Homework 2 : Statistical Data Mining II (EAS 507)

Shubham Sharma, Person No.: 50290293, Class No.: 43

April 1, 2019

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 Eu- clidean distance, after scaling the variables to have standard de- viation 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.

The given dataset *USArrests* is part of the base *R* package which has data of 50 states in alphabetical order. The dataset has four variables and all of them have very different means as well as variance.

The figure 1 shows the dendogram for hierarchical clustering with complete linkage and Euclidean distance. For three disctinct clusters, the dendogram needs to be cut at a height of 110. The cluster distribution for different states is shown in Figure 2.

The figure 3 shows the dendogram for hierarchical clustering with complete linkage and Euclidean distance after scaling. The cluster distribution for different states is shown in Figure 4 after scaling has been done. Figure 5 shows a comparison of the clustering results before and after scaling. Thus, we see that the results before and after scaling differ. The rand index for the two cases come out to be 0.69 indicating a difference in clustering.

As has already been discussed, in this case *Assault* has maximum mean as well as variance and thus will have more weightage in deciding the clusters. Thus, scaling should be done in this case since a good clustering is defined by the degree of the closeness of data points within the cluster. The intuition behind this is that since the clustering algorithms require a definition of distance, if we do not scale the data, we give attributes which have larger magnitudes more importance such as *Assault* in this case. Thus, after scaling the clusters obtained are more accurate.

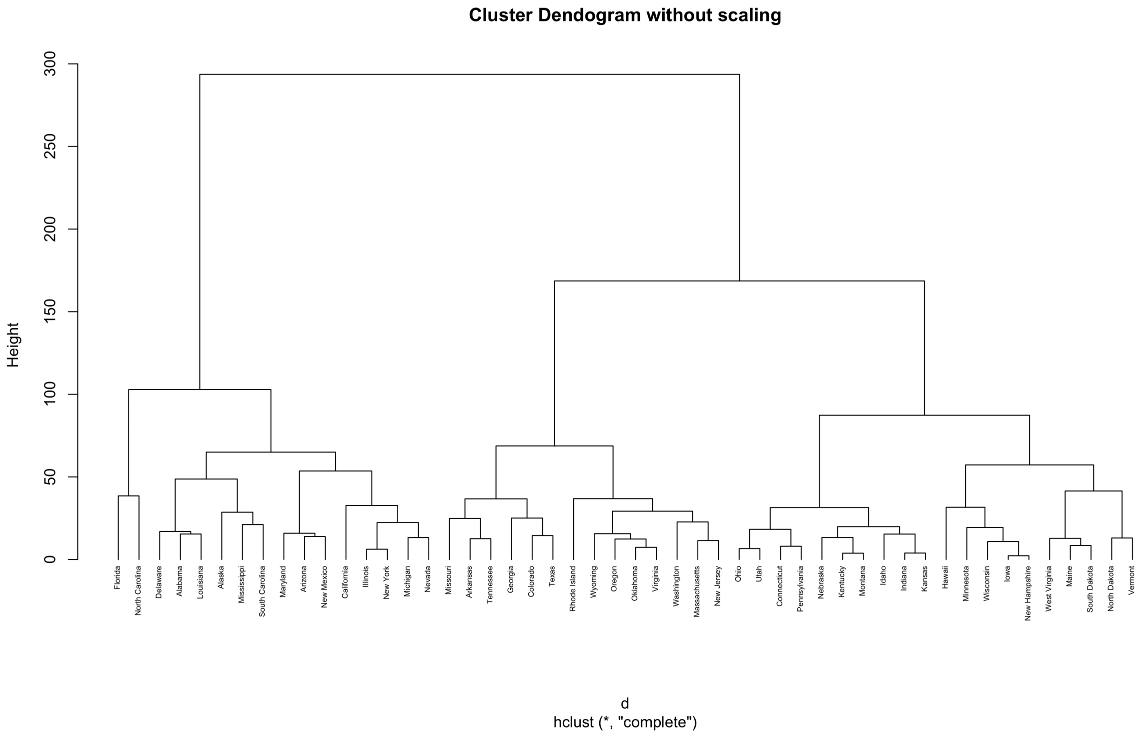
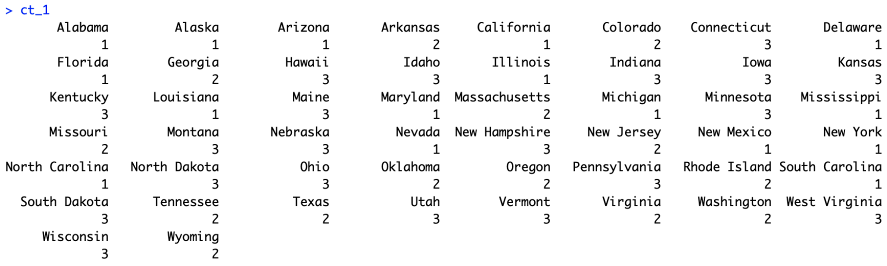


