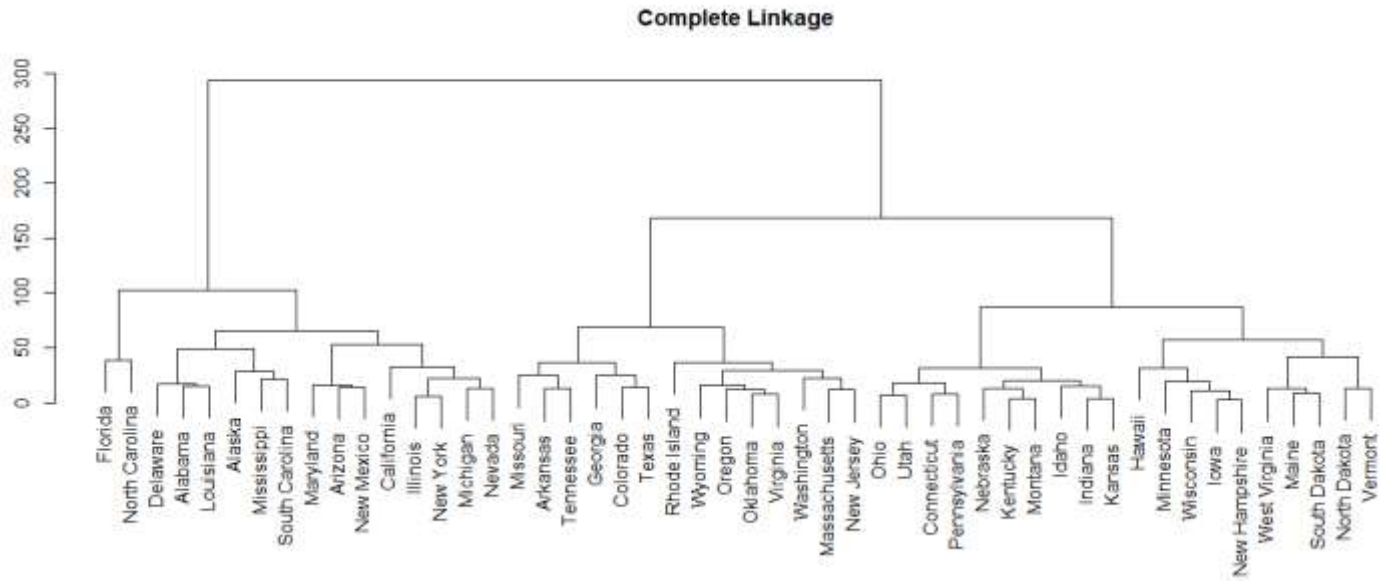
Statistical Data Mining II- Project 2

Problem 1

In this problem we had to the "USArrests" data and perform hierarchical clustering on the states.

a) In this part we performed hierarchical clustering of the observations using complete linkage. Euclidean distance was used as the dissimilarity measure. We clustered the states as shown below in the cluster dendrogram.



b) In this part we had to cut the dendrogram at a height that results in k three distinct clusters. The data below shows the result with the city names with their corresponding clusters:

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

The data below shows the grouping the states with their clusters:

Cluster 1:

[1] Alabama	Alaska	Arizona	California	Delaware
[6] Florida	Illinois	Louisiana	Maryland	Michigan
[11] Mississippi	Nevada	New Mexico	New York	North Carolina
[16] South Carolina				

Cluster 2:

```
 [1] Arkansas      Colorado      Georgia       Massachusetts Missouri
 [6] New Jersey    Oklahoma      Oregon        Rhode Island  Tennessee
[11] Texas         Virginia      Washington    Wyoming
```

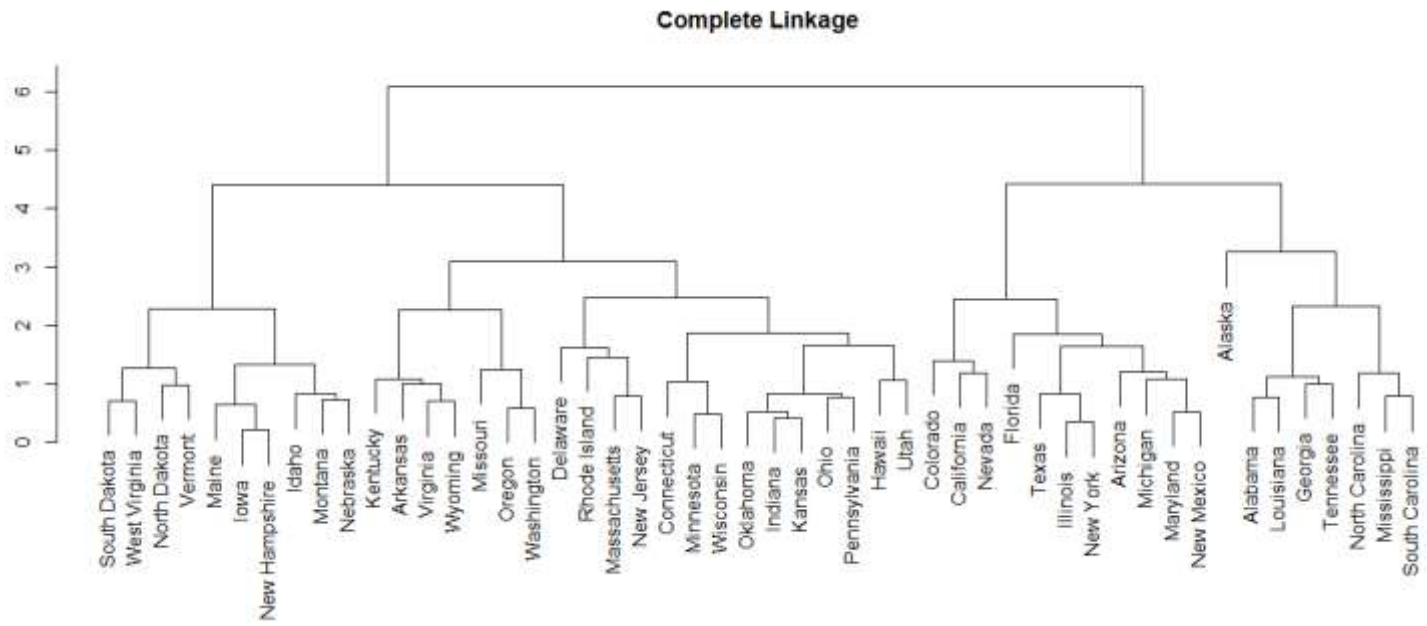
Cluster 3:

```
 [1] Connecticut   Hawaii        Idaho         Indiana       Iowa
 [6] Kansas        Kentucky      Maine         Minnesota     Montana
[11] Nebraska      New Hampshire North Dakota  Ohio          Pennsylvania
[16] South Dakota  Utah          Vermont       West Virginia Wisconsin
```

