Stat 897 DAA 12

Penn State

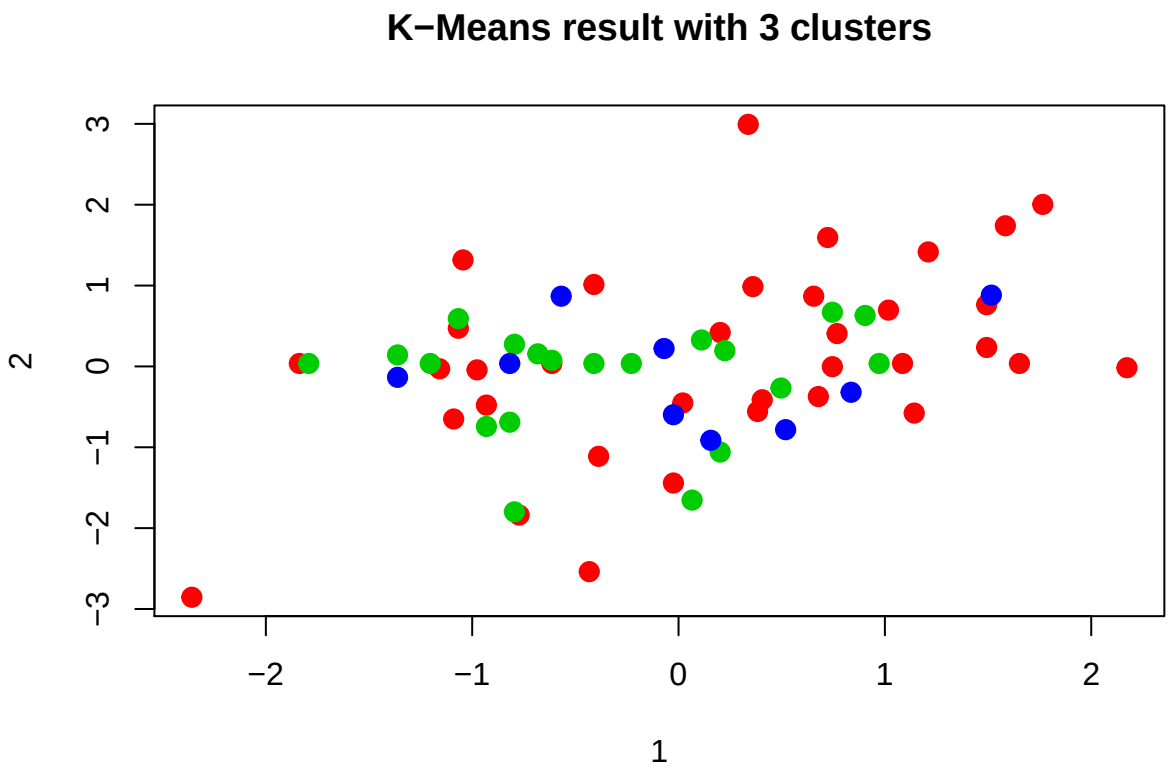
December 3, 2017

In this assignment we will use the `NCI60` data found in the `ISLR` library.

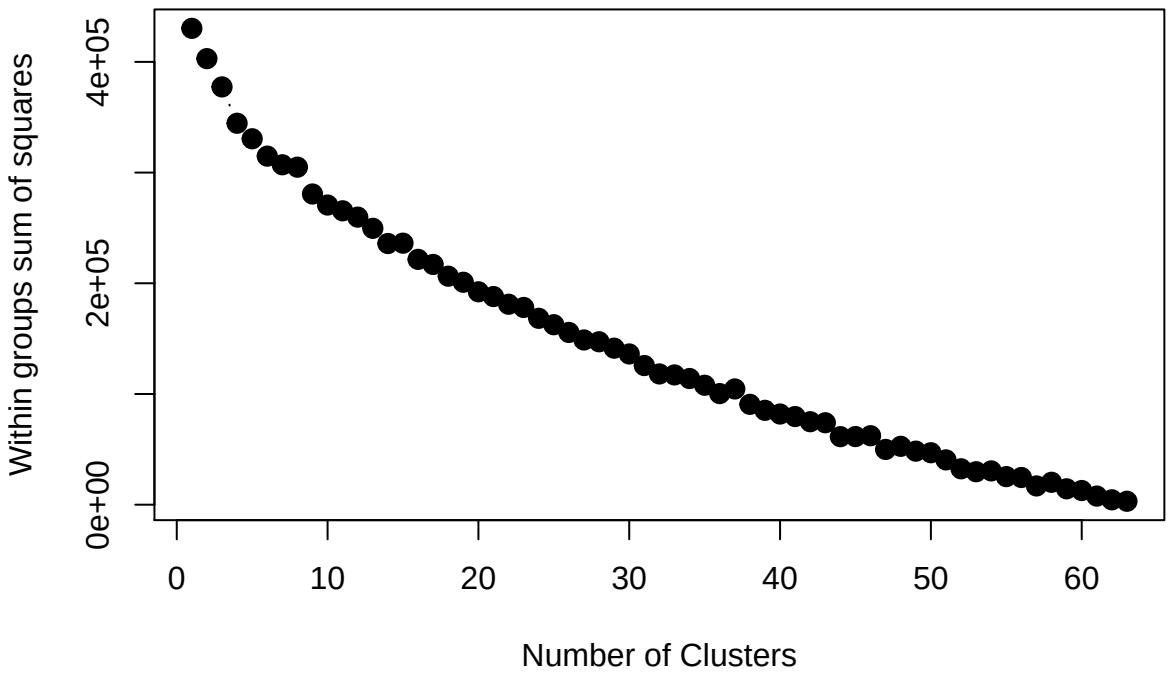
```
library(ISLR)
library(fpc)

nci.labs = NCI60$labs  #labels - for checking later
nci.data = NCI60$data

sd.data=scale(nci.data)
# euclidean distance
data.dist=dist(sd.data)
```

(a) Run k-means clustering on the data using k = 3. Next, use the elbow method to choose an optimal number of clusters (based on total within sums of squares). Is there a clear choice? What is a potential way to choose the optimal k when the elbow is visually ambiguous? (Note: this is an open-ended question. I'm not looking for a specific answer, but for you to use your intuition.)

```
set.seed(5)
km.3.orig = kmeans(sd.data, 3, nstart=100)

# Plot results
plot(sd.data, col =(km.3.orig$cluster +1) , main="K-Means result with 3 clusters", pch=20,
     cex=2)
```




```
nc = nrow(nci.data)-1
wss <- rep(0,nc)

set.seed(5)
for (i in 1:nc) wss[i] <- sum(kmeans(sd.data,
                                     centers=i)$withinss)
plot(1:nc, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares",
     pch=20, cex=2)
```