Figure 1: Dendogram for hierarchical clustering with complete linkage and Euclidean distance before scaling



```
> ct_1
       Alabama         Alaska        Arizona       Arkansas     California       Colorado    Connecticut       Delaware 
             1              1              1              2              1              2              3              1 
       Florida        Georgia         Hawaii          Idaho       Illinois        Indiana           Iowa         Kansas 
             1              2              3              3              1              3              3              3 
      Kentucky      Louisiana          Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi 
             3              1              3              1              2              1              3              1 
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey     New Mexico       New York 
             2              3              3              1              3              2              1              1 
North Carolina   North Dakota           Ohio       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
             1              3              3              2              2              3              2              1 
  South Dakota      Tennessee          Texas           Utah        Vermont       Virginia     Washington  West Virginia 
             3              2              2              3              3              2              2              3 
     Wisconsin        Wyoming 
             3              2 
```

Figure 2: Cluster distribution for different states before scaling

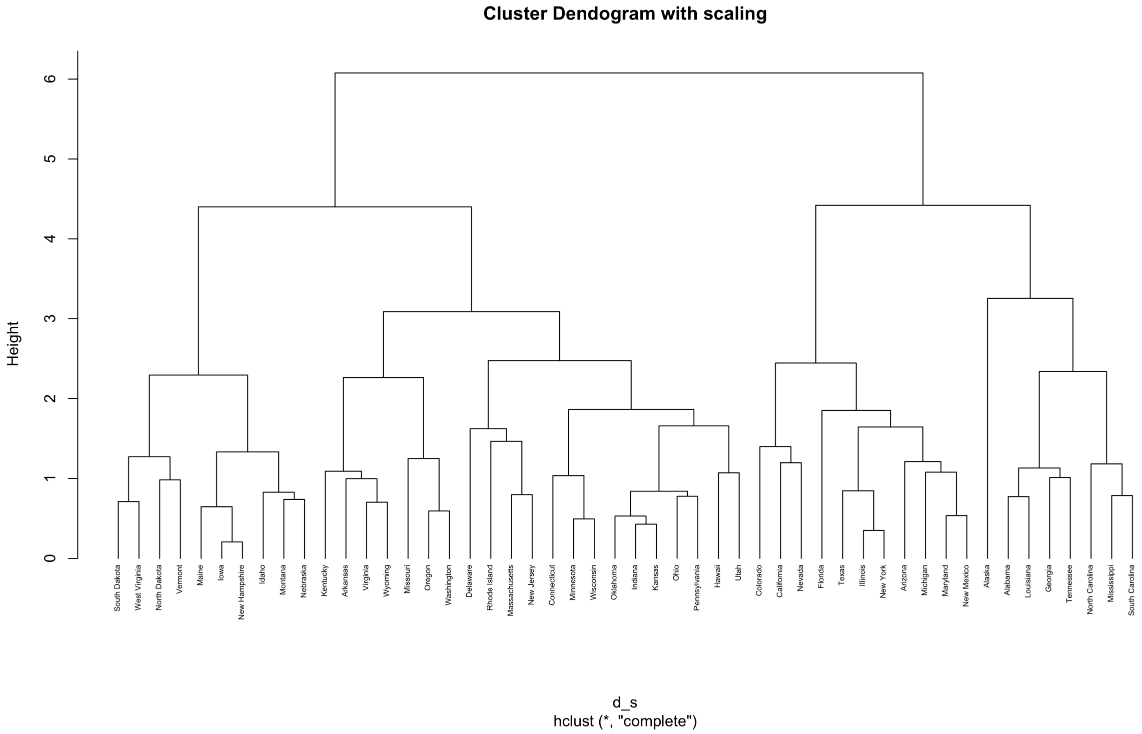
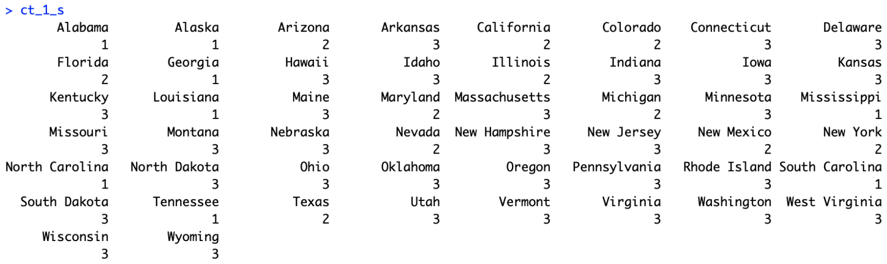
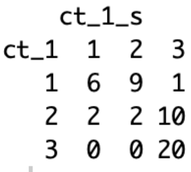


Figure 3: Dendogram for hierarchical clustering with complete linkage and Euclidean distance after scaling



```
> ct_1_s
       Alabama         Alaska        Arizona       Arkansas     California       Colorado    Connecticut       Delaware 
             1              1              2              3              2              2              3              3 
       Florida        Georgia         Hawaii          Idaho       Illinois        Indiana           Iowa         Kansas 
             2              1              3              3              2              3              3              3 
      Kentucky      Louisiana          Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi 
             3              1              3              2              3              2              3              1 
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey     New Mexico       New York 
             3              3              3              2              3              3              2              2 
North Carolina   North Dakota           Ohio       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
             1              3              3              3              3              3              3              1 
  South Dakota      Tennessee          Texas           Utah        Vermont       Virginia     Washington  West Virginia 
             3              1              2              3              3              3              3              3 
     Wisconsin        Wyoming 
             3              3 
```

Figure 4: Cluster distribution for different states after scaling



```
     ct_1_s
ct_1  1  2  3
   1  6  9  1
   2  2  2 10
   3  0  0 20
```

Figure 5: Table to compare clustering results before and after scaling, ct_1: before scaling, ct_1_s: after scaling

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.

