Homework 12

Alex Soupir

April 13, 2020

Packages: ISLR, ggplot2, ggdendro, matrixStats

Collaborators:

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

1. Question 10.7.8 pg 416

In Section 10.2.3, a formula for calculating PVE was given in Equation 10.8. We also saw that the PVE can be obtained using the `sdev` output of the `prcomp()` function.

On the 'USArrests data, calculate PVE in two ways:

(a) Using the `sdev` output of the prcomp() function, as was done in Section 10.2.3.

```
## Scaled arrests data - Mean:
```

```
##        Murder       Assault      UrbanPop          Rape
## -7.663087e-17  1.112408e-16 -4.332808e-16  8.942391e-17
```

```
##
## Scaled arrests data - variance:
```

```
##   Murder  Assault UrbanPop     Rape
##        1        1        1        1
```

```
##
## Proportion of Variance Explained (PVE):
```

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

(b) By applying Equation 10.8 directly. That is, use the `prcomp()` function to compute the principal component loadings. Then, use those loadings in Equation 10.8 to obtain the PVE.

$rotation contains the corresponding pricipal component loadings. Need to multiply the data by the loadings, then sum the columns and rows divided by the total sum of the squared data x. (science.smith.edu)

```
## Proportion of Variance Explained using Equation 10.8:
```

```
##        PC1        PC2        PC3        PC4
## 0.62006039 0.24744129 0.08914080 0.04335752
```

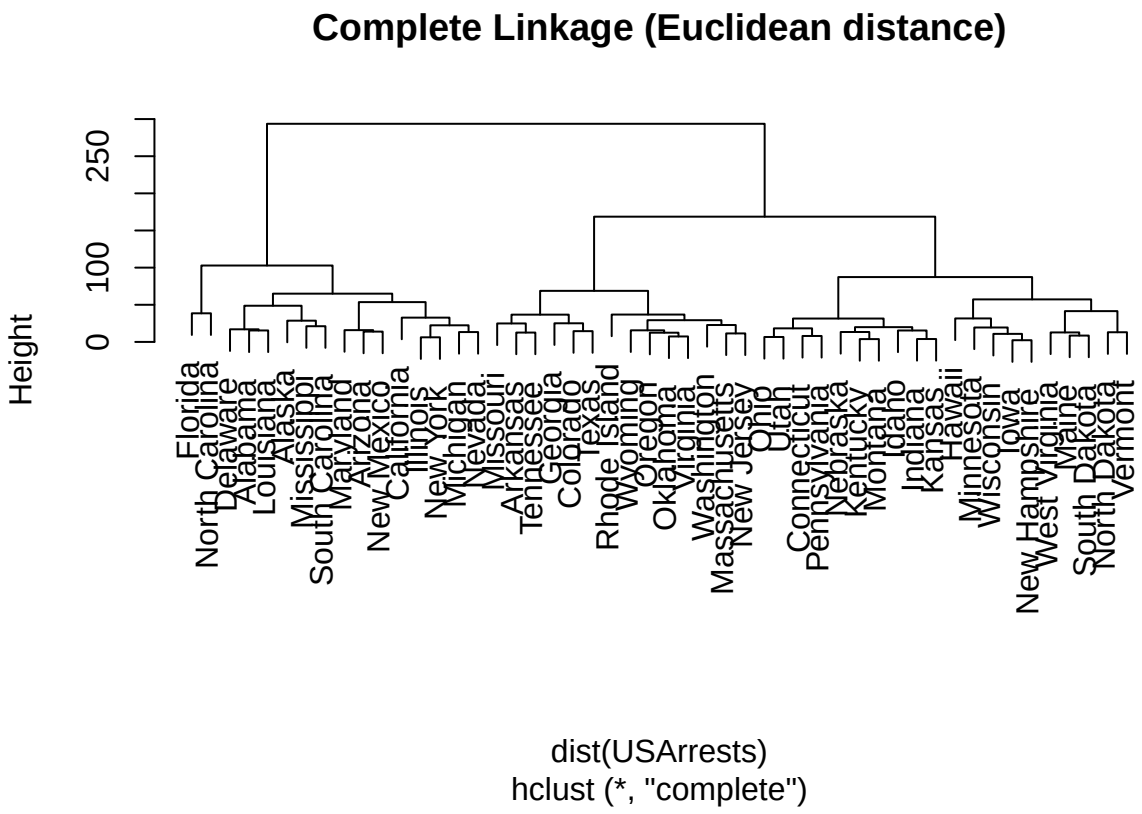
These two approaches should give the same results.

