CS5100J – Data Analysis : Assignment 3

Task 1 and Task 2: Implement the agglomerative clustering with the following linkage: single, complete, average and centroid. Apply your program to the NCI microarray .

The output from this dataset for Single, Complete, Average and Centroid linkages, displayed as below. This shows level wise merging of clusters as well as the height and labels at each level for the first 30 levels:

Level	Single	Compete	Average	Centroid
0	[50] to [51], height = 38.230333, Labels : BREAST to MCF7D-repro	[50] to [51], height = 38.230333, Labels : BREAST to MCF7D-repro	[50] to [51], height = 38.230333, Labels : BREAST to MCF7D-repro	[50] to [51], height = 38.230333, Labels : BREAST to MCF7D-repro
1	[49] to [50,51], height = 38.596042, Labels : MCF7A-repro to BREAST,MCF7D-repro	[49] to [50,51], height = 38.596042, Labels : MCF7A-repro to BREAST,MCF7D-repro	[49] to [50,51], height = 38.596042, Labels : MCF7A-repro to BREAST,MCF7D-repro	[49] to [50,51], height = 38.596042, Labels : MCF7A-repro to BREAST,MCF7D-repro
2	[57] to [58], height = 39.105625, Labels : BREAST to BREAST	[57] to [58], height = 39.105625, Labels : BREAST to BREAST	[57] to [58], height = 39.105625, Labels : BREAST to BREAST	[57] to [58], height = 39.105625, Labels : BREAST to BREAST
3	[21] to [22], height = 45.151581, Labels : UNKNOWN to OVARIAN	[21] to [22], height = 45.151581, Labels : UNKNOWN to OVARIAN	[21] to [22], height = 45.151581, Labels : UNKNOWN to OVARIAN	[21] to [22], height = 45.151581, Labels : UNKNOWN to OVARIAN
4	[35] to [36], height = 45.353381, Labels : K562B-repro to K562A-repro	[35] to [36], height = 45.353381, Labels : K562B-repro to K562A-repro	[35] to [36], height = 45.353381, Labels : K562B-repro to K562A-repro	[35] to [36], height = 45.353381, Labels : K562B-repro to K562A-repro
5	[35,36] to [37], height = 45.442952, Labels : K562B-repro,K562A-repro to LEUKEMIA	[35,36] to [37], height = 45.442952, Labels : K562B-repro,K562A-repro to LEUKEMIA	[35,36] to [37], height = 45.442952, Labels : K562B-repro,K562A-repro to LEUKEMIA	[35,36] to [37], height = 45.442952, Labels : K562B-repro,K562A-repro to LEUKEMIA
6	[1] to [2], height = 51.438231, Labels : CNS to CNS	[1] to [2], height = 51.438231, Labels : CNS to CNS	[1] to [2], height = 51.438231, Labels : CNS to CNS	[1] to [2], height = 51.438231, Labels : CNS to CNS
7	[61] to [62], height = 56.780154, Labels : MELANOMA to MELANOMA	[61] to [62], height = 56.780154, Labels : MELANOMA to MELANOMA	[61] to [62], height = 56.780154, Labels : MELANOMA to MELANOMA	[61] to [62], height = 56.780154, Labels : MELANOMA to MELANOMA
8	[12] to [13], height = 57.917264, Labels : RENAL to RENAL	[12] to [13], height = 57.917264, Labels : RENAL to RENAL	[12] to [13], height = 57.917264, Labels : RENAL to RENAL	[12] to [13], height = 57.917264, Labels : RENAL to RENAL
9	[12,13] to [14], height = 60.350523, Labels : RENAL,RENAL to RENAL	[12,13] to [14], height = 60.350523, Labels : RENAL,RENAL to RENAL	[12,13] to [14], height = 60.350523, Labels : RENAL,RENAL to RENAL	[12,13] to [14], height = 60.350523, Labels : RENAL,RENAL to RENAL
10	[42] to [44], height = 60.496507, Labels : COLON to COLON	[42] to [44], height = 60.496507, Labels : COLON to COLON	[42] to [44], height = 60.496507, Labels : COLON to COLON	[12,13,14] to [15], height = 55.713942, Labels : RENAL,RENAL,RENAL to RENAL
11	[12,13,14] to [17], height = 60.804411, Labels : RENAL,RENAL,RENAL to RENAL	[39] to [40], height = 61.554253, Labels : LEUKEMIA to LEUKEMIA	[39] to [40], height = 61.554253, Labels : LEUKEMIA to LEUKEMIA	[12,13,14,15] to [16], height = 52.202651, Labels : RENAL,RENAL,RENAL,RENAL to RENAL