The table below shows the number of states corresponding to each cluster, which was done for the variable grouping/clustering i.e. the states:

clusters	1	2	3
1	16	0	0
2	0	14	0
3	0	0	20

c) In this part we had to hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one, as shown below in the cluster dendrogram.



We cut the dendrogram at a height that results in k three distinct clusters and below are the city names with their corresponding clusters.

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

Grouping the states with their clusters, we get:

Cluster 1:

```
[1] Alabama        Alaska         Arizona        California     Delaware
[6] Florida        Illinois       Louisiana      Maryland       Michigan
[11] Mississippi   Nevada         New Mexico     New York       North Carolina
[16] South Carolina
```

Cluster 2:

```
 [1] Arkansas      Connecticut   Delaware      Hawaii        Idaho
 [6] Indiana       Iowa          Kansas        Kentucky      Maine
[11] Massachusetts Minnesota     Missouri      Montana       Nebraska
[16] New Hampshire New Jersey    North Dakota  Ohio          Oklahoma
[21] Oregon        Pennsylvania  Rhode Island  South Dakota  Utah
[26] Vermont       Virginia      Washington    West Virginia Wisconsin
[31] Wyoming
```

Cluster 3:

```
 [1] Connecticut   Hawaii        Idaho         Indiana       Iowa
 [6] Kansas        Kentucky      Maine         Minnesota     Montana
[11] Nebraska      New Hampshire North Dakota  Ohio          Pennsylvania
[16] South Dakota  Utah          Vermont       West Virginia Wisconsin
```

The table below shows the number of states corresponding to each cluster:

```
clusters  1  2  3
       1  6  9  1
       2  2  2 10
       3  0  0 20
```

d) The effect of scaling the variables changes the hierarchical clustering obtained for the variables. The height of fusion, as measured on the vertical axis changes after the scaling as shown on the complete hierarchical clustering dendogram. Thus, the observations that fuse at the very bottom of the tree which are quite similar to each other, and the observations that fuse close to the top of the tree which are quite different, change after scaling.

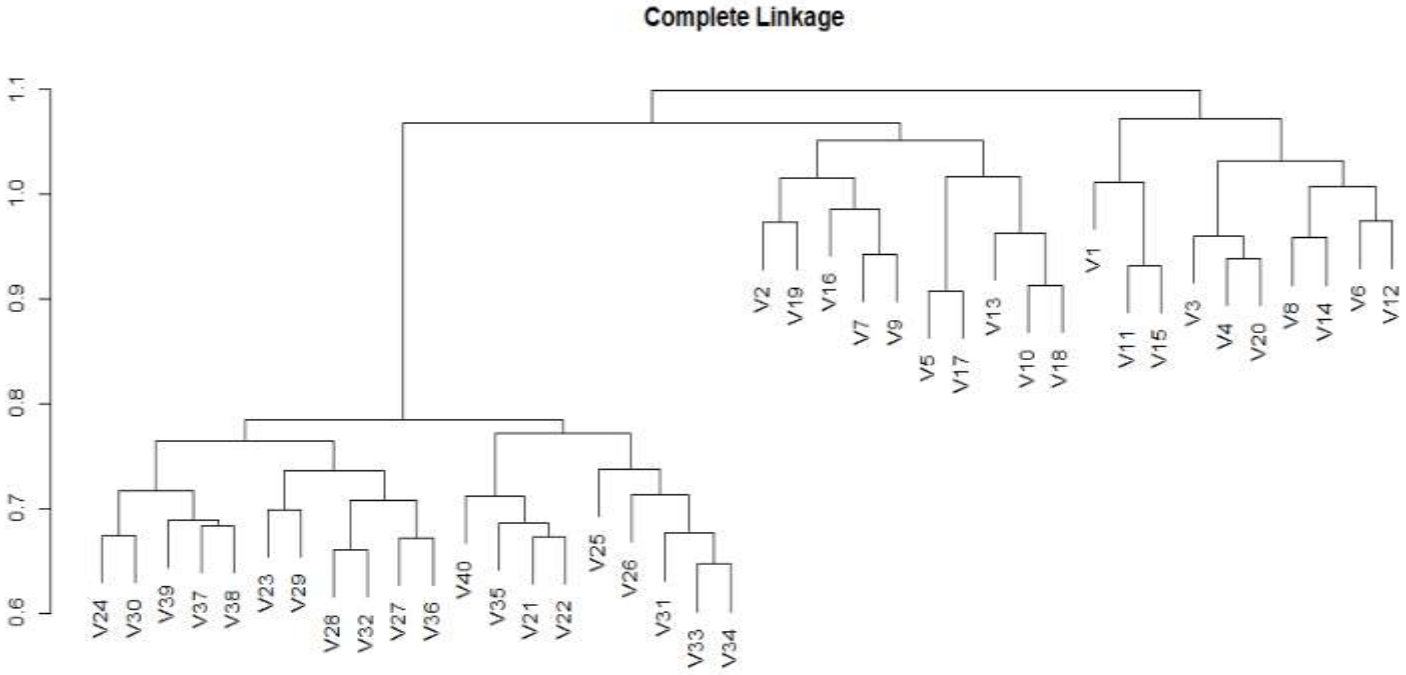
Also, in the original cluster the US states fall in the 3 distinct clusters, indicating the cluster purity, whereas with the scaling the states become spread out/mixed over three different clusters. The variables be scaled before the inter-observation dissimilarities are