The given dataset namely the gene expresion data set have 1000 observations(genes) and 40 variables(tissue samples). We observe that mean and variance for all the features lie nearby. The first 20 samples are from healthy patients whereas the latter are from diseased patients.

The dendograms for Complete, Single and Average linkage are shown in figures 6, 7 and 8 respectively.

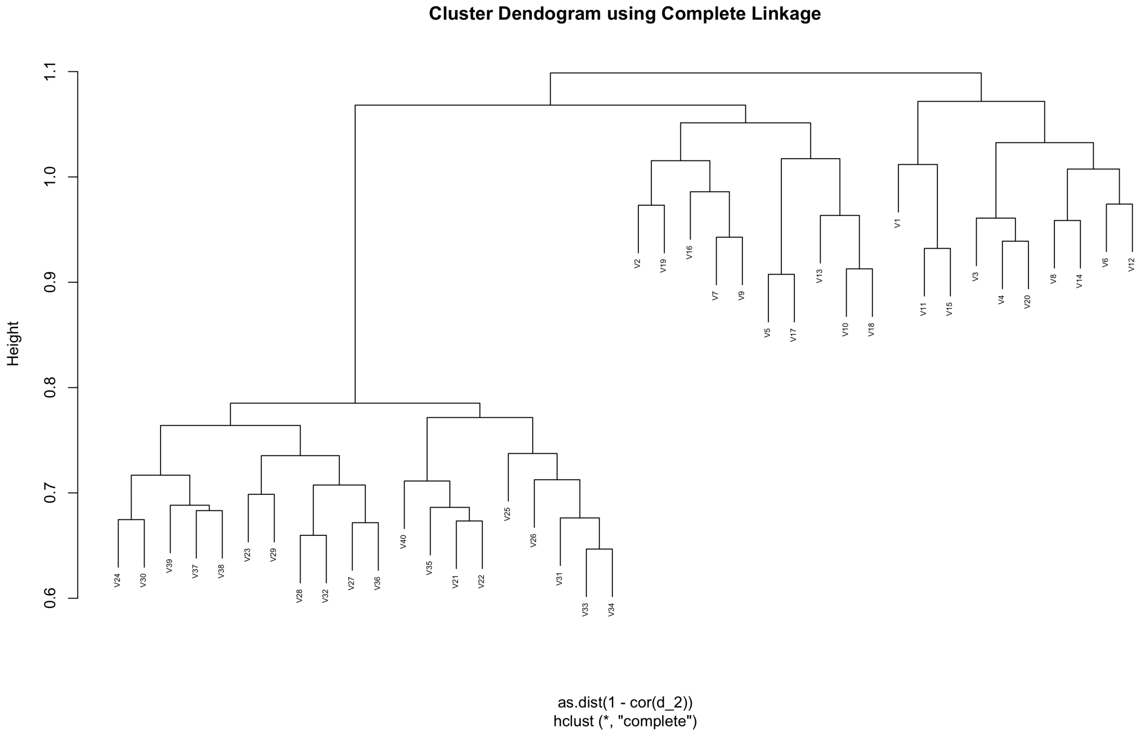


Figure 6: Dendogram for Complete Linkage

The genes separate the samples into the two groups for Single and Complete Linkage but not for Average linkage as is clear from the graph. Thus, the results depend on the type of linkage used.

Principal Component Analysis is applied to identify which genes differ the most across the two groups. To get the maximum variance in the genes we apply PCA to genes and thus we transpose the matrix to make the genes as our column. We sum up the principal components loadings(absloute value) for each gene as it characterizes the weight of each gene and the genes which have more sum value will differ most across the two groups. Thus, the top 10 genes obtained that differ most across the two are 865, 68, 911, 428, 624, 11, 524, 980, 803 and 25.