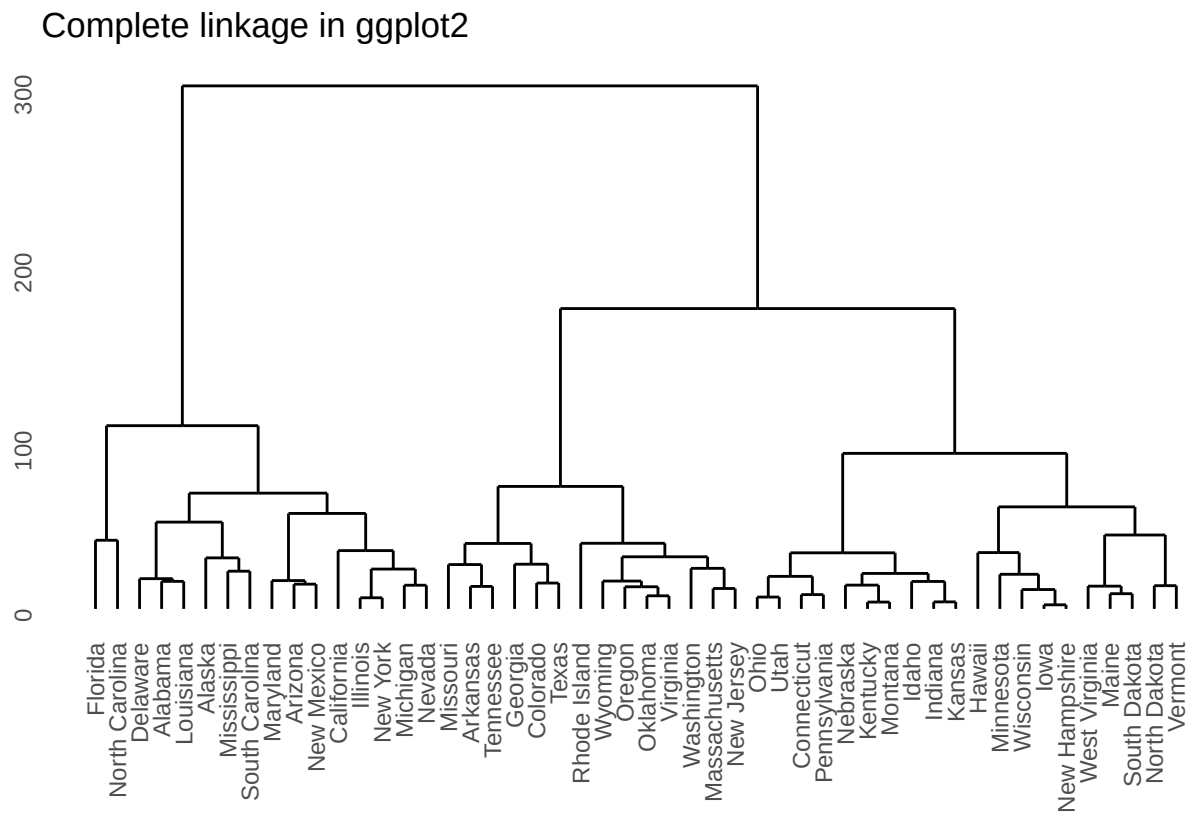
Hint: You will only obtain the same results in (a) and (b) if the same data is used in both cases. For instance, if in (a) you performed `prcomp()` using centered and scaled variables, then you must center and scale the variables before applying Equation 10.3 (? 10.8?) in (b).

