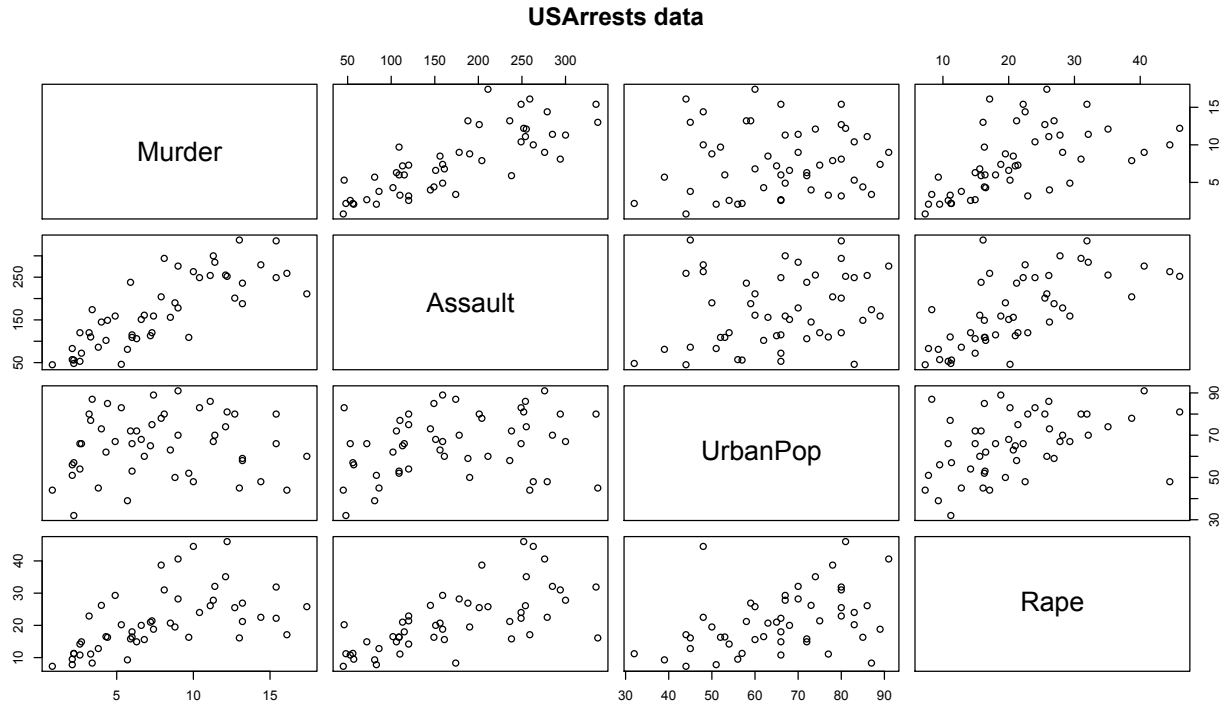
Homework 2 Report

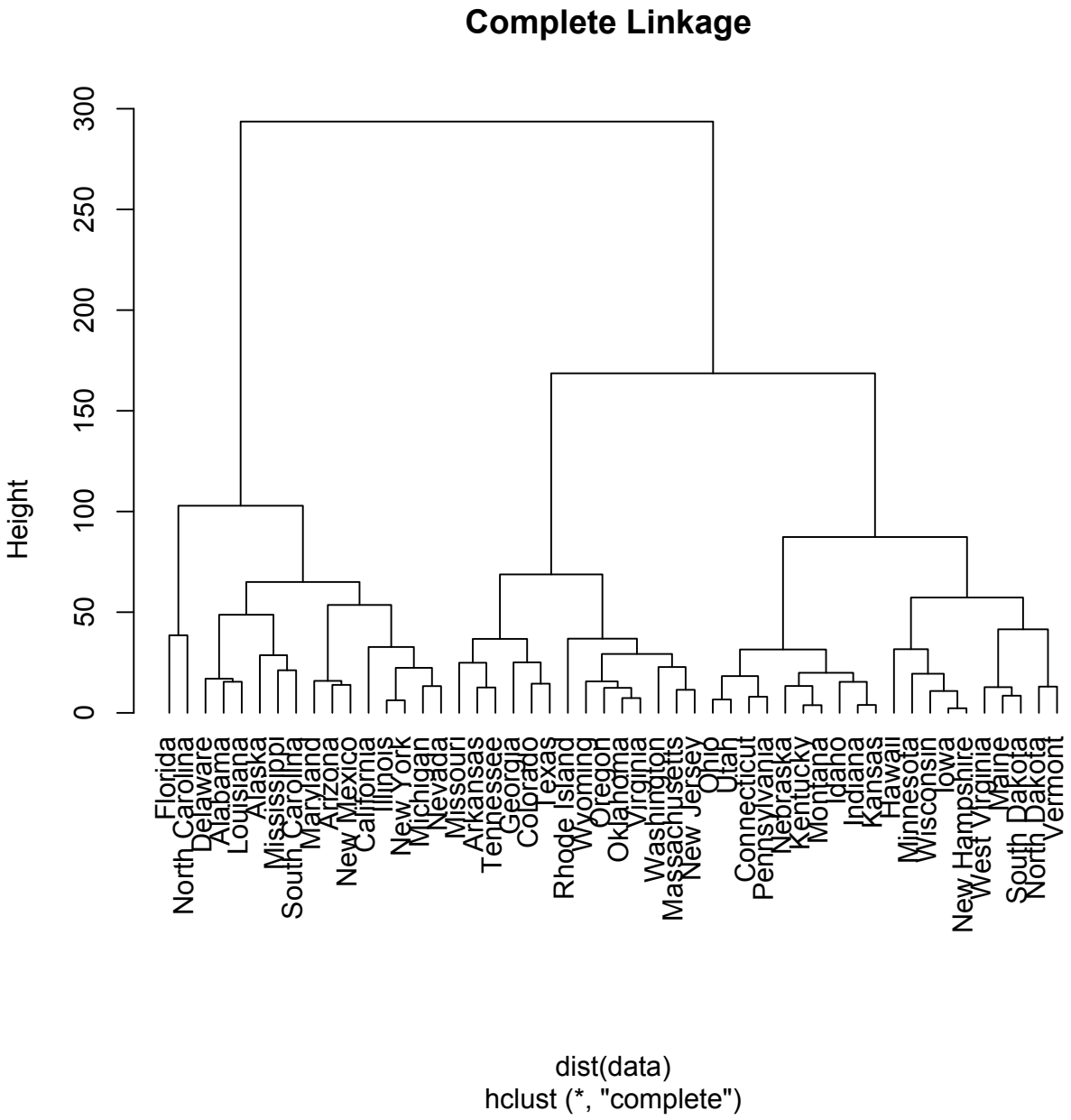
Name: Akshay Adlakha (akshayad)
Person#: 50317479