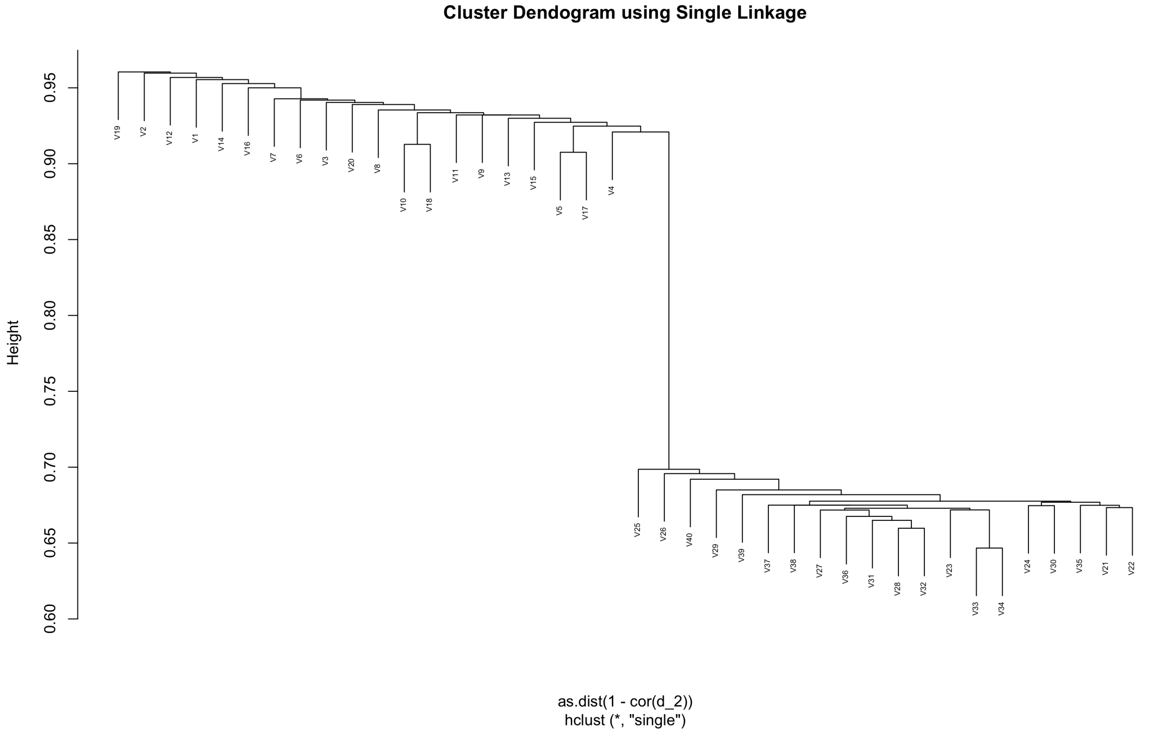


Figure 7: Dendogram for Single Linkage

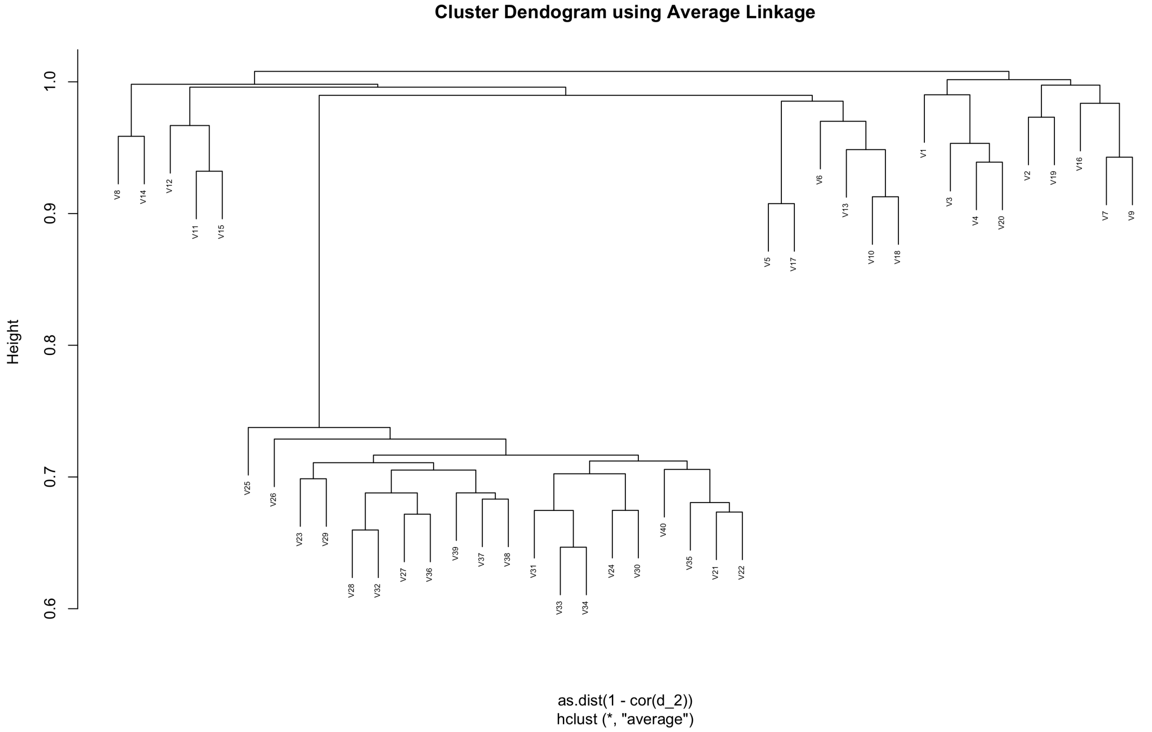


Figure 8: Dendogram for Average Linkage

Question 3: Access the data primate.scapulae (on UB learns).

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?.

The primate scapulae dataset has 105 observations of 11 variables. A majority of values for the feature *gamma* is NA and thus this feature is ommitted.

The dendograms for Single linkage, Average linkage and Complete linkage are shown in figures 9, 10 and 11 respectively.

