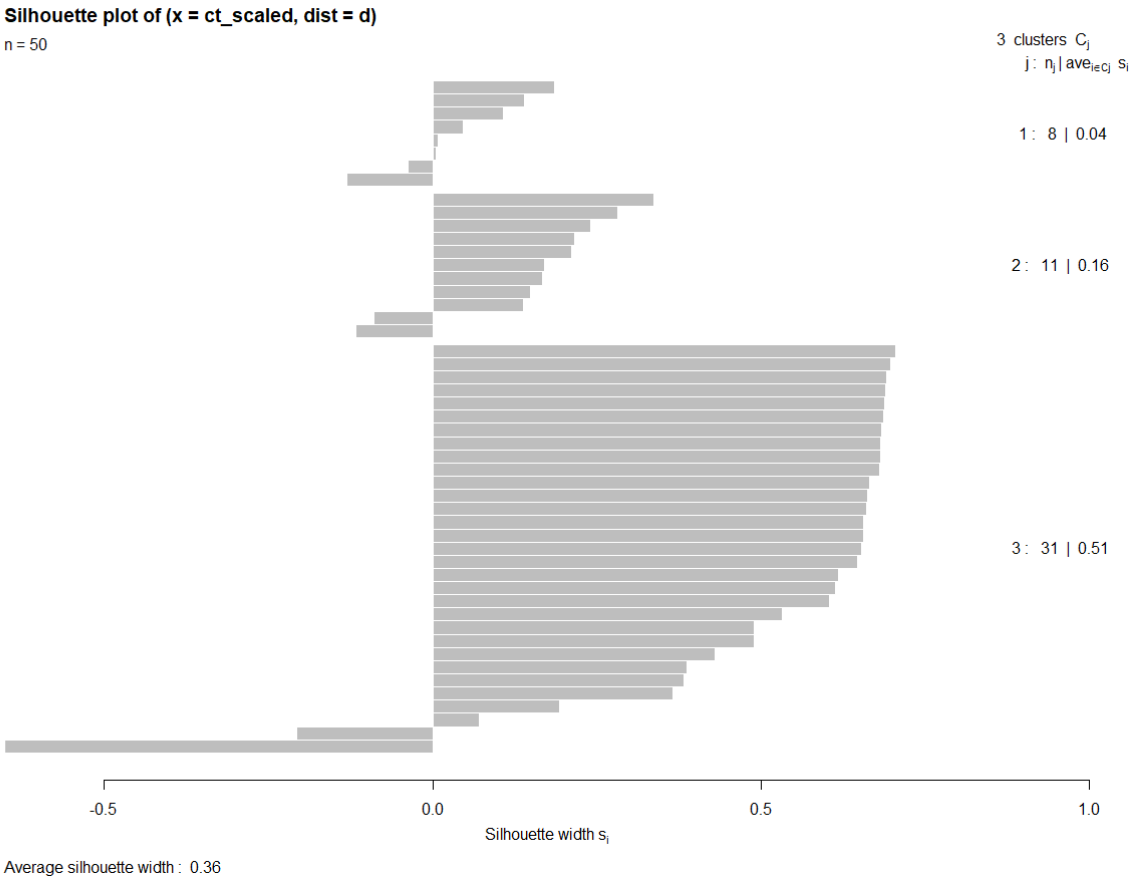


Fig: Silhouette after Scaling (3 clusters)

Before and after scaling comparison of classificaitons of states:

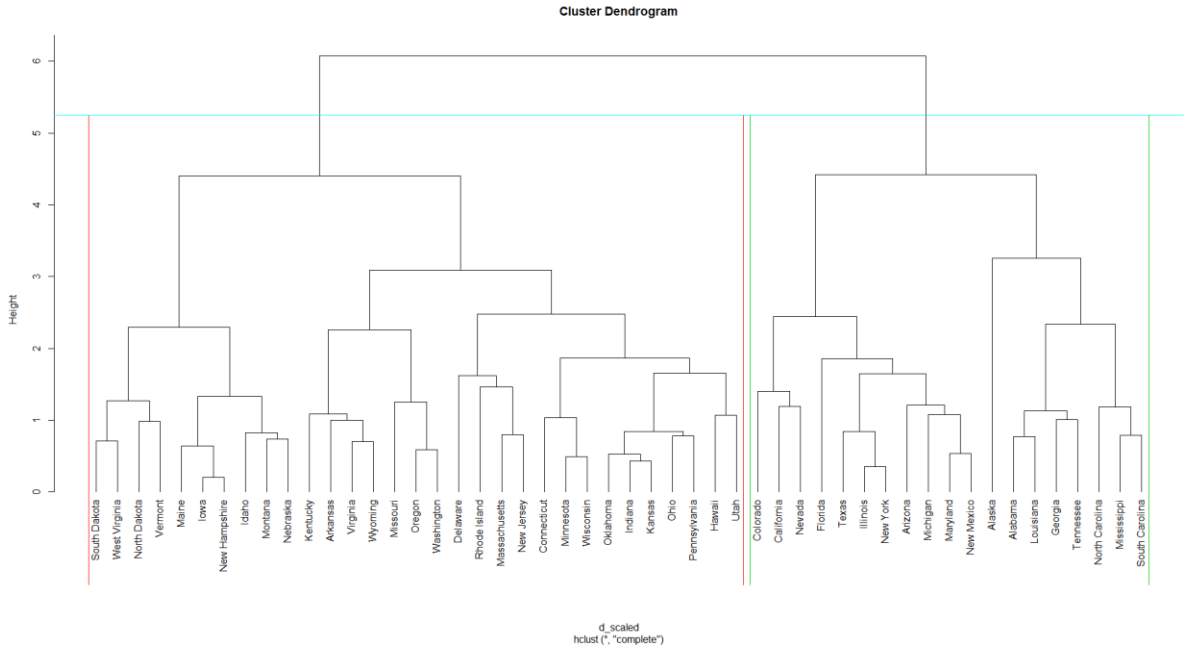
State	No Scaling	After Scaling
Alabama	1	1
Alaska	1	1
Arizona	1	2
Arkansas	2	3
California	1	2
Colorado	2	2
Connecticut	3	3
Delaware	1	3
Florida	1	2
Georgia	2	1
Hawaii	3	3
Idaho	3	3
Illinois	1	2
Indiana	3	3
Iowa	3	3
Kansas	3	3
Kentucky	3	3
Louisiana	1	1

Maine	3	3
Maryland	1	2
Massachusetts	2	3
Michigan	1	2
Minnesota	3	3
Mississippi	1	1
Missouri	2	3
Montana	3	3
Nebraska	3	3
Nevada	1	2
New Hampshire	3	3
New Jersey	2	3
New Mexico	1	2
New York	1	2
North Carolina	1	1
North Dakota	3	3
Ohio	3	3
Oklahoma	2	3
Oregon	2	3
Pennsylvania	3	3
Rhode Island	2	3
South Carolina	1	1
South Dakota	3	3
Tennessee	2	1
Texas	2	2
Utah	3	3
Vermont	3	3
Virginia	2	3
Washington	2	3
West Virginia	3	3
Wisconsin	3	3
Wyoming	2	3

		Clustering WITH Scaling		
		1	2	3
Clustering NO scaling	1	6	9	1
	2	2	2	10
	3	0	0	10

Changes:

The scaled datset is not performing well using 3 clusters as can be observed from the dendrogram and silhouette above. So, I used fviz_nbclust() function with silhouette method to find the optimal number of clusters and the result was 2. The resulting dendrogram for 2 clusters is shown below.



d) <u>Justification</u>

The data should be scaled before performing clustering since the variables might be of different orders that bias the clustering algorithm. Scaling brings all the features/variables to a common scale and this makes sure that all the variables are given equal importance during clustering.

Question 2

On the book website, www.StatLearning.com, there is a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.
(a) Load in the data using read.csv(). You will need to select header=F.
(b) Apply hierarchical clustering to the samples using correlation based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?
(c) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question, and apply it here.

Solution:

b) The results do depend on the type of linkage used. The dendrograms for different linkages are shown below.

