Note: The merge saves the new cluster name as the min of the new cluster in the datamatrix

For Single Linkage:

```
> clust.single <- hc(data, k, 'single')
[1] "level: 0"
[1] "height: 38.2303326650995"
[1] "merge: 50 51"
[1] "level: 1"
[1] "height: 38.5960416615816"
[1] "merge: 49 50"
[1] "level: 2"
[1] "height: 39.1056247853175"
[1] "merge: 57 58"
[1] "level: 3"
[1] "height: 45.1515809571716"
[1] "merge: 21 22"
[1] "level: 4"
[1] "height: 45.3533812944176"
[1] "merge: 35 36"
[1] "level: 5"
[1] "height: 45.4429522845579"
[1] "merge: 37 35"
[1] "level: 6"
[1] "height: 51.43823072875"
[1] "merge: 1 2"
[1] "level: 7"
[1] "height: 56.7801537322295"
[1] "merge: 61 62"
[1] "level: 8"
[1] "height: 56.9683108656075"
[1] "merge: 60 61"
[1] "level: 9"
[1] "height: 57.9172637924558"
[1] "merge: 12 13"
[1] "level: 10"
[1] "height: 60.3505227083712"
[1] "merge: 14 12"
[1] "level: 11"
[1] "height: 60.4965065701272"
[1] "merge: 42 44"
[1] "level: 12"
[1] "height: 60.8044109958134"
[1] "merge: 17 14"
[1] "level: 13"
[1] "height: 61.5542532619369"
[1] "merge: 39 40"
```

[1] "level: 14"
[1] "height: 61.6375036165801"
[1] "merge: 31 32"
[1] "level: 15"
[1] "height: 61.9292807028045"
[1] "merge: 46 42"
[1] "level: 16"
[1] "height: 62.1780594004175"
[1] "merge: 15 16"
[1] "level: 17"
[1] "height: 62.4294728842052"
[1] "merge: 17 14"
[1] "level: 18"
[1] "height: 63.2641409776857"
[1] "merge: 45 46"
[1] "level: 19"
[1] "height: 63.4286314852814"
[1] "merge: 30 45"
[1] "level: 20"
[1] "height: 63.5760759705679"
[1] "merge: 11 17"
[1] "level: 21"
[1] "height: 63.7458778886246"
[1] "merge: 64 60"
[1] "level: 22"
[1] "height: 64.1922273243777"
[1] "merge: 31 32"
[1] "level: 23"
[1] "height: 64.3726220563971"
[1] "merge: 59 64"
[1] "level: 24"
[1] "height: 65.1289210215624"
[1] "merge: 33 31"
[1] "level: 25"
[1] "height: 65.2984922609882"
[1] "merge: 24 33"
[1] "level: 26"
[1] "height: 65.4727079007854"
[1] "merge: 29 24"
[1] "level: 27"
[1] "height: 65.700670543825"
[1] "merge: 9 11"
[1] "level: 28"
[1] "height: 65.7710667950191"
[1] "merge: 29 24"
[1] "level: 29"
[1] "height: 65.9120453185545"
[1] "merge: 27 29"

[1] "level: 30"
[1] "height: 65.9381464424553"
[1] "merge: 3 1"
[1] "level: 31"
[1] "height: 66.4870807234393"
[1] "merge: 63 59"
[1] "level: 32"
[1] "height: 66.9087797281851"
[1] "merge: 57 58"
[1] "level: 33"
[1] "height: 67.8378924790861"
[1] "merge: 27 29"
[1] "level: 34"
[1] "height: 68.1658753895546"
[1] "merge: 43 27"
[1] "level: 35"
[1] "height: 68.2051174169133"
[1] "merge: 52 49"
[1] "level: 36"
[1] "height: 68.5800971122978"
[1] "merge: 56 57"
[1] "level: 37"
[1] "height: 68.8130841601359"
[1] "merge: 53 43"
[1] "level: 38"
[1] "height: 69.0272629179475"
[1] "merge: 48 53"
[1] "level: 39"
[1] "height: 69.9845584067099"
[1] "merge: 52 49"
[1] "level: 40"
[1] "height: 69.9896654512773"
[1] "merge: 21 22"
[1] "level: 41"
[1] "height: 70.5172849516472"
[1] "merge: 47 21"
[1] "level: 42"
[1] "height: 70.9375961915901"
[1] "merge: 28 47"
[1] "level: 43"
[1] "height: 71.2739011542286"
[1] "merge: 23 28"
[1] "level: 44"
[1] "height: 71.7447056962072"
[1] "merge: 4 23"
[1] "level: 45"
[1] "height: 72.5517251083991"
[1] "merge: 26 4"