2. Question 10.7.9 pg 416

Consider the `USArrests` data. We will now perform hierarchical clustering on the states.

(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.





(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

```
## States in cluster 1:

##        Alabama         Alaska        Arizona     California       Delaware
##              1              1              1              1              1
##        Florida       Illinois      Louisiana       Maryland       Michigan
##              1              1              1              1              1
##    Mississippi         Nevada     New Mexico       New York North Carolina
##              1              1              1              1              1
## South Carolina
##              1

##
## States in cluster 2:

##       Arkansas       Colorado        Georgia Massachusetts       Missouri
##              2              2              2              2              2
##     New Jersey       Oklahoma         Oregon   Rhode Island     Tennessee
##              2              2              2              2              2
##          Texas       Virginia     Washington        Wyoming
##              2              2              2              2

##
## States in cluster 3:

##    Connecticut         Hawaii          Idaho        Indiana           Iowa
##              3              3              3              3              3
```

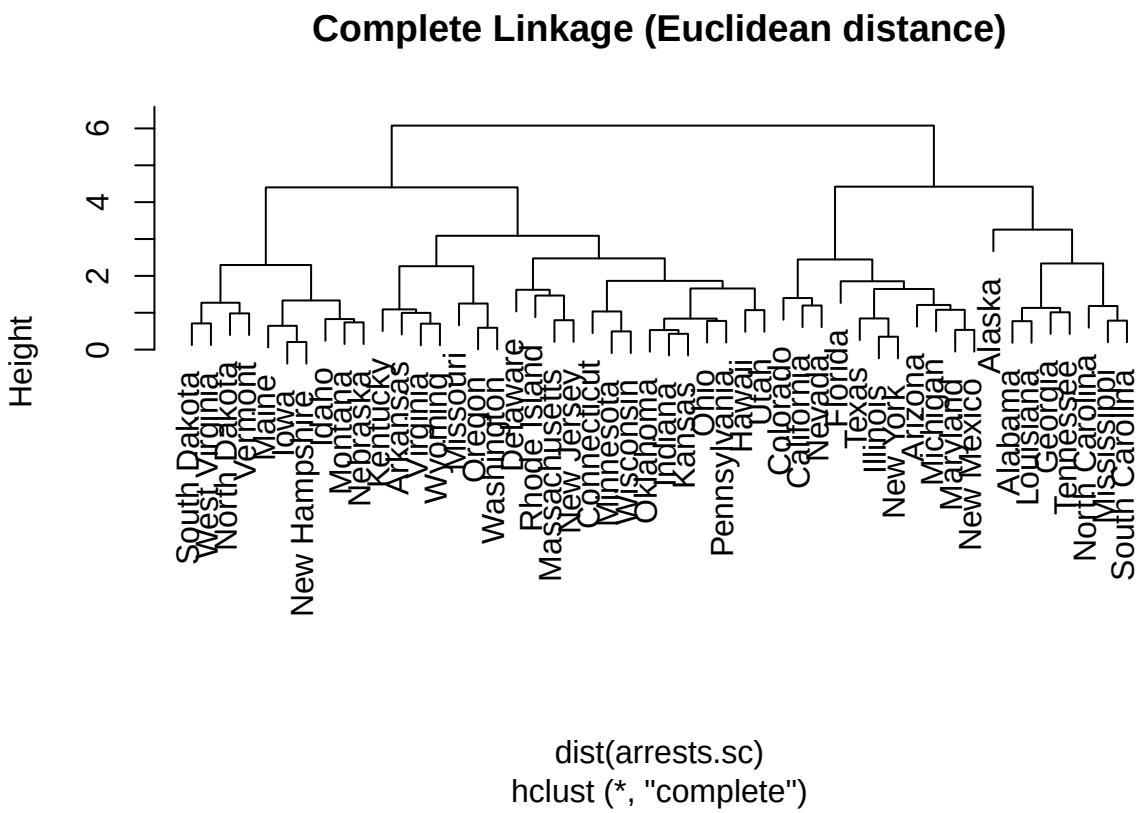
```
##        Kansas      Kentucky         Maine     Minnesota       Montana
##             3             3             3             3             3
##      Nebraska New Hampshire  North Dakota          Ohio  Pennsylvania
##             3             3             3             3             3
##  South Dakota          Utah       Vermont West Virginia     Wisconsin
##             3             3             3             3             3
```