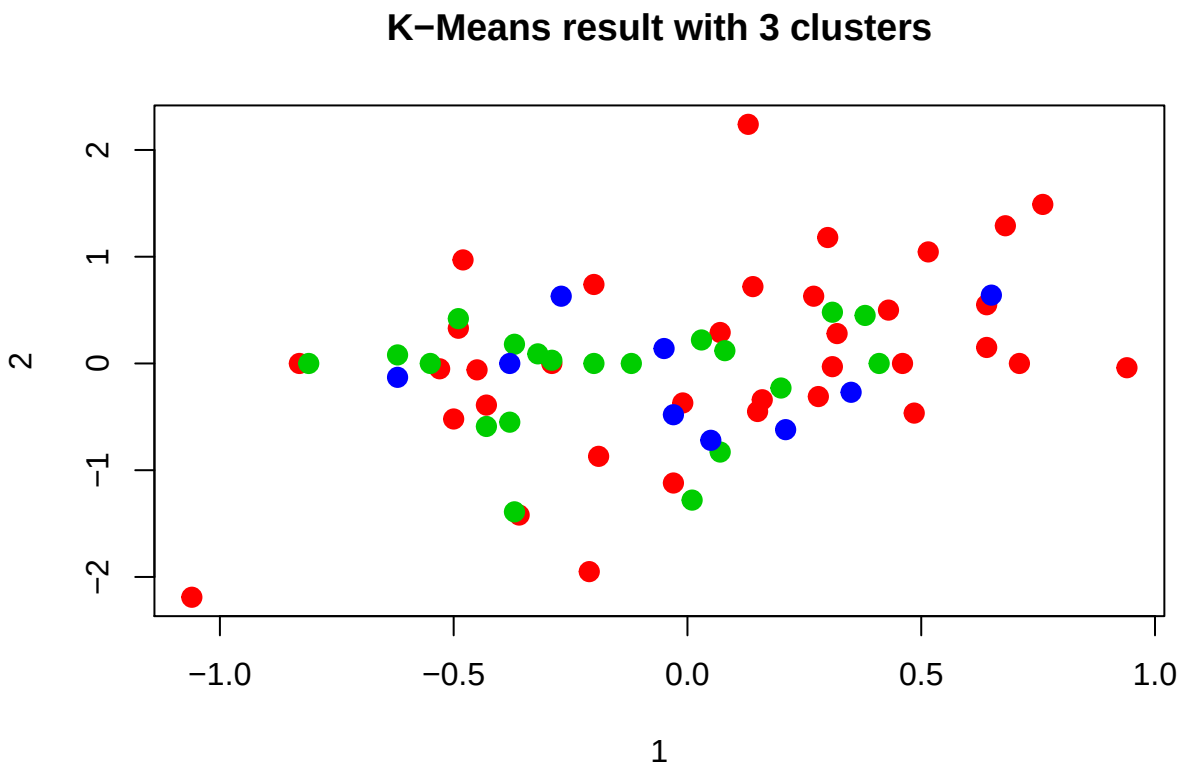


There is no clear choice and this speaks to the problem of the elbow/scree plots as it relates to finding a relatively sharp break/turn - this requires subjectivity and ambiguity, especially where there are either no clear breaks or two or more apparent breaks. In our plot we see a constant descent and there is no clear elbow formation. Maybe in this case we can rely on a relatively higher jump in parameter reduction before it starts to decrease at a more or less similar rate. For instance in our plot we see that there is a jump in the reduction of wss as we go from 8 to 9 clusters. After this the reduction is at a more or less similar rate.

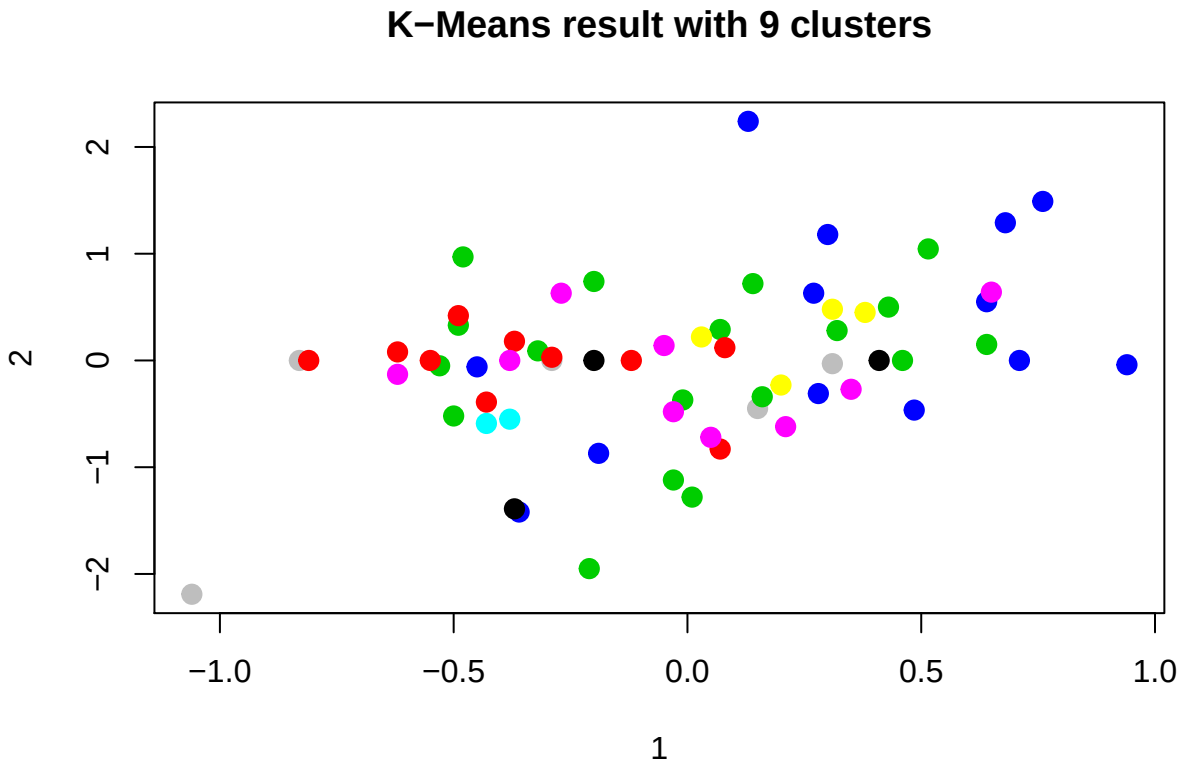
Let's go with a preferred selection of 9 clusters for the subsequent steps.

(b) Tabulate the clusters for k = 3 against the clusters using your optimal k. What do you observe?

```
set.seed(5)
km.9.orig = kmeans(sd.data, 9, nstart=100)

# Plot results
plot(nci.data, col =(km.3.orig$cluster +1) , main="K-Means result with 3 clusters",
     pch=20, cex=2)
```




```
plot(nci.data, col =(km.9.orig$cluster +1) , main="K-Means result with 9 clusters",
     pch=20, cex=2)
```




```
table(km.3.orig$cluster, km.9.orig$cluster)
```

```
##
##      1  2  3  4  5  6  7  8  9
##   1  0 16 13  0  0  0  5  0  1
##   2  3  2  0  2  0  4  0  3  6
##   3  0  0  0  0  9  0  0  0  0
```

```
table(km.3.orig$cluster, nci.labs)
```

```
##    nci.labs
##     BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##   1      3   5     0           0           0        0           0
##   2      2   0     7           1           1        6           1
##   3      2   0     0           0           0        0           0
##    nci.labs
##     MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##   1           0        1     8       6        2     9       1
##   2           1        0     1       0        0     0       0
##   3           0        7     0       0        0     0       0
```

```
table(km.9.orig$cluster, nci.labs)
```

```
##    nci.labs
##     BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##   1      0   0     0           0           0        3           0
##   2      0   0     1           0           0        0           0
##   3      2   3     0           0           0        0           0
```

```
##   4      0   0     0           0           0        2           0
##   5      2   0     0           0           0        0           0
##   6      2   0     0           0           0        0           1
##   7      1   2     0           0           0        0           0
##   8      0   0     0           1           1        1           0
##   9      0   0     6           0           0        0           0
##    nci.labs
##     MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##   1           0        0     0       0        0     0       0
##   2           0        1     5       5        2     3       1
##   3           0        0     2       1        0     5       0
##   4           0        0     0       0        0     0       0
##   5           0        7     0       0        0     0       0
##   6           1        0     0       0        0     0       0
##   7           0        0     1       0        0     1       0
##   8           0        0     0       0        0     0       0
##   9           0        0     1       0        0     0       0
```

```
sapply(list(kmeans_3 = km.3.orig$cluster, kmeans_9 = km.9.orig$cluster),
    function(c) cluster.stats(data.dist, c)[c("within.cluster.ss")])
```

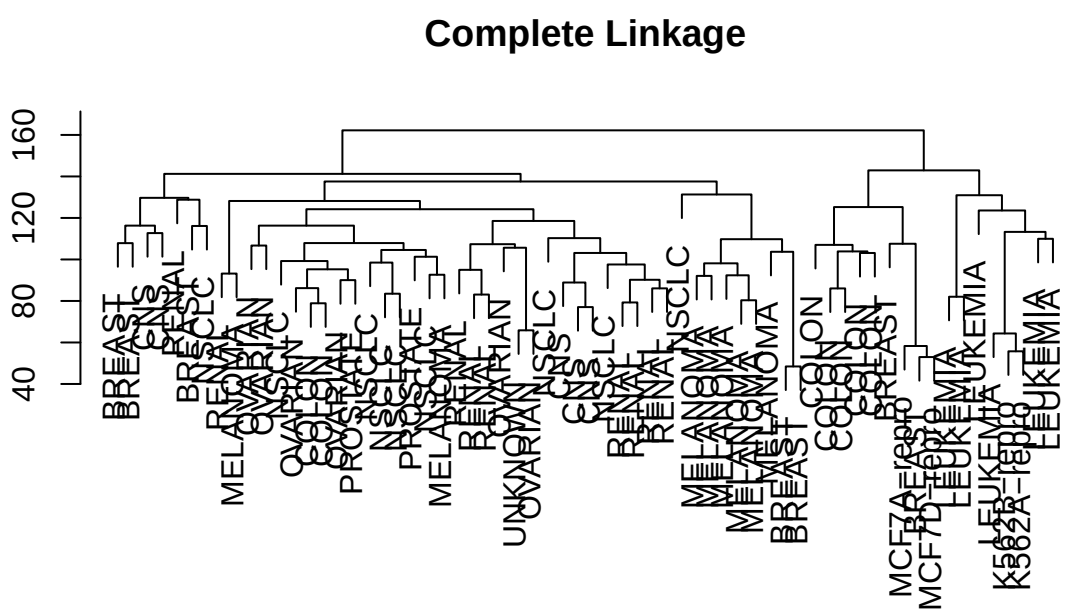
```
## $kmeans_3.within.cluster.ss
## [1] 366350.6
##
## $kmeans_9.within.cluster.ss
## [1] 279306.9
```