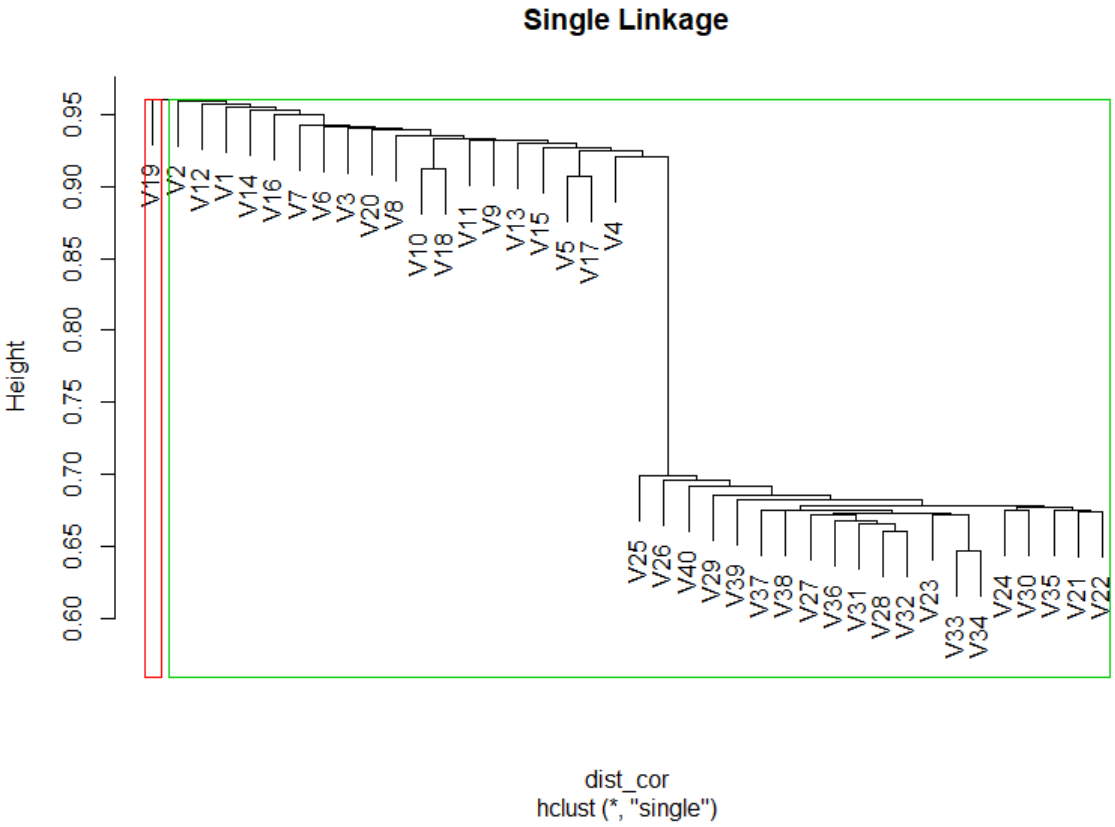


Fig: Single Linkage Dendrogram

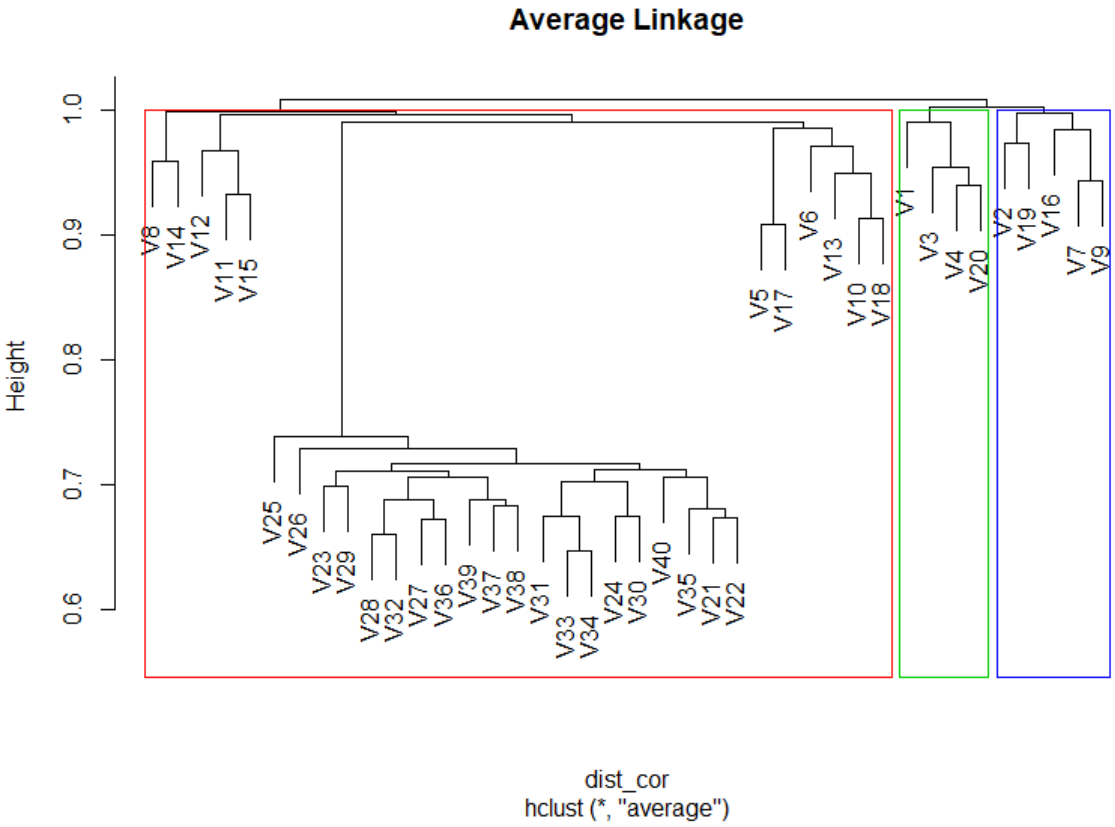


Fig: Average Linkage Dendrogram

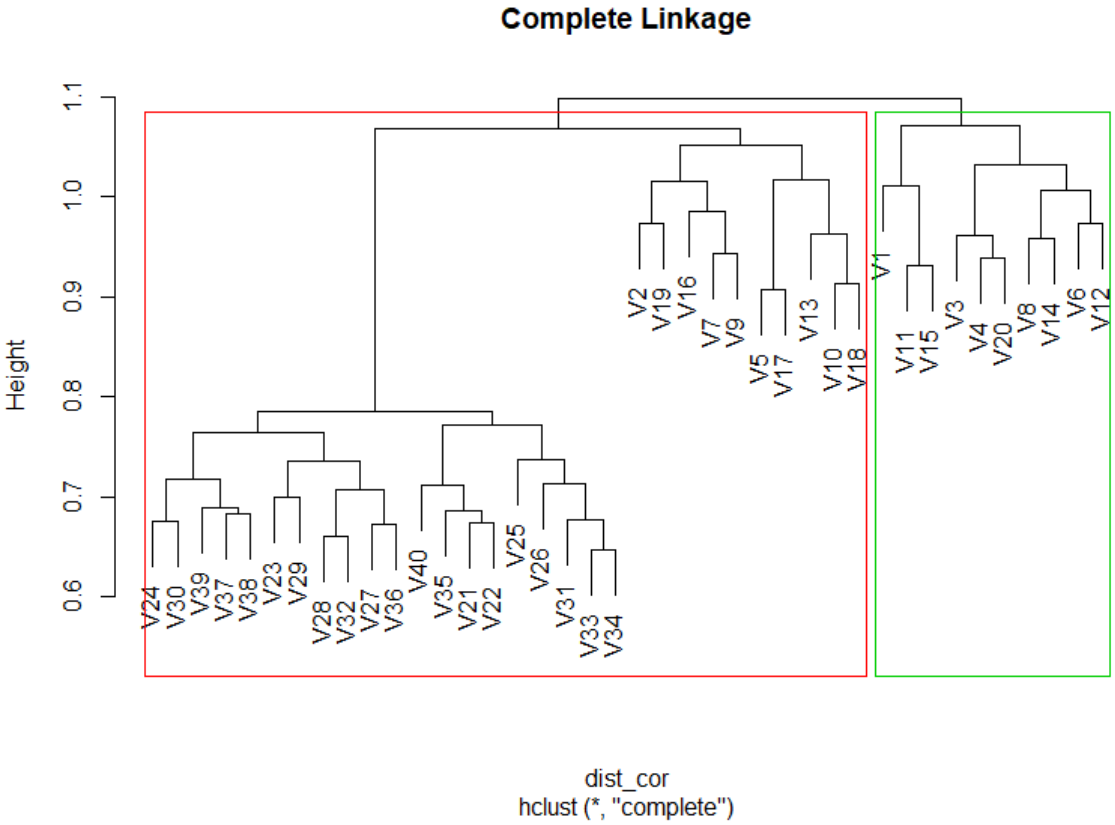


Fig: Complete Linkage Dendrogram

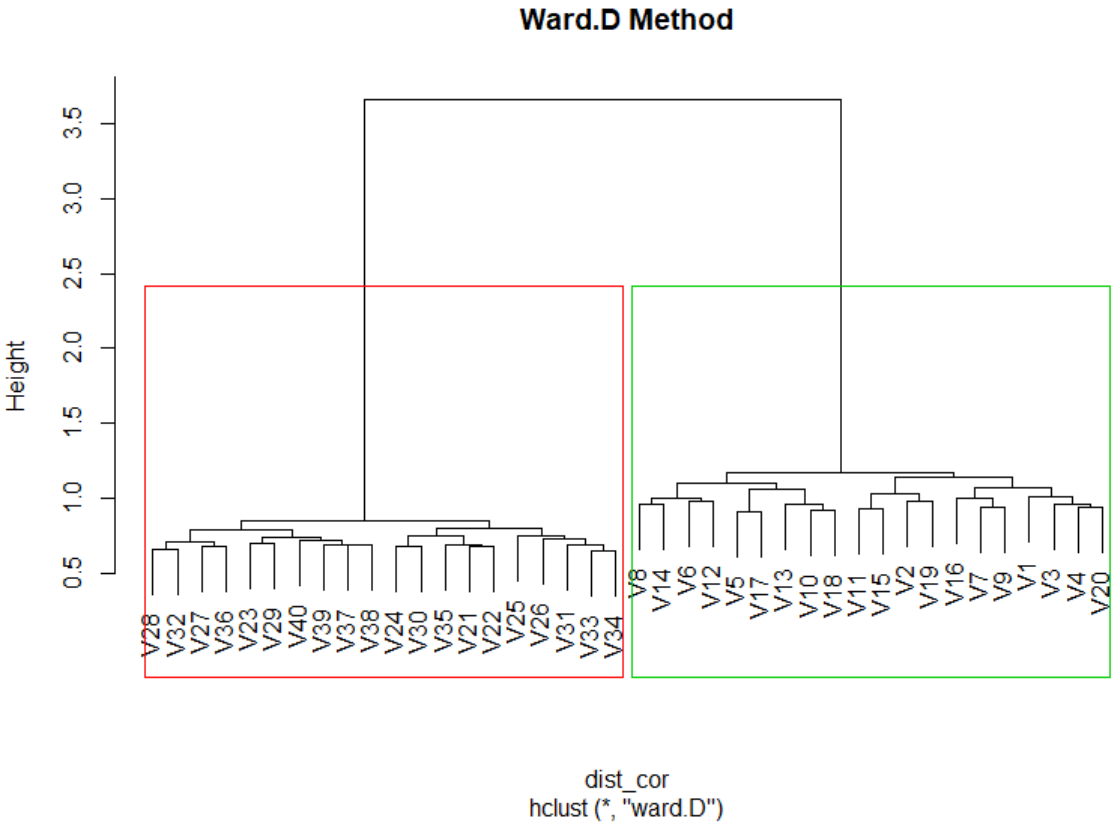


Fig: Ward.D Dendrogram

The results obtained clearly differ based on the type of linkage used.
The complete, single, ward.D methods give two clusters while average linkage gives three clusters.

