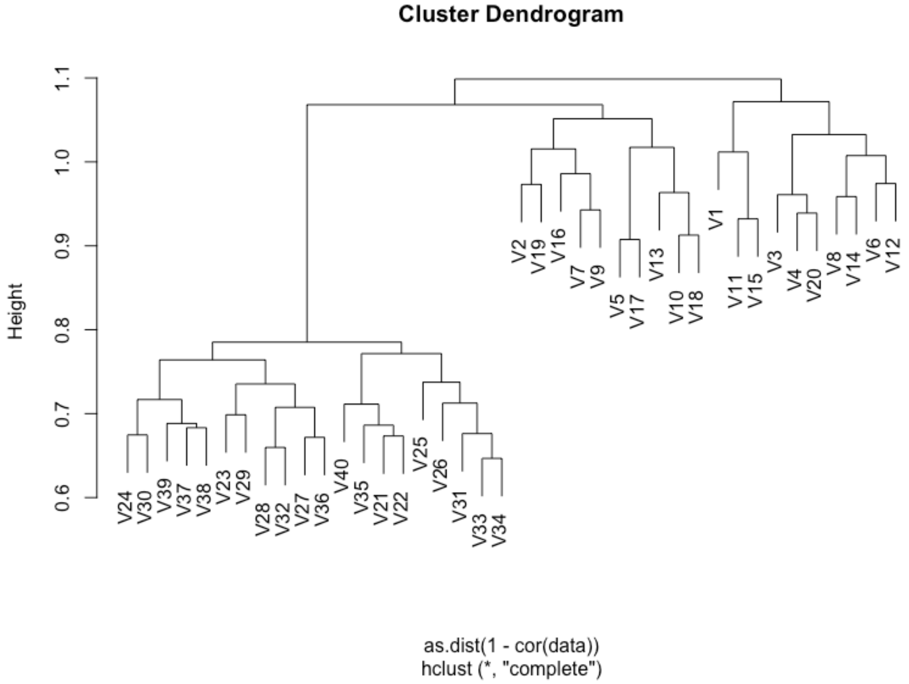
Problem 3:

a) Loaded the data.

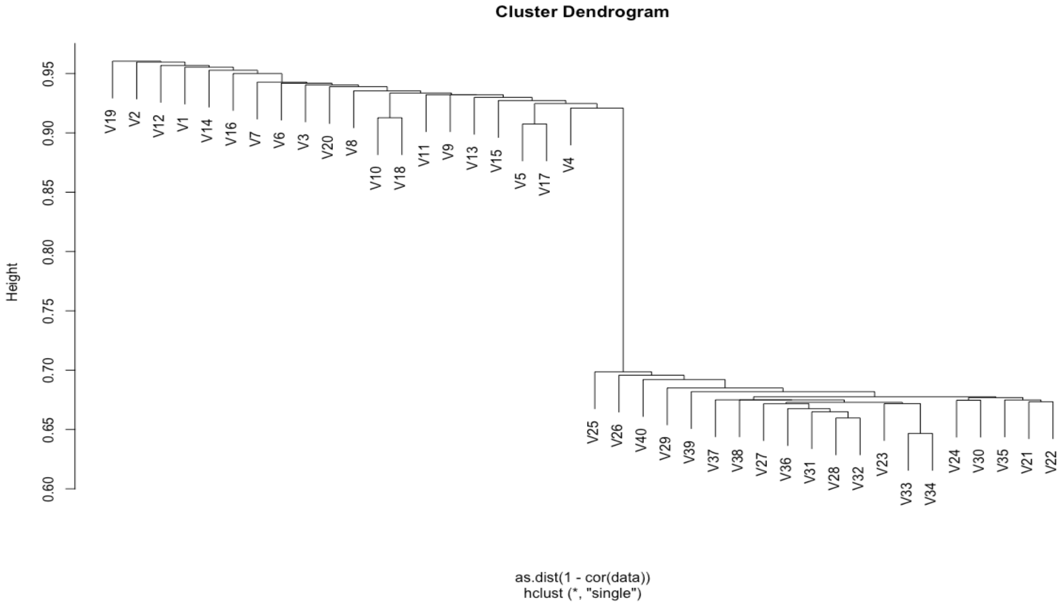
```
> data <- read.csv("~/DS/STA DM2/HW2/Ch10Ex11.csv", header = FALSE)
```

b) Hierarchical clustering to the samples using correlation-based distance:

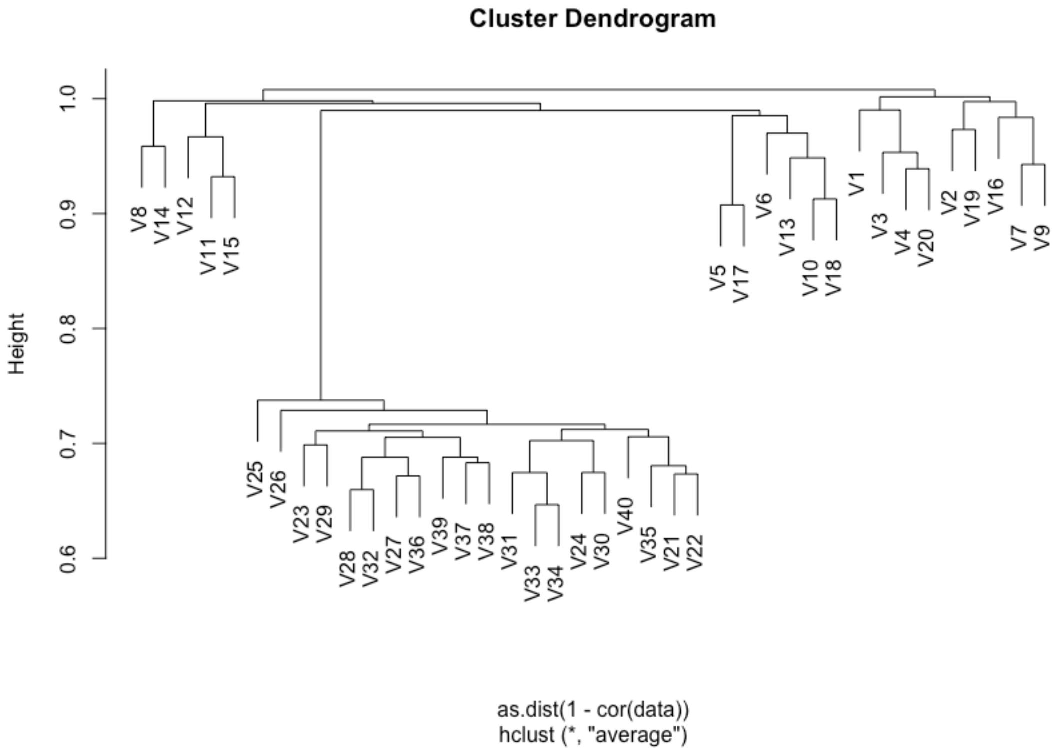
- Complete Linkage Dendrogram:



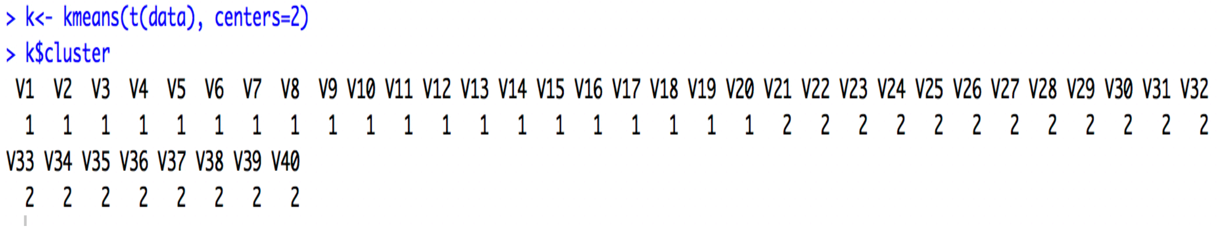
- Single Linkage Dendrogram:



- Average Linkage Dendrogram:



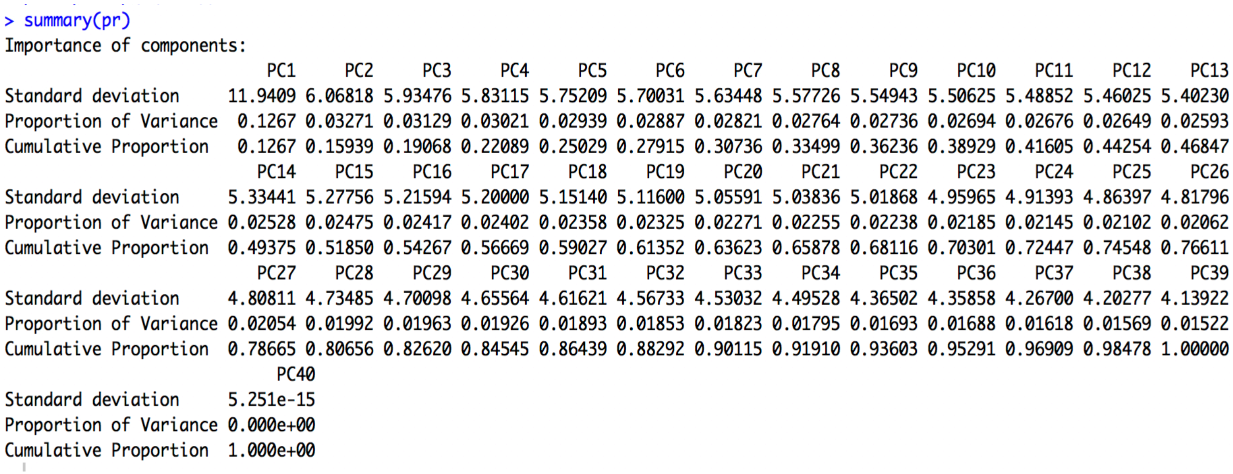
- K-means:



```
> k<- kmeans(t(data), centers=2)
> k$cluster
 V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2   2 
V33 V34 V35 V36 V37 V38 V39 V40 
  2   2   2   2   2   2   2   2 
```

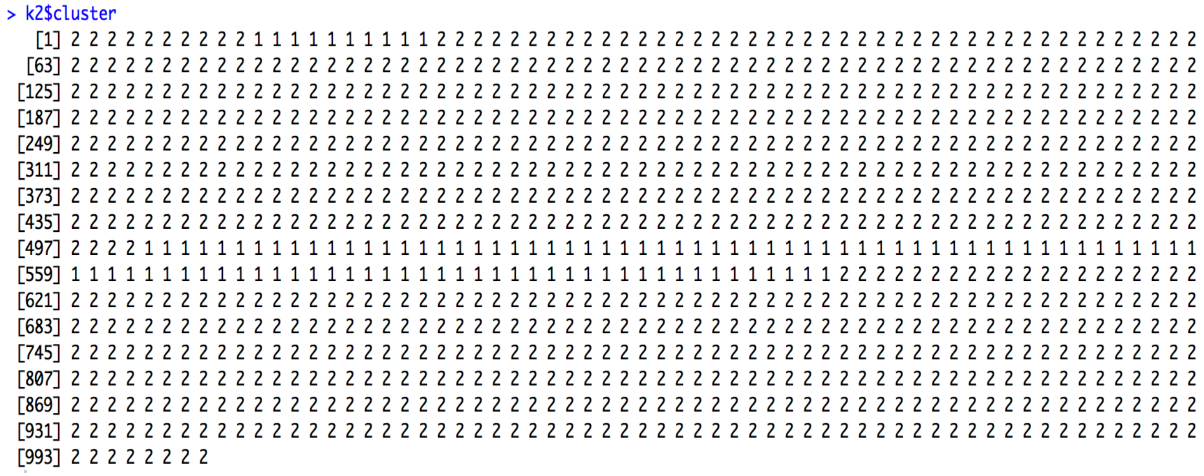
- We get pretty different results when using different linkage methods as we obtain two clusters for complete and single linkages but three clusters for average linkage.
- But K-Means was able to correctly separate the two groups.

