Assignment 3 – Report

Task 1:

Used two functions hclust & seq_order, it calculates Euclidean distances between the observations and then perform clustering on basis of the linkage selection. There were four linkage selection options: single, complete, average & centroid.

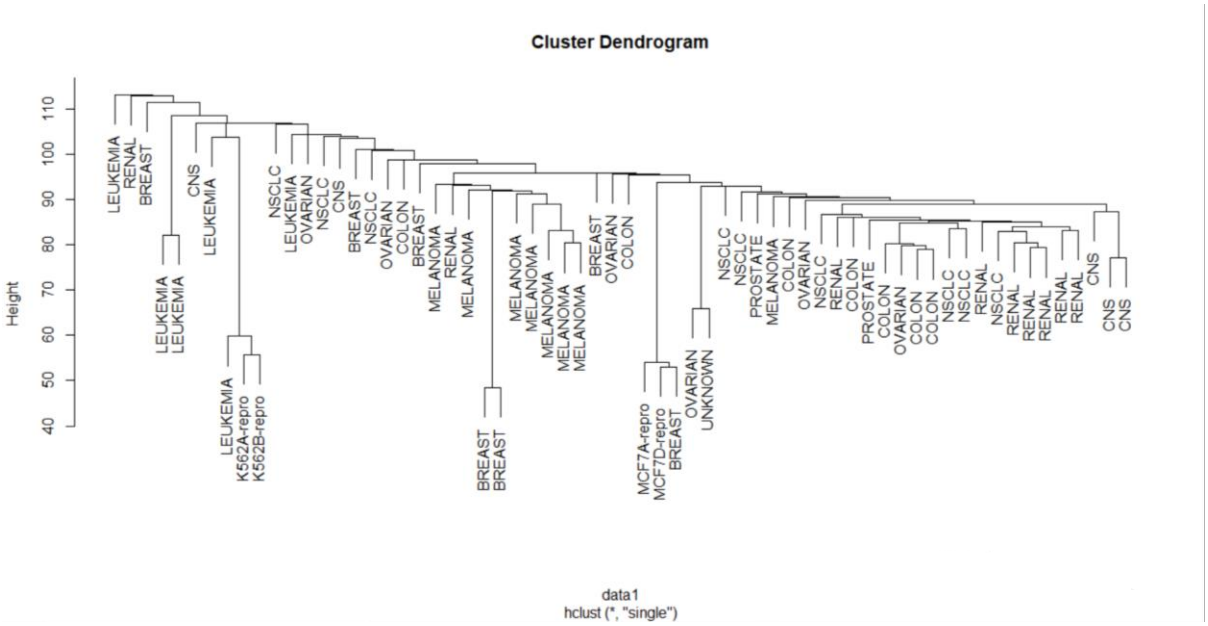
Task 2:

Scaled & Applied NCI microarray dataset to above functions.

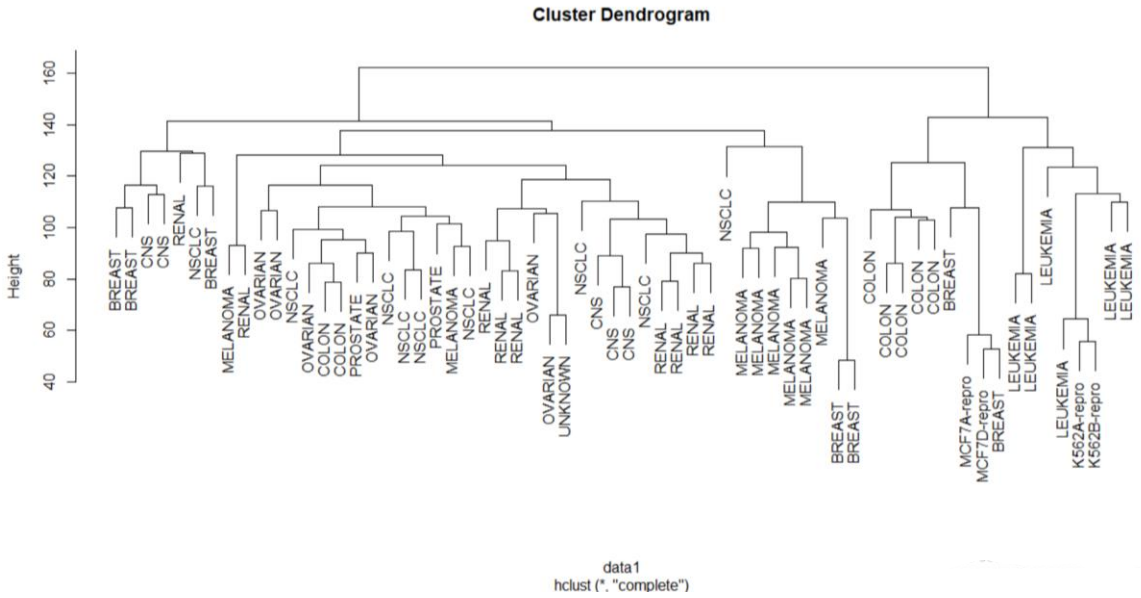
Task 3:

Let's first have a look at the plots of different linkage functions.

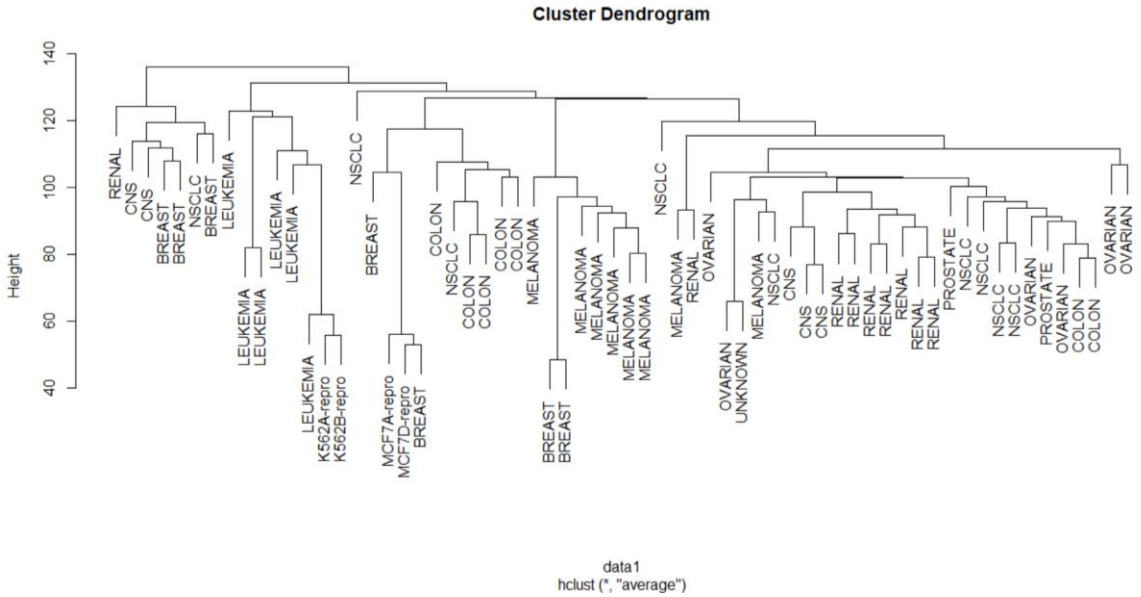
1. Single linkage



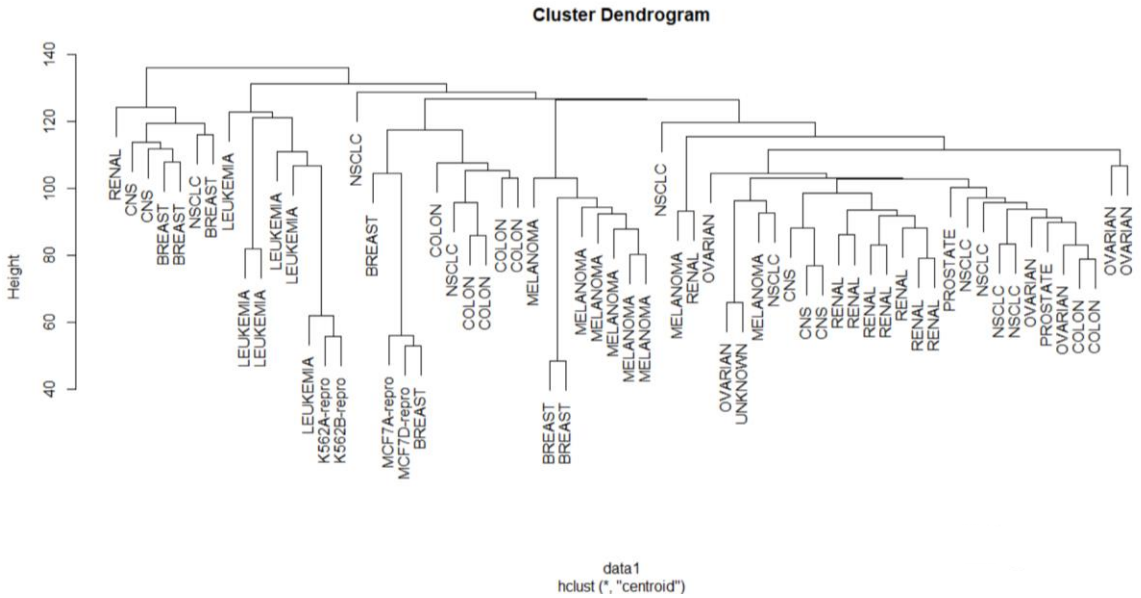
2. Complete linkage



3. Average linkage



4. Centroid linkage



We can see that single linkage is measuring the closest points where as complete linkage measures the furthest. Interestingly average linkage results are somewhere in between of single linkage and complete linkage. Centroid linkage has a similar plot to average linkage. Choice of linkage heavily affects the results.

In the above cluster dendrogram's the vertical scale represents height. Position of fusion is what signifies the (dis)similarity between various samples. More the height at which fusion is, less similar the samples are which means the height at which there is a cut affects the clusters.

I tried various K values and see the no. of sample that belong to what cluster:

K =2		
Linkage	Clusters	
	1	2
single	63	1
complete	47	17
average	57	7
centroid	57	7

K = 4				
Linkage	Clusters			
	1	2	3	4
single	61	1	1	1
complete	40	7	8	9
average	48	7	8	1
centroid	48	7	8	1

K = 6						
Linkage	Clusters					
	1	2	3	4	5	6
single	58	1	1	1	2	1