We see the following results - Cluster 3 of 3-cluster and cluster 5 of 9-cluster match (2 Breast, 7 Melanoma)

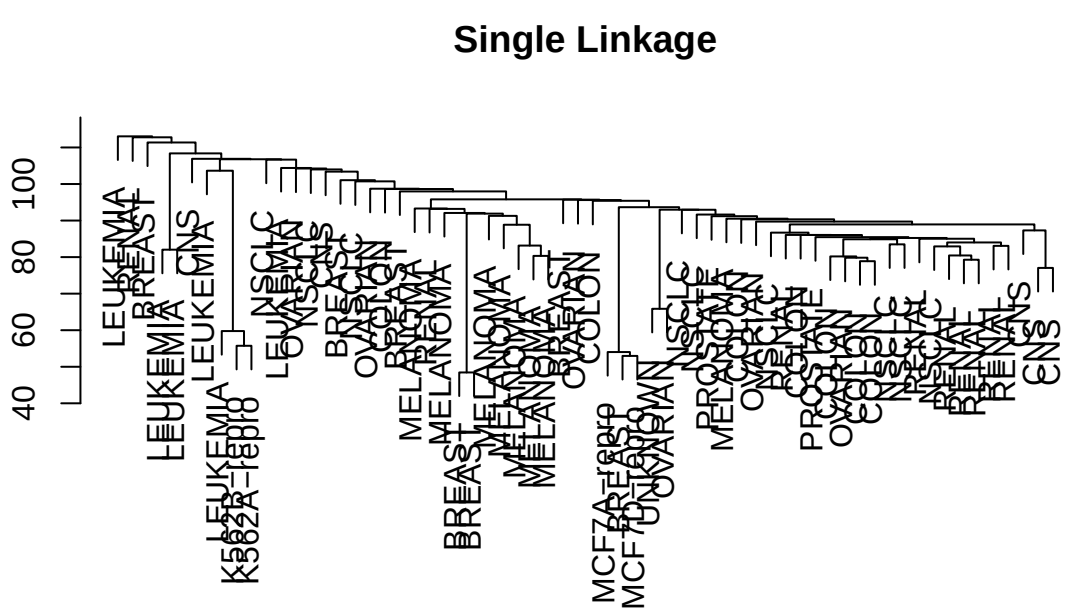
- 3-cluster has most of its elements in cluster 1 (35 out of 64). 9-cluster is more balanced.
- None of the others cluster match. Cluster 1 and 2 in 3-cluster has 35 and 20 elements. The second 9-cluster breaks those into many new clusters
- 3-cluster configuration withinss: 201441.0 127760.0 37149.6 Total withinss: 366350.6
- 9-cluster configuration withinss: 13817.715 87504.903 64637.753 3364.531 37149.601 10326.094 28944.928 3605.765 29955.611 Total withinss: 279306.9
- 9-cluster has a much less withinss than the 3-cluster

(c) Now perform hierarchical clustering using both single and complete clustering. Plot the dendograms.

```
hc.complete.orig=hclust(data.dist, method="complete")
hc.single.orig=hclust(data.dist, method="single")

plot(hc.complete.orig, labels = nci.labs, main = "Complete Linkage", xlab = "",
    sub = "", ylab = "")
```




```
plot(hc.single.orig, labels = nci.labs, main = "Single Linkage",
    xlab = "", sub = "", ylab = "")
```



We see that the complete linkage is giving us a more balanced cluster. Single linkage as expected tends to yield extended clusters to which single leaves are fused one by one.

(d) Cut the trees to obtain the number of clusters you found optimal for kmeans. Tabulate the clusters for both single and complete versus the kmeans clusters. What do you observe? Based on the dendograms, does cutting the trees at this point make sense?

```
hc.complete.orig.clusters = cutree(hc.complete.orig, 9)
hc.single.orig.clusters =cutree(hc.single.orig, 9)

table(hc.complete.orig.clusters, km.9.orig$cluster)
```

```
##
## hc.complete.orig.clusters  1  2  3  4  5  6  7  8  9
##                         1  0 17 11  0  1  0  0  0  2
##                         2  0  0  2  0  0  0  2  0  0
##                         3  0  0  0  0  0  0  2  0  0
##                         4  0  0  0  0  0  0  1  0  0
##                         5  3  0  0  0  0  0  0  3  0
##                         6  0  0  0  2  0  0  0  0  0
##                         7  0  0  0  0  0  4  0  0  5
##                         8  0  1  0  0  0  0  0  0  0
##                         9  0  0  0  0  8  0  0  0  0
```

```
table(hc.single.orig.clusters, km.9.orig$cluster)
```

```
##
## hc.single.orig.clusters  1  2  3  4  5  6  7  8  9
##                       1  0 17 13  0  9  4  2  0  7
##                       2  0  0  0  0  0  0  1  0  0
##                       3  0  0  0  0  0  0  1  0  0
##                       4  0  0  0  0  0  0  1  0  0
##                       5  1  0  0  0  0  0  0  0  0
##                       6  1  0  0  0  0  0  0  3  0
##                       7  0  0  0  2  0  0  0  0  0
##                       8  1  0  0  0  0  0  0  0  0
##                       9  0  1  0  0  0  0  0  0  0
```

```
table(hc.complete.orig.clusters, nci.labs)
```

```
##                           nci.labs
## hc.complete.orig.clusters BREAST CNS COLON K562A-repro K562B-repro
##                         1      0   3     2           0           0
##                         2      2   2     0           0           0
##                         3      1   0     0           0           0
##                         4      0   0     0           0           0
##                         5      0   0     0           1           1
##                         6      0   0     0           0           0
##                         7      2   0     5           0           0
##                         8      0   0     0           0           0
##                         9      2   0     0           0           0
##                           nci.labs
## hc.complete.orig.clusters LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA NSCLC
##                         1        0           0           0        2     7
##                         2        0           0           0        0     0
##                         3        0           0           0        0     1
##                         4        0           0           0        0     0
##                         5        4           0           0        0     0
##                         6        2           0           0        0     0
##                         7        0           1           1        0     0
##                         8        0           0           0        0     1
##                         9        0           0           0        6     0
##                           nci.labs
## hc.complete.orig.clusters OVARIAN PROSTATE RENAL UNKNOWN
##                         1       6        2     8       1
##                         2       0        0     0       0
##                         3       0        0     0       0
##                         4       0        0     1       0
##                         5       0        0     0       0
##                         6       0        0     0       0
##                         7       0        0     0       0
##                         8       0        0     0       0
##                         9       0        0     0       0
```

```
table(hc.single.orig.clusters, nci.labs)
```