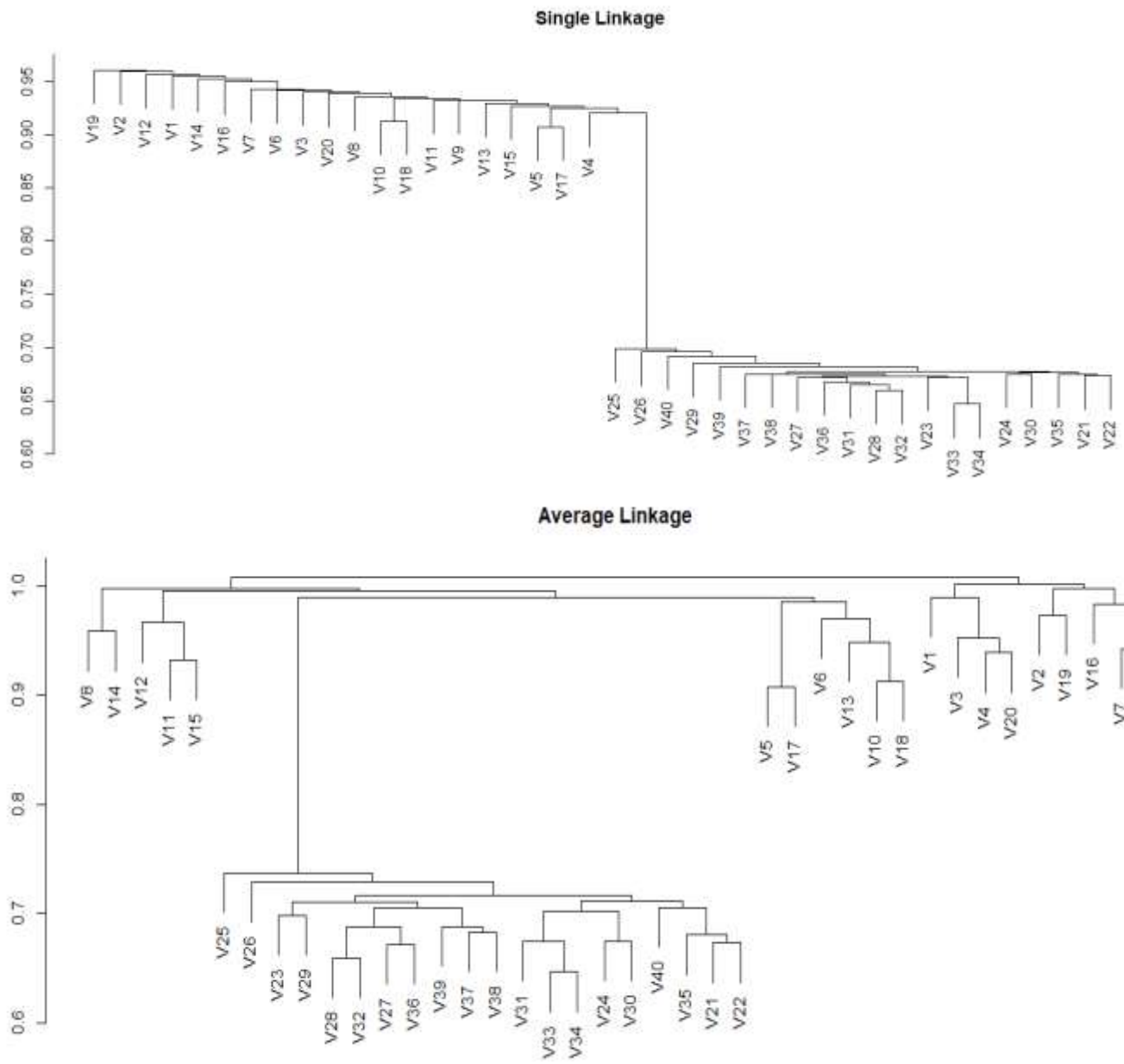
computed when the data variables have large means and large variance or outliers. This is because then the clusters that we observe would be driven by those variables. This will lead to poor fitting clusters. This step of standardizing the variables to have standard deviation one before performing clustering, should be performed if we want each observation to be on the same scale.