Cluster 1: Alabama, Alaska, Arizona, Delaware, Florida, Hawaii, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, and South Carolina.

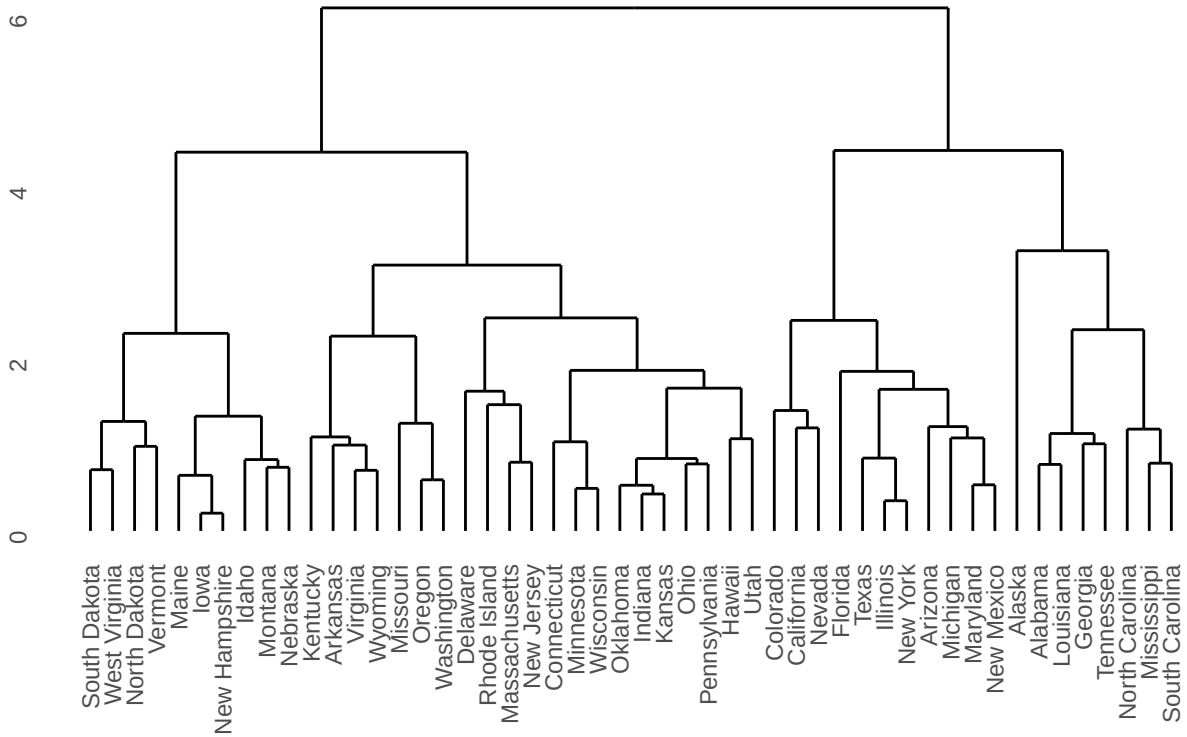
Cluster 2: Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennesee, Texas, Virginia, Washinton, and Wyoming.