[1] "level: 46"
[1] "height: 72.8503483188853"
[1] "merge: 54 26"
[1] "level: 47"
[1] "height: 73.5608302937134"
[1] "merge: 25 54"
[1] "level: 48"
[1] "height: 74.0046569144778"
[1] "merge: 6 25"
[1] "level: 49"
[1] "height: 74.1766469037468"
[1] "merge: 7 6"

The Cluster Number of each observation:
 14 14 14 14 1 14 14 2 14 3 14 14 14 14 14 14 14 4 5 6 14 14 14 14 14 14 14 14 14 14 14
14 14 7 11 11 11 8 12 12 9 14 14 14 14 14 14 14 14 14 14 14 14 10 13 13 13 13 13 13 13
13 13

For Complete Linkage:

> clust.complete <- hc(data, k, 'complete')
[1] "level: 0"
[1] "height: 38.2303326650995"
[1] "merge: 50 51"
[1] "level: 1"
[1] "height: 39.1056247853175"
[1] "merge: 57 58"
[1] "level: 2"
[1] "height: 39.9998987588587"
[1] "merge: 49 50"
[1] "level: 3"
[1] "height: 45.1515809571716"
[1] "merge: 21 22"
[1] "level: 4"
[1] "height: 45.3533812944176"
[1] "merge: 35 36"
[1] "level: 5"
[1] "height: 48.0116281889191"
[1] "merge: 37 35"
[1] "level: 6"
[1] "height: 51.43823072875"
[1] "merge: 1 2"
[1] "level: 7"
[1] "height: 56.7801537322295"
[1] "merge: 61 62"
[1] "level: 8"

[1] "height: 57.9172637924558"
[1] "merge: 12 13"
[1] "level: 9"
[1] "height: 60.4965065701272"
[1] "merge: 42 44"
[1] "level: 10"
[1] "height: 61.5542532619369"
[1] "merge: 39 40"
[1] "level: 11"
[1] "height: 61.6375036165801"
[1] "merge: 31 32"
[1] "level: 12"
[1] "height: 62.1780594004175"
[1] "merge: 15 16"
[1] "level: 13"
[1] "height: 62.7816325750382"
[1] "merge: 14 15"
[1] "level: 14"
[1] "height: 63.2641409776857"
[1] "merge: 45 46"
[1] "level: 15"
[1] "height: 63.6744111200614"
[1] "merge: 60 61"
[1] "level: 16"
[1] "height: 65.5144960970407"
[1] "merge: 44 45"
[1] "level: 17"
[1] "height: 65.5915789808415"
[1] "merge: 13 14"
[1] "level: 18"
[1] "height: 65.7641353618998"
[1] "merge: 29 30"
[1] "level: 19"
[1] "height: 65.4727079007854"
[1] "merge: 42 29"
[1] "level: 20"
[1] "height: 66.2310987408852"
[1] "merge: 59 60"
[1] "level: 21"
[1] "height: 67.0731309277801"
[1] "merge: 11 12"
[1] "level: 22"
[1] "height: 67.2853859414947"
[1] "merge: 14 11"
[1] "level: 23"
[1] "height: 67.8378924790861"
[1] "merge: 3 9"
[1] "level: 24"