Problem 1: In this problem, I was given a USArrests dataset. I had to perform hierarchical clustering on the states.

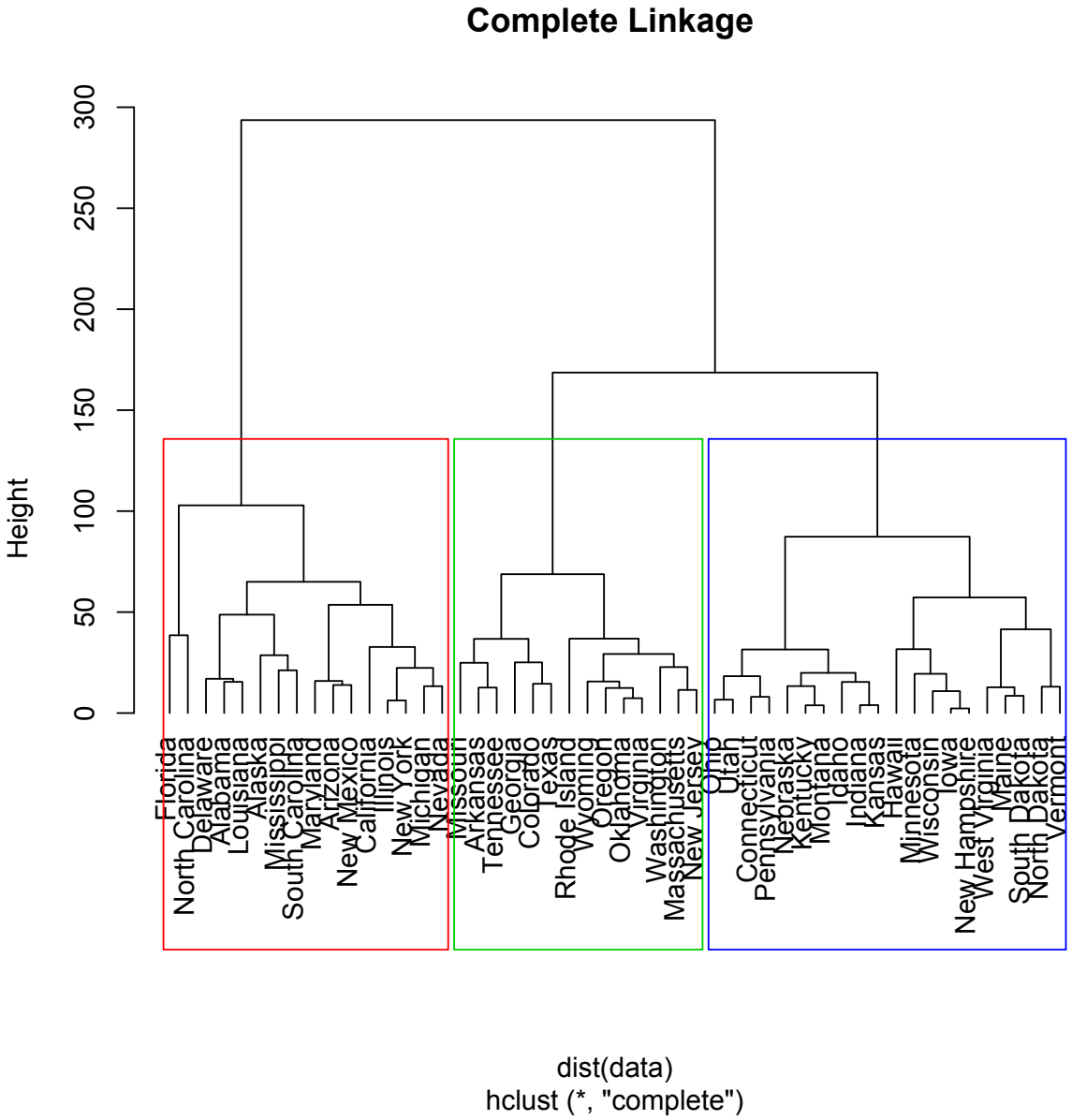
After loading the data, did a sanity check by checking NA values. And, visualized it through pairs method. Here is the result:



Then, did hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Here is the hierarchical clustering structure:

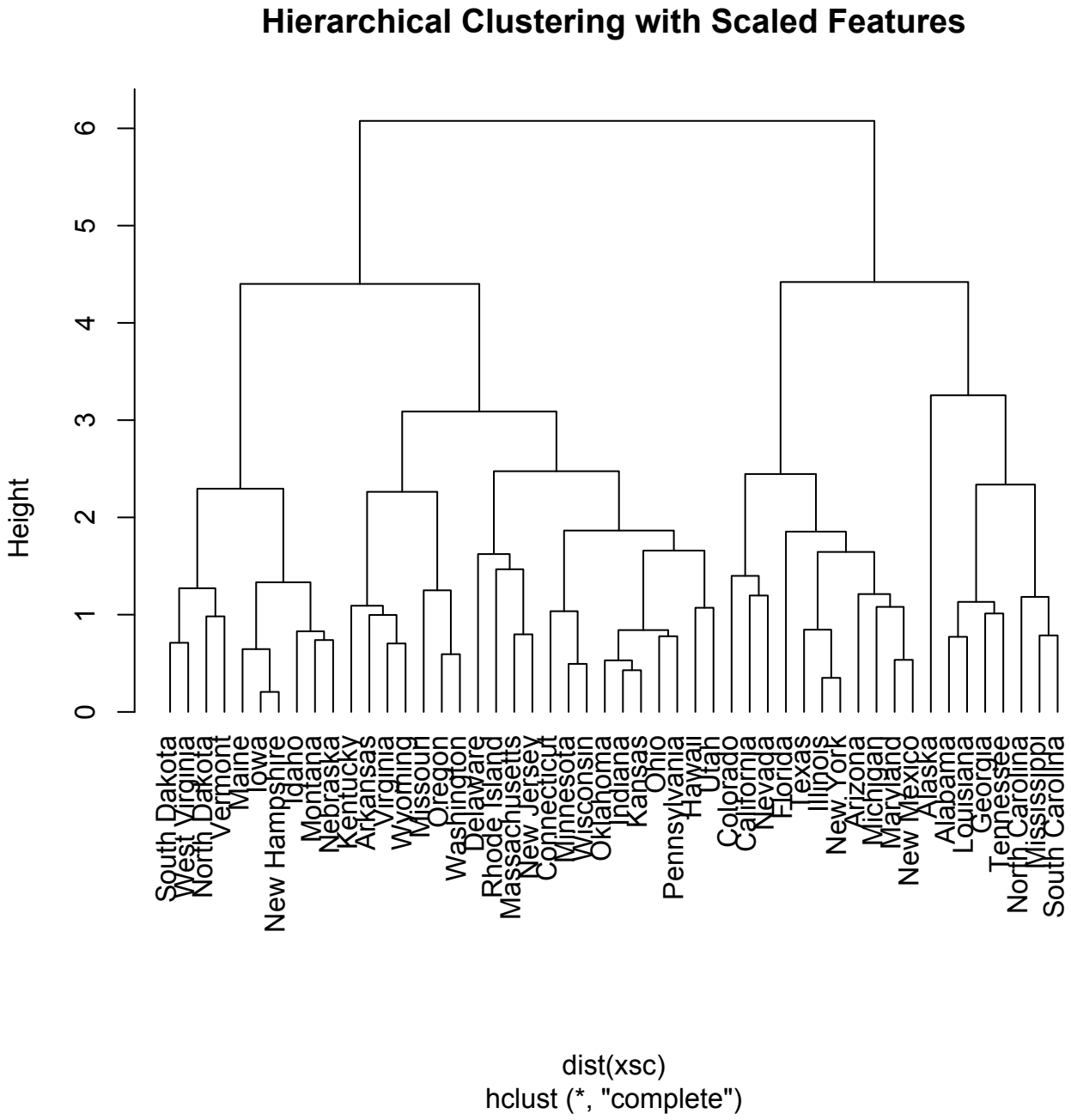


From the above structure, we can see that cutting the tree horizontally at height of 150 would result in three clusters. I get the following result:

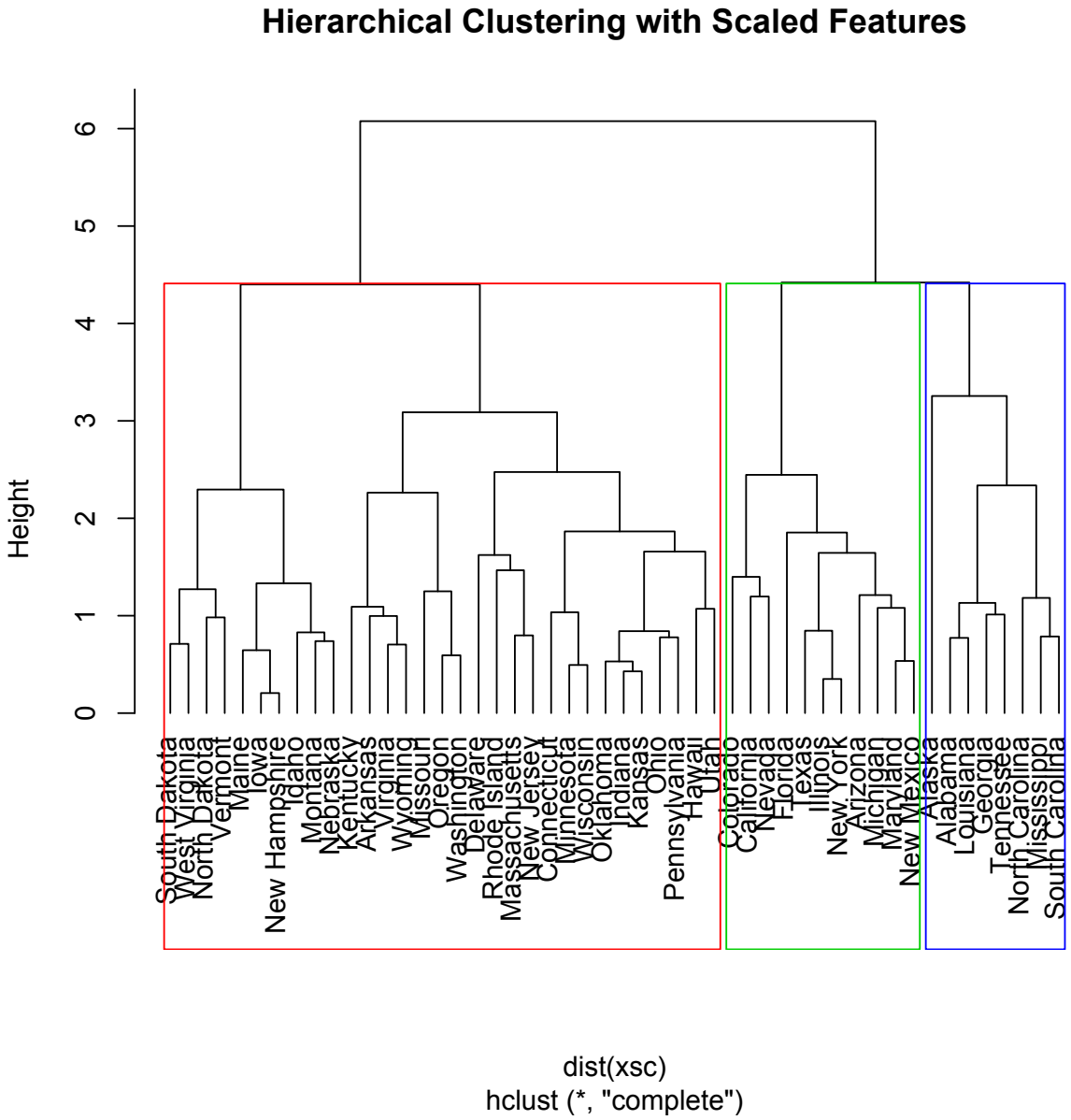


The states are clustered into three different clusters. As we can see that the states are grouped with different colors to distinguish that which state belongs to which cluster.

Then, the variables are scaled using the scale function to have the standard deviation one. Hierarchical clustering is then performed on this scaled data. I get the following result:



Then, the scaled features cluster is cut into three different clusters.



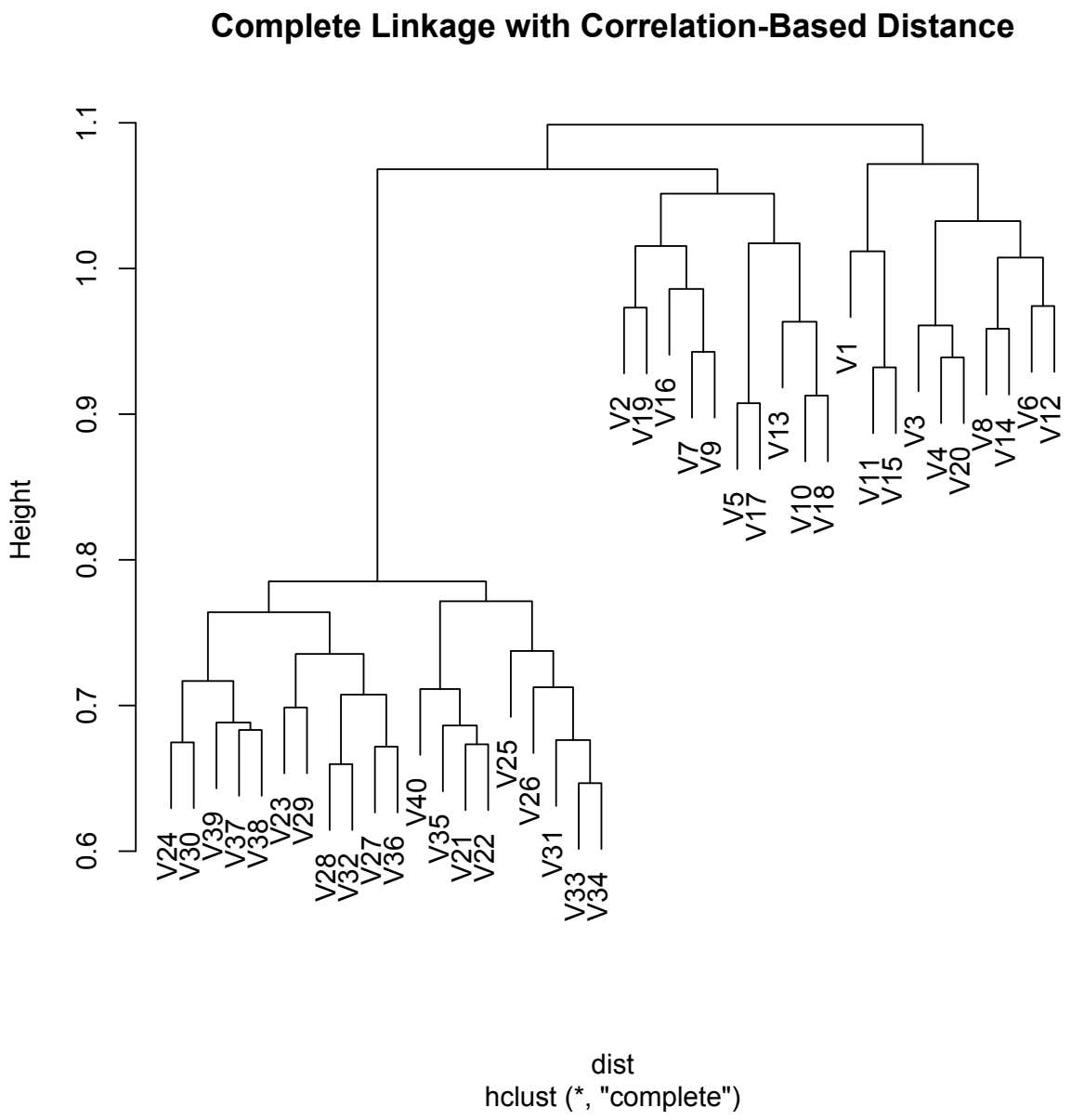
We can see that the number of states in one of the clusters has increased by scaling the variables. Thus, scaling made a significant impact on the clustering.