Problem 2

In this part of the problem we had to load the data using read.csv() and we selected header=F. The data 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) i) In this part we had to apply hierarchical clustering to the samples using correlation based distance, as a dissimilarity distance and plot the dendrogram, as shown below. Correlation-based distance considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance. Thus, the dissimilarity measure is very important, as it has a strong effect on the resulting dendrogram.





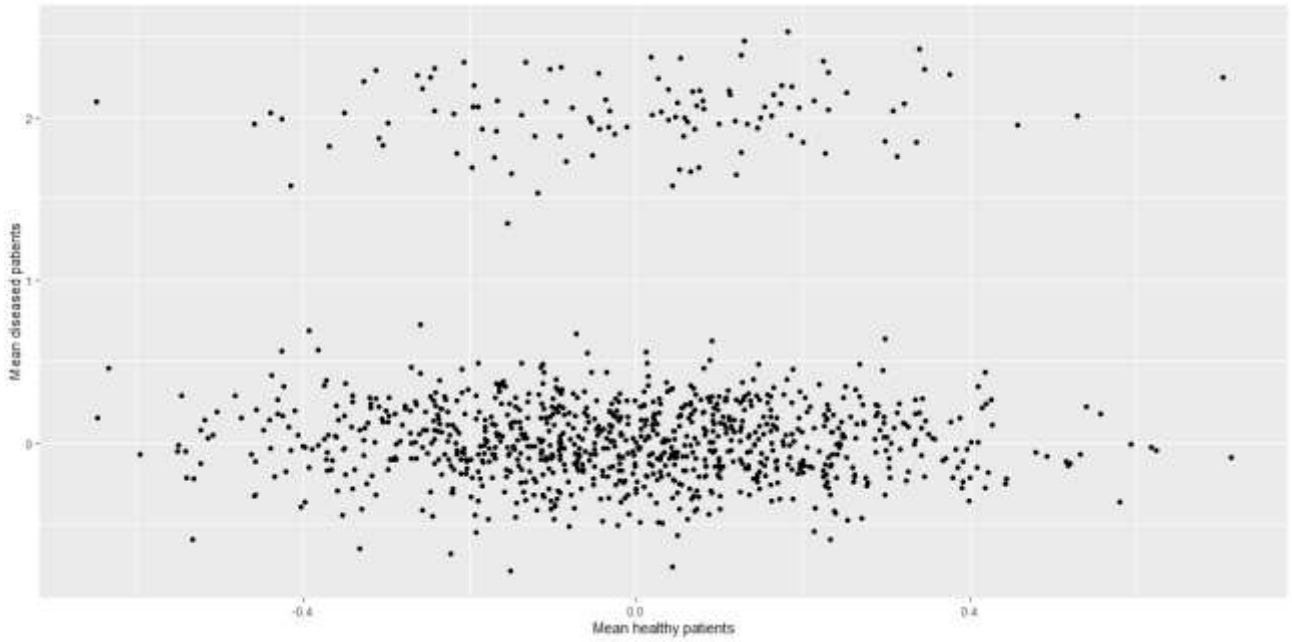
b) ii) The genes separate the samples into the two groups depending on the cutting of the dendrogram height and the type of linkage used.

From the complete linkage dendrogram, if we cut at a height of 1.08 we obtain 2 clusters, but not of the two groups (healthy and diseased. For the average linkage dendrogram if we cut at height of 1.02 we obtain 2 distinct clusters, but not into the two groups which are desired. The single linkage yields trailing clusters as shown in the dendrogram above leads to chaining and groups we do not see groups merging/ clusters, as there are very large clusters onto which individual observations attach one-by-one, thus it is difficult to determine the clustering. On the

other hand, complete and average linkage tend to yield more balanced, attractive and evenly sized clusters.

Thus, resulting number of clusters does depend on the type of linkage used. Since, different dendrogram are obtained for the type of linkage used. Further cuts can be made as one descends the dendrogram in order to obtain any number of clusters.

b) In this part our collaborator wants to know which genes differ the most across the two groups. To solve this problem we determined the means of the healthy and the diseased patients and plotted as shown below.