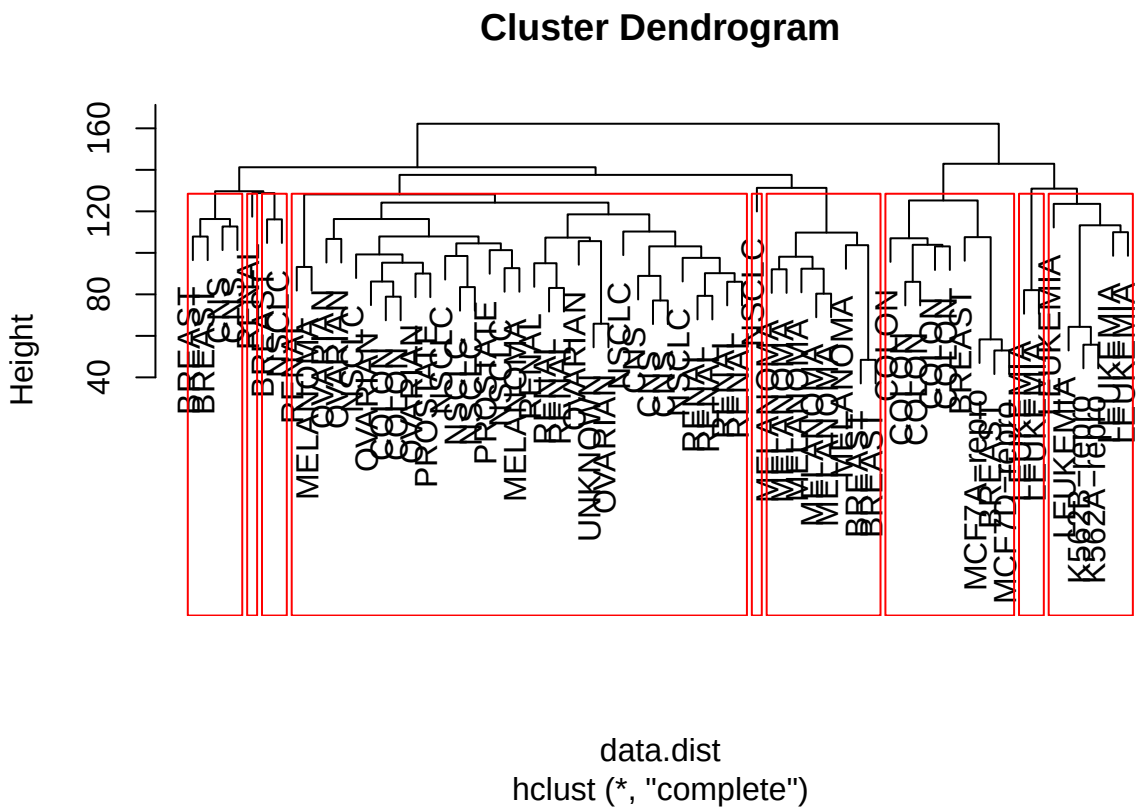
```
##                         nci.labs
## hc.single.orig.clusters BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA
##                       1      6   4     7           0           0        0
```

```
##                       2      0   1     0           0           0        0
##                       3      1   0     0           0           0        0
##                       4      0   0     0           0           0        0
##                       5      0   0     0           0           0        1
##                       6      0   0     0           1           1        2
##                       7      0   0     0           0           0        2
##                       8      0   0     0           0           0        1
##                       9      0   0     0           0           0        0
##                        nci.labs
## hc.single.orig.clusters MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN
##                       1           1           1        8     8       6
##                       2           0           0        0     0       0
##                       3           0           0        0     0       0
##                       4           0           0        0     0       0
##                       5           0           0        0     0       0
##                       6           0           0        0     0       0
##                       7           0           0        0     0       0
##                       8           0           0        0     0       0
##                       9           0           0        0     1       0
##                        nci.labs
## hc.single.orig.clusters PROSTATE RENAL UNKNOWN
##                       1        2     8       1
##                       2        0     0       0
##                       3        0     0       0
##                       4        0     1       0
##                       5        0     0       0
##                       6        0     0       0
##                       7        0     0       0
##                       8        0     0       0
##                       9        0     0       0
```

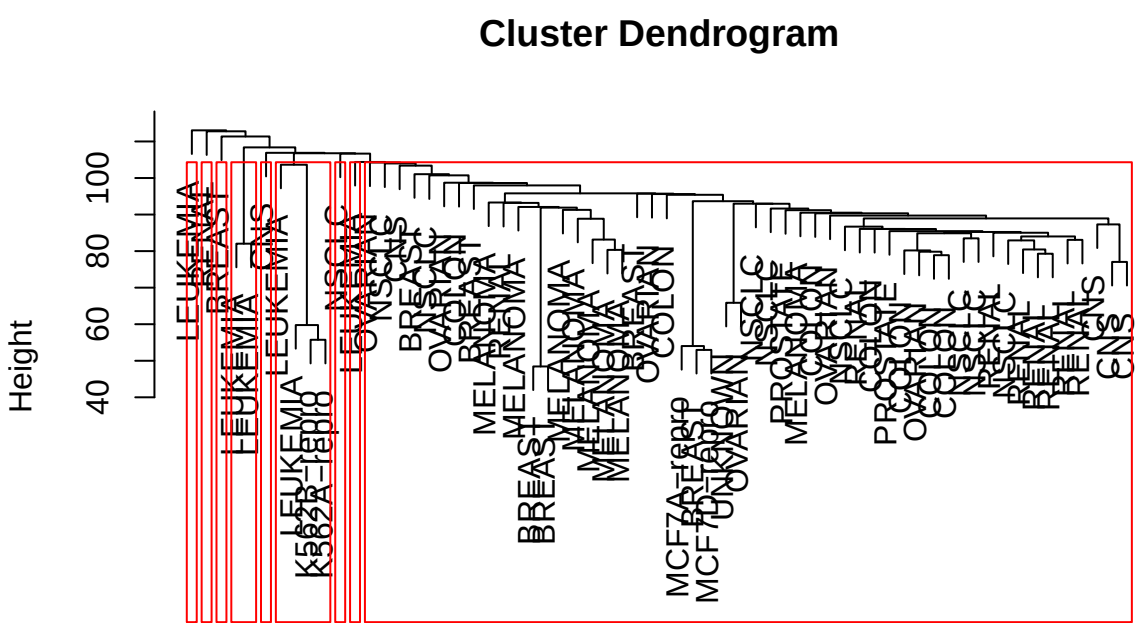
```
table(km.9.orig$cluster, nci.labs)
```

```
##    nci.labs
##     BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##   1      0   0     0           0           0        3           0
##   2      0   0     1           0           0        0           0
##   3      2   3     0           0           0        0           0
##   4      0   0     0           0           0        2           0
##   5      2   0     0           0           0        0           0
##   6      2   0     0           0           0        0           1
##   7      1   2     0           0           0        0           0
##   8      0   0     0           1           1        1           0
##   9      0   0     6           0           0        0           0
##    nci.labs
##     MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##   1           0        0     0       0        0     0       0
##   2           0        1     5       5        2     3       1
##   3           0        0     2       1        0     5       0
##   4           0        0     0       0        0     0       0
##   5           0        7     0       0        0     0       0
##   6           1        0     0       0        0     0       0
##   7           0        0     1       0        0     1       0
##   8           0        0     0       0        0     0       0
##   9           0        0     1       0        0     0       0
```

```
plot(hc.complete.orig, labels = nci.labs)
rect.hclust(hc.complete.orig, k=9, border="red")
```




```
plot(hc.single.orig, labels = nci.labs)
rect.hclust(hc.single.orig, k=9, border="red")
```




```
sapply(list(kmeans_9 = km.9.orig$cluster, hc_single_9 = hc.single.orig.clusters,
            hc_complete_9 = hc.complete.orig.clusters),
    function(c) cluster.stats(data.dist, c)[c("within.cluster.ss")])
```

```
## $kmeans_9.within.cluster.ss
## [1] 279306.9
##
## $hc_single_9.within.cluster.ss
## [1] 331462.8
##
## $hc_complete_9.within.cluster.ss
## [1] 295426.8
```

Compare 9 cut hierarchical cluster (complete) and k-means 9-cluster - 9-hc-complete has most of the data elements in cluster 1 (31 / 64) - Cluster results don't match with the k-means 9-cluster output

Compare 9 cut hierarchical cluster (single) and k-means 9-cluster - 9-hc-single has most of the data elements in cluster 1 (52/64) - The clusters dont match - The 9-hc-single is very unbalanced.

In summary for both the hierarchical clusters we see that one cluster has most of the elements. This doesn't seem to be an appropriate cut. The problem is more severe when using single linkage. The k-means within cluster ss is the lowest.

(e) Repeat parts (c) and (d) using a different distance measure (than euclidean). Give a reason for your choice. What differences (if any) do you see when you tabulate the results?