So, it is important to scale the variables first if the feature range is much greater than others, then the distance would be bigger and impact on the overall distance is high. Each variable will in effect be given equal importance in the hierarchical clustering if the variables scaling is done to have standard deviation one before the inter-observation dissimilarities are computed.

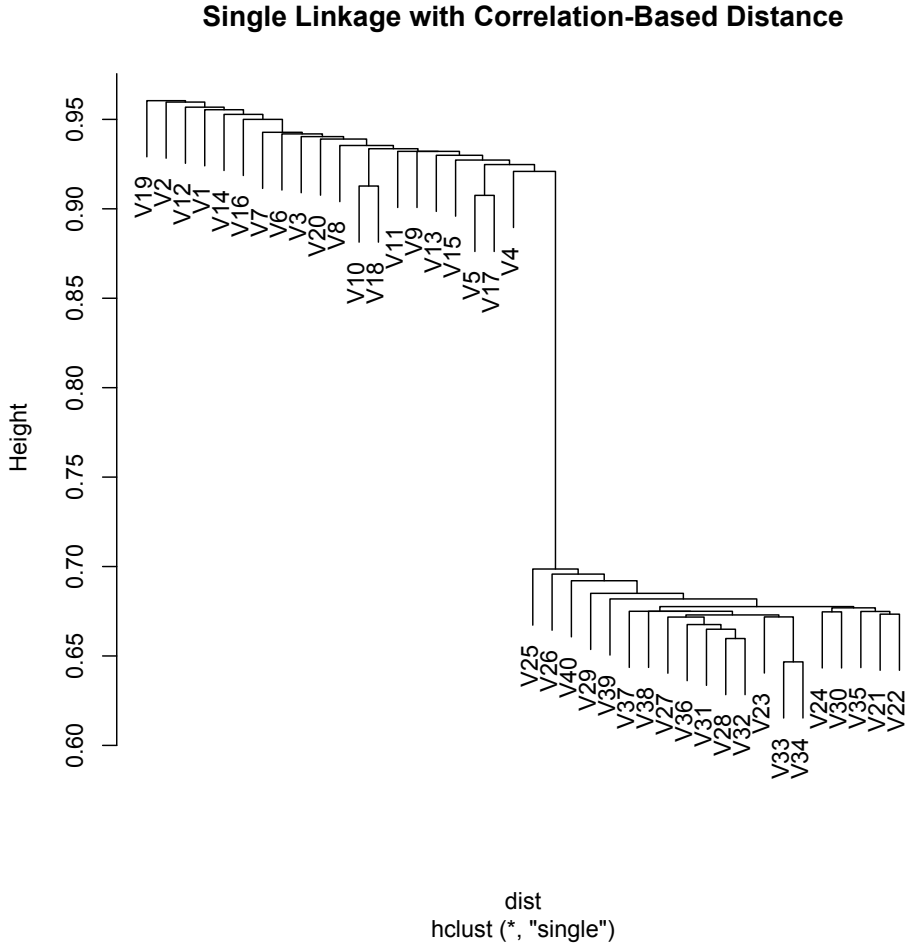
Problem 2: In this problem, I was given a dataset of gene expression from StatsLearning. I had to perform hierarchical clustering using correlation based distance.

Loaded the data using read.csv and specified argument header=F. Then did a sanity check and explored the data.

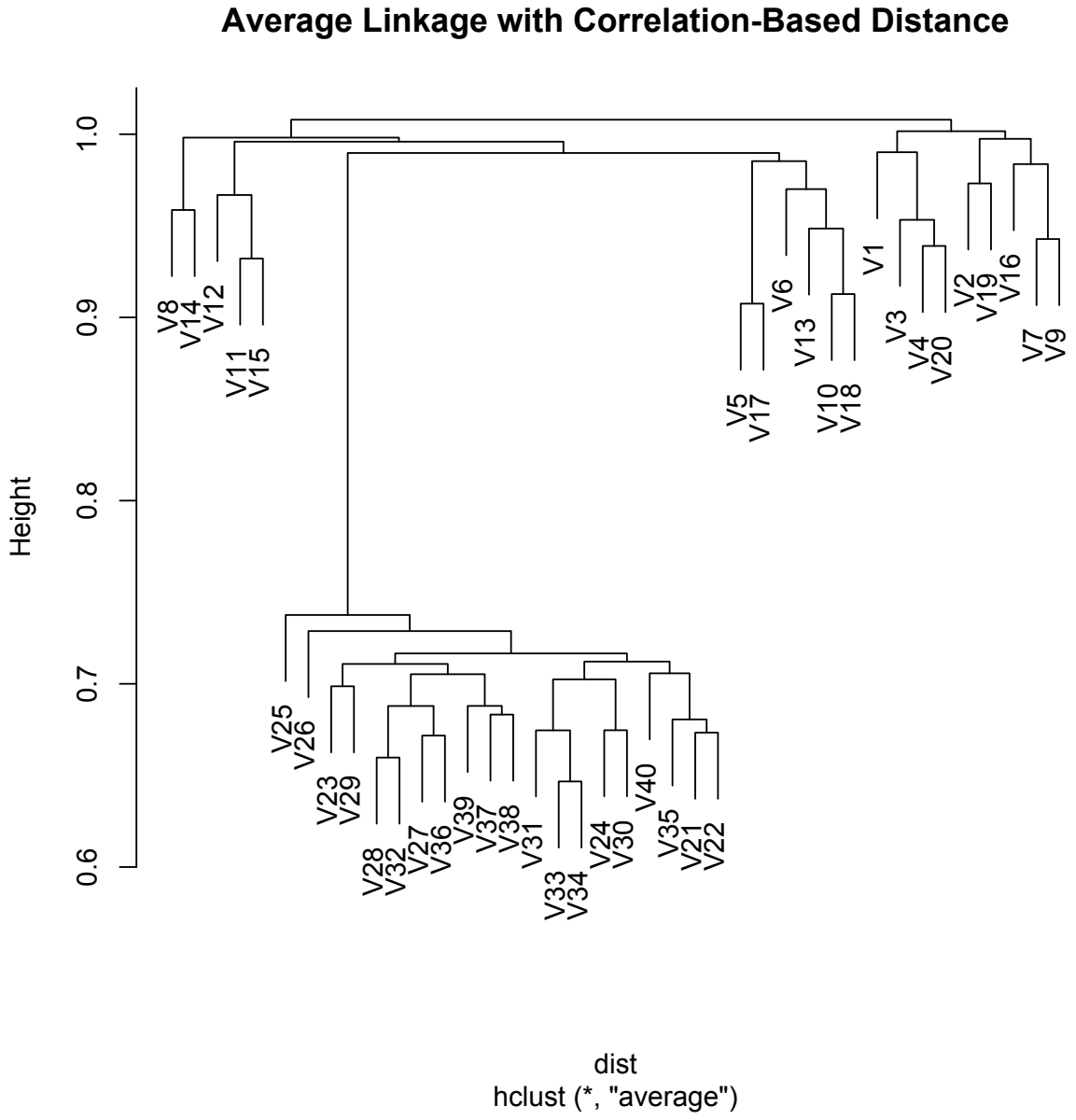
Then did hierarchical clustering using correlation based distance with complete linkage. Here is the result:



Then created hierarchical clustering using correlation based distance with single linkage. Got the following result:



Then created hierarchical clustering using correlation based distance with average linkage. Got the following result:



The samples get separated into two groups of first 20 patients and next 20 patients with complete and single. We get different structures when different linkages are used. When complete and single linkage used, we get two clusters but with average linkage we get three clusters.

c) We can use PCA to know which genes differ the most across the two groups. And, the absolute values of the total loadings for each gene as it characterizes the weight of each gene.
After implementing PCA, rotation gave the Eigen values of the features. Then took the absolute value of the total loadings for each gene to see the most different genes across two groups.

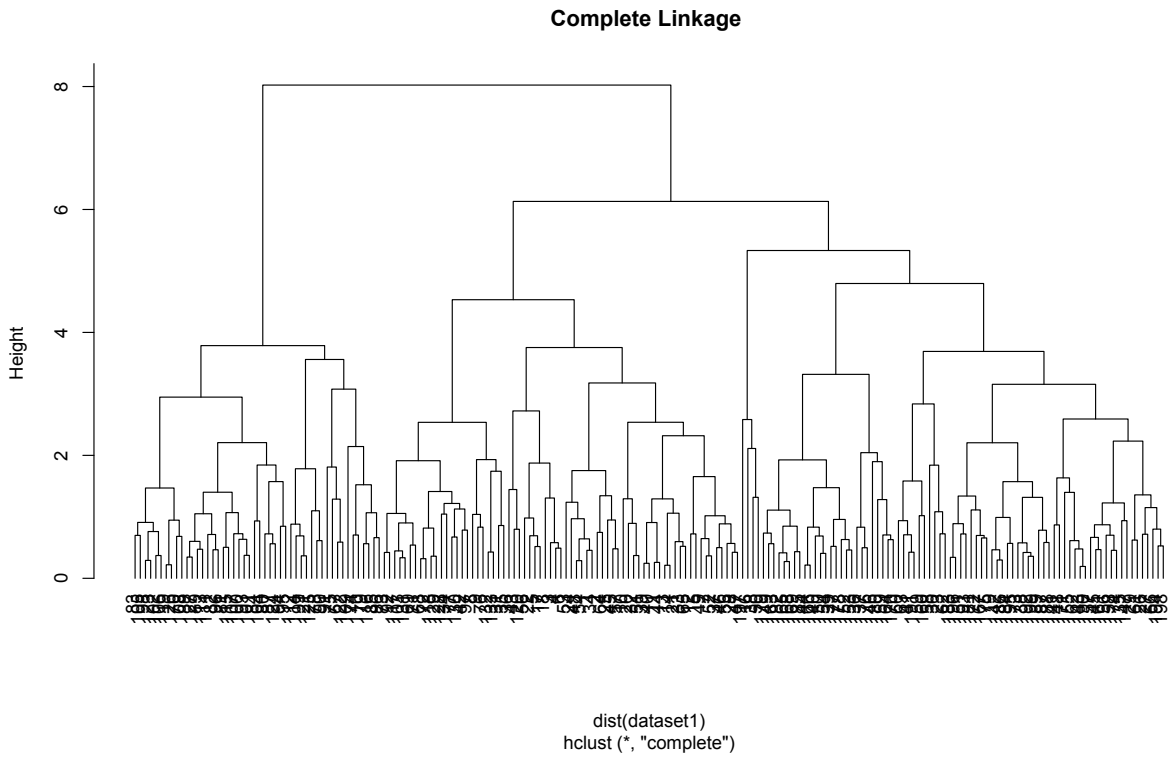
The following are the 10 most different genes:

```
pcs <- order(abs(cumsum), decreasing = TRUE)
pcs[1:10]
[1] 865  68 911 428 624  11 524 803 980 822
```

