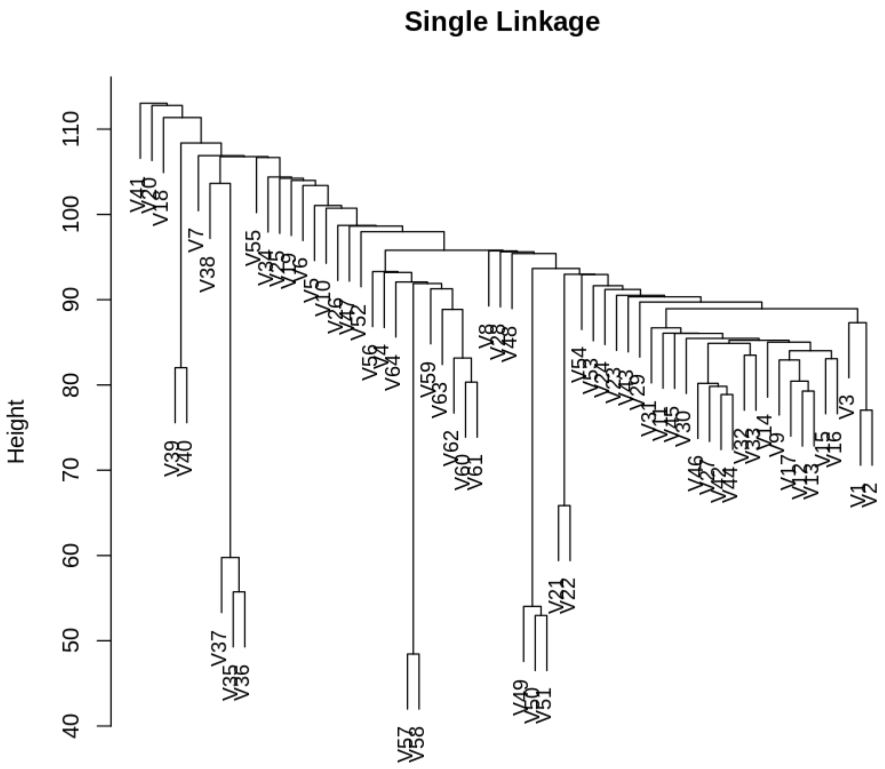
Report for Assignment 3

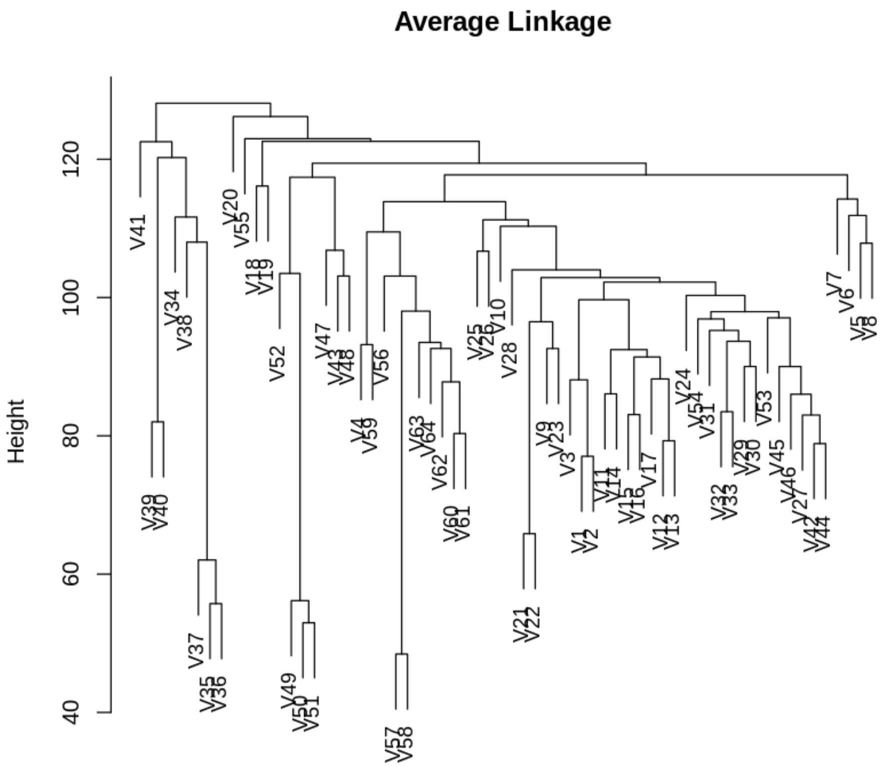
Task 1: I have implemented all the four linkage clustering algorithms as instructed in the task. To implement Single, Average, Complete Linkages, which are based on pair-wise distances, I have pre-calculated all the distances in a matrix. I have then exploited the sparsity in form of Infinity values or NaN values to find the nearest cluster or sample. This saved the trouble of calculating distances for inter and intra-cluster separately. After, finding the desired distance, I have put that entry to NaN (infinity for Single Linkage). I have used the similar trick to change unwanted distances to NaN (infinity for Single Linkage). I have created a matrix of nrow by nrow, which stores cluster assignment for every iteration in its column. The first column of the matrix, cluster_matrix, has initilisation state where each observation is assigned as a cluster. There is a vector variable of name branch_length and it stores the height/ dissimilarity at each iteration.