12	[39] to [40], height = 61.554253, Labels : LEUKEMIA to LEUKEMIA	[31] to [32], height = 61.637504, Labels : NSCLC to NSCLC	[31] to [32], height = 61.637504, Labels : NSCLC to NSCLC	[12,13,14,15,16] to [17], height = 56.037061, Labels : RENAL,RENAL,RENAL,RENAL,RENAL to RENAL
13	[31] to [32], height = 61.637504, Labels : NSCLC to NSCLC	[42,44] to [46], height = 61.929281, Labels : COLON,COLON to COLON	[42,44] to [46], height = 61.929281, Labels : COLON,COLON to COLON	[12,13,14,15,16,17] to [32], height = 57.002349, Labels : RENAL,RENAL,RENAL,RENAL,RENAL,RENAL to NSCLC
14	[42,44] to [46], height = 61.929281, Labels : COLON,COLON to COLON	[15] to [16], height = 62.178059, Labels : RENAL to RENAL	[15] to [16], height = 62.178059, Labels : RENAL to RENAL	[11] to [12,13,14,15,16,17,32], height = 55.598321, Labels : RENAL to RENAL,RENAL,RENAL,RENAL,RENAL,RENAL,NSCLC
15	[15] to [16], height = 62.178059, Labels : RENAL to RENAL	[60] to [61,62], height = 63.674411, Labels : MELANOMA to MELANOMA,MELANOMA	[60] to [61,62], height = 63.674411, Labels : MELANOMA to MELANOMA,MELANOMA	[11,12,13,14,15,16,17,32] to [30], height = 56.404105, Labels : RENAL,RENAL,RENAL,RENAL,RENAL,RENAL,RENAL,NSCLC to PROSTATE
16	[12,13,14,17] to [15,16], height = 62.429473, Labels: RENAL,RENAL,RENAL,RENAL to RENAL,RENAL	[30] to [31,32], height = 64.192227, Labels : PROSTATE to NSCLC,NSCLC	[12,13,14] to [15,16], height = 64.010526, Labels : RENAL,RENAL,RENAL to RENAL,RENAL	[9] to [11,30,12,13,14,15,16,17,32], height = 57.221442, Labels : NSCLC to RENAL,PROSTATE,RENAL,RENAL,RENAL,RENAL,RENAL,RENAL,NSCLC
17	[42,44,46] to [45], height = 63.264141, Labels : COLON,COLON,COLON to COLON	[11] to [12,13,14], height = 65.282644, Labels : RENAL to RENAL,RENAL,RENAL	[30] to [31,32], height = 64.192227, Labels : PROSTATE to NSCLC,NSCLC	[3] to [9,11,30,12,13,14,15,16,17,32], height = 57.575104, Labels : CNS to NSCLC,RENAL,PROSTATE,RENAL,RENAL,RENAL,RENAL,RENAL,RENAL,NSCLC
18	[11] to [12,13,14,17,15,16], height = 63.576076, Labels : RENAL to RENAL,RENAL,RENAL,RENAL,RENAL,RENAL	[42,44,46] to [45], height = 65.514496, Labels : COLON,COLON,COLON to COLON	[42,44,46] to [45], height = 64.389319, Labels : COLON,COLON,COLON to COLON	[42] to [44], height = 60.496507, Labels : COLON to COLON
19	[60] to [61,62], height = 63.674411, Labels : MELANOMA to MELANOMA,MELANOMA	[15,16] to [17], height = 65.858255, Labels : RENAL,RENAL to RENAL	[60,61,62] to [64], height = 65.044272, Labels : MELANOMA,MELANOMA, MELANOMA to MELANOMA	[39] to [40], height = 61.554253, Labels : LEUKEMIA to LEUKEMIA
20	[60,61,62] to [64], height = 63.745878, Labels : MELANOMA,MELANOMA,MELANOMA to MELANOMA	[59] to [60,61,62], height = 66.231099, Labels : MELANOMA to MELANOMA,MELANOMA,MELANOMA	[12,13,14,15,16] to [17], height = 66.244626, Labels : RENAL,RENAL,RENAL,RENAL,RENAL to RENAL	[42,44] to [46], height = 61.929281, Labels : COLON,COLON to COLON[42,44] to [46], height = 61.929281, Labels : COLON,COLON to COLON
21	[30] to [31,32], height = 64.192227, Labels : PROSTATE to NSCLC,NSCLC	[59,60] to [64], height = 66.342666, Labels : MELANOMA,MELANOMA to MELANOMA	[59] to [60,64,61,62], height = 66.715384, Labels : MELANOMA to MELANOMA,MELANOMA, MELANOMA,MELANOMA	[42,44,46] to [45], height = 56.466286, Labels : COLON,COLON,COLON to COLON