Cluster 3: Connecticut, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, North Dakota, Ohio, Pennsylvania, South Dakota, Utah, West Virginia, and Wisconsin.

(c) Hierarchically cluster the states using complete linkage and Euclidean distance, *after scaling the variables to have standard deviation one.*



Complete linkage in ggplot2




```
## States in cluster 1:

##        Alabama         Alaska        Georgia      Louisiana    Mississippi
##              1              1              1              1              1
## North Carolina South Carolina      Tennessee
##              1              1              1

##
## States in cluster 2:

##    Arizona California   Colorado    Florida   Illinois   Maryland   Michigan
##          2          2          2          2          2          2          2
##     Nevada New Mexico   New York      Texas
##          2          2          2          2

##
## States in cluster 3:

##      Arkansas   Connecticut      Delaware        Hawaii         Idaho
##             3             3             3             3             3
##       Indiana          Iowa        Kansas      Kentucky         Maine
##             3             3             3             3             3
## Massachusetts     Minnesota      Missouri       Montana      Nebraska
##             3             3             3             3             3
## New Hampshire    New Jersey  North Dakota          Ohio      Oklahoma
##             3             3             3             3             3
##        Oregon  Pennsylvania  Rhode Island  South Dakota          Utah
##             3             3             3             3             3
##       Vermont      Virginia    Washington West Virginia     Wisconsin
```

```
##            3             3             3             3             3
##      Wyoming
##            3
```

(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

The scaling first changes the height of the dendrogram a great deal, and also changes what are being clustered closest together after using `cutree()`. I think in this case it makes sense to scale the data first. This is because, as Dr. Saunders said in lecture, scaling is the most useful when the variables being used have different units or scales. In this case, as in lecture, `UrbanPop` is a percent rather than per 100,000 like the other 3 variables.

3. Question 10.7.11 pg 417

On the book website, www.StatLearning.com, there is a gene expression data set (`Ch10Ex11.csv`) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

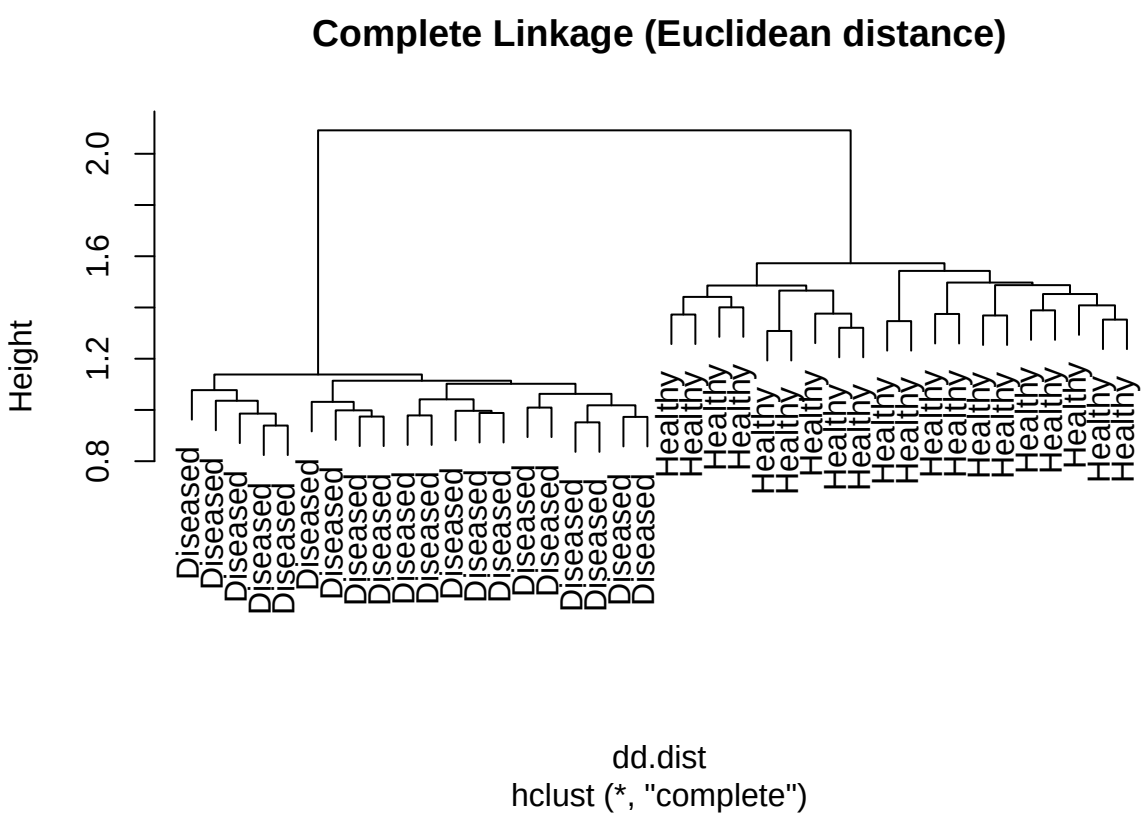
(a) Load the data using `read.csv()`. You will need to select `header=F`.