[1] "height: 68.2658071011412"
[1] "merge: 13 3"
[1] "level: 25"
[1] "height: 68.6060228178099"
[1] "merge: 63 64"
[1] "level: 26"
[1] "height: 66.4870807234393"
[1] "merge: 61 63"
[1] "level: 27"
[1] "height: 67.8364566889944"
[1] "merge: 60 61"
[1] "level: 28"
[1] "height: 69.9340961767304"
[1] "merge: 43 42"
[1] "level: 29"
[1] "height: 69.9932141350108"
[1] "merge: 32 42"
[1] "level: 30"
[1] "height: 70.6227323071042"
[1] "merge: 24 32"
[1] "level: 31"
[1] "height: 70.9375961915901"
[1] "merge: 27 28"
[1] "level: 32"
[1] "height: 71.0063981802724"
[1] "merge: 62 60"
[1] "level: 33"
[1] "height: 71.9980110029875"
[1] "merge: 13 24"
[1] "level: 34"
[1] "height: 72.1485772779284"
[1] "merge: 59 62"
[1] "level: 35"
[1] "height: 72.1587132629337"
[1] "merge: 27 13"
[1] "level: 36"
[1] "height: 72.1896185800297"
[1] "merge: 30 27"
[1] "level: 37"
[1] "height: 72.4335523219322"
[1] "merge: 16 14"
[1] "level: 38"
[1] "height: 72.4335523219322"
[1] "merge: 12 13"
[1] "level: 39"
[1] "height: 72.4455541130647"
[1] "merge: 32 30"
[1] "level: 40"

[1] "height: 72.4537784221494"
[1] "merge: 14 13"
[1] "level: 41"
[1] "height: 72.4050459848564"
[1] "merge: 32 14"
[1] "level: 42"
[1] "height: 72.7029181841542"
[1] "merge: 29 32"
[1] "level: 43"
[1] "height: 72.7368219240676"
[1] "merge: 17 12"
[1] "level: 44"
[1] "height: 73.1037517040576"
[1] "merge: 13 29"
[1] "level: 45"
[1] "height: 73.310163551949"
[1] "merge: 30 32"
[1] "level: 46"
[1] "height: 73.5608302937134"
[1] "merge: 25 26"
[1] "level: 47"
[1] "height: 73.6766982322482"
[1] "merge: 53 31"
[1] "level: 48"
[1] "height: 71.3957229835243"
[1] "merge: 30 53"
[1] "level: 49"
[1] "height: 73.401850673772"
[1] "merge: 42 30"

The Cluster Number of each observation:
 25 25 35 1 2 3 4 5 35 6 32 33 35 35 33 33 33 7 8 9 23 23 10 31 36 36 34 34 34 37 37 37 11 12 24 24 24 13 26 26 14 37 29 29 27 27 15 16 22 22 22 17 37 18 19 20 21 21 30 30 30 30 30 30

For Average Linkage:

 clust.average <- hc(data, k, 'average')
[1] "level: 0"
[1] "height: 38.2303326650995"
[1] "merge: 50 51"
[1] "level: 1"
[1] "height: 39.1056247853175"
[1] "merge: 57 58"
[1] "level: 2"
[1] "height: 39.2979702102202"

[1] "merge: 49 50"
[1] "level: 3"
[1] "height: 45.1515809571716"
[1] "merge: 21 22"
[1] "level: 4"
[1] "height: 45.3533812944176"
[1] "merge: 35 36"
[1] "level: 5"
[1] "height: 46.7272902367385"
[1] "merge: 37 35"
[1] "level: 6"
[1] "height: 51.43823072875"
[1] "merge: 1 2"
[1] "level: 7"
[1] "height: 56.7801537322295"
[1] "merge: 61 62"
[1] "level: 8"
[1] "height: 57.9172637924558"
[1] "merge: 12 13"
[1] "level: 9"
[1] "height: 60.3213609928344"
[1] "merge: 60 61"
[1] "level: 10"
[1] "height: 60.4965065701272"
[1] "merge: 42 44"
[1] "level: 11"
[1] "height: 61.5542532619369"
[1] "merge: 39 40"
[1] "level: 12"
[1] "height: 61.6375036165801"
[1] "merge: 31 32"
[1] "level: 13"
[1] "height: 62.1780594004175"
[1] "merge: 15 16"
[1] "level: 14"
[1] "height: 62.6055527296217"
[1] "merge: 14 15"
[1] "level: 15"
[1] "height: 63.2641409776857"
[1] "merge: 45 46"
[1] "level: 16"
[1] "height: 63.2922679539899"
[1] "merge: 13 14"
[1] "level: 17"
[1] "height: 63.7218883999226"
[1] "merge: 44 45"
[1] "level: 18"
[1] "height: 65.0349372746071"