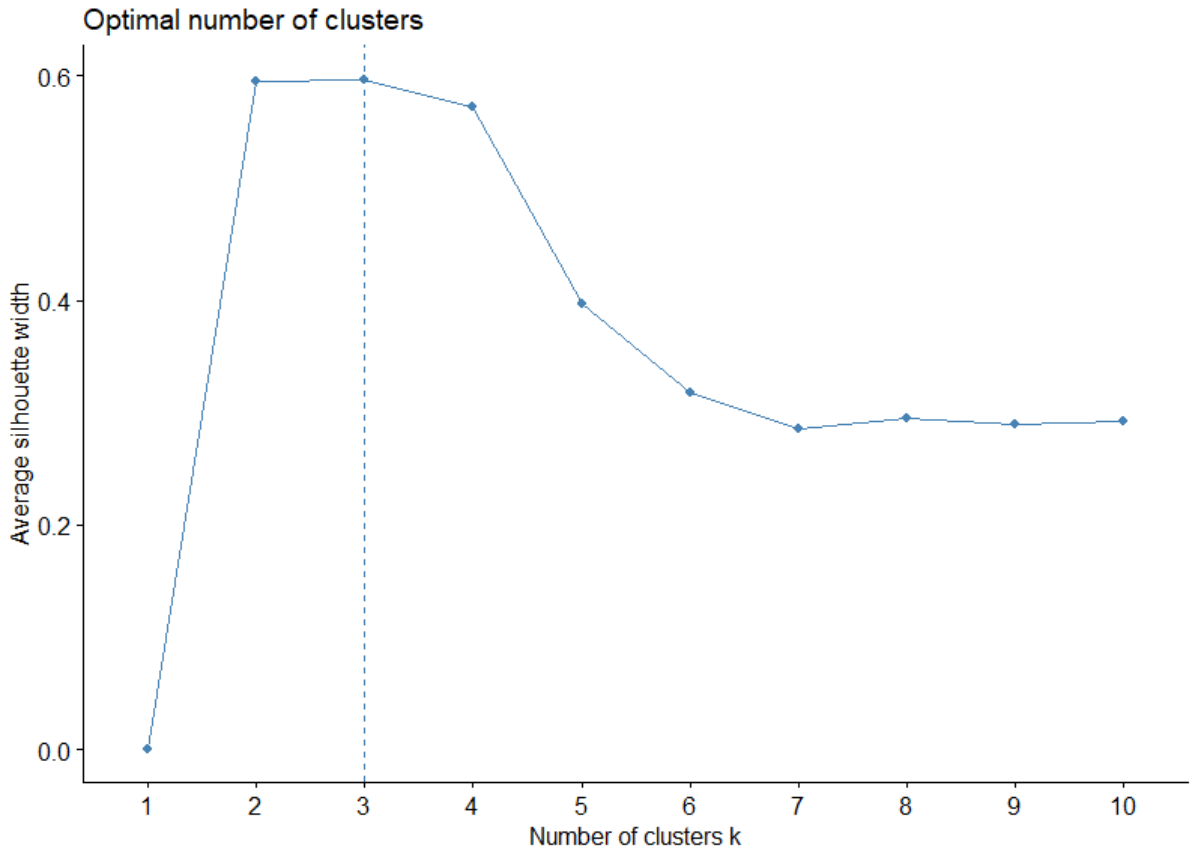
d) We can use principal component analysis to answer the question. We will arrange in decreasing order the total weight given to each gene for different principal components.
The top 10 in the order will be the genes, which differ a lot in the two group
The gene number 865,68,911,428,624,,11,524,803,980,822 are the 10 most different genes across the group.

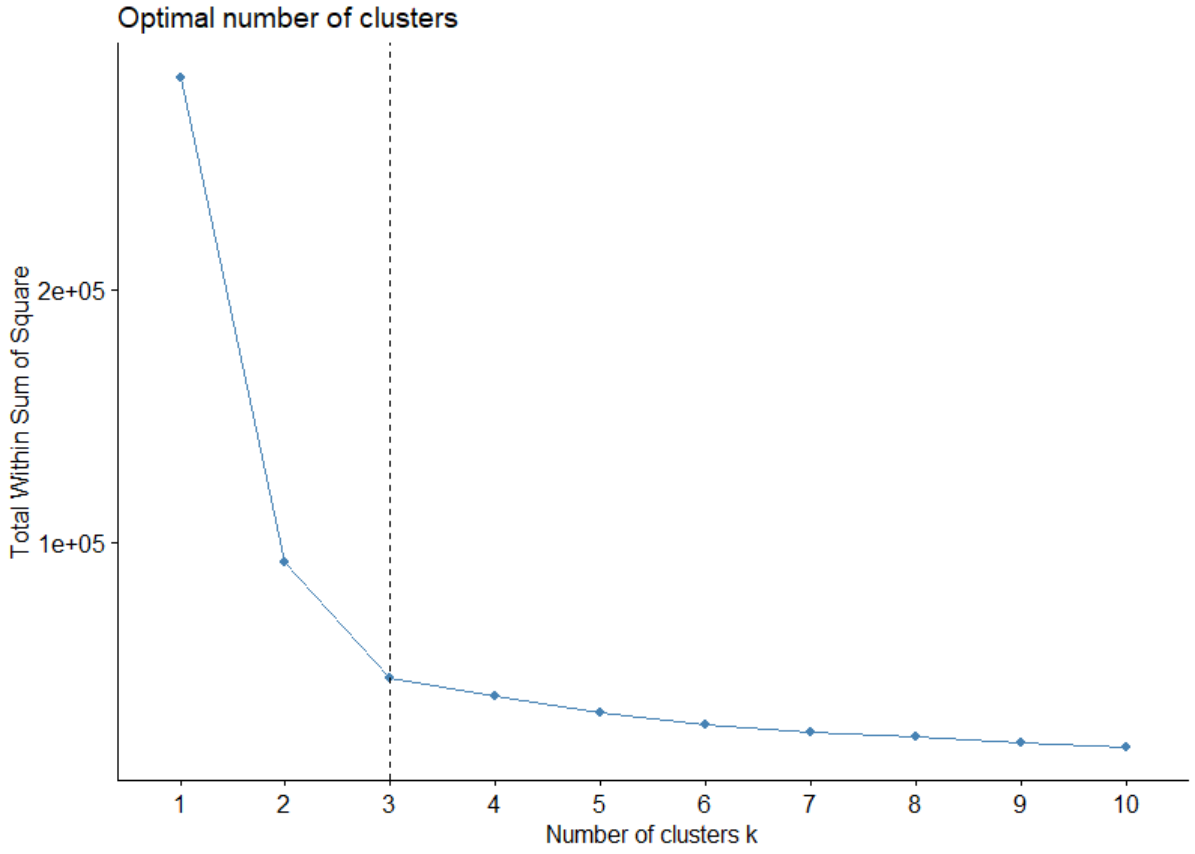
Question 3

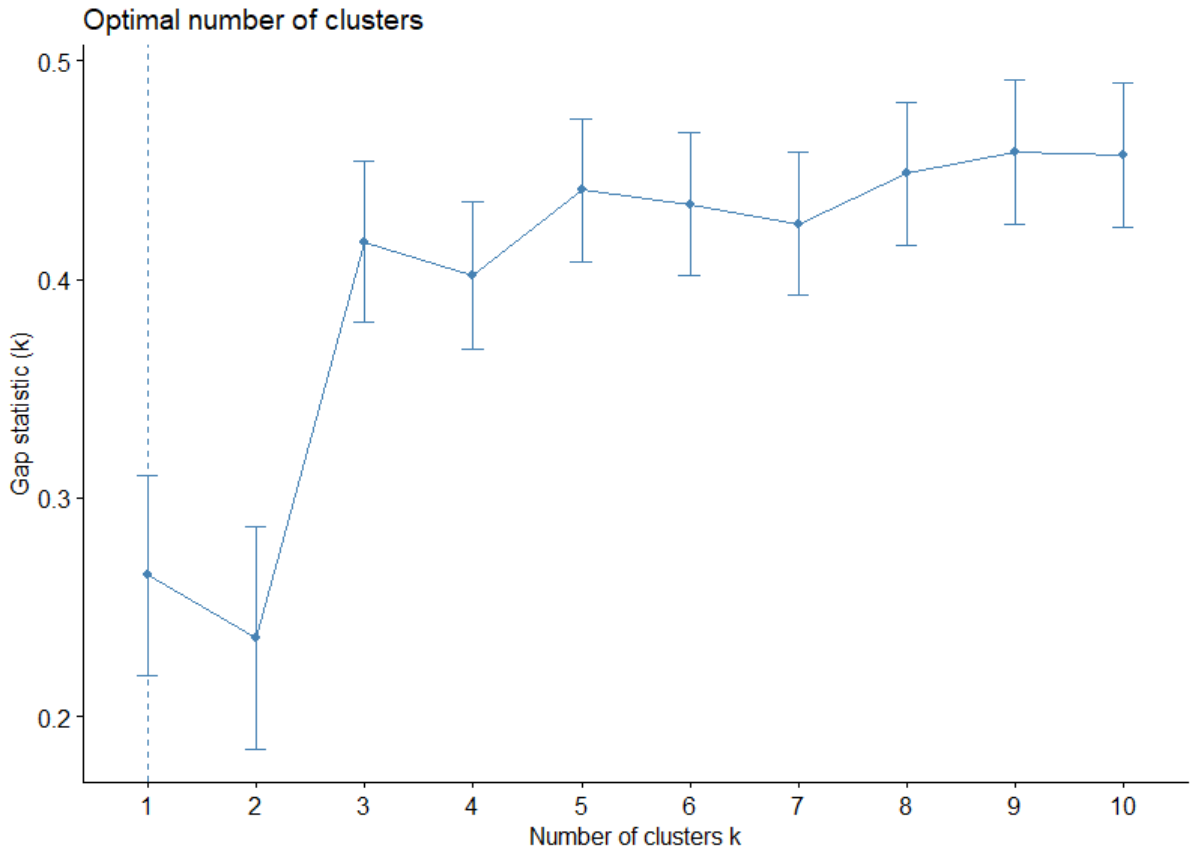
a) Cluster the data based on single-linkage, average linkage, and complete-linkage agglomerative hierarchical clustering. Decide on the groupings, and justify it, for all three methods. Calculate the misclassification rate. Which method performed the best and which method performed the worst? Was the result in line with your expectations?
b) Cluster the data based on K-means or K-medoids. Use an analytical technique to justify your choice in "k". How did the performance compare to the hierarchical clustering of part a? Which did you feel was a better method for this data?