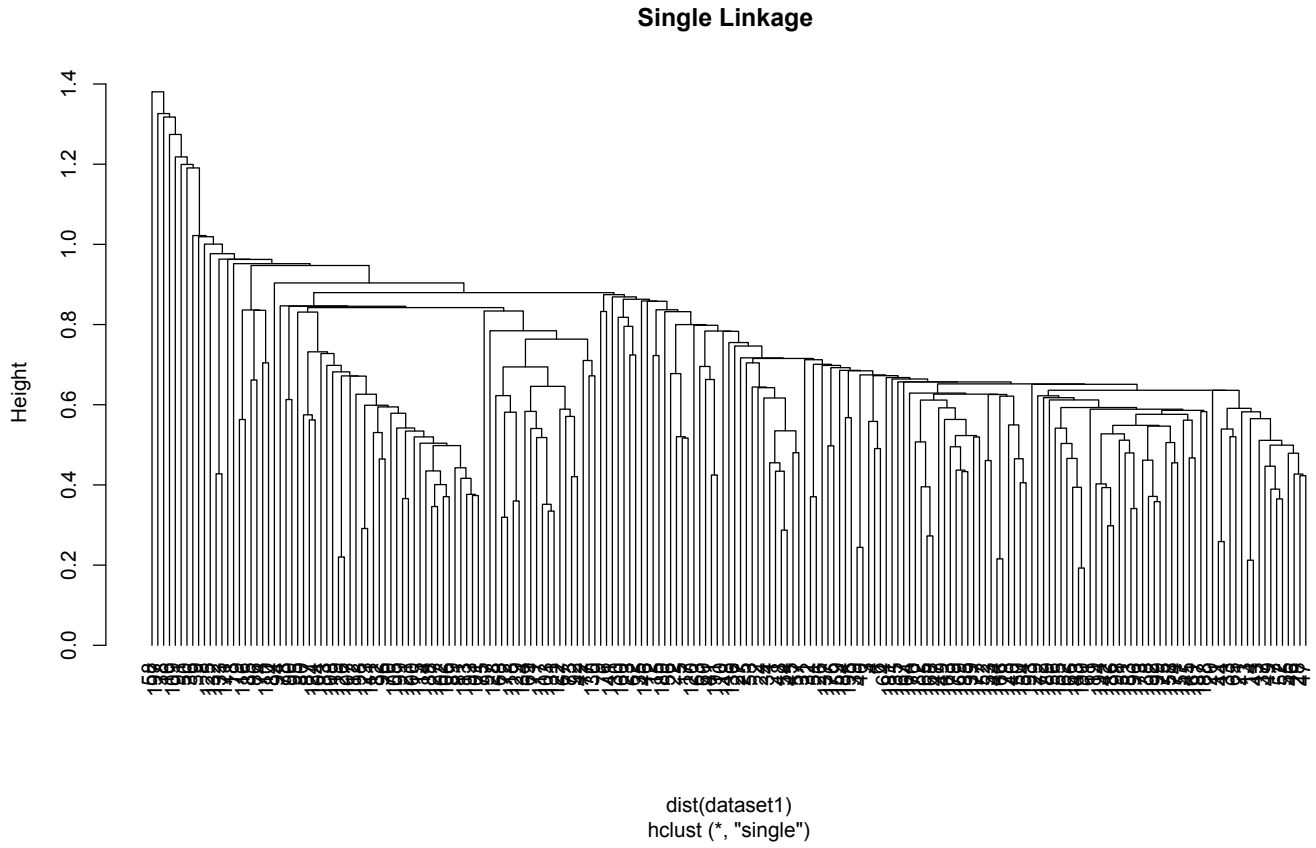
Problem 3: In this problem, I was given a seeds dataset and had to perform the single-linkage, average linkage, and complete-linkage agglomerative hierarchical clustering.

After loading the data, did a sanity check by checking missing values and explored the data. Removed the seed group column to perform clustering. And, scaled the data to have standard deviation one.

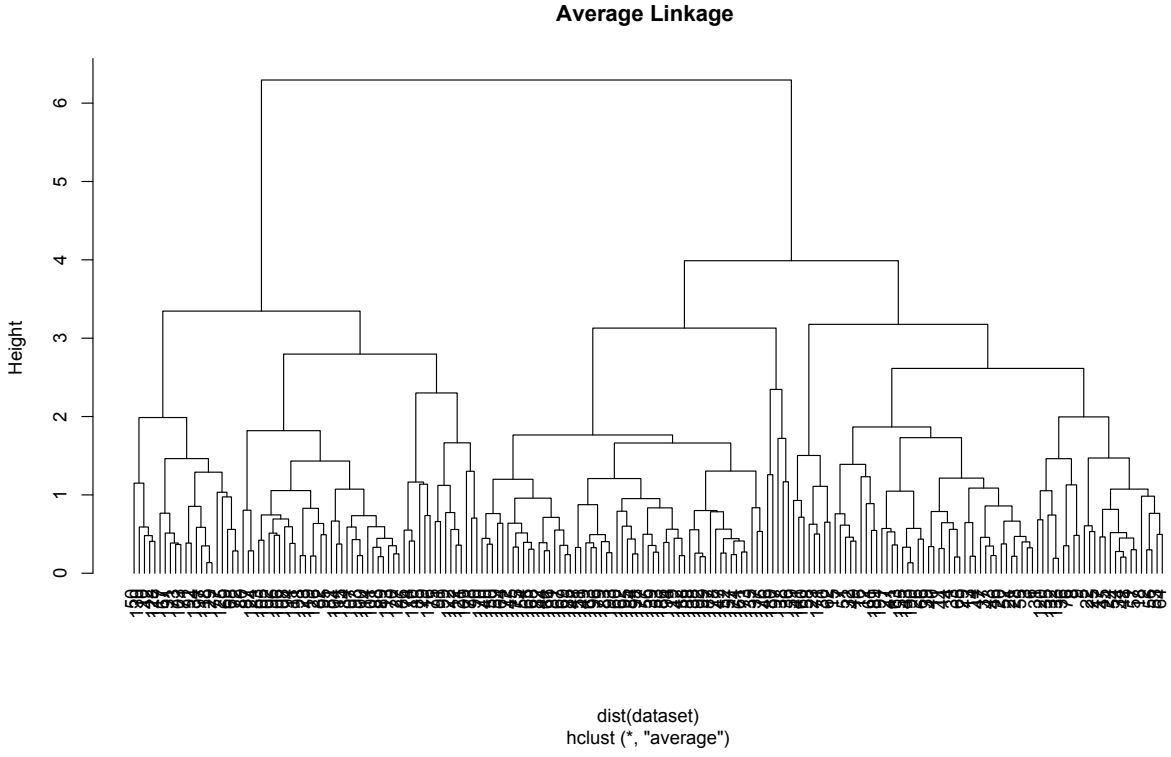
Then, clustered data based on complete linkage.