22	[59] to [60,64,61,62], height = 64.372622, Labels : MELANOMA to MELANOMA,MELANOMA,MELANOMA,MELANOMA	[3] to [9], height = 67.837892, Labels : CNS to NSCLC	[3] to [9], height = 67.837892, Labels : CNS to NSCLC	[42,44,46,45] to [53], height = 62.435723, Labels : COLON,COLON,COLON,COLON to NSCLC
23	[30,31,32] to [33], height = 65.128921, Labels : PROSTATE,NSCLC,NSCLC to NSCLC	[59,60,64] to [63], height = 68.606023, Labels : MELANOMA,MELANOMA,MELANOMA to MELANOMA	[29] to [30,31,32], height = 67.878675, Labels : OVARIAN to PROSTATE,NSCLC,NSCLC	[42,44,46,45,53] to [48], height = 62.177763, Labels : COLON,COLON,COLON,COLON,NSCLC to COLON
24	[9] to [11,12,13,14,17,15,16], height = 65.700671, Labels : NSCLC to RENAL,RENAL,RENAL,RENAL,RENAL,RENAL,RENAL	[29] to [30,31,32], height = 69.993214, Labels : OVARIAN to PROSTATE,NSCLC,NSCLC	[24] to [29,30], height = 67.632242, Labels : PROSTATE to OVARIAN,PROSTATE	[29] to [42,44,46,45,53,48], height = 61.609040, Labels : OVARIAN to COLON,COLON,COLON,COLON,NSCLC,COLON
25	9,11] to [30,33,31,32], height = 66.410872, Labels : NSCLC,RENAL to PROSTATE,NSCLC,NSCLC,NSCLC	[24] to [29,30], height = 67.632242, Labels : PROSTATE to OVARIAN,PROSTATE	[11] to [12,13,14,15,16,17], height = 69.763865, Labels : RENAL to RENAL,RENAL,RENAL,RENAL,RENAL,RENAL	[24] to [29,42,44,46,45,53,48], height = 59.485371, Labels : PROSTATE to OVARIAN,COLON,COLON,COLON,COLON,NSCLC,COLON
26	[57,58] to [59,60,64,61,62], height = 67.678752, Labels : BREAST,BREAST to MELANOMA,MELANOMA,MELANOMA,MELANOMA,MELANOMA	[27] to [28], height = 70.937596, Labels : OVARIAN to OVARIAN	[59,60,64,61,62] to [63], height = 70.586999, Labels : MELANOMA,MELANOMA,MELANOMA,MELANOMA,MELANOMA to MELANOMA	[60] to [61,62], height = 63.674411, Labels : MELANOMA to MELANOMA,MELANOMA
27	[42,44,46,45] to [43], height = 68.165875, Labels : COLON,COLON,COLON,COLON to COLON	[3,9] to [23], height = 71.273901, Labels : CNS,NSCLC to MELANOMA	[27] to [28], height = 70.937596, Labels : OVARIAN to OVARIAN	[60,61,62] to [64], height = 56.734782, Labels : MELANOMA,MELANOMA,MELANOMA to MELANOMA
28	[49,50,51] to [52], height = 68.205117, Labels : MCF7A-repro,BREAST,MCF7D-repro to BREAST	[25] to [26], height = 73.560830, Labels : OVARIAN to OVARIAN	[3,9] to [24,29,30], height = 70.989513, Labels : CNS,NSCLC to PROSTATE,OVARIAN,PROSTATE	[59] to [60,64,61,62], height = 58.656174, Labels : MELANOMA to MELANOMA,MELANOMA,MELANOMA,MELANOMA
29	[57,58,59,60,64,61,62] to [63], height = 68.606023, Labels : BREAST,BREAST,MELANOMA,MELANOMA,MELANOMA,MELANOMA,MELANOMA to MELANOMA	[56] to [57,58], height = 74.010797, Labels : MELANOMA to BREAST,BREAST	[49,50,51] to [52], height = 71.506071, Labels : MCF7A-repro,BREAST,MCF7D-repro to BREAST	[59,60,64,61,62] to [63], height = 59.522113, Labels : MELANOMA,MELANOMA,MELANOMA,MELANOMA,MELANOMA to MELANOMA
30	[42,44,46,45,43] to [53], height = 68.813084, Labels : COLON,COLON,COLON,COLON,COLON to NSCLC	[3,9,23] to [24,29,30], height = 74.312891, Labels : CNS,NSCLC,MELANOMA to PROSTATE,OVARIAN,PROSTATE	[42,44,46,45] to [53], height = 72.381506, Labels : COLON,COLON,COLON,COLON to NSCLC	[57,58] to [59,63,60,64,61,62], height = 61.701218, Labels : BREAST,BREAST to MELANOMA,MELANOMA,MELANOMA,MELANOMA,MELANOMA,MELANOMA