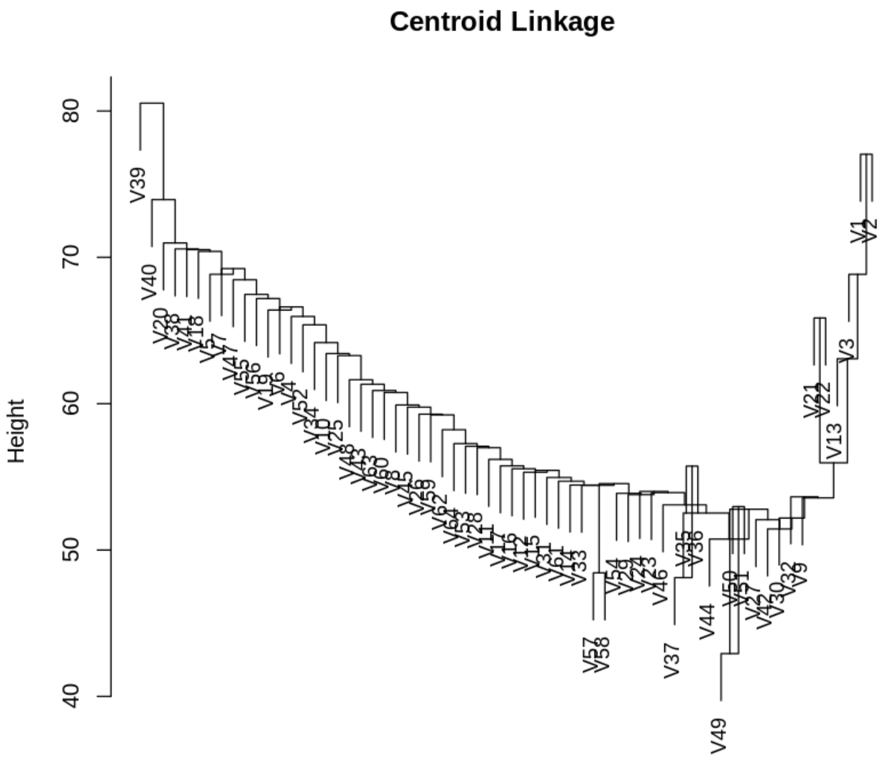
Choosing a column of the matrix, cluster_matrix n+1-k gives the cluster assignment equivalent to cutting a dendrogram at k. where n is the number of columns