Then, clustered data based on single linkage.



Then, clustered data based on average linkage.



Then, I computed rand index and adjusted rand index to compare the performance of all three methods used. Here is the result:

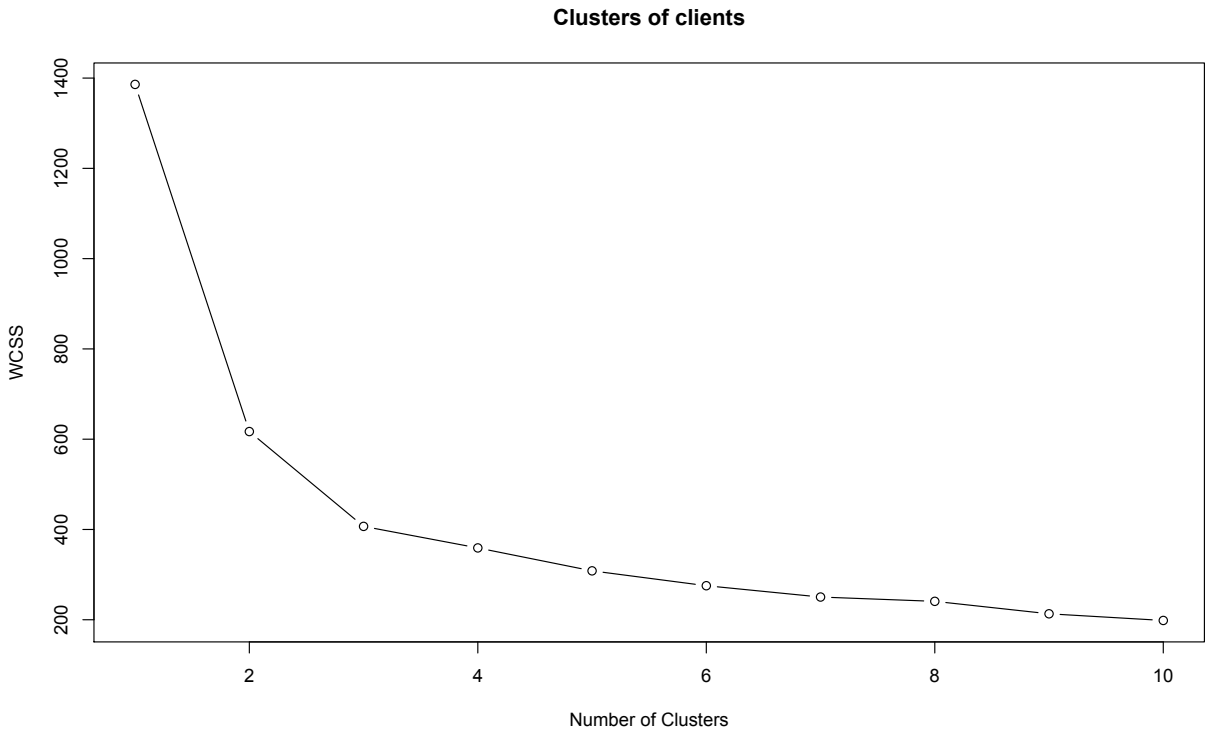
```
> table
                          Complete       Single      Average
Rand Index           0.8031572001 0.3370387290 0.8764529719
Adjusted Rand Index 0.5599256850 0.0002118046 0.7208841642
```

From the above table, the rand index and adjusted rand index for Average Linkage came out to be the maximum. The rand index value range between 0 and 1. A value 0 indicates that the clusterings do not agree on any pair of points and 1 indicates that the data clusterings are exactly the same.
So, the result of Average linkage is better.

Average linkage with euclidean distance performs the best among these three methods and single linkage method with euclidean distance performs worst.

After implementing these three methods, I was expecting the complete linkage would perform the best among these three methods. But the average linkage method came to be the best one.

b) Then clustered data based on k-means algorithm. This algorithm requires to have the value of k beforehand. To know the value of k, I used the elbow method. Here is the result:



From above plot, interpreted the value of k. The value of k is chosen where the curve is bending. So, we have k=3.
Then, used the k-means method with k=3. And, calculated the rand index and adjusted rand index to compare the result of part a) and part b). The same I computed for k-medoids algorithm with 3 clusters.

Comparison result of these algorithms with hierarchical clustering are given below:

```
> table
                       Complete       Single      Average      K-Means    K-medoids
Rand Index          0.8031572001 0.3370387290 0.8764529719 0.9123902340 0.8863001878
Adjusted Rand Index 0.5599256850 0.0002118046 0.7208841642 0.8021194264 0.7436104534
```

Here I got the maximum rand index for K-means algorithm. So, it performs better among all these algorithms.
Hierarchical clustering can be done by interpreting the appropriate numbers of clusters from the dendrogram. It is flexible as we don't need to have k beforehand and has less assumptions. However, since we knew the number of clusters already, so would have chosen k-means algorithms to cluster the data.