Determining Optimal Number of Clusters:

I used fviz_nbclust() with different methods to find the optimal number of clusters. I have chosen 3 clusters to be the best representation for the data. The plots are shown below.


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Optimal number of clusters
Total Within Sum of Square
1e+05
2e+05
1
2
3
4
5
6
7
8
9
10
Number of clusters k



Single linkage

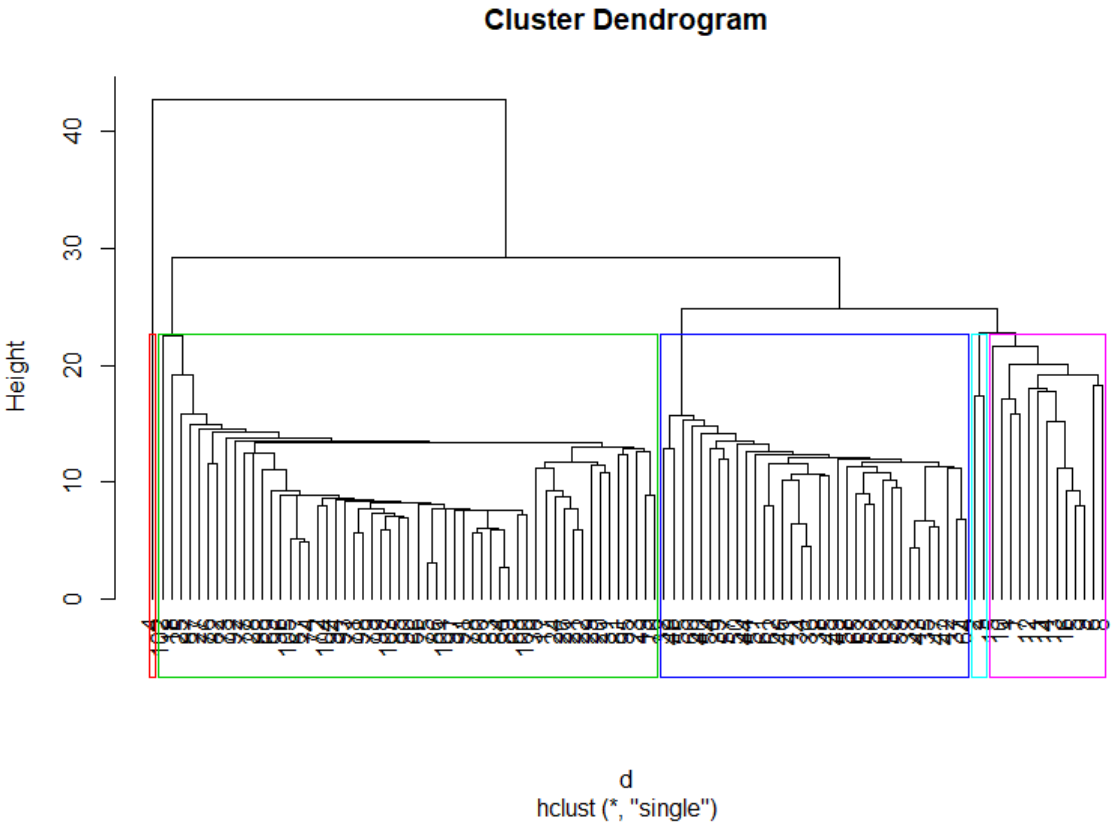
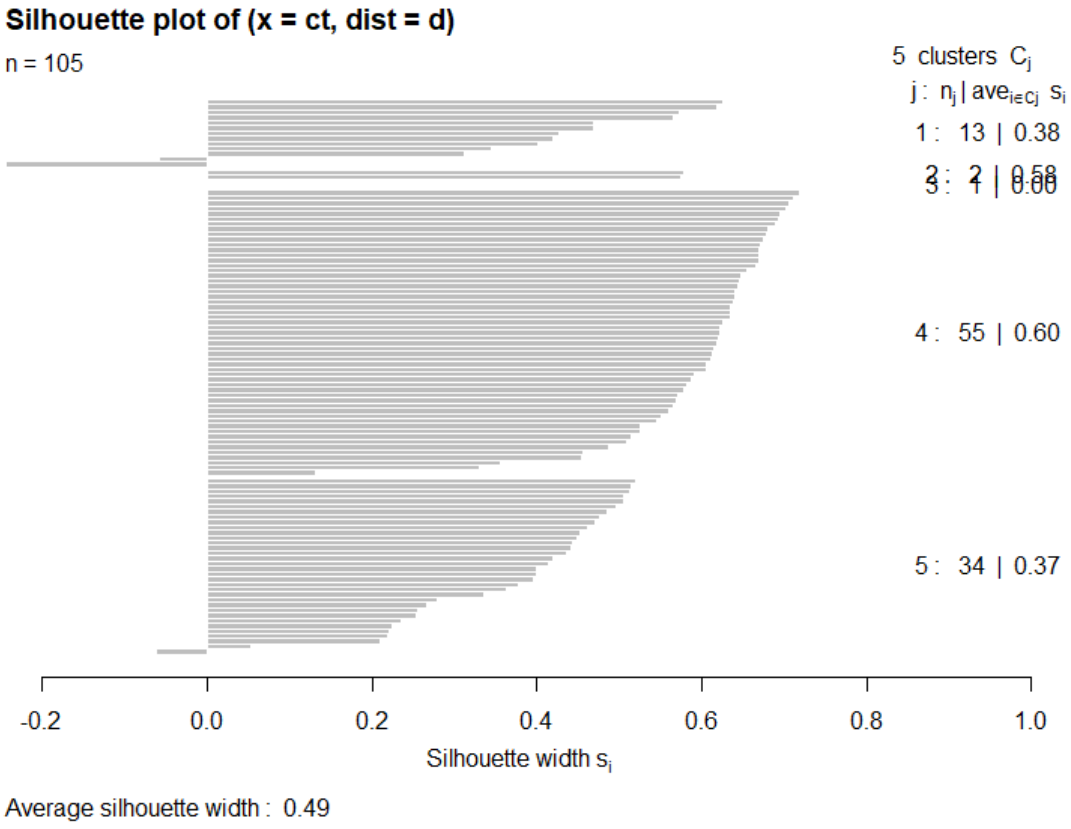


Fig: Single Linkage with 5 clusters

From the above dendrogram, it looks ideal for three cluster when cut near 26. However, the number of classes in the actual data are 5, so let's look at the silhouette for 5 clusters and if it is indeed doesn't look good, we can look at the 3 clusters solution.

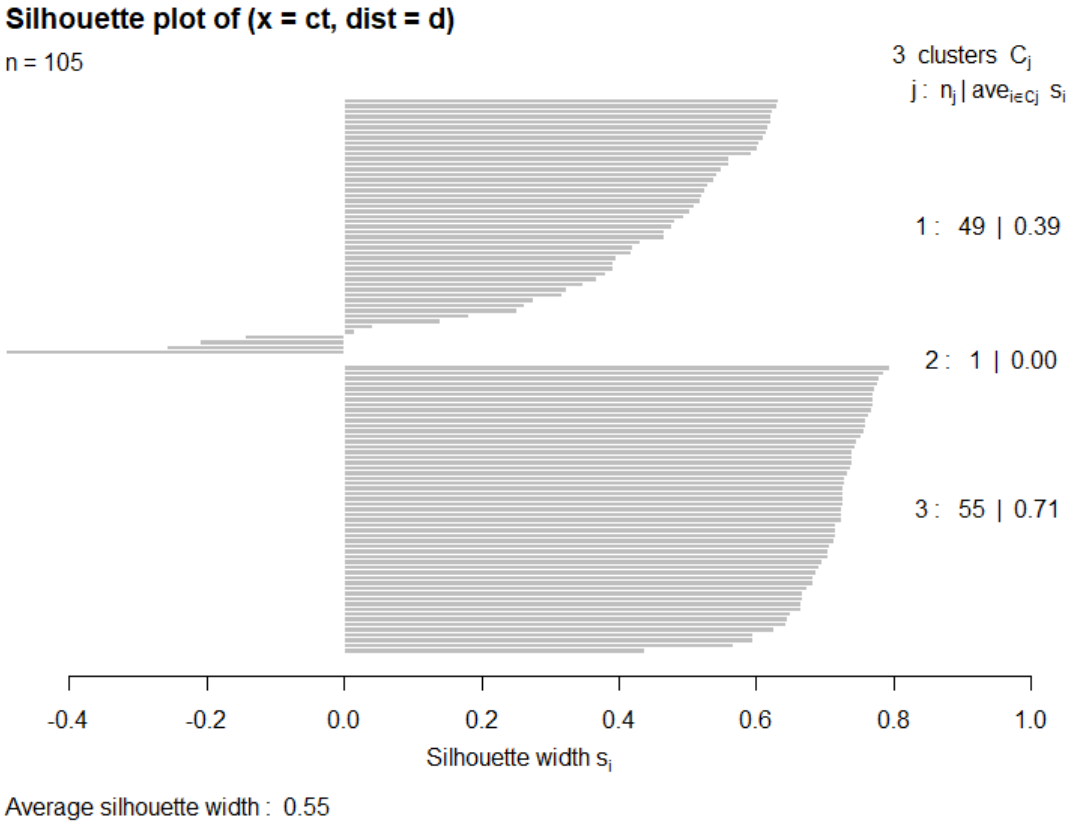
*Silhouette for **5 clusters**:*

The silhouette does not look good because there are negative values and so forth. Even though the number of classes in the actual data are 5 this looks bad. Average silhouette width of 0.49 for 5 clusters whereas for 3 clusters it is 0.55.



Silhouette for 3 clusters:

The groups are a bit uneven but an average width of Silhouette is 0.55, which is better than for 5 clusters.