We will try Correlation-based distance. This might work better because of the following rationale: 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. In this use-case when correlation-based distance is used, then genes with similar values (e.g. gene A and B has same value for Leukemia, Renal cancer) will be clustered together. Therefore, for this application, correlation-based distance may be a better choice.

```
# Though part c doesn't have k-means,
# doing this here to see if anything changes with k-means
data.dist.cor=as.dist(1- cor(t(nci.data)))

set.seed(5)
km.9.cor.orig = kmeans(data.dist.cor, 9, nstart=100)
table(km.9.orig$cluster, km.9.cor.orig$cluster)
```

```
##
##    1 2 3 4 5 6 7 8 9
##  1 0 0 0 0 0 3 0 0 0
##  2 3 0 0 4 0 0 8 1 2
##  3 4 0 0 1 6 0 2 0 0
##  4 0 0 0 0 0 2 0 0 0
##  5 0 0 9 0 0 0 0 0 0
##  6 0 4 0 0 0 0 0 0 0
##  7 0 0 0 2 3 0 0 0 0
##  8 0 0 0 0 0 3 0 0 0
##  9 0 0 0 0 0 0 1 6 0
```

```
table(km.9.orig$cluster, nci.labs)
```

```
##    nci.labs
##     BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##  1      0   0     0           0           0        3           0
##  2      0   0     1           0           0        0           0
##  3      2   3     0           0           0        0           0
##  4      0   0     0           0           0        2           0
##  5      2   0     0           0           0        0           0
##  6      2   0     0           0           0        0           1
##  7      1   2     0           0           0        0           0
##  8      0   0     0           1           1        1           0
##  9      0   0     6           0           0        0           0
##    nci.labs
##     MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##  1           0        0     0       0        0     0       0
##  2           0        1     5       5        2     3       1
##  3           0        0     2       1        0     5       0
##  4           0        0     0       0        0     0       0
##  5           0        7     0       0        0     0       0
##  6           1        0     0       0        0     0       0
##  7           0        0     1       0        0     1       0
##  8           0        0     0       0        0     0       0
##  9           0        0     1       0        0     0       0
```

```
table(km.9.cor.orig$cluster, nci.labs)
```

```
##    nci.labs
##     BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##   1      0   0     0           0           0        0           0
##   2      2   0     0           0           0        0           1
##   3      2   0     0           0           0        0           0
##   4      1   0     0           0           0        0           0
##   5      2   5     0           0           0        0           0
##   6      0   0     0           1           1        6           0
##   7      0   0     0           0           0        0           0
##   8      0   0     7           0           0        0           0
##   9      0   0     0           0           0        0           0
##    nci.labs
##     MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##   1           0        0     0       0        0     7       0
##   2           1        0     0       0        0     0       0
##   3           0        7     0       0        0     0       0
##   4           0        1     2       1        0     1       1
##   5           0        0     1       0        0     1       0
##   6           0        0     0       0        0     0       0
##   7           0        0     4       5        2     0       0
##   8           0        0     0       0        0     0       0
##   9           0        0     2       0        0     0       0
```

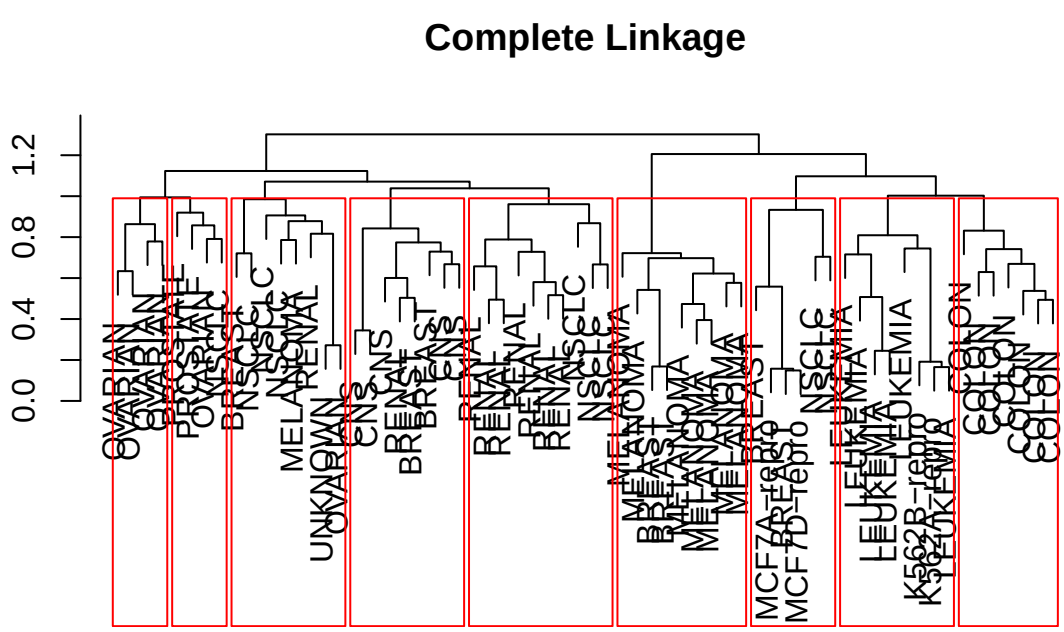
```
km.9.cor.orig$tot.withinss
```

```
## [1] 41.91002
```

Observations: - The clusters are different (few similarities) but the distributions seem to be similar. - All cases of leukemia fall in a single cluster when using corelation based distance. In the case of euclidean that is not the case. Same applies to CNS, COLON. - Overall since we see higher proportion of one type of cancers falling into a single cluster, we can say that we see slightly better performance when using corelation based distance in the k-means procedure. - There is a huge reduction in the total within ss - this is desirable

Hierarchical clustering:

```
set.seed(5)
hc.complete.cor.orig=hclust(data.dist.cor, method="complete")
hc.single.cor.orig=hclust(data.dist.cor, method="single")

plot(hc.complete.cor.orig, labels = nci.labs, main = "Complete Linkage", xlab = "",
    sub = "", ylab = "")
rect.hclust(hc.complete.cor.orig, k=9, border="red")
```




```
plot(hc.single.cor.orig, labels = nci.labs, main = "Single Linkage",
    xlab = "", sub = "", ylab = "")
rect.hclust(hc.single.cor.orig, k=9, border="red")
```

Single Linkage


```
hc.complete.cor.orig.clusters = cutree(hc.complete.cor.orig, 9)
hc.single.cor.orig.clusters =cutree(hc.single.cor.orig, 9)

table(hc.complete.cor.orig.clusters, km.9.cor.orig$cluster)
```

```
##
## hc.complete.cor.orig.clusters 1 2 3 4 5 6 7 8 9
##                             1 0 0 0 0 8 0 0 0 0
##                             2 0 0 0 7 1 0 0 0 0
##                             3 7 0 0 0 0 0 3 0 0
##                             4 0 0 0 0 0 0 0 4 0
##                             5 0 0 0 0 0 0 0 4 0
##                             6 0 0 0 0 0 0 8 0 0
##                             7 0 0 0 0 0 0 0 0 7
##                             8 0 4 0 0 0 0 0 0 2
##                             9 0 0 9 0 0 0 0 0 0
```

```
table(hc.single.cor.orig.clusters, km.9.cor.orig$cluster)
```

```
##
## hc.single.cor.orig.clusters 1 2 3 4 5 6 7 8 9
##                           1 7 0 9 6 9 0 3 0 0
##                           2 0 0 0 1 0 0 0 0 0
##                           3 0 0 0 0 0 0 1 0 0
##                           4 0 0 0 0 0 0 3 0 0
##                           5 0 0 0 0 0 0 1 0 0
##                           6 0 0 0 0 0 0 1 0 0
```

```
##                           7 0 0 0 0 0 0 1 0 0
##                           8 0 4 0 0 0 8 1 7 0
##                           9 0 0 0 0 0 0 0 0 2
```

```
table(hc.complete.cor.orig.clusters, nci.labs)
```

```
##                              nci.labs
## hc.complete.cor.orig.clusters BREAST CNS COLON K562A-repro K562B-repro
##                             1      2   5     0           0           0
##                             2      1   0     0           0           0
##                             3      0   0     0           0           0
##                             4      0   0     0           0           0
##                             5      0   0     0           0           0
##                             6      0   0     0           1           1
##                             7      0   0     7           0           0
##                             8      2   0     0           0           0
##                             9      2   0     0           0           0
##                              nci.labs
## hc.complete.cor.orig.clusters LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA
##                             1        0           0           0        0
##                             2        0           0           0        1
##                             3        0           0           0        0
##                             4        0           0           0        0
##                             5        0           0           0        0
##                             6        6           0           0        0
##                             7        0           0           0        0
##                             8        0           1           1        0
##                             9        0           0           0        7
##                              nci.labs
## hc.complete.cor.orig.clusters NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##                             1     0       0        0     1       0
##                             2     3       1        0     1       1
##                             3     3       0        0     7       0
##                             4     1       1        2     0       0
##                             5     0       4        0     0       0
##                             6     0       0        0     0       0
##                             7     0       0        0     0       0
##                             8     2       0        0     0       0
##                             9     0       0        0     0       0
```

```
table(hc.single.cor.orig.clusters, nci.labs)
```

```
##                              nci.labs
## hc.single.cor.orig.clusters BREAST CNS COLON K562A-repro K562B-repro
##                            1      5   5     0           0           0
##                            2      0   0     0           0           0
##                            3      0   0     0           0           0
##                            4      0   0     0           0           0
##                            5      0   0     0           0           0
##                            6      0   0     0           0           0
##                            7      0   0     0           0           0
##                            8      2   0     7           1           1
##                            9      0   0     0           0           0
##                             nci.labs
## hc.single.cor.orig.clusters LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA
```

```
##                             1        0           0           0        7
##                             2        0           0           0        1
##                             3        0           0           0        0
##                             4        0           0           0        0
##                             5        0           0           0        0
##                             6        0           0           0        0
##                             7        0           0           0        0
##                             8        6           1           1        0
##                             9        0           0           0        0
##                              nci.labs
## hc.single.cor.orig.clusters NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##                             1     6       1        0     9       1
##                             2     0       0        0     0       0
##                             3     0       0        1     0       0
##                             4     0       3        0     0       0
##                             5     0       1        0     0       0
##                             6     0       1        0     0       0
##                             7     0       0        1     0       0
##                             8     1       0        0     0       0
##                             9     2       0        0     0       0
```

```
sapply(list(kmeans_9_cor = km.9.cor.orig$cluster,
            hc_single_9_cor = hc.single.cor.orig.clusters,
            hc_complete_9_cor = hc.complete.cor.orig.clusters),
    function(c) cluster.stats(data.dist.cor, c)[c("within.cluster.ss")])
```

```
## $kmeans_9_cor.within.cluster.ss
## [1] 13.9959
##
## $hc_single_9_cor.within.cluster.ss
## [1] 21.05816
##
## $hc_complete_9_cor.within.cluster.ss
## [1] 14.33262
```