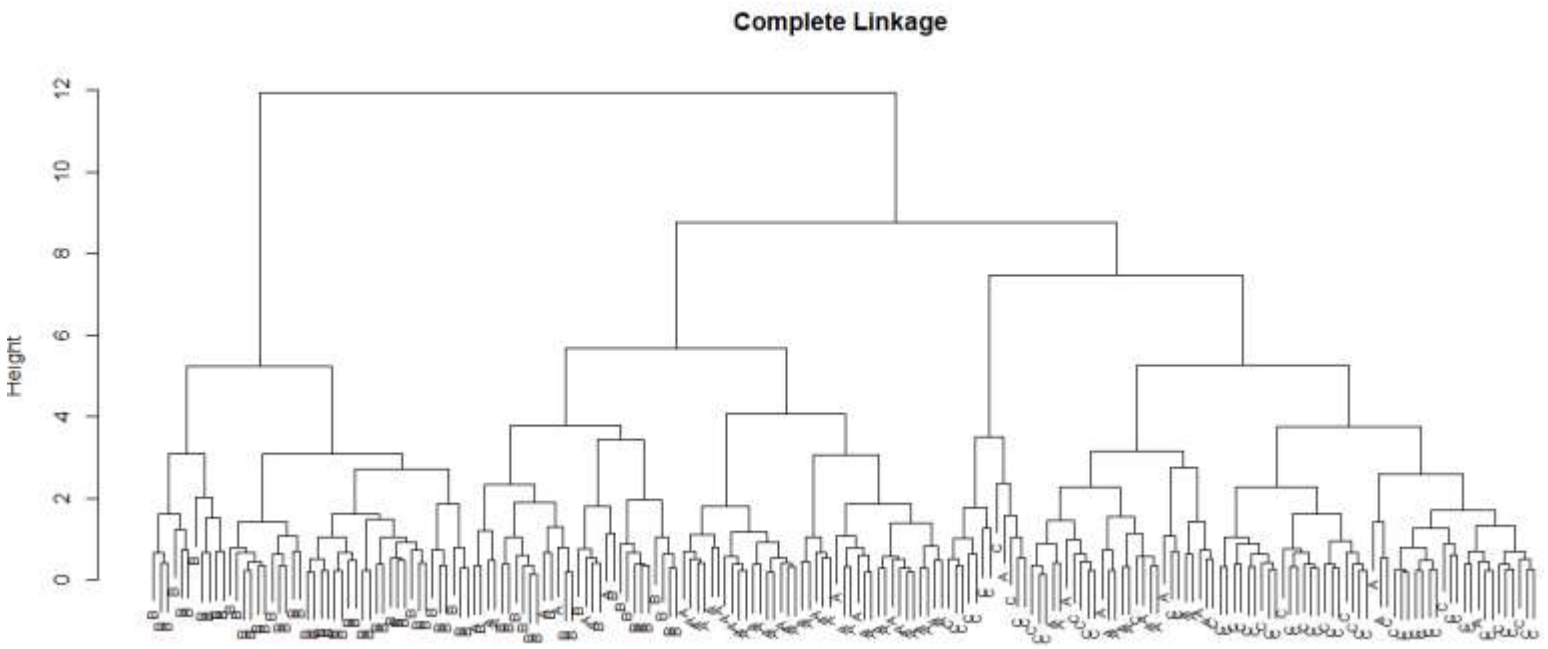
To determine which genes differ the most across the two groups, I took the absolute difference between the mean healthy and the mean diseased patient. Below are the ten genes with most difference between them, with the top being the most difference to the bottom gene being the least different.

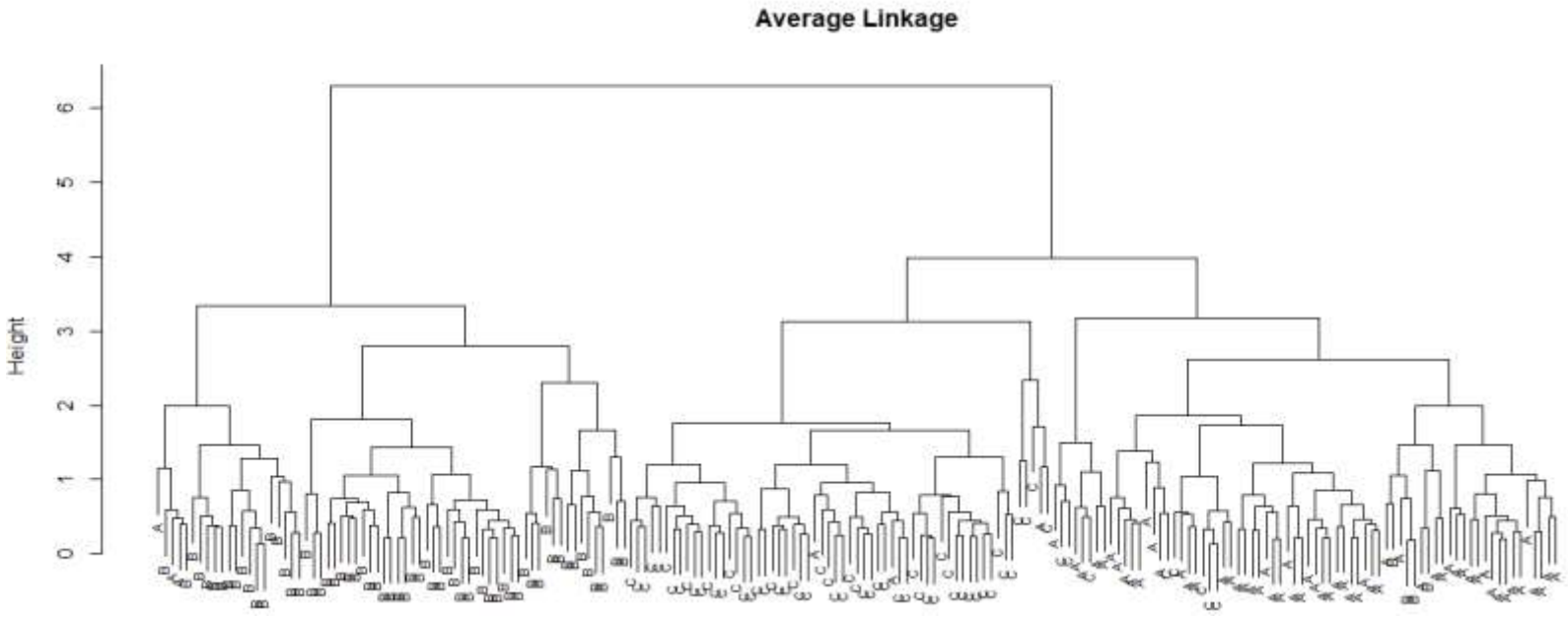
Gene	Difference
600	2.747577
584	2.601985
549	2.550757
540	2.545174
502	2.544461
568	2.519418
582	2.496084
565	2.470820
562	2.465549
554	2.436718