Confusion Matrix and Accuracy for 5 Classes:

As the number of actual classes in the data are 5, using the 5 cluster predictions to obtain the accuracy and misclassification rate. I used 'confusionMatrix' function from 'caret' packages to obtain these statistics.

```
          Reference
Prediction  1  2  3  4  5
         1 13  0  0  0  0
         2  2  0  0  0  0
         3  1  0  0  0  0
         4  0 15  0  0 40
         5  0  0 20 14  0

Overall Statistics

               Accuracy : 0.1238
```

As we can see accuracy is just: **12.38%,** so misclassification rate**: 87.62%**

Confusion Matrix and Accuracy for 3 Classes:

```
          Reference
Prediction  1  2  3  4  5
         1 15  0 20 14  0
         2  1  0  0  0  0
```

```
     3  0 15  0  0 40
     4  0  0  0  0  0
     5  0  0  0  0  0

Overall Statistics

          Accuracy : 0.1429
```

Accuracy: **14.29%** and misclassification rate: **85.71%**

Average Linkage

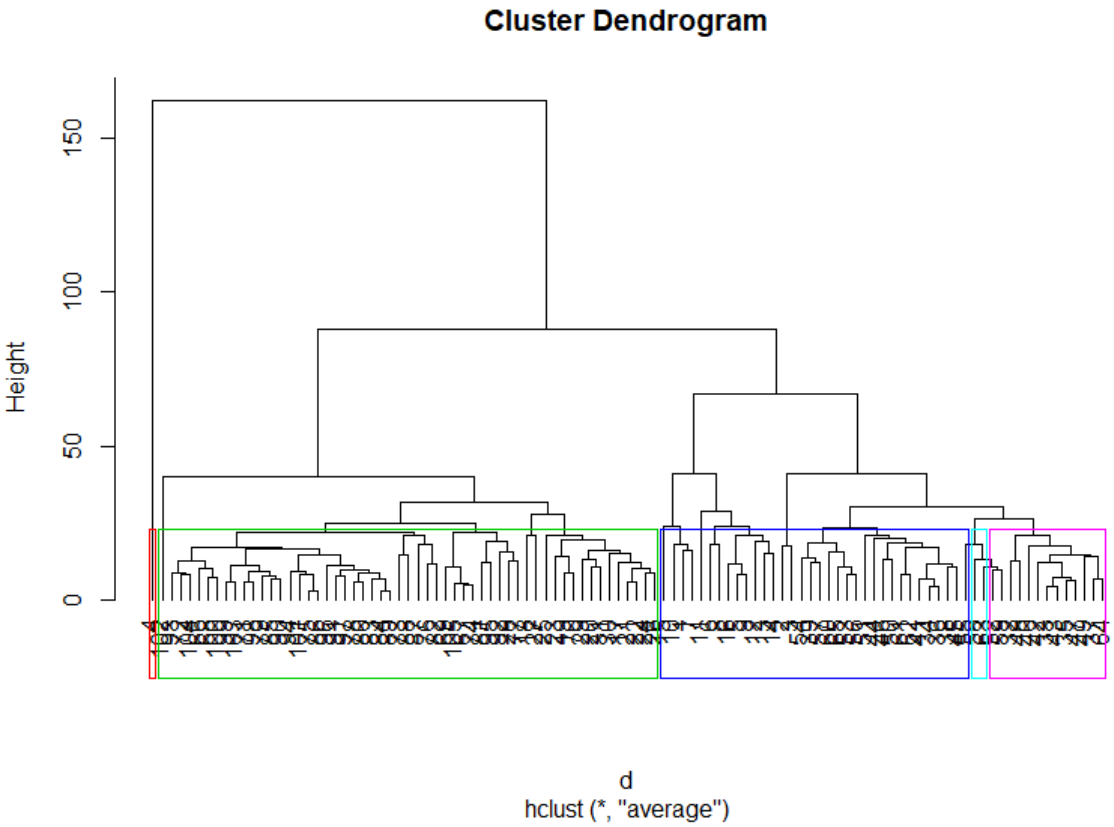
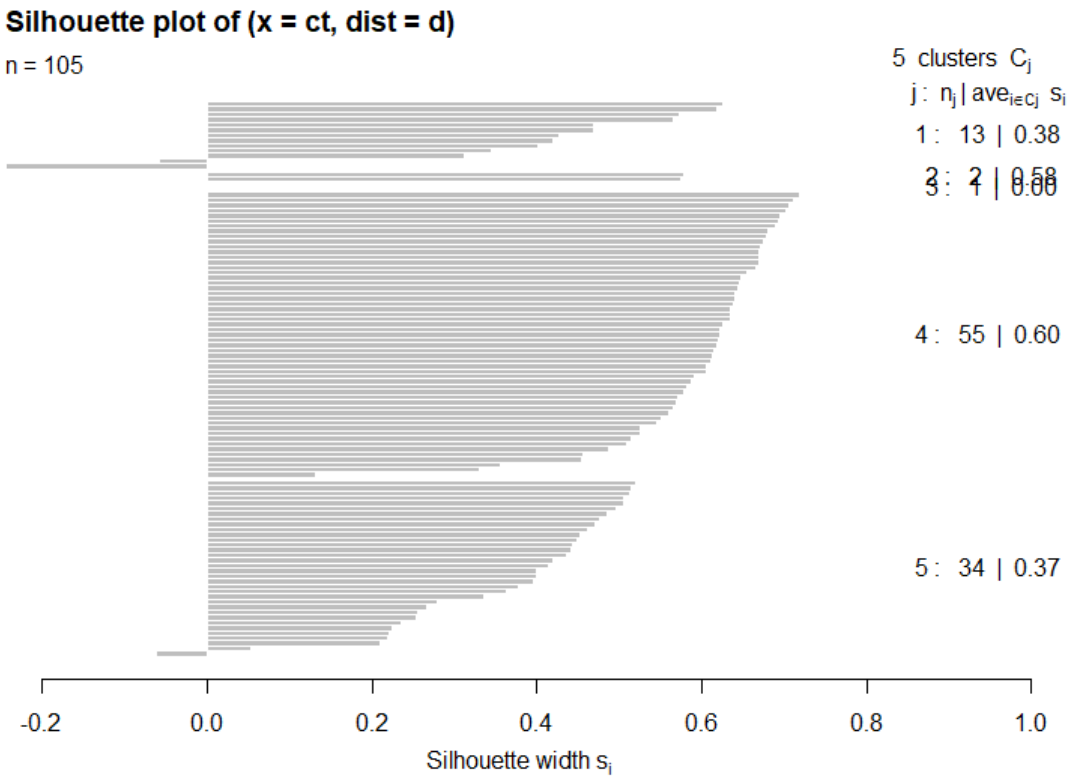


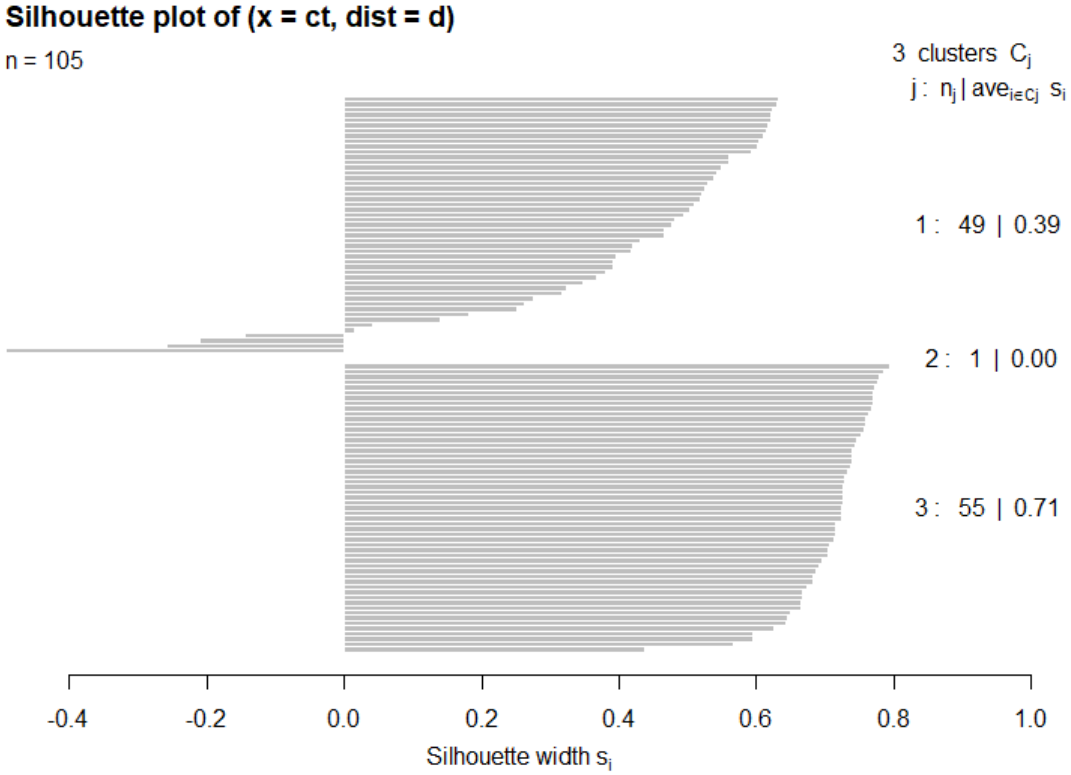
Fig: Average Linkage Dendogram

5 Cluster Silhouette for Average Linkage:



Average silhouette width : 0.49

3 Cluster Silhouette for Average Linkage:



Average silhouette width : 0.55

Confusion Matrix and Misclassification rate for 5 Clusters (Single Linkage):

```
          Reference
Prediction  1  2  3  4  5
         1 13  0  0  0  0
         2  2  0  0  0  0
         3  1  0  0  0  0
         4  0 15  0  0 40
         5  0  0 20 14  0

Overall Statistics

               Accuracy : 0.1238
```