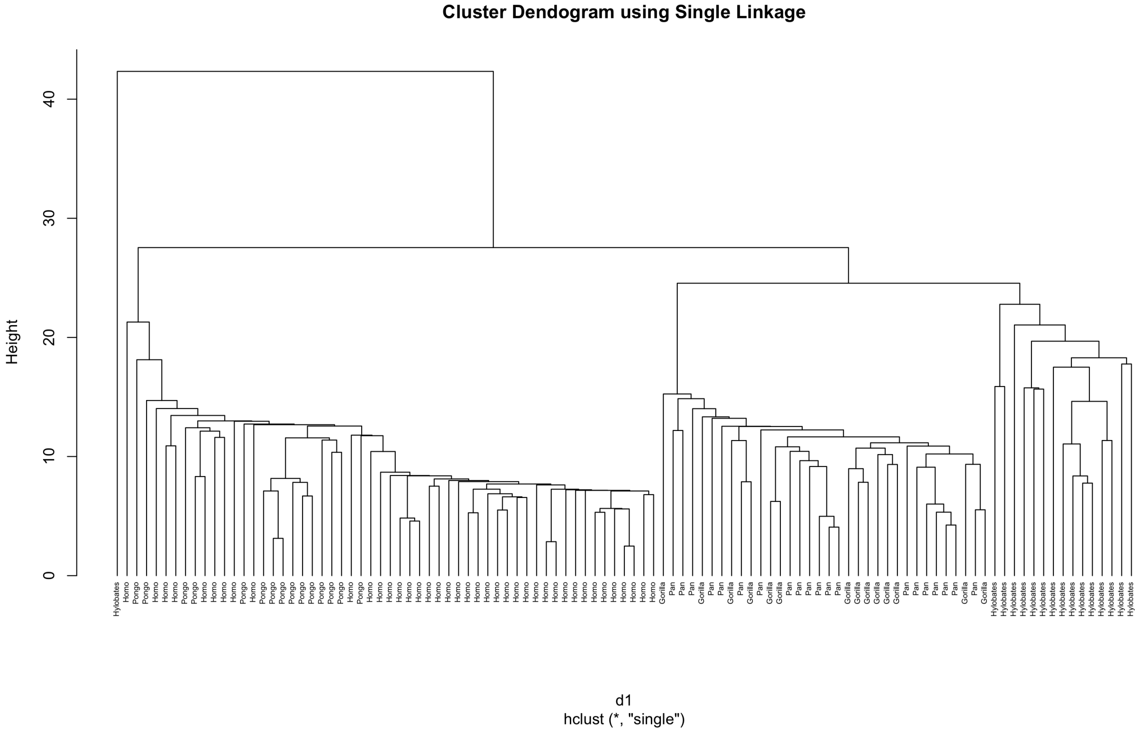


Figure 9: Dendogram for Single Linkage

The groupings, i.e. the value of k is decided by checking and taking the maximum average silhouette width and the corresponding k value for each method. We find that k comes out to be 4 for single, average as well as complete linkage with average silhouette width as 0.572, 0.589 and 0.580. On looking at the rand and the adjusted rand index values, we find that for single they are 0.83 and 0.74, for average: 0.81 and 0.66 and for complete: 0.81 and 0.63 respectively. Thus, we see that Single linkage performed slightly better than the other two methods in this case. However, if we look at the misclassification rates we find that they turn out to be erroneous since the number of clusters are different from the original class labels and there is no guarantee that the labels assigned to clusters by the algorithm correspond to the original class labels i.e. we might get a shuffle in the cluster labels after applying this algorithm which would be mislabled data as compared to original class labels.

K-medoids were applied and according to the adjusted rand index value, k=5 is the optimal choice for which we get the maximum value of adjusted rand index: 0.767. (which tells us the degree of similarity between the result of the clustering algorithm and the original class label taking into account that random chance will cause some objects to occupy the same clusters).