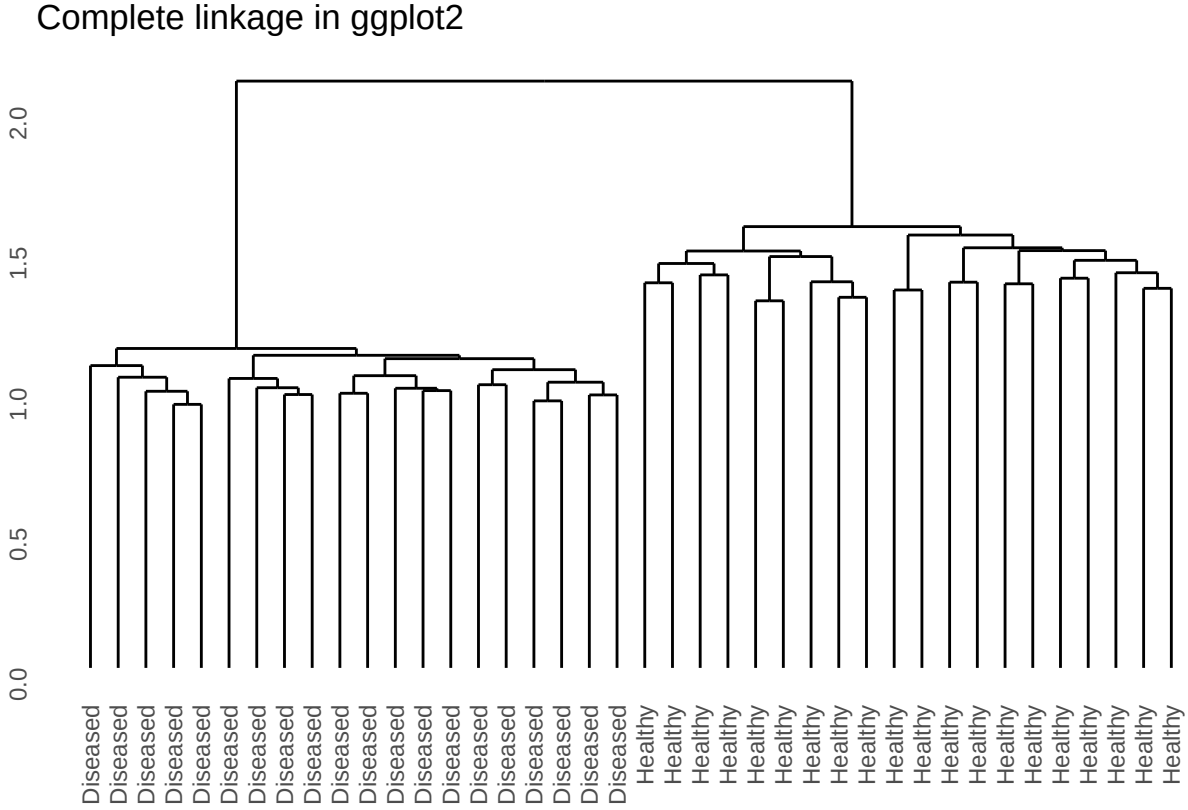
Here I also added the column names as they are described in the question with the first 20 being "Healthy" and the second 20 being "Diseased".