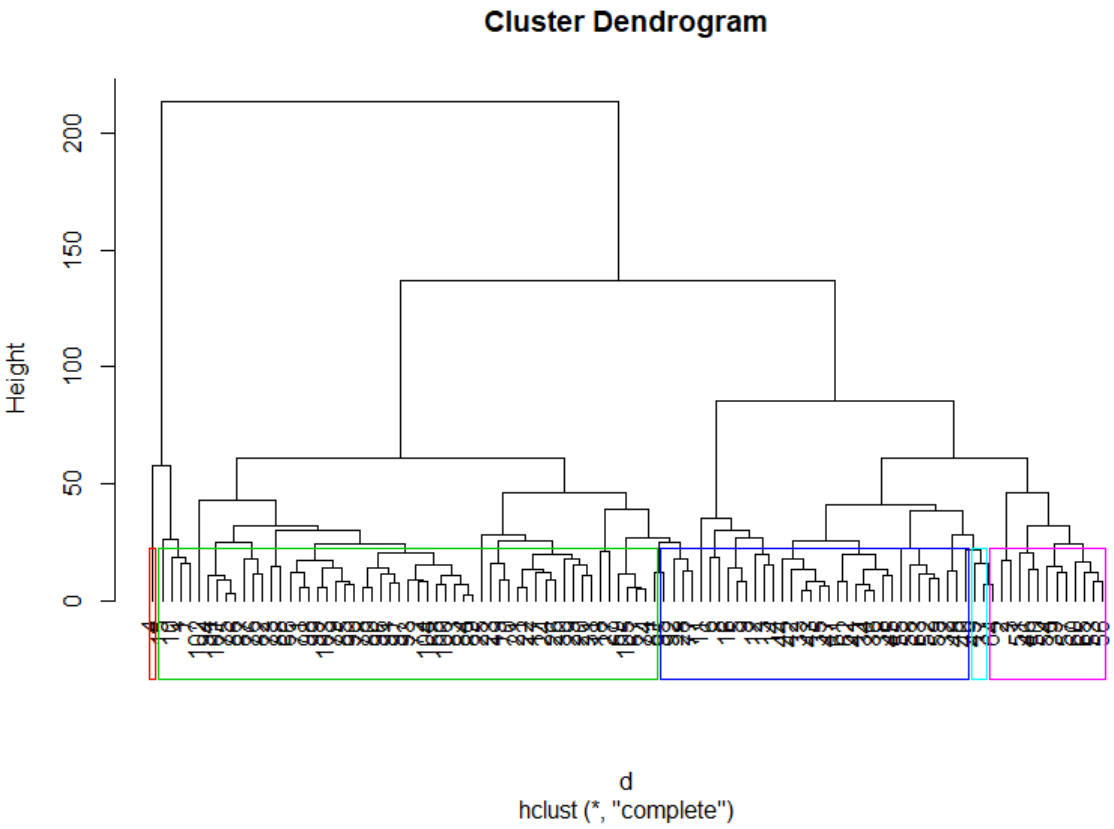
As we can see, accuracy is just: **12.38%,** so misclassification rate: **87.62%**

Confusion Matrix and Misclassification rate for 3 Clusters (Single Linkage):

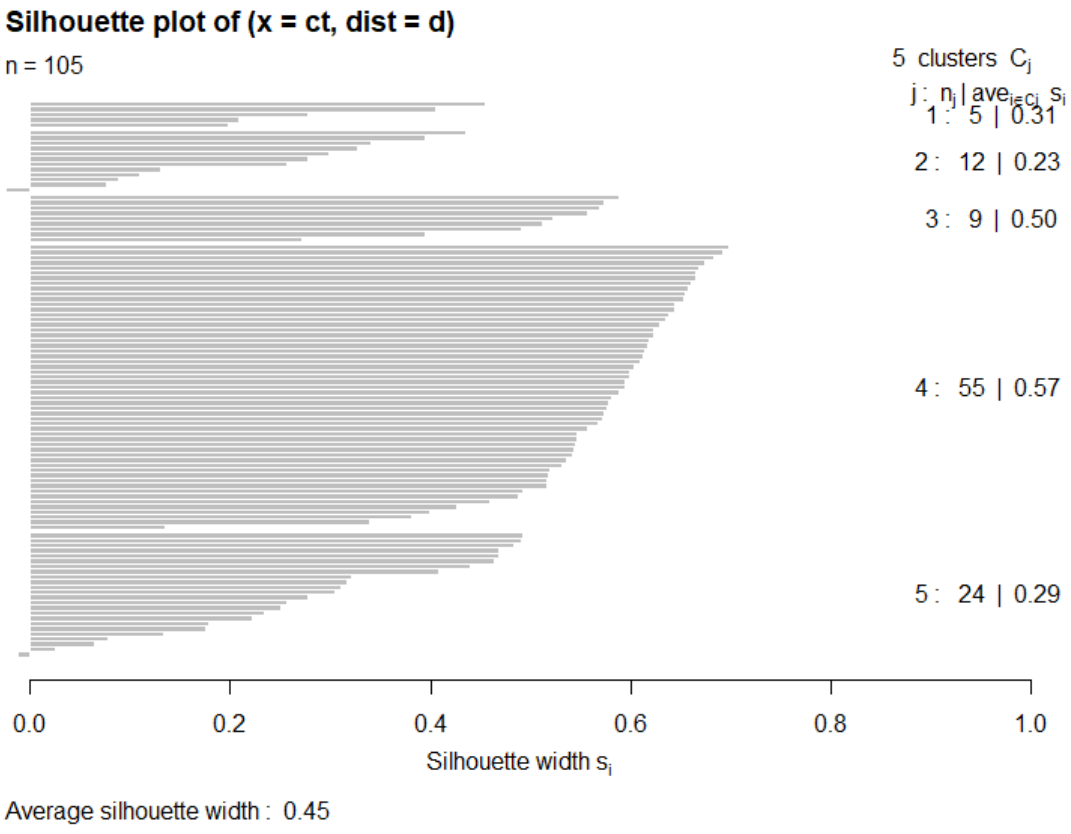
```
          Reference
Prediction  1  2  3  4  5
         1 15  0 20 14  0
         2  1  0  0  0  0
         3  0 15  0  0 40
         4  0  0  0  0  0
         5  0  0  0  0  0

Overall Statistics

               Accuracy : 0.1429
```

Accuracy: **14.29%;** Misclassification rate: **85.71%**

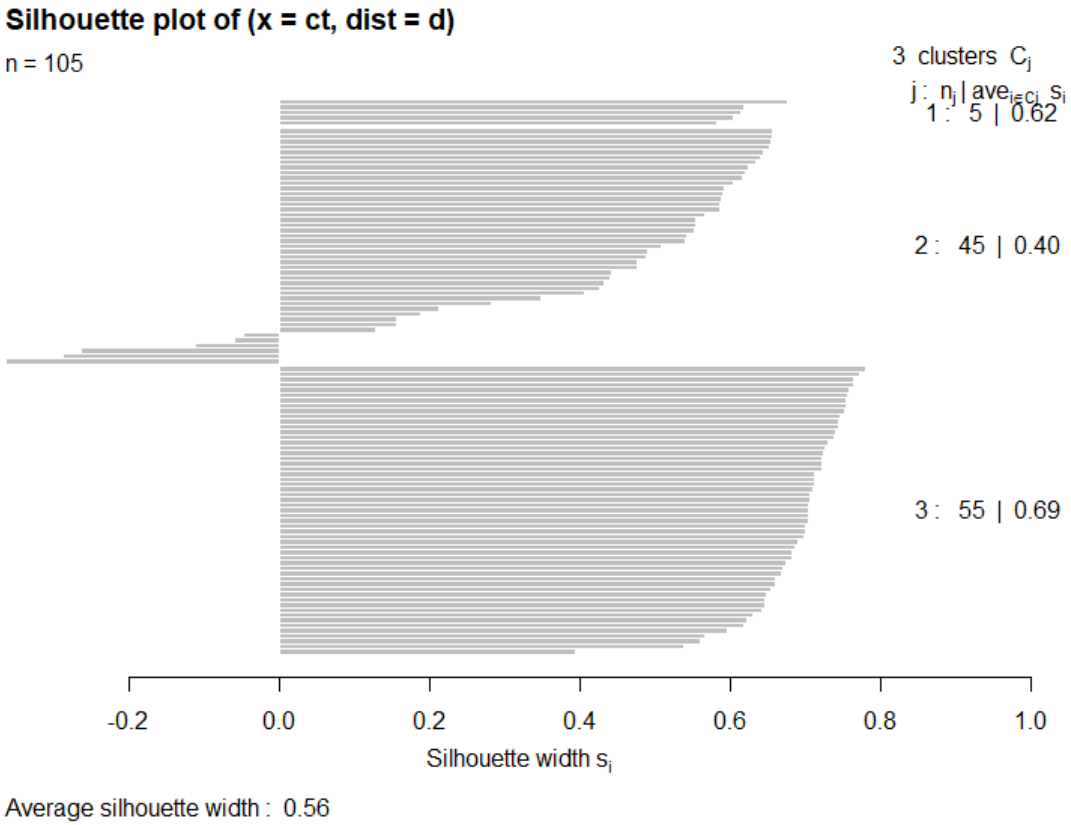
Complete Linkage



Silhouette for 5 clusters Complete Linkage:



Silhouette for 3 clusters for Complete Linkage:



Confusion Matrix and Misclassification rate for 5 clusters (Complete Linkge):

```
      Reference
Prediction  1  2  3  4  5
      1  5  0  0  0  0
      2  2  0  4  6  0
      3  9  0  0  0  0
      4  0 15  0  0 40
      5  0  0 16  8  0

Overall Statistics

          Accuracy : 0.0476
```

As we can see, accuracy is just: **4.76%,** so misclassification rate: **95.24%**

Confusion Matrix and Misclassification rate for 3 clusters (Complete Linkge):

```
      Reference
Prediction  1  2  3  4  5
      1  5  0  0  0  0
      2 11  0 20 14  0
      3  0 15  0  0 40
      4  0  0  0  0  0
      5  0  0  0  0  0
```

Overall Statistics

Accuracy : 0.0476

Accuracy: **4.76%;** Misclassification rate: **95.24%**

Single and Average methods have same performance and perform better than Complete-linkage agglomerative clustering. Complete linkage method is the worst of the three.