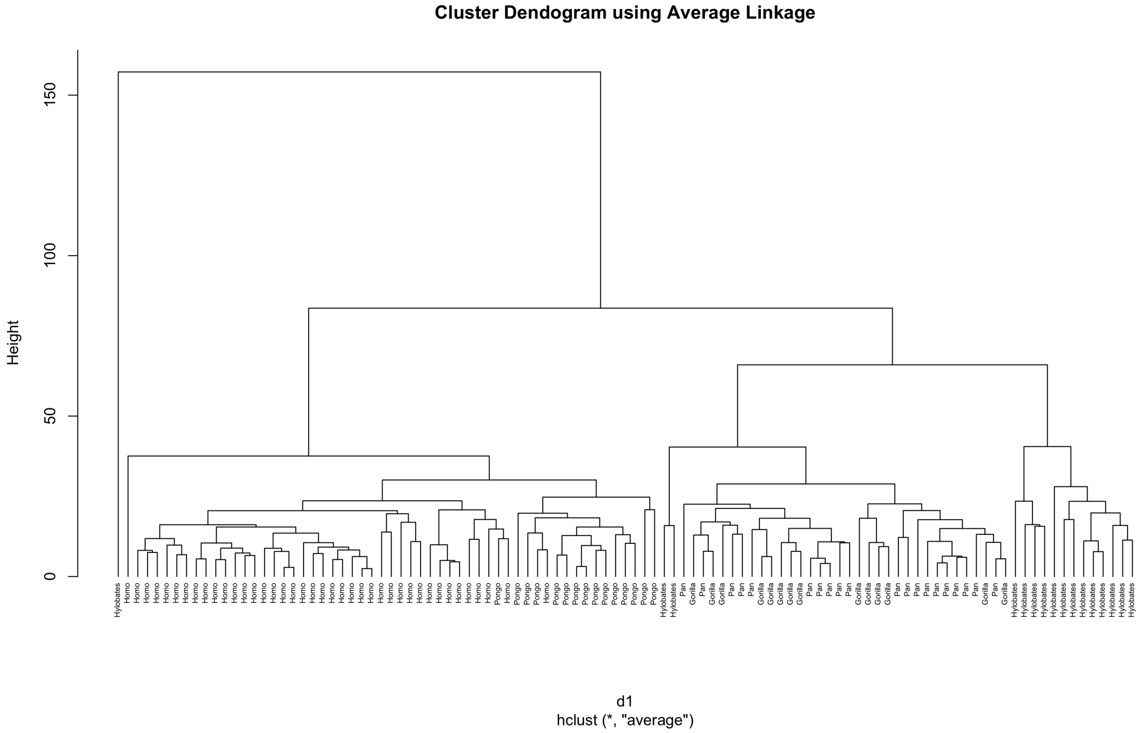


Figure 10: Dendogram for Average Linkage

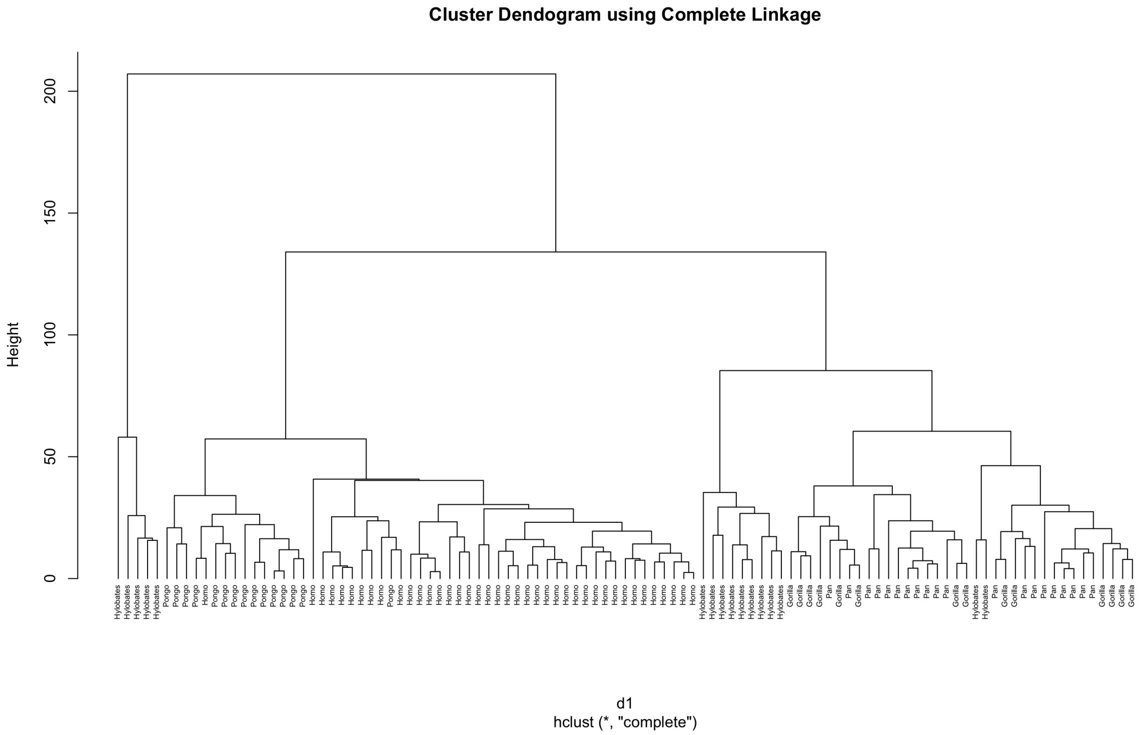


Figure 11: Dendogram for Complete Linkage

According the gap statistic plot in figure 12, we also conclude that k = 5 is optimal since we get maximum gap value considering minimum k, within one standard error of the neighbouring value at k = 5.

We can see that the K-medoids performed slightly better than the heirarchical clustering method if we look at the rand index and the adjusted rand index values. K-medoids did not do a good job in classifying the observations into correct clusters as compared the heirarchical clustering although it was able to predict the correct clusters better unlike heirarchical clustering. Henceforth, K-medoids is a better method for this data.

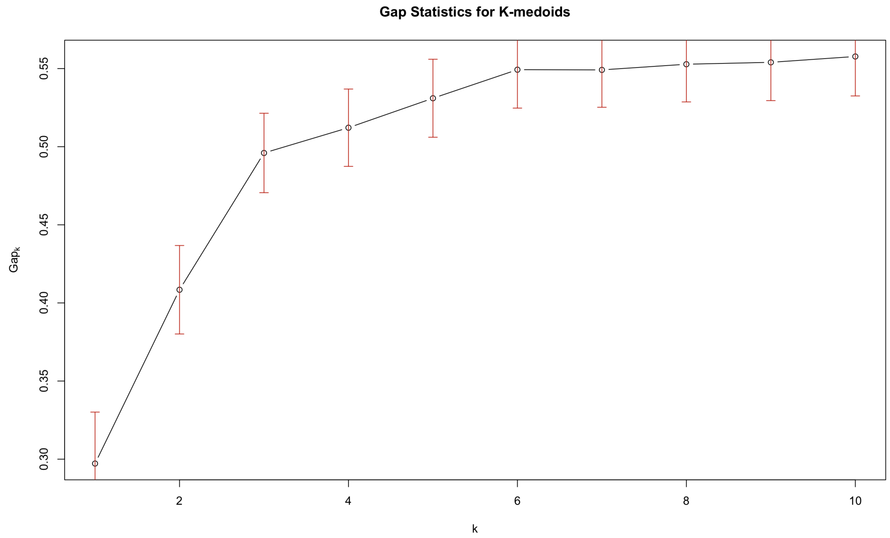


Figure 12: Gap Statistic for K-medoids

Question 4: Run a batch-SOM analysis on the Wisconsin Breast-Cancer data. Describe how well the SOM methods cluster the tumor cases into benign and malignant. Compute the U-matrix and discuss its representation for these data.