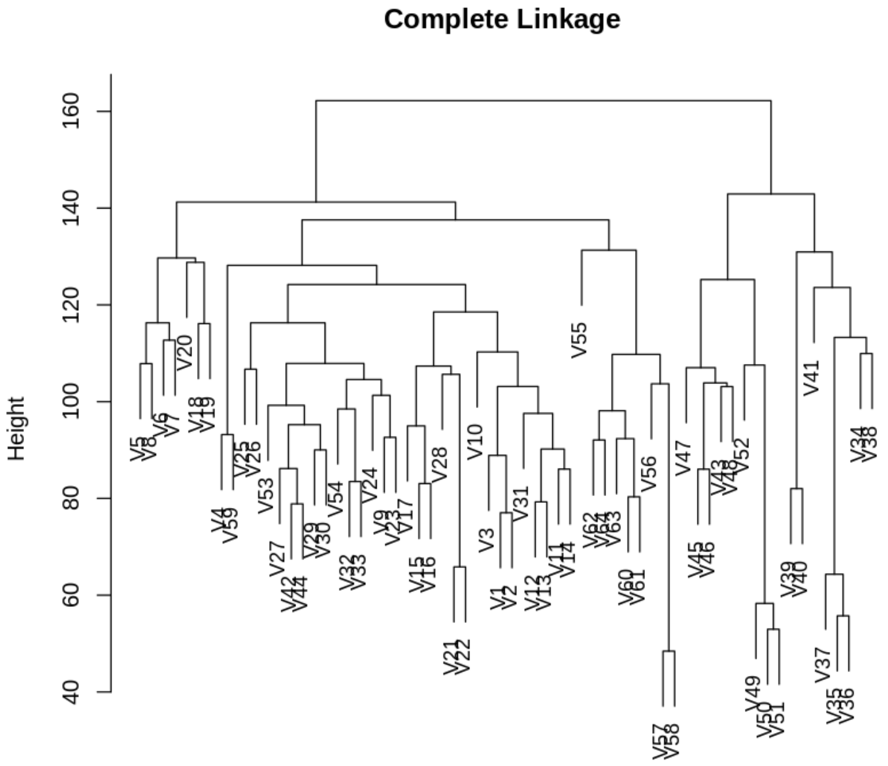
Task 2: All linkage algorithms are applied on the given NCI microarray dataset. The data is preprocessed by using scale function in R.

Task 3: Different linkage functions provide different performance on a dataset. Each linkage function can provide a different insight to the data. Typically, single linkage tends to form trailing clusters i.e., very large clusters onto which individual observation attach one-by-one. Centroid linkage, on the other hand, can lead to undesirable inversions. Complete linkage and Average linkage tend to form more balanced, attractive clusters. Therefore, Complete linkage and Average linkage are amongst the most used linkage functions.


Single Linkage
Height
110
100
90
80
70
60
50
40
V41
V20
V18
V39
V40
V7
V38
V37
V35
V36
V55
V34
V25
V19
V6
V5
V10
V26
V47
V52
V56
V4
V64
V57
V58
V59
V63
V62
V60
V61
V8
V28
V48
V49
V50
V51
V21
V22
V54
V53
V24
V23
V43
V29
V31
V11
V45
V30
V46
V27
V42
V44
V32
V33
V14
V9
V17
V12
V13
V15
V16
V3
V1
V2


Average Linkage
Height
40
60
80
100
120
V41
V39
V40
V34
V38
V37
V35
V36
V20
V55
V18
V19
V52
V49
V50
V51
V47
V43
V48
V4
V59
V56
V57
V58
V63
V64
V62
V60
V61
V25
V26
V10
V28
V21
V22
V9
V23
V3
V1
V2
V11
V14
V15
V16
V17
V12
V13
V24
V54
V31
V32
V33
V29
V30
V53
V45
V46
V27
V42
V44
V7
V6
V5
V8


Centroid Linkage
Height
80
70
60
50
40
V39
V40
V20
V38
V41
V18
V5
V7
V47
V55
V56
V19
V6
V4
V52
V34
V10
V25
V43
V63
V60
V45
V26
V59
V62
V64
V53
V28
V11
V17
V16
V12
V15
V31
V61
V14
V33
V57
V58
V54
V29
V24
V23
V46
V37
V35
V36
V44
V49
V50
V51
V27
V42
V30
V32
V9
V21
V22
V13
V3
V1
V2