We make the following observations: - Complete linkage provides a much more balanced tree. - Single linkage is not balanced and tends to yield extended clusters to which single leaves are fused one by one. - Single linkage leads to a tree that has almost all the data in 2 clusters and remaining clusters are scarcely populated. - In summary the 9 size tree seems relatively more appropriate for the complete linkage tree. The single linkage leads to an inappropriate tree. - Lowest within cluster ss is for kmeans followed by complete linkage.

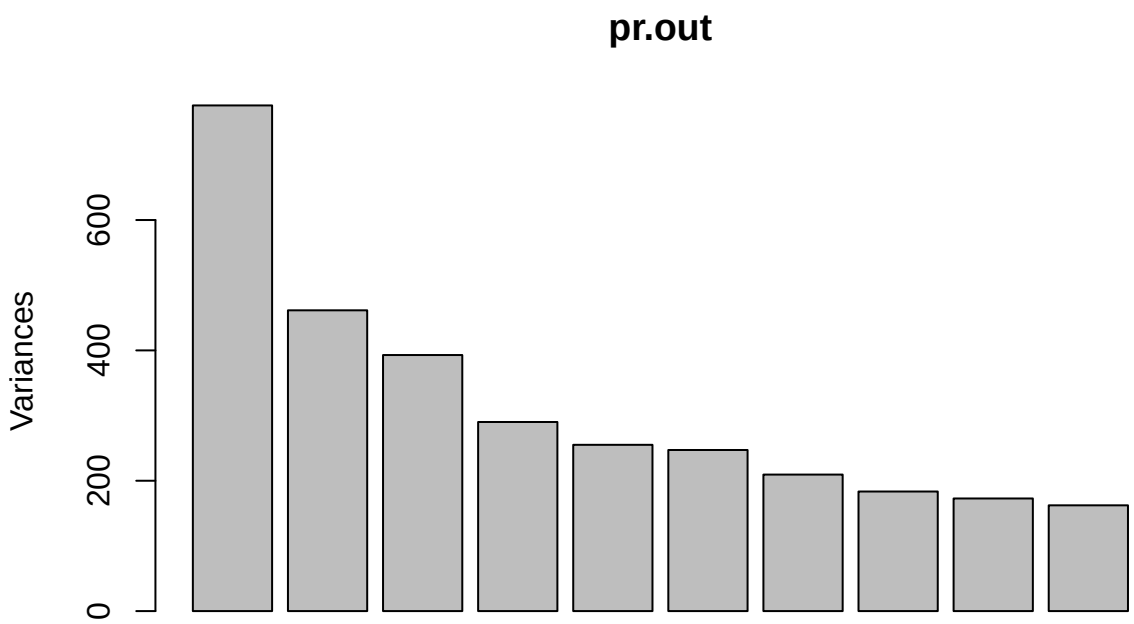
(f) Using PCA, pull out a number of principal components for the NCI60 data. Explain your choice of number of PCs.

```
pr.out = prcomp(nci.data, scale = T)
summary(pr.out)
```

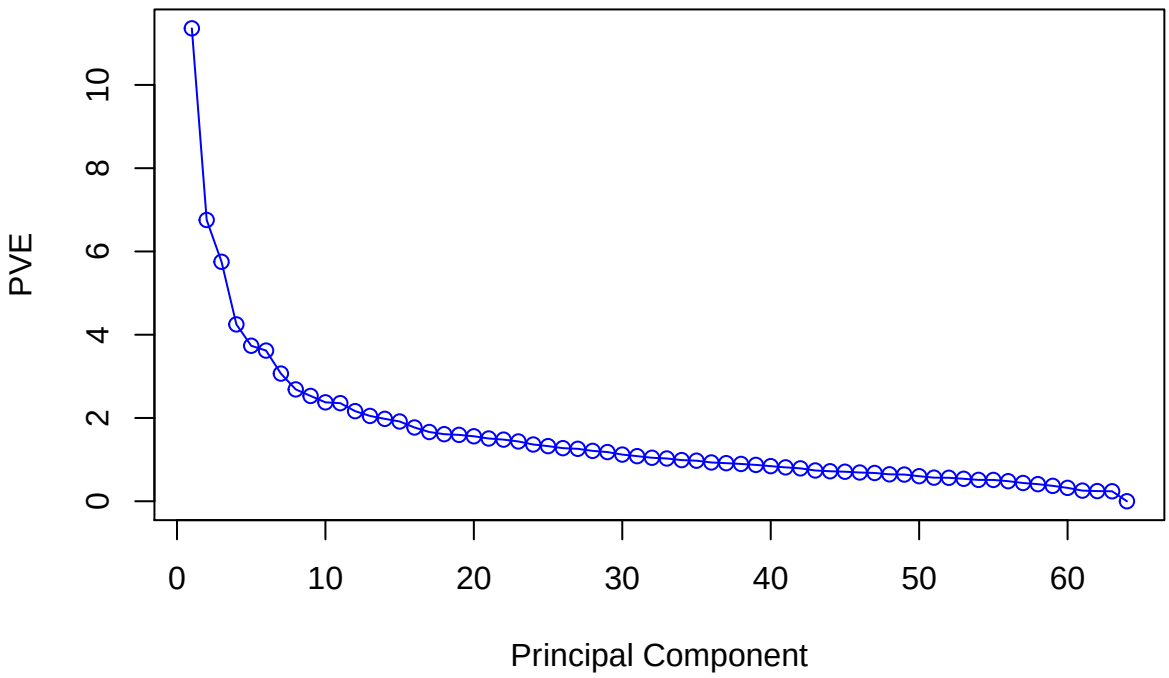
```
## Importance of components:
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     27.8535 21.48136 19.82046 17.03256 15.97181
## Proportion of Variance  0.1136  0.06756  0.05752  0.04248  0.03735
## Cumulative Proportion   0.1136  0.18115  0.23867  0.28115  0.31850
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     15.72108 14.47145 13.54427 13.14400 12.73860
```

```
## Proportion of Variance  0.03619  0.03066  0.02686  0.02529  0.02376
## Cumulative Proportion   0.35468  0.38534  0.41220  0.43750  0.46126
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     12.68672 12.15769 11.83019 11.62554 11.43779
## Proportion of Variance  0.02357  0.02164  0.02049  0.01979  0.01915
## Cumulative Proportion   0.48482  0.50646  0.52695  0.54674  0.56590
##                            PC16     PC17     PC18     PC19    PC20
## Standard deviation     11.00051 10.65666 10.48880 10.43518 10.3219
## Proportion of Variance  0.01772  0.01663  0.01611  0.01594  0.0156
## Cumulative Proportion   0.58361  0.60024  0.61635  0.63229  0.6479
##                            PC21    PC22    PC23    PC24    PC25    PC26
## Standard deviation     10.14608 10.0544 9.90265 9.64766 9.50764 9.33253
## Proportion of Variance  0.01507  0.0148 0.01436 0.01363 0.01324 0.01275
## Cumulative Proportion   0.66296  0.6778 0.69212 0.70575 0.71899 0.73174
##                            PC27   PC28    PC29    PC30    PC31    PC32
## Standard deviation     9.27320 9.0900 8.98117 8.75003 8.59962 8.44738
## Proportion of Variance 0.01259 0.0121 0.01181 0.01121 0.01083 0.01045
## Cumulative Proportion  0.74433 0.7564 0.76824 0.77945 0.79027 0.80072
##                            PC33    PC34    PC35    PC36    PC37    PC38
## Standard deviation     8.37305 8.21579 8.15731 7.97465 7.90446 7.82127
## Proportion of Variance 0.01026 0.00988 0.00974 0.00931 0.00915 0.00896
## Cumulative Proportion  0.81099 0.82087 0.83061 0.83992 0.84907 0.85803
##                            PC39    PC40    PC41   PC42    PC43   PC44
## Standard deviation     7.72156 7.58603 7.45619 7.3444 7.10449 7.0131
## Proportion of Variance 0.00873 0.00843 0.00814 0.0079 0.00739 0.0072
## Cumulative Proportion  0.86676 0.87518 0.88332 0.8912 0.89861 0.9058
##                            PC45   PC46    PC47    PC48    PC49    PC50
## Standard deviation     6.95839 6.8663 6.80744 6.64763 6.61607 6.40793
## Proportion of Variance 0.00709 0.0069 0.00678 0.00647 0.00641 0.00601
## Cumulative Proportion  0.91290 0.9198 0.92659 0.93306 0.93947 0.94548
##                            PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     6.21984 6.20326 6.06706 5.91805 5.91233 5.73539
## Proportion of Variance 0.00566 0.00563 0.00539 0.00513 0.00512 0.00482
## Cumulative Proportion  0.95114 0.95678 0.96216 0.96729 0.97241 0.97723
##                            PC57   PC58    PC59    PC60    PC61    PC62
## Standard deviation     5.47261 5.2921 5.02117 4.68398 4.17567 4.08212
## Proportion of Variance 0.00438 0.0041 0.00369 0.00321 0.00255 0.00244
## Cumulative Proportion  0.98161 0.9857 0.98940 0.99262 0.99517 0.99761
##                            PC63      PC64
## Standard deviation     4.04124 2.148e-14
## Proportion of Variance 0.00239 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00
```

```
plot(pr.out)
```




```
pve = 100 * pr.out$sdev^2/sum(pr.out$sdev^2)
plot(pve, type = "o", ylab = "PVE", xlab = "Principal Component", col = "blue")
```