The given Wisconsin Breast-Cancer dataset has 569 obs. of 32 variables. The id of the observations were removed and the true label was added. The question requires us to apply the batch Self Organizing Map (SOM) and see how well it clusters the cancer data into two clusters i.e. Malignant and Benign. There are 357 records for beningn and 212 for malignant tumours. A SOM grid of 3x3 was created and a rectangular topology was used to run batch analysis for 2000 times. The observations are plotted in figure 13. On the top right in this figure we see the iteration graph that shows the change in distribution. We observe that after 2000 iterations the graph line becomes stagnant and we observe little change in the distribution of SOM. The count plot shows the count of samples in each node. The neighbour distance graph (figure 14) tells us about the distance of each point from other points i.e. neighbours

We observe that the analysis divided the data into 3 x 3 grids accurately since the mapping and counts of the points are well distributed among the grids. The distance of the neighbours is a certain constant from the neighbour distance plot. The analysis reduced the data by mapping it into 9 categories and the dendogram obtained is shown in figure 15.

When we make two clusters from the dendogram and compare the results to the mappings, we get the following plots as shown in figure 16. As seen from the above mapping plot, the two distinct clusters that were obtained from hierarchical clustering are clearly separated here. The method of clustering adopted and the anomalies in the data could have caused the misclassifications. The U2 matrix picture shows that the clustering is near perfect and the ratio of benign to malignant is almost same as true values.