Expected and Actual

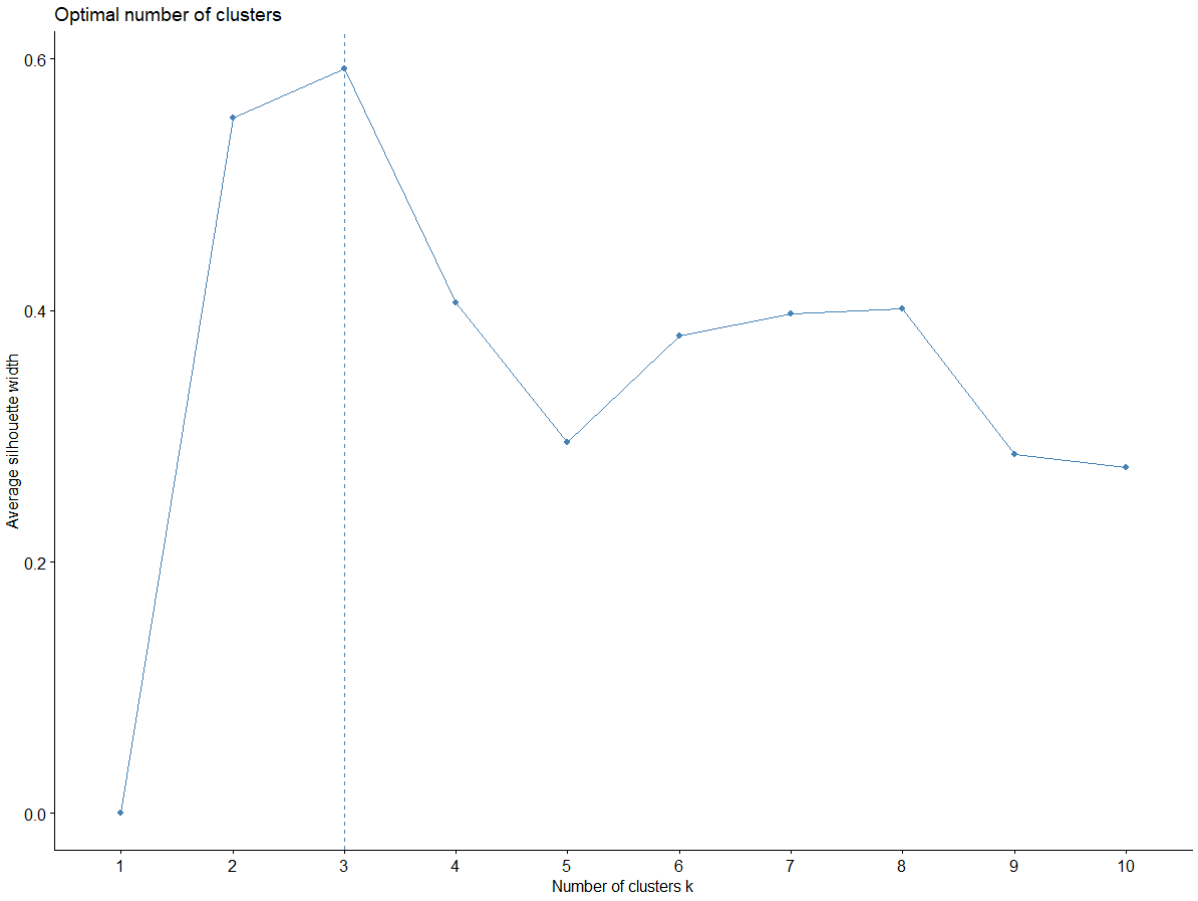
I expected the Single-linkage to perform better. Because Single Linkage only requires that, a single dissimilarity be small for two groups to be considered close together, irrespective of the other observation dissimilarities between the groups. It will therefore have a tendency to combine, at relatively low thresholds, observations linked by a series of close intermediate observations. Average linkage method performance was expected to be a bit poor than single linkage as it takes average distance into account rather than closest.

Whereas, the Complete linkage does the exact opposite. Two groups are considered close only if all the observations in their union are relatively similar.

I expected single-linkage to perform better because from the silhouettes we saw above, the dataset even though it has 5 classes was better classified using 3 clusters. So in this case, we would require a method which would easily group clusters together and that method is single-linkage.

K-means:

Silhouette method to find optimum number of clusters for K-means:



From the graph, the optimal number of clusters for the dataset is 3 even though there are 5 classes in the dataset.

Accuracy and misclassification for 3-clusters (k-means):

```
          Reference
Prediction  1  2  3  4  5
         1 13  0  0  0  0
         2  0 15  0  0  0
         3  3  0 20 14  0
         4  0  0  0  0  0
         5  0  0  0  0  0

Overall Statistics

               Accuracy : 0.7385
```

Accuracy of **73.85%** and misclassification rate of: **26.15%**

Accuracy and misclassification for 5-clusters (k-means):

```
          Reference
Prediction  1  2  3  4  5
         1  2  0  8  7  0
         2  9  0  0  0  0
         3  0 15  0  0  0
         4  5  0  0  0  0
         5  0  0 12  7  0

Overall Statistics

               Accuracy : 0.0308
```

Accuracy is **3.08%** and misclassification rate is: **96.92%**

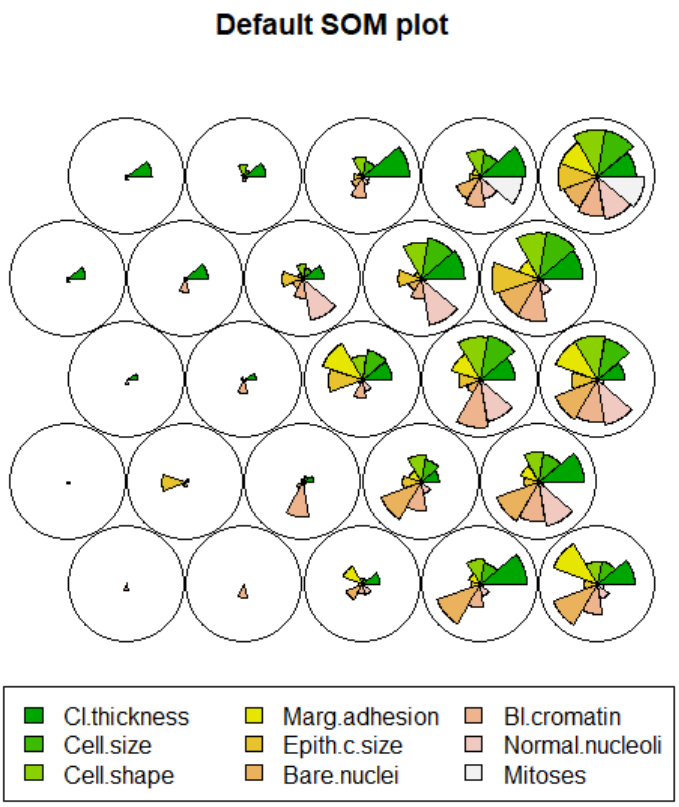
When we observe the values for 5 classes, k-means performs worse than any method of heirarchichal clustering. However, for 3 classes k-means performs the best.

Question 4

Run a batch-SOM analysis on the Wisconsin Breast-Cancer data. Describe how well the SOM methods cluster the tumour cases into benign and malignant. Compute the U-matrix and discuss its representation for these data.

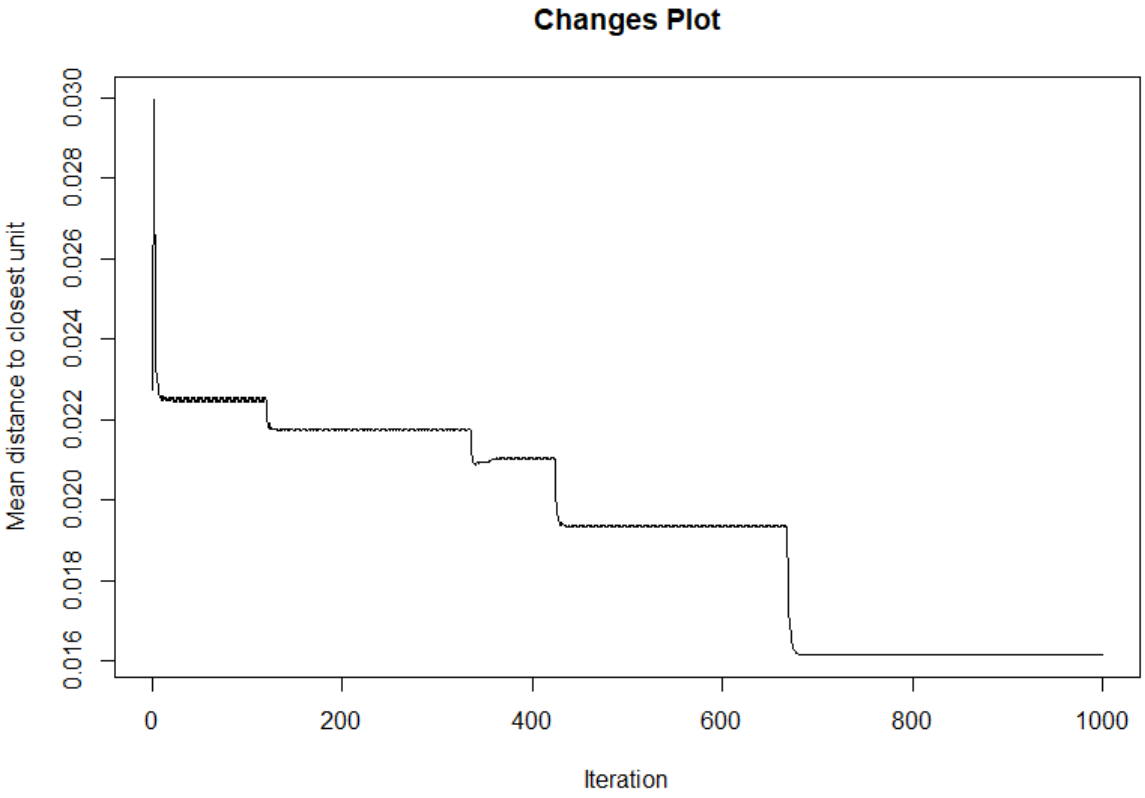
Solution:

The default SOM Plot



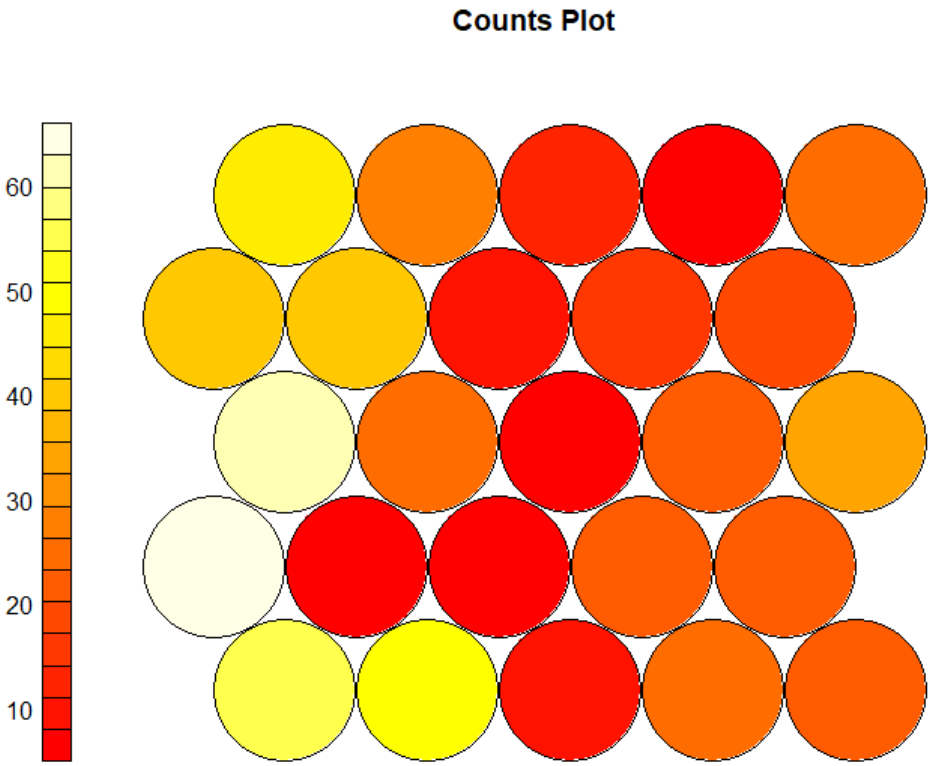
Changes Plot

As we can see, we are able to converge at 1000.



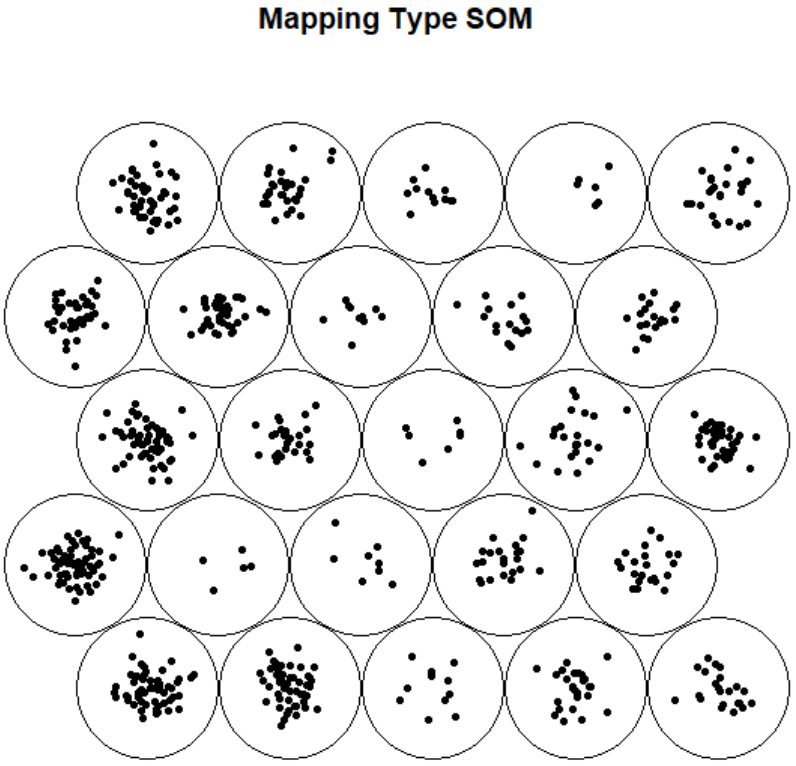
Counts Plot

The 'count' type SOM creates a heatmap based on the number of observations/cases assigned to each cell.



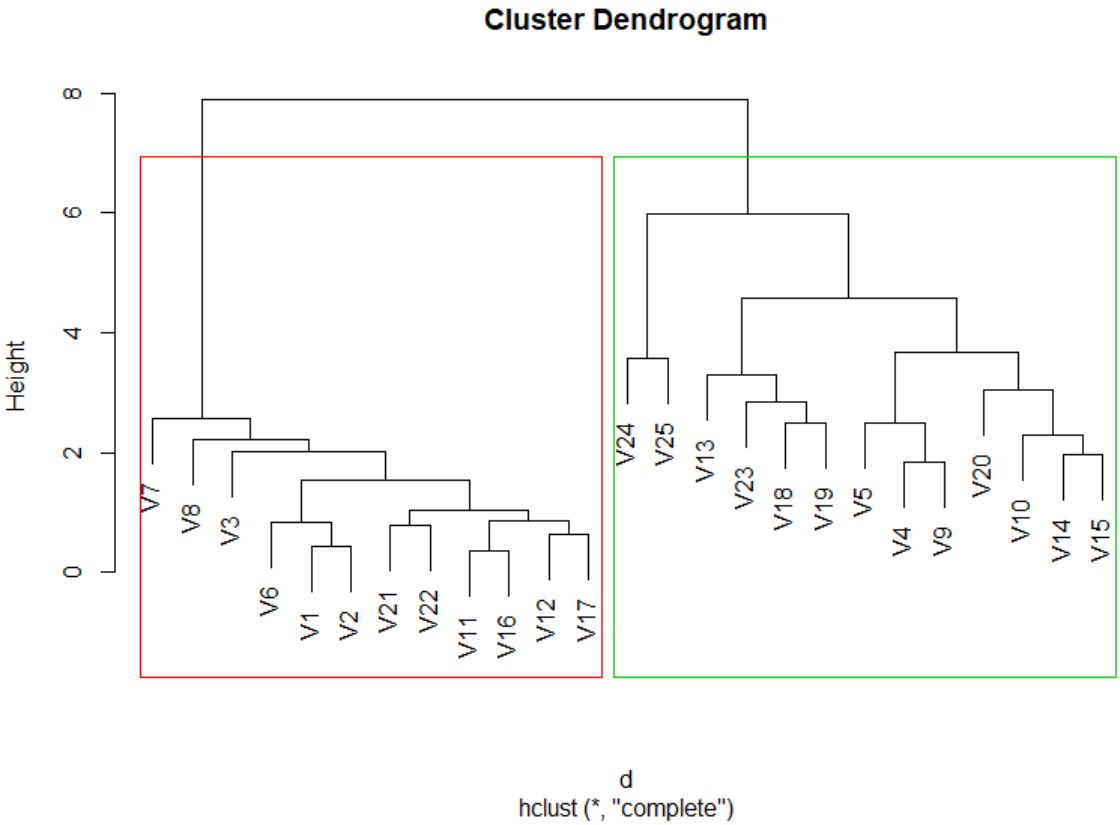
Mapping Plot

The observations/cases are plotted on this map based on how close their stat lines are to the representative vectors.



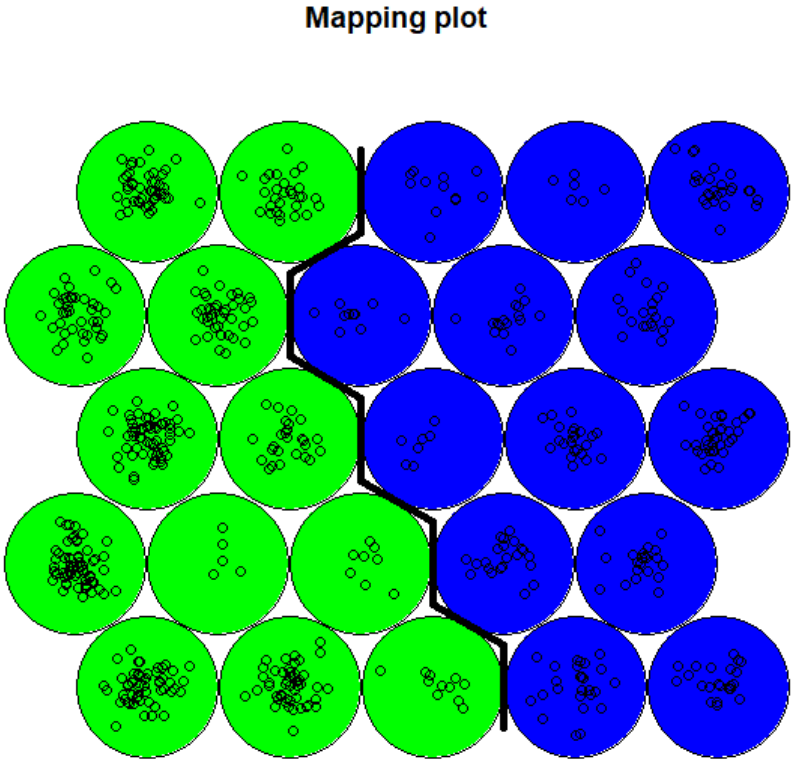
Mapping Distance:
When we plot with type = ‘dist.neighbours’, the cells are coloured depending on the overall distance to their nearest neighbours, which allows us to visualize how far apart different features are in the higher dimensional space. It gives similar information to U-matrix. The ‘dark green’ are the closest to each other on the lower right hand side. While, the ‘light green colours’ are further apart and are dissimilar from the ‘dark green’ ones are separated by a good distance. The ‘white’ ones are very different from their neighbours.

Clustering Dendrogram:



Clustering of the Observations:

The results of the clustering are visualized using the SOM plot function and added the hierarchical clustering to get the below plot. We can see that clusters are well separated as well.



Analysing Accuracy of SOM

I used **xyf** function to implement supervised SOM on the dataset. Predict function was used to make predictions and then accuracy was computed on the 'test' dataset created from the original dataset.

The obtained accuracy was 91.57%. Therefore, the SOM method is able to cluster the tumor cases into benign and malignant with very good accuracy.

```
Confusion Matrix and Statistics

           Reference
Prediction  benign malignant
  benign        46         2
  malignant      5        30

               Accuracy : 0.9157
```