We see an elbow at about 8 components - at that level we have explained 41.2% of variance. This doesn't seem too much but the plot and the table indicates that the remaining components have lower marginal contribution to the overall variance.

(g) Using these PCs, repeat kmeans clustering and hierarchical clustering (with your preferred distance function and both linkage methods). Compare the results to the results for the corresponding method on the original data.

We will use corelation based distance.

K-Means

```
data.dist.pc.cor=as.dist(1- cor(t(pr.out$x[, 1:8])))

set.seed(5)
km.9.cor.pca = kmeans(data.dist.pc.cor, 9, nstart=100)
km.9.cor.pca$tot.withinss
```

```
## [1] 146.7237
```

```
table(km.9.cor.orig$cluster, km.9.cor.pca$cluster)
```

```
##
##    1 2 3 4 5 6 7 8 9
##  1 0 3 4 0 0 0 0 0 0
##  2 0 0 0 0 0 0 0 0 4
##  3 0 0 0 0 0 9 0 0 0
```

```
##   4 0 1 1 3 2 0 0 0 0
##   5 0 6 0 0 3 0 0 0 0
##   6 8 0 0 0 0 0 0 0 0
##   7 0 1 2 0 1 0 6 1 0
##   8 0 0 0 0 0 0 0 7 0
##   9 0 0 0 2 0 0 0 0 0
```

```
table(km.9.cor.orig$cluster, nci.labs)
```

```
##    nci.labs
##     BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##   1      0   0     0           0           0        0           0
##   2      2   0     0           0           0        0           1
##   3      2   0     0           0           0        0           0
##   4      1   0     0           0           0        0           0
##   5      2   5     0           0           0        0           0
##   6      0   0     0           1           1        6           0
##   7      0   0     0           0           0        0           0
##   8      0   0     7           0           0        0           0
##   9      0   0     0           0           0        0           0
##    nci.labs
##     MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##   1           0        0     0       0        0     7       0
##   2           1        0     0       0        0     0       0
##   3           0        7     0       0        0     0       0
##   4           0        1     2       1        0     1       1
##   5           0        0     1       0        0     1       0
##   6           0        0     0       0        0     0       0
##   7           0        0     4       5        2     0       0
##   8           0        0     0       0        0     0       0
##   9           0        0     2       0        0     0       0
```

```
table(km.9.cor.pca$cluster, nci.labs)
```

```
##    nci.labs
##     BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##   1      0   0     0           1           1        6           0
##   2      2   3     0           0           0        0           0
##   3      0   0     0           0           0        0           0
##   4      0   0     0           0           0        0           0
##   5      1   2     0           0           0        0           0
##   6      2   0     0           0           0        0           0
##   7      0   0     0           0           0        0           0
##   8      0   0     7           0           0        0           0
##   9      2   0     0           0           0        0           1
##    nci.labs
##     MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##   1           0        0     0       0        0     0       0
##   2           0        0     2       0        0     4       0
##   3           0        0     2       0        1     4       0
##   4           0        1     2       1        0     0       1
##   5           0        0     2       0        0     1       0
##   6           0        7     0       0        0     0       0
##   7           0        0     0       5        1     0       0
##   8           0        0     1       0        0     0       0
```

```
##   9           1        0     0       0        0     0       0
```

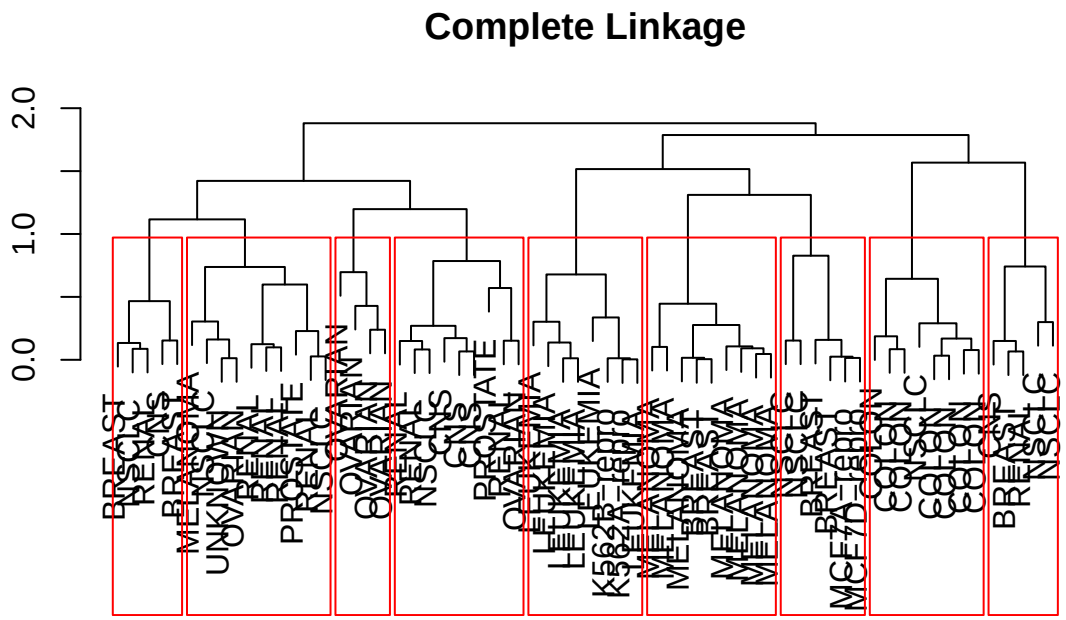
```
km.9.cor.pca$tot.withinss
```

```
## [1] 146.7237
```

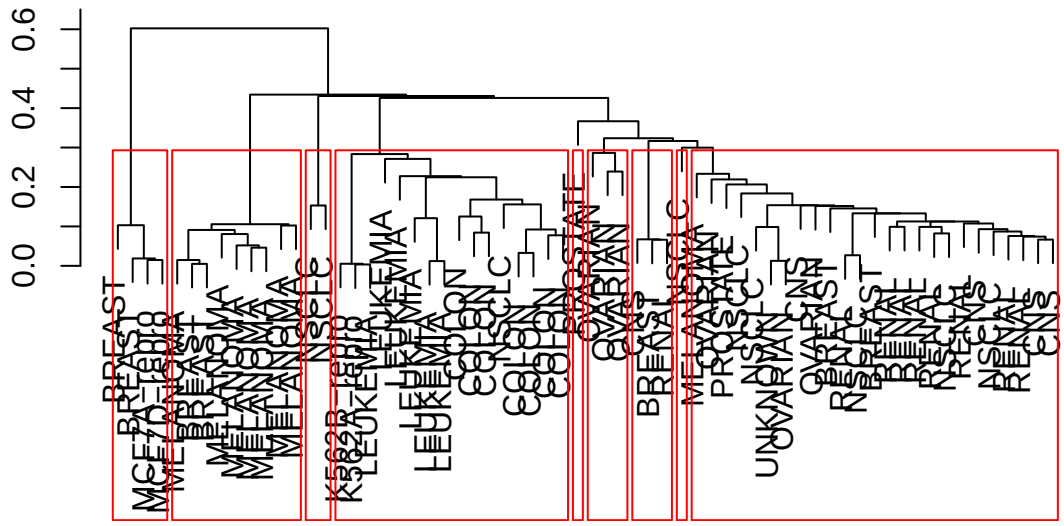
Observations: - 9-kmeans-cor-orig-cluster cluster # 2 is same as 9-kmeans-cor-pca-cluster cluster # 9 (Breast:2, MCF7A-repro:1, MCF7D-repro:1) - Most clusters are different (few similarities) but the distributions seem to be similar. - We see that the total within ss increases from the kmeans cluster we got with all the data when using corelation based distance. However it is much lower than the values that we obtained with euclidean distance and using the original data.