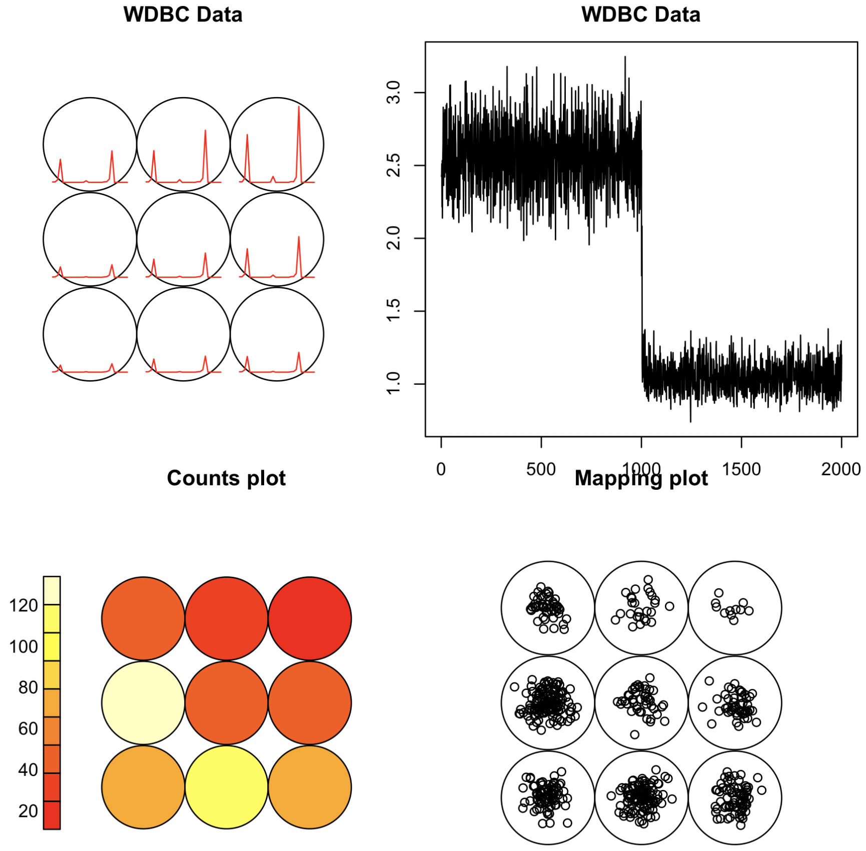


Figure 13: Various Plots

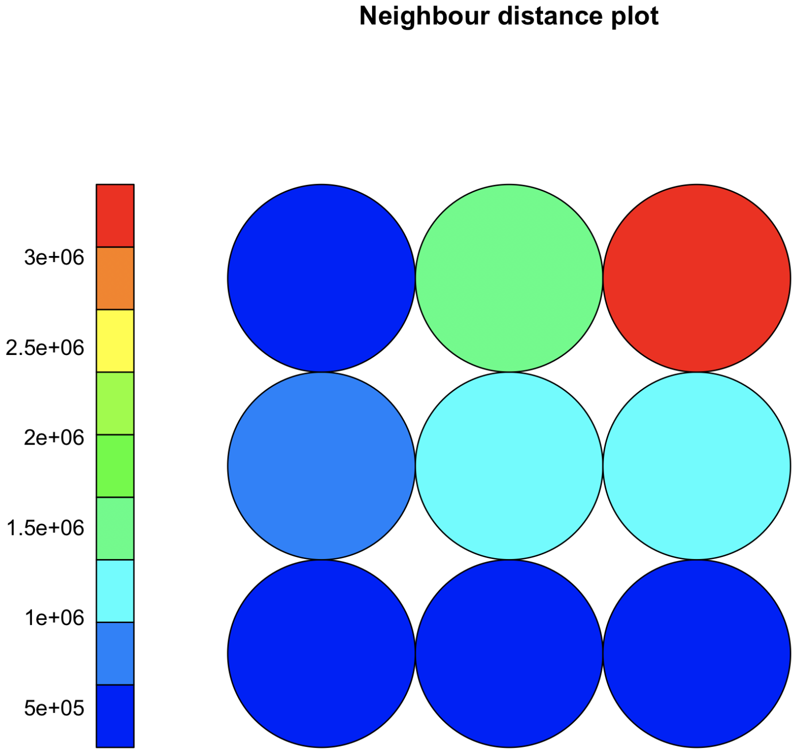


Figure 14: Neighbour Distance plot

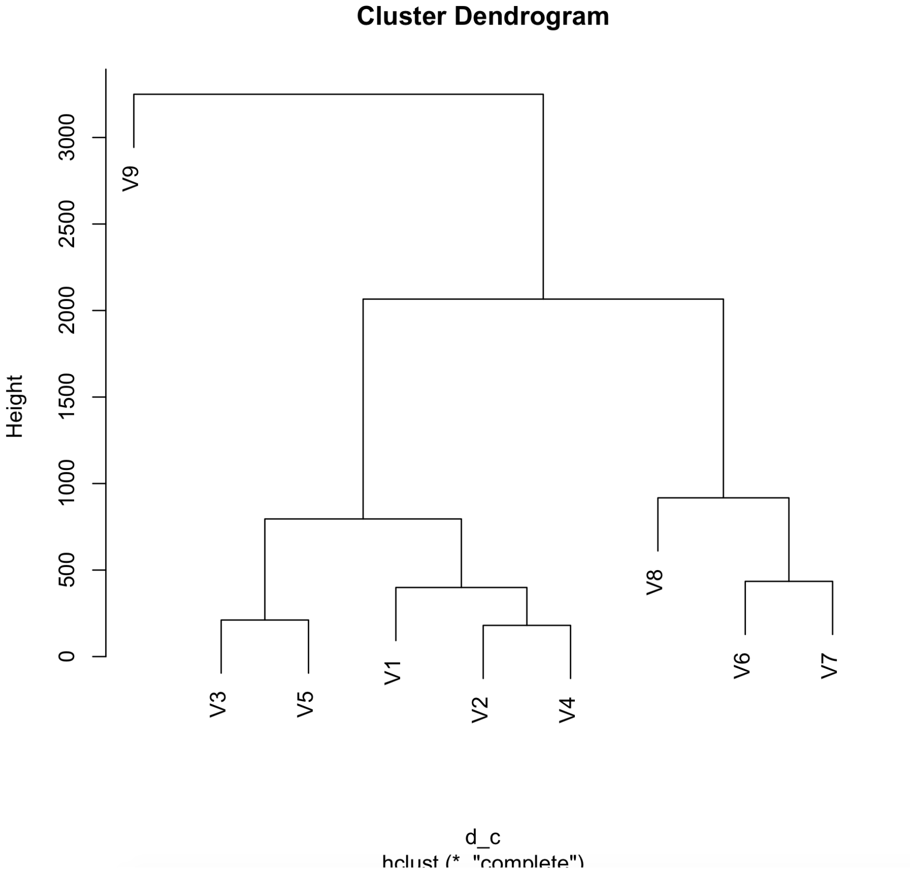


Figure 15: Cluster Dendogram

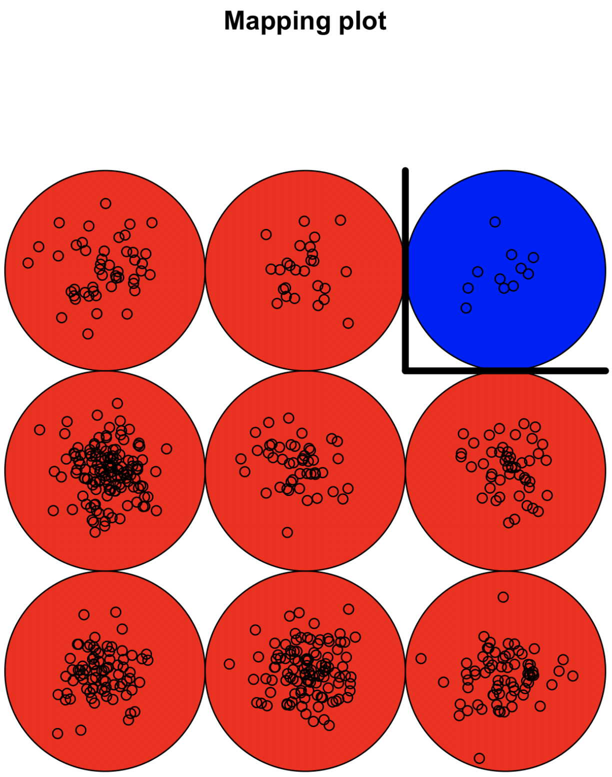


Figure 16: Mapping Plot