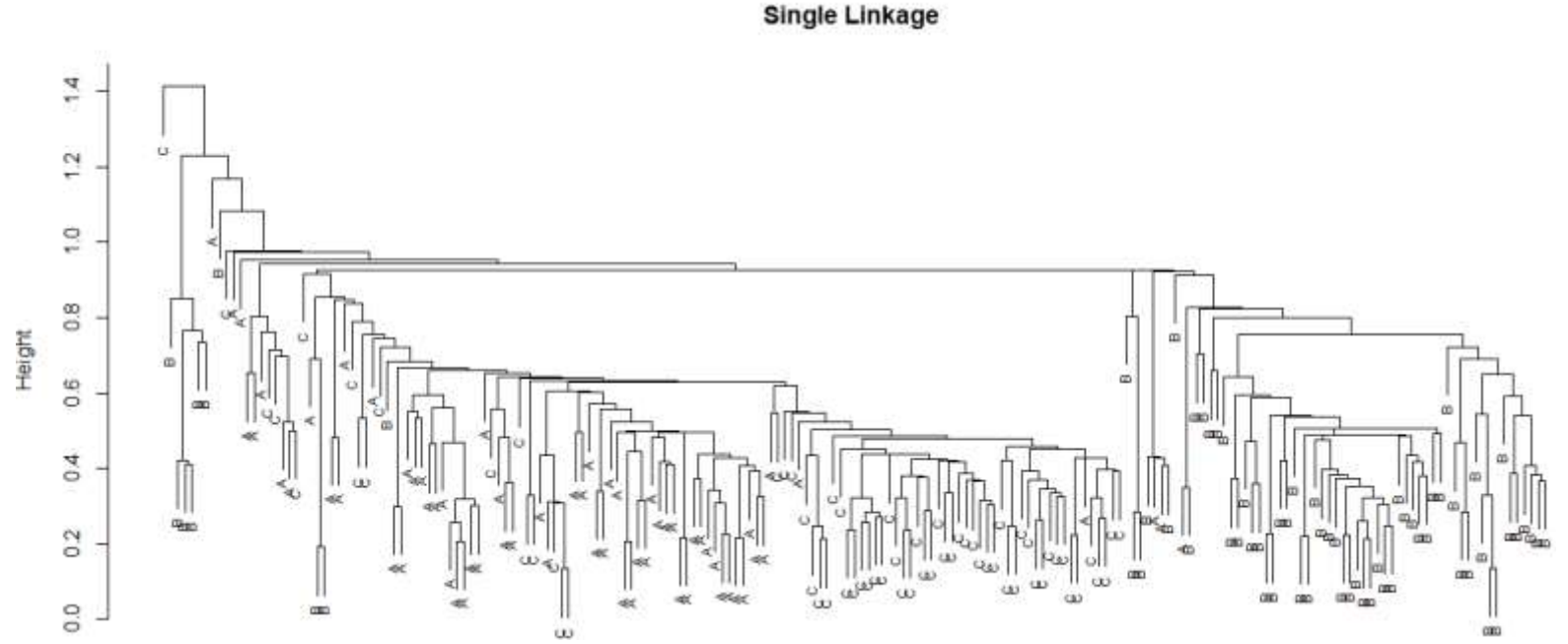
Problem 3

In this problem we had to use the “seeds data”. The data contains the geometrical properties of kernels belonging to three different varieties of wheat (seed group).

a) We had to cluster the data based on single-linkage, average linkage, and complete-linkage agglomerative hierarchical clustering. The results are shown below in the hierarchical clustering dendograms. We had to remove the “seed group” column, which is the true label (A,B,C) to perform the clustering.







To decide on the groupings for all three methods, adjusted rand index was used. The justification for using this technique is that the adjusted Rand index is the corrected-for-chance version of the Rand index. It's calculated using the formula below:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

The reason it is used is because the Rand index, which is measure of the similarity between two clusters, tends to be optimistic for small dataset and can be misleading. It also measure performance based on values between 0 and +1, but the adjusted Rand index gives negative values if the index is less than the expected index. Both strategies use true labels of the data and are used if we want to compare 2 data clustering.

The Adjusted Rand index technique uses the output of the 3 hierarchical clustering algorithms to select the number of clusters in seed data set.

Adjusted Rand index for complete linkage hierarchical clustering:

k=2	k=3	k=4	k=5	k=6
0.6530775	0.5200210	0.4623355	0.5020200	0.5007037

Adjusted Rand index for average linkage hierarchical clustering:

k=2	k=3	k=4	k=5	k=6
0.8703981	0.7482942	0.6562729	0.6407289	0.6177764

Adjusted Rand index for single linkage hierarchical clustering:

k=2	k=3	k=4	k=5	k=6
0.519117824	0.001509422	0.001208842	0.001487219	0.001396953

All three methods select k=2 to be the number of clusters. This was the number of grouping for each of the three methods, given k=2 gave the highest adjusted rand index for the k clusters in each linkage method. From the three method of linkages, we see that the average linkage hierarchical clustering performed the best as it has the highest value of adjusted rand index 0.870 while the single linkage method performed the worst, as it has the lowest value of adjusted rand index, 0.519. However, the choice of linkage certainly does affect the results obtained, as all three methods estimate the number of clusters to be 2.