As I wrote earlier, the single linkage function has merged the some very small clusters which are very different than the others. Centroid linkage function has a very weird dendrogram structure due to the inversion. Both linkage functions seem to perform poorly on NCI dataset. Average linkage is more balanced than the other two, but it is also merging few outlier pairs to the bigger clusters. The dendrogram of Complete Linkage is much more balanced and attractive than the others for the NCI dataset, and looks to perform the best on our dataset. However, we can easily see that none of the clustering methods are perfect for our dataset, but they still provide very useful information.

Task 4: Applied K-Means to NCI dataset with various values of K. With small values of K, the performance of the dataset seems to be poor, which should be expected since we have 15 different labels. As we increase the value of K the performance begins to increase. However, since our dataset/labels are not well separable the performance begins to decrease particularly if we choose K >7. For K = 15, many of the observation under the same labelled and classified under the same cluster for smaller value of K, moves to another clusters. K562A-repro, K562B-repro are similar labels and therefore are classified under same cluster, same follows with

MCF7A-repro and MCF7D-repro. There are observations of Breast cancer which are distributed under different clusters and with higher values of K, their distribution gets wider.

Task 5: Since the Complete linkage function looked the most balanced and attractive, I am comparing the agglomerative Hierarchical with Complete linkage with K-means. I plot the table function with K = 4 for both the clustering types.

```
              hc_cluster

 km_cluster         1 2 3 4

                 1  11  0  0  9      2   9  0  0 0

                 3   0  0  8  0

                 4  20  7  0  0
```

This table shows how different the distribution can be with different clustering algorithms. If we compare both with K = 4, we see that Cluster 2 in K-means clustering is similar to cluster 3 in hierarchical clustering, but other clusters are different., for example cluster 4 in K-means clustering contains a subset of the observations assigned to cluster 1 by hierarchical clustering, as well as all of the observations assigned to cluster 2 by hierarchical clustering. For our dataset, k-means seems to be working better than the hierarchical clustering. For the dataset, the hierarchical clustering seems to be putting most of the observations in one cluster including the one of different labels, but K-means is spreading it to other clusters too.

Task 6: There is no single way of choosing the number of K. There are many approaches, and their usage depends on the problem at hand. One of the simplest methods is to plot a dendrogram and then decide on the optimal number of K. However, using a dendrogram can be very subjective. There are few other approaches to find the optimal number of K below:

1) Elbow method- It's based on the idea to define clusters such that total intra-cluster variation or total within cluster sum of squares (WSS) is minimised. WSS measures the compactness of a cluster and we want it to be as minimum as possible. The Elbow method looks at the total WSS as a function of the number of clusters. We aim to choose a number of clusters so that adding another cluster doesn't improve much better the total WSS. We use a plot to visualise it and choose the number of clusters at the location of an elbow.

2) Average silhouette method- This method measures the quality of a clustering. It identifies how well each observation lies within its cluster. Higher average silhouette width is an indicator of good clustering. This method computes the average silhouette of observations for different values of k. The optimal number of clusters is chosen such that it maximises the average silhouette over a range of possible values for k.

3) Gap Statistic- The gap statistic compares the total within intra-cluster variation for different values of k. The optimal number of k is the value that maximize the gap statistic (i.e, that yields the largest gap statistic). This defines how far the clustering structure is from the random uniform distribution of points.

4) The Sum of Squares method- Another clustering validation method would be to choose the optimal number of clusters by minimizing the within-cluster sum of squares (a measure of how tight each cluster is) and maximizing the between-cluster sum of squares (a measure of how well separated each cluster is from the others).

5) Assigning P-value to the formed clusters- We can assign P-value to the formed clusters and discard the cluster if the P-value of the formed cluster is more than we want.