[1] "merge: 64 60"
[1] "level: 19"
[1] "height: 65.3405086404369"
[1] "merge: 42 44"
[1] "level: 20"
[1] "height: 65.7641353618998"
[1] "merge: 29 30"
[1] "level: 21"
[1] "height: 64.4506696930334"
[1] "merge: 42 29"
[1] "level: 22"
[1] "height: 66.0896514472333"
[1] "merge: 59 64"
[1] "level: 23"
[1] "height: 66.177887622959"
[1] "merge: 11 12"
[1] "level: 24"
[1] "height: 63.7373282068113"
[1] "merge: 14 11"
[1] "level: 25"
[1] "height: 66.7310307783423"
[1] "merge: 12 13"
[1] "level: 26"
[1] "height: 66.8819162231878"
[1] "merge: 32 42"
[1] "level: 27"
[1] "height: 67.4626986394462"
[1] "merge: 3 1"
[1] "level: 28"
[1] "height: 67.8514721429925"
[1] "merge: 15 14"
[1] "level: 29"
[1] "height: 66.8814367541476"
[1] "merge: 16 15"
[1] "level: 30"
[1] "height: 68.4153126166358"
[1] "merge: 24 32"
[1] "level: 31"
[1] "height: 68.6216008333031"
[1] "merge: 11 12"
[1] "level: 32"
[1] "height: 69.0499857831425"
[1] "merge: 43 42"
[1] "level: 33"
[1] "height: 67.4687471969467"
[1] "merge: 46 43"
[1] "level: 34"
[1] "height: 68.425008875365"

[1] "merge: 45 46"
[1] "level: 35"
[1] "height: 69.2169071376889"
[1] "merge: 63 59"
[1] "level: 36"
[1] "height: 69.3915668461588"
[1] "merge: 13 24"
[1] "level: 37"
[1] "height: 69.9086691837626"
[1] "merge: 32 11"
[1] "level: 38"
[1] "height: 70.4102002631917"
[1] "merge: 17 16"
[1] "level: 39"
[1] "height: 70.5500385560074"
[1] "merge: 27 13"
[1] "level: 40"
[1] "height: 70.6173061026467"
[1] "merge: 33 31"
[1] "level: 41"
[1] "height: 69.5652302315369"
[1] "merge: 30 33"
[1] "level: 42"
[1] "height: 69.5952108474373"
[1] "merge: 13 30"
[1] "level: 43"
[1] "height: 69.4720946647401"
[1] "merge: 14 13"
[1] "level: 44"
[1] "height: 70.1540598520635"
[1] "merge: 11 14"
[1] "level: 45"
[1] "height: 70.6531406874718"
[1] "merge: 32 17"
[1] "level: 46"
[1] "height: 70.8847275531081"
[1] "merge: 31 27"
[1] "level: 47"
[1] "height: 70.9299210806495"
[1] "merge: 16 11"
[1] "level: 48"
[1] "height: 71.2739011542286"
[1] "merge: 9 23"
[1] "level: 49"
[1] "height: 69.476108996489"
[1] "merge: 13 9"

The Cluster Number of each observation:

29 29 29 1 2 3 4 5 36 6 35 33 36 35 33 35 33 7 8 9 25 25 36 34 10 11 34 12 34 35 35 35 35
13 26 26 26 14 27 27 15 34 30 30 30 30 16 17 24 24 24 18 19 20 21 22 23 23 31 31 31 31
31 31

For Centroid Linkage:

 clust.centroid <- hc(data, k, 'centroid')
[1] "level: 0"
[1] "height: 38.2303326650995"
[1] "merge: 50 51"
[1] "level: 1"
[1] "height: 0.00273189738092078"
[1] "merge: 54 50"
[1] "level: 2"
[1] "height: 0.000910632460306919"
[1] "merge: 50 51"
[1] "level: 3"
[1] "height: 0.000364252984122782"
[1] "merge: 54 50"
[1] "level: 4"
[1] "height: 0.000136594869046072"
[1] "merge: 50 51"
[1] "level: 5"
[1] "height: 5.25364880946573e-05"
[1] "merge: 54 50"
[1] "level: 6"
[1] "height: 2.00139002265176e-05"
[1] "merge: 50 51"
[1] "level: 7"
[1] "height: 7.65237361598226e-06"
[1] "merge: 54 50"
[1] "level: 8"
[1] "height: 2.9218153806132e-06"
[1] "merge: 50 51"
[1] "level: 9"
[1] "height: 1.1161991341109e-06"
[1] "merge: 54 50"
[1] "level: 10"
[1] "height: 4.26326058105414e-07"
[1] "merge: 50 51"
[1] "level: 11"
[1] "height: 1.62845575885084e-07"
[1] "merge: 54 50"
[1] "level: 12"
[1] "height: 6.22009627346265e-08"
[1] "merge: 50 51"
[1] "level: 13"
[1] "height: 2.37587284498963e-08"

[1] "merge: 54 50"
[1] "level: 14"
[1] "height: 9.07501573541025e-09"
[1] "merge: 50 51"
[1] "level: 15"
[1] "height: 3.46634888501174e-09"
[1] "merge: 54 50"
[1] "level: 16"
[1] "height: 1.32402644750784e-09"
[1] "merge: 50 51"
[1] "level: 17"
[1] "height: 5.05731886923932e-10"
[1] "merge: 54 50"
[1] "level: 18"
[1] "height: 1.93171752899124e-10"
[1] "merge: 50 51"
[1] "level: 19"
[1] "height: 7.3778844145167e-11"
[1] "merge: 54 50"

Discuss the performance of hierarchical agglomerative clustering when using different linkage functions.

Answer :
     It is observable that each linkage function gives different clustering and the performance of these can be found using the original label dataset given.

For single Linkage: The cluster formed for is too large for some values that gives a wrong prediction of clusters given the original values.
For complete Linkage: The clusters formed are too sparse and may give smaller clusters leaving out the actual points required in the cluster.
For Average Linkage: it performs better than the first two but still misses out few points to include in the cluster.
For Centroid Linkage: This one performs the best among all 4 linkages but takes more time computationally.

Apply the R function kmeans() to the above NCI microarray data set with different K and discuss its performance.

Answer:
     It is clear that when the K is too low it has most of the points under the same clusters and the performance in such cases is pretty bad. On the other hand, when the K is too high, each of the points start forming their own cluster and hence we get poor performance if compared to the original labels.

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

Answer:

k=14
>kmeanOutput <- kmeans(data,k)
> as.vector(kmeanOutput$cluster)
 [1] 5 5 13 13 13 13 1 13 1 8 7 7 7 7 7 7 7 1 1 1 1 1 1 4 4 4 4 4
[29] 4 4 4 4 4 14 9 9 9 14 12 12 14 4 11 11 11 11 11 11 3 3 3 6 4 2 2 10
[57] 10 10 10 10 10 10 10 10

>clust.average <- hc(data, k, 'average')
The Cluster Number of each observation:
 29 29 29 1 2 3 4 5 36 6 35 33 36 35 33 35 33 7 8 9 25 25 36 34 10 11 34 12 34 35 35 35 35
13 26 26 26 14 27 27 15 34 30 30 30 30 16 17 24 24 24 18 19 20 21 22 23 23 31 31 31 31
31 31

For the optimal number of cluster i.e. k=14, HAC outperforms kmean by creating clusters more precisely with few outliers.

As agglomerative hierarchical clustering is done one cluster at time it is easier to divide the cluster at any given height but in kmeans, this is done simultaneously creating multiple branches with same heights. Cutting down these branches is not an easy task in kmeans as compared to AHC.

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

A trivial way to choose the number of clusters is to use an empirical method where the number of clusters is taken as the square root of N/2 $(\sqrt{\frac{N}{2}})$, where N is the number of observations.

The best way to find out the value of k without the prior knowledge of domain, is to use either elbow method or gap statistic.
In elbow method, Sum of within-cluster variance is calculated for each value of K and plotting the K values and their corresponding sum of within-cluster variance helps in finding the number of clusters. Initially, Error measure decreases with increase in cluster number. After a particular point, Error measure starts flattening. This point where it starts flattening can be chosen as the value of K.