<u>Task 3 :</u> <u>Discuss the performance of hierarchical agglomerative clustering</u>

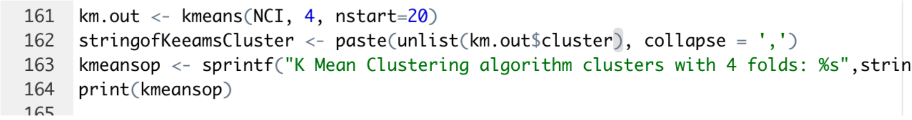
We see that all the fours linkages work differently when it comes to implementation. The Single linkage works of the minimum distances of the points. It computes distance between each of the points and takes the minimum. In contrast to that, complete linkage takes the max distance while making the cluster. Average linkage takes the mean(average) between the points to form the clusters.

The functionality of centroid is different because its first takes the centroid of the points and then forms the clusters.

<u>Task 4 :</u> <u>Apply the R function kmeans() to the above NCI microarray data set with different K and discuss its performance.</u>

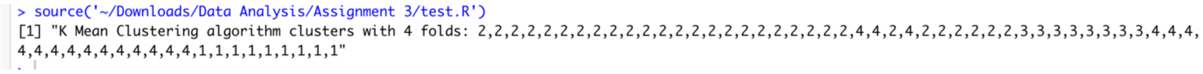
Screenshot of the code:

K = 4



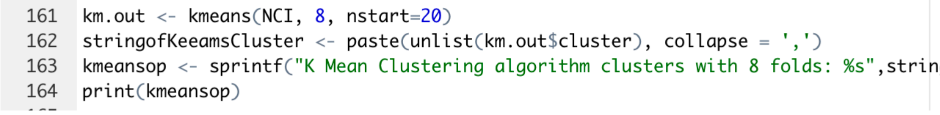
```
km.out <- kmeans(NCI, 4, nstart=20)
stringofKeeamsCluster <- paste(unlist(km.out$cluster), collapse = ',')
kmeansop <- sprintf("K Mean Clustering algorithm clusters with 4 folds: %s",stringofKeeamsCluster)
print(kmeansop)
```

Output:



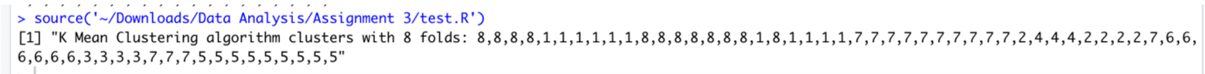
```
> source('~/Downloads/Data Analysis/Assignment 3/test.R')
[1] "K Mean Clustering algorithm clusters with 4 folds: 2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,4,4,2,4,2,2,2,2,2,2,3,3,3,3,3,3,3,3,4,4,4,
4,4,4,4,4,4,4,4,4,4,4,1,1,1,1,1,1,1,1,1"
```

K = 8



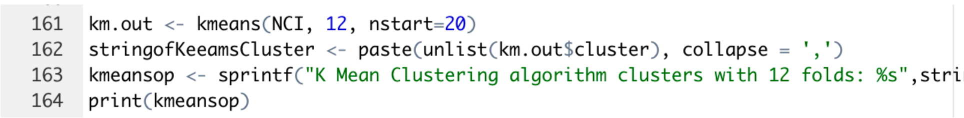
```
km.out <- kmeans(NCI, 8, nstart=20)
stringofKeeamsCluster <- paste(unlist(km.out$cluster), collapse = ',')
kmeansop <- sprintf("K Mean Clustering algorithm clusters with 8 folds: %s",stringofKeeamsCluster)
print(kmeansop)
```

Output:



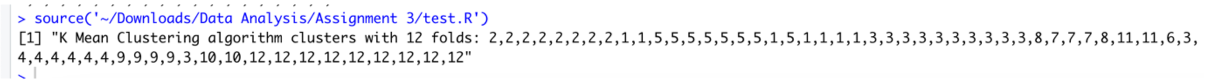
```
> source('~/Downloads/Data Analysis/Assignment 3/test.R')
[1] "K Mean Clustering algorithm clusters with 8 folds: 8,8,8,8,1,1,1,1,1,1,8,8,8,8,8,8,8,1,8,1,1,1,1,7,7,7,7,7,7,7,7,7,7,2,4,4,4,2,2,2,2,2,7,6,6,
6,6,6,6,3,3,3,3,7,7,7,5,5,5,5,5,5,5,5,5"
```

K = 12



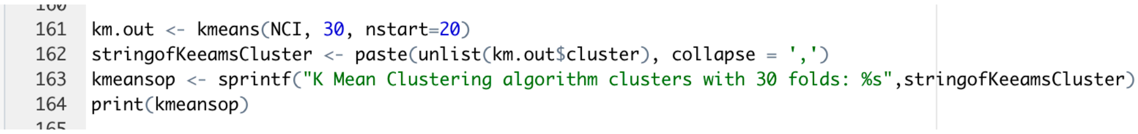
```
km.out <- kmeans(NCI, 12, nstart=20)
stringofKeeamsCluster <- paste(unlist(km.out$cluster), collapse = ',')
kmeansop <- sprintf("K Mean Clustering algorithm clusters with 12 folds: %s",stringofKeeamsCluster)
print(kmeansop)
```

Output:



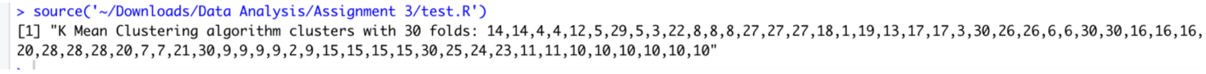
```
> source('~/Downloads/Data Analysis/Assignment 3/test.R')
[1] "K Mean Clustering algorithm clusters with 12 folds: 2,2,2,2,2,2,2,2,1,1,5,5,5,5,5,5,5,1,5,1,1,1,1,3,3,3,3,3,3,3,3,3,3,8,7,7,7,8,11,11,6,3,
4,4,4,4,4,4,9,9,9,9,3,10,10,12,12,12,12,12,12,12,12,12"
```