The single linkage yields trailing clusters as shown in the dendogram above, i.e. very large clusters onto which individual observations attach one-by-one. This leads to chaining and groups do not see groups merging. On the other hand, complete and average linkage tend to yield more balanced, attractive clusters. They give symmetric, broad shoulders, good clean groups in the dendogram. For this reason, complete and average linkage are generally preferred to single linkage. This goes with our expectation as complete and average linkage tend to yield evenly sized clusters whereas single linkage tends to yield extended clusters to which single leaves are fused one by one.

b) In this part we had to cluster the data based on K-means or K-medoids. I decided to cluster based on k-mediods.

This is the summary of the k-medoids for the seed data set:

```
$pamobject
Medoids:
      ID  Area Perimeter Compactness Length.Kernel Width.Kernel Asymmetry Length.Kernel.Grove
[1,]  19 12.72     13.57      0.8686         5.226        3.049     4.102              4.914
[2,] 117 18.14     16.12      0.8772         6.059        3.563     3.619              6.011
Clustering vector:
  [1] 1 1 1 1 2 1 1 2 2 1 1 1 1 1 1 2 1 1 1 1 2 1 1 2 1 1 1 1 1 2 1 1 1 1 2 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 2
 [69] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1
[137] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Objective function:
   build     swap
2.307739 1.998831

Available components:
 [1] "medoids"    "id.med"     "clustering" "objective"  "isolation"  "clusinfo"   "silinfo"    "diss"       "call"       "data"

$nc
[1] 2

$crit
 [1] 0.0000000 0.5238727 0.4687289 0.4182288 0.3816176 0.3310709 0.3397493 0.3409435 0.3559635 0.3355913
```

From the summary we can see that optimal number of clusters the system provide is k= 2.
The reuslts for the cluster of the obersavations using kmediods can be seen below:

```
    A  B  C
1  53  0 65
2  13 68  0
```

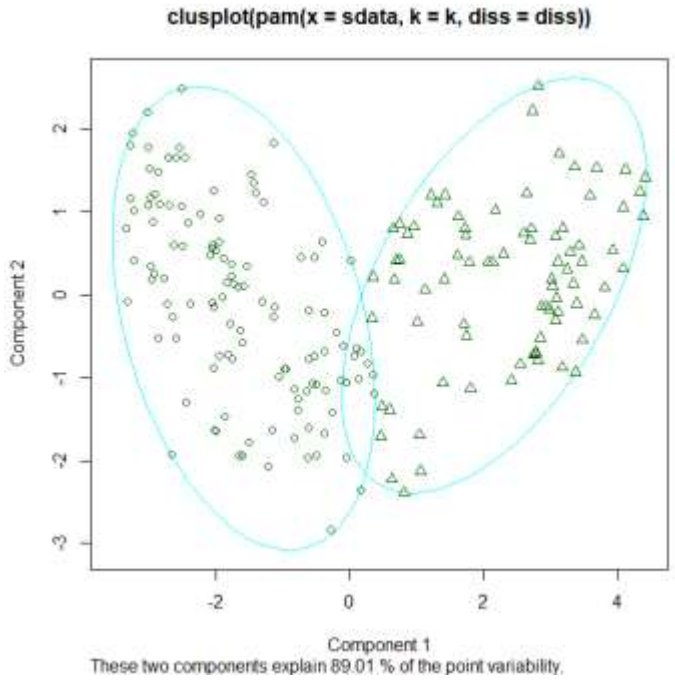
From the table above we can see that A observations belong to both cluster 1 and cluster 2, however B and C belong to one pure cluster, with B belonging to cluster 2 and C belonging o cluster 1.

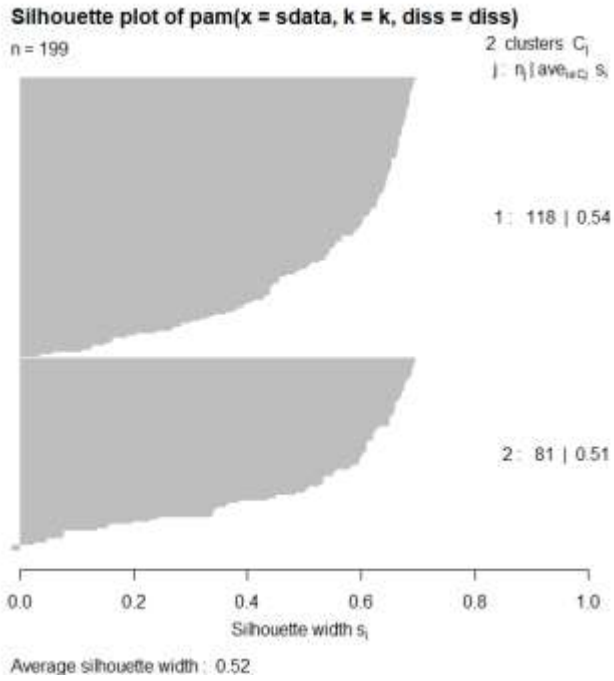
Given that the number of cluster was k=2, we plotted the cluster plot and the silhouette analytical technique to validate the value of k, as shown below.