complete	31	7	8	9	1	8
average	30	7	8	10	1	8
centroid	30	7	8	10	1	8

Task4 & 5:

Applied k means function to NCI dataset for different values of K (2,4,6,8,11). Checked what no. of samples belong to which cluster.

K=2

K clusters	1	2
no. of samples	35	29

K=4

K clusters	1	2	3	4
no. of samples	8	11	20	25

K=6

K clusters	1	2	3	4	5	6
no. of samples	3	10	15	13	19	4

K=8

K clusters	1	2	3	4	5	6	7	8
no. of samples	12	15	2	9	11	6	2	7

K=11

K clusters	1	2	3	4	5	6	7	8
no. of samples	7	3	5	9	1	8	1	10

K clusters	9	10	11
no. of samples	11	3	6

Total sum of squares for k11 = 430290.

Let’s compare a kmeans vs complete linkage for k =4

K=4

K clusters	1	2	3	4
(K means)	8	11	20	25
(complete linkage)	40	7	8	9

We can see how (dis)similarly both the functions perform the same task for k = 4. K-means & HC with cuts in dendrogram to obtain same no. of cluster can produce pretty different results.

K-means tries to find the global optimal where as hierarchical agglomerative clustering is a good way to visualize and get different results using different linkage options. K-means and hierarchical clustering both does the same job but differently, so we just need to understand our aim and choose accordingly.