c) To know which group differs across the diseased patients and healthy patients we can look at the loading vectors outputted from PCA to see which genes are used to describe the variance the most:



```
> summary(pr)
Importance of components:
                           PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13
Standard deviation     11.9409 6.06818 5.93476 5.83115 5.75209 5.70031 5.63448 5.57726 5.54943 5.50625 5.48852 5.46025 5.40230
Proportion of Variance  0.1267 0.03271 0.03129 0.03021 0.02939 0.02887 0.02821 0.02764 0.02736 0.02694 0.02676 0.02649 0.02593
Cumulative Proportion   0.1267 0.15939 0.19068 0.22089 0.25029 0.27915 0.30736 0.33499 0.36236 0.38929 0.41605 0.44254 0.46847
                          PC14    PC15    PC16    PC17    PC18    PC19    PC20    PC21    PC22    PC23    PC24    PC25    PC26
Standard deviation     5.33441 5.27756 5.21594 5.20000 5.15140 5.11600 5.05591 5.03836 5.01868 4.95965 4.91393 4.86397 4.81796
Proportion of Variance 0.02528 0.02475 0.02417 0.02402 0.02358 0.02325 0.02271 0.02255 0.02238 0.02185 0.02145 0.02102 0.02062
Cumulative Proportion  0.49375 0.51850 0.54267 0.56669 0.59027 0.61352 0.63623 0.65878 0.68116 0.70301 0.72447 0.74548 0.76611
                          PC27    PC28    PC29    PC30    PC31    PC32    PC33    PC34    PC35    PC36    PC37    PC38    PC39
Standard deviation     4.80811 4.73485 4.70098 4.65564 4.61621 4.56733 4.53032 4.49528 4.36502 4.35858 4.26700 4.20277 4.13922
Proportion of Variance 0.02054 0.01992 0.01963 0.01926 0.01893 0.01853 0.01823 0.01795 0.01693 0.01688 0.01618 0.01569 0.01522
Cumulative Proportion  0.78665 0.80656 0.82620 0.84545 0.86439 0.88292 0.90115 0.91910 0.93603 0.95291 0.96909 0.98478 1.00000
                            PC40
Standard deviation     5.251e-15
Proportion of Variance 0.000e+00
Cumulative Proportion  1.000e+00
```

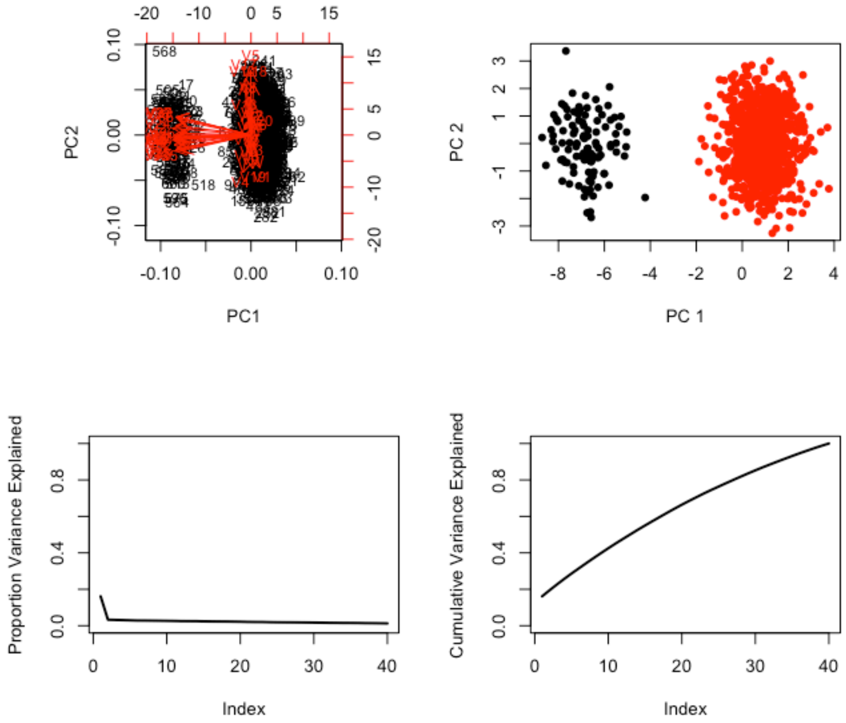
- Running K-means with 2 clusters will allow us to identify the genes that have different expression values:



```
> k2$cluster
   [1] 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
  [63] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [125] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [187] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [249] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [311] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [373] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [435] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [497] 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [559] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [621] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [683] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [745] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [807] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [869] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [931] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [993] 2 2 2 2 2 2 2 2
```

- The result of K-Means with 2 clusters suggests than the genes 11-20 and 500 - 600 differ the most between the 2 groups.

- This is confirmed by running principal components. The 1st principal component explains ~ 20% of the variation in the data and separates the 2 groups. Coloring the points by the clusters generated by K-Means shows that the K-Means clusters the 2 groups correctly.


PC2
PC1
V5
PC 2
PC 1
Proportion Variance Explained
Index
Cumulative Variance Explained
Index