Hierarchical clustering:

```
set.seed(5)
hc.complete.cor.pca=hclust(data.dist.pc.cor, method="complete")
hc.single.cor.pca=hclust(data.dist.pc.cor, method="single")

plot(hc.complete.cor.pca, labels = nci.labs, main = "Complete Linkage", xlab = "",
    sub = "", ylab = "")
rect.hclust(hc.complete.cor.pca, k=9, border="red")
```




```
plot(hc.single.cor.pca, labels = nci.labs, main = "Single Linkage",
    xlab = "", sub = "", ylab = "")
rect.hclust(hc.single.cor.pca, k=9, border="red")
```




```
hc.complete.cor.pca.clusters = cutree(hc.complete.cor.pca, 9)
hc.single.cor.pca.clusters =cutree(hc.single.cor.pca, 9)

table(hc.complete.cor.pca.clusters, km.9.cor.pca$cluster)
```

```
##
## hc.complete.cor.pca.clusters 1 2 3 4 5 6 7 8 9
##                            1 0 7 0 0 0 0 2 0 0
##                            2 0 4 0 0 1 0 0 0 0
##                            3 0 0 0 0 5 0 0 0 0
##                            4 0 0 7 3 0 0 0 0 0
##                            5 0 0 0 0 0 0 0 4 0 0
##                            6 8 0 0 0 0 0 0 0 0 0
##                            7 0 0 0 0 0 0 0 8 0
##                            8 0 0 0 2 0 0 0 0 4
##                            9 0 0 0 0 0 0 9 0 0 0
```

```
table(hc.single.cor.pca.clusters, km.9.cor.pca$cluster)
```

```
##
## hc.single.cor.pca.clusters  1  2  3  4  5  6  7  8  9
##                          1  0 11  7  3  2  0  2  0  0
##                          2  0  0  0  0  3  0  0  0  0
##                          3  0  0  0  0  0  0  1  0  0
##                          4  0  0  0  0  0  0  3  0  0
##                          5  0  0  0  0  1  0  0  0  0
##                          6  8  0  0  0  0  0  0  8  0
```

```
##                          7  0  0  0  0  0  0  0  0  4
##                          8  0  0  0  2  0  0  0  0  0
##                          9  0  0  0  0  0  9  0  0  0
```

```
table(hc.complete.cor.pca.clusters, nci.labs)
```

```
##                            nci.labs
## hc.complete.cor.pca.clusters BREAST CNS COLON K562A-repro K562B-repro
##                            1      0   3     0           0           0
##                            2      2   1     0           0           0
##                            3      1   1     0           0           0
##                            4      0   0     0           0           0
##                            5      0   0     0           0           0
##                            6      0   0     0           1           1
##                            7      0   0     7           0           0
##                            8      2   0     0           0           0
##                            9      2   0     0           0           0
##                            nci.labs
## hc.complete.cor.pca.clusters LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA
##                            1        0           0           0        0
##                            2        0           0           0        0
##                            3        0           0           0        0
##                            4        0           0           0        1
##                            5        0           0           0        0
##                            6        6           0           0        0
##                            7        0           0           0        0
##                            8        0           1           1        0
##                            9        0           0           0        7
##                            nci.labs
## hc.complete.cor.pca.clusters NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##                            1     1       1        1     3       0
##                            2     1       0        0     1       0
##                            3     2       0        0     1       0
##                            4     2       1        1     4       1
##                            5     0       4        0     0       0
##                            6     0       0        0     0       0
##                            7     1       0        0     0       0
##                            8     2       0        0     0       0
##                            9     0       0        0     0       0
```

```
table(hc.single.cor.pca.clusters, nci.labs)
```

```
##                           nci.labs
## hc.single.cor.pca.clusters BREAST CNS COLON K562A-repro K562B-repro
##                           1      2   4     0           0           0
##                           2      1   1     0           0           0
##                           3      0   0     0           0           0
##                           4      0   0     0           0           0
##                           5      0   0     0           0           0
##                           6      0   0     7           1           1
##                           7      2   0     0           0           0
##                           8      0   0     0           0           0
##                           9      2   0     0           0           0
##                            nci.labs
## hc.single.cor.pca.clusters LEUKEMIA MCF7A-repro MCF7D-repro MELANOMA NSCLC
```

```
##                           1        0           0           0        1     5
##                           2        0           0           0        0     0
##                           3        0           0           0        0     0
##                           4        0           0           0        0     0
##                           5        0           0           0        0     1
##                           6        6           0           0        0     1
##                           7        0           1           1        0     0
##                           8        0           0           0        0     2
##                           9        0           0           0        7     0
##                            nci.labs
## hc.single.cor.pca.clusters OVARIAN PROSTATE RENAL UNKNOWN
##                           1       3        1     8       1
##                           2       0        0     1       0
##                           3       0        1     0       0
##                           4       3        0     0       0
##                           5       0        0     0       0
##                           6       0        0     0       0
##                           7       0        0     0       0
##                           8       0        0     0       0
##                           9       0        0     0       0
```

```
table(hc.complete.cor.pca.clusters, hc.complete.cor.orig.clusters)
```

```
##                             hc.complete.cor.orig.clusters
## hc.complete.cor.pca.clusters 1 2 3 4 5 6 7 8 9
##                            1 4 0 3 1 1 0 0 0 0
##                            2 3 1 1 0 0 0 0 0 0
##                            3 1 3 1 0 0 0 0 0 0
##                            4 0 4 5 1 0 0 0 0 0
##                            5 0 0 0 1 3 0 0 0 0
##                            6 0 0 0 0 0 8 0 0 0
##                            7 0 0 0 1 0 0 7 0 0
##                            8 0 0 0 0 0 0 0 6 0
##                            9 0 0 0 0 0 0 0 0 9
```

```
sapply(list(kmeans_9_cor.pca = km.9.cor.pca$cluster,
            hc_single_9_cor_pca = hc.single.cor.pca.clusters,
            hc_complete_9_cor_pca = hc.complete.cor.pca.clusters),
    function(c) cluster.stats(data.dist.pc.cor, c)[c("within.cluster.ss")])
```

```
## $kmeans_9_cor.pca.within.cluster.ss
## [1] 4.230982
##
## $hc_single_9_cor_pca.within.cluster.ss
## [1] 10.37182
##
## $hc_complete_9_cor_pca.within.cluster.ss
## [1] 4.145854
```

We make the following observations: - Complete linkage provides a balanced tree. - Though single linkage is not as balanced as complete linkage, it is much better than the previous iteration of single linkage with original data. So there is improvement. - The complete linkage clusters with original and PCA data (both using corelation based distance) are similar though visually the PCA looks better and also gives lower within cluster ss.

(h) For which method(s) do you observe a substantial change in the clusters based on the original data versus the PC? Which method(s) don't change as much?

The summary of results: - K-Means with 9-clusters, PCA, cor distance (IS ALMOST SAME AS BUT NOT AS GOOD AS) K-Means with 9-clusters, original, cor distance

- K-Means with 9-clusters, PCA, cor distance (IS MUCH BETTER THAN) K-Means with 9-clusters, original, euclidean distance
- Hierchical with complete linkage, 9-clusters, PCA, cor distance (IS ALMOST SAME AS) Hierchical with complete linkage, 9-clusters, original, cor distance
- Hierchical with single linkage, 9-clusters, PCA, cor distance (IS BETTER THAN) Hierchical with single linkage, 9-clusters, original, cor distance