K = 30



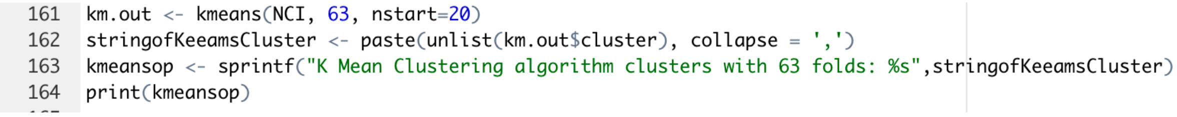
```
km.out <- kmeans(NCI, 30, nstart=20)
stringofKeeamsCluster <- paste(unlist(km.out$cluster), collapse = ',')
kmeansop <- sprintf("K Mean Clustering algorithm clusters with 30 folds: %s",stringofKeeamsCluster)
print(kmeansop)
```

Output:



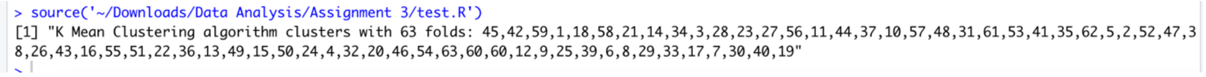
```
> source('~/Downloads/Data Analysis/Assignment 3/test.R')
[1] "K Mean Clustering algorithm clusters with 30 folds: 14,14,4,4,12,5,29,5,3,22,8,8,8,27,27,27,18,1,19,13,17,17,3,30,26,26,6,6,30,30,16,16,16,
20,28,28,28,20,7,7,21,30,9,9,9,9,2,9,15,15,15,15,30,25,24,23,11,11,10,10,10,10,10,10"
```

K = 64



```
km.out <- kmeans(NCI, 63, nstart=20)
stringofKeeamsCluster <- paste(unlist(km.out$cluster), collapse = ',')
kmeansop <- sprintf("K Mean Clustering algorithm clusters with 63 folds: %s",stringofKeeamsCluster)
print(kmeansop)
```

Output:



```
> source('~/Downloads/Data Analysis/Assignment 3/test.R')
[1] "K Mean Clustering algorithm clusters with 63 folds: 45,42,59,1,18,58,21,14,34,3,28,23,27,56,11,44,37,10,57,48,31,61,53,41,35,62,5,2,52,47,3
8,26,43,16,55,51,22,36,13,49,15,50,24,4,32,20,46,54,63,60,60,12,9,25,39,6,8,29,33,17,7,30,40,19"
```

We see that the clustered output are not upto the mark for lower values for K. For example, K = 4, as we see there are only 4 clusters that are formed. There might be a lot of dissimilarity between the points within one cluster.

Also, for very high value of K, there are too many clusters formed. There needs to be a balance between the number of clusters which are to be formed and the number of data points. Obviously K should be greater that the number of labels.

Task 5 : Compare and contrast the performance of K-means and hierarchical agglomerative clustering.

- First of all, K-means takes divides the whole dataset into the number of cluster defined, so this needs prior analysis of how many clusters are needed, whereas hierarchical agglomerative clustering forms different number of clusters at each level. The tree formed can be cut at any level depending on the number of clusters needed.
- Hierarchical agglomerative clustering is not suitable for large dataset as it takes a lot of time as witnessed in the NCI dataset. On the other hand, K-means can easily be implemented for a large dataset.
- As witnessed, Unlike K-means, the hierarchical agglomerative clustering takes a bottom up approach, starting with one dataset in each cluster and merges them gradually on different levels. K-means on the other hand takes the whole dataset in every iteration.

Task 6: Optional: Discuss how to choose the number of clusters in the K-means and hierarchical agglomerative clustering.

As discussed previously, for K-means, clustered output are not upto the mark for lower values for K. For example, K = 4, as we see there are only 4 clusters that are formed. There might be a lot of dissimilarity between the points within one cluster.

Also, for very high value of K, there are too many clusters formed. There needs to be a balance between the number of clusters which are to be formed and the number of data points. Obviously K should be greater that the number of labels.

For hierarchical agglomerative clustering, one can go with the flow as the tree formations builds and cut the tree whenever one is satisfied with the number of clusters and the datapoints in the clusters.

The dendrograms helps visualize the clusters at each level and can be use to analyse the stopping criteria for hierarchical agglomerative clustering.