The silhouette analytical method is where each cluster is represented by a silhouette, which is based on the comparison of its tightness and separation. This silhouette shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters. The entire clustering is displayed by combining the silhouette into a single plot, to allow overview of the data configuration.

The s(i) is computed for each point in the dataset using the formula below to tell us what the k should be, where a(i) represents the within cluster dissimilarity and b(i) represents the between cluster dissimilarity.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$



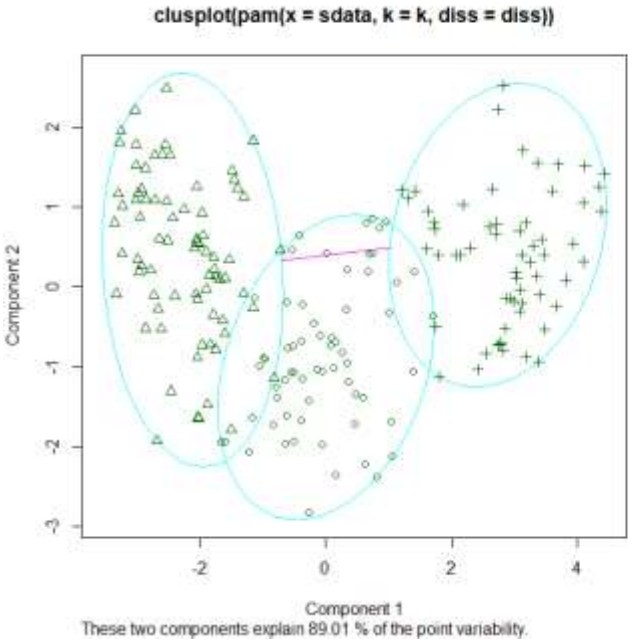


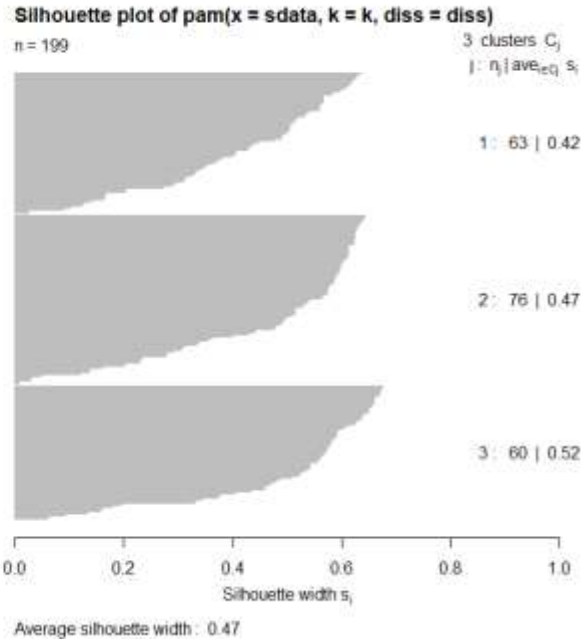
From the cluster plot and from the silhouette plot shape above, we can see that the observations fit in 2 clusters well. There is an outlier which does not fit well in cluster 2, as it has a negative silhouette value s(i). To select the appropriate number of clusters k, we look at the average silhouette plot value which is 0.52. This value indicate that the clusters are a reasonable structure given it falls in the range of 0.51-0.70.

To verify, we tried a cluster of k=3 using the k-medoids clustering algorithm.

	A	B	C
1	54	9	0
2	11	0	65
3	1	59	0

From the table above we can see that A and B observations belong to a mixture of clusters. A belongs to cluster 1, 2,3 and B belongs to cluster 1 and 3, however, observations C belong to one pure cluster which is cluster 2.





From the cluster plot, we see that the observations are clustered well, but they overlap between the three clusters. From the silhouette plot we can see that the observations fit well in the three clusters. However, the average silhouette width for the entire data is 0.47, which is interpreted as a weak and artificial structure. Also, this value is less than the silhouette clustering with k=2, suggesting k=2 is better choice.

The adjusted rand index was determined for the k-medoids clustering algorithm for different numbers of clusters, as shown below.

k=2	k=3	k=4	k=5	k=6
0.7742362	0.7195739	0.6684703	0.5797120	0.4889719

Based on the adjusted rand index above, we can see that the cluster of 2 has the highest value, indicating a best choice for the cluster number for the seed data set.

The performance of the k-medoids clustering compared to the hierarchical clustering of part a was not better than the average hierarchical clustering, as for k=2 it had and adjusted rand index of 0.87, however it was better than the complete and single hierarchical clustering as they had an adjusted rand index of 0.65 and 0.52, respectively.

Hierarchical clustering is a better method for the clustering of this data. K- medoids clustering requires us to pre-specify the number of clusters K. Hierarchical clustering does not require that we commit to a particular choice of K. Hierarchical clustering has an added advantage in that it results in an attractive tree-based representation of the observations, called a dendrogram, which allows us to view at once the clusterings obtained for each possible number of clusters, from 1 to n. Also, ee can easily draw conclusions about the similarity of two observations based on the location on the vertical axis and we can look at the dendrogram and select by eye a

sensible number of clusters, based on the heights of the fusion and the number of clusters desired. Furthermore, hierarchical clustering does not use class labels of the dataset, whereas k-medoids does.