(b) Apply hierarchical clustering to the samples using correlation-based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the linkage used?

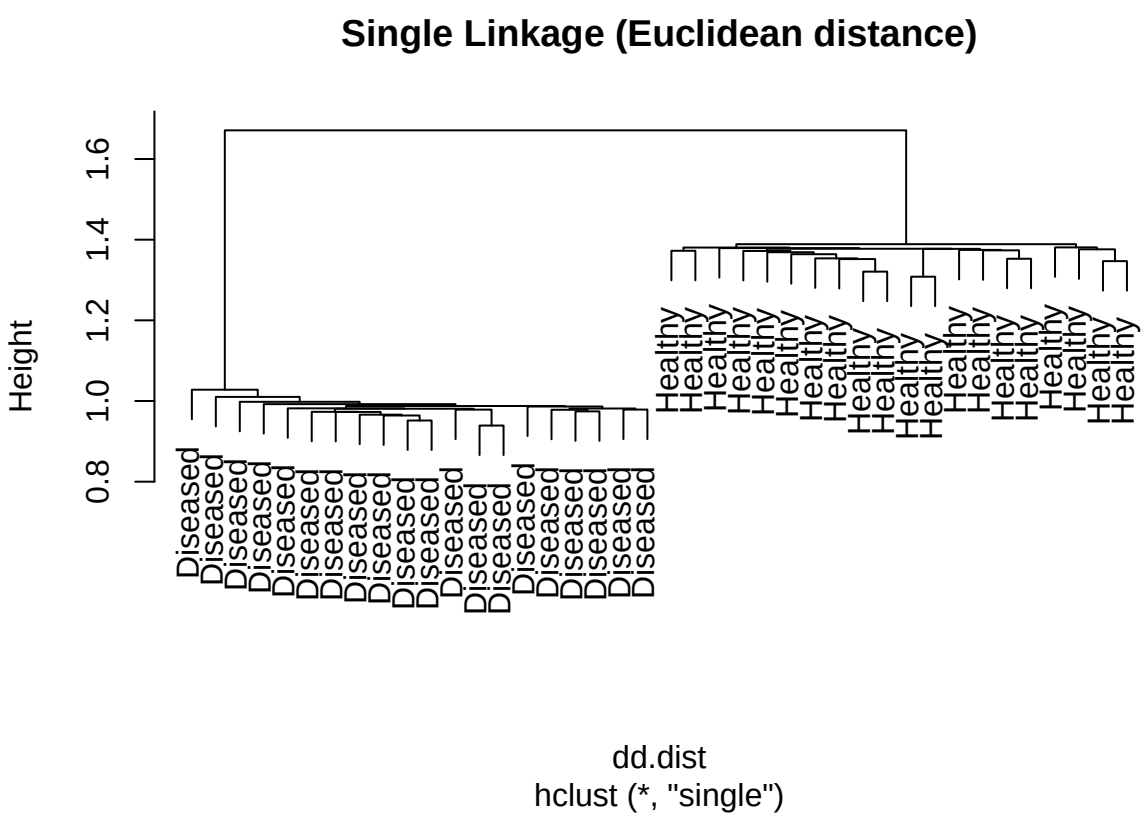
linkage is the `method`

The hierarchical clustering does separate the samples into 2 very distinct clusters, which appear to cluster like samples together i.e. healthy with healthy and diseased with diseased, but not healthy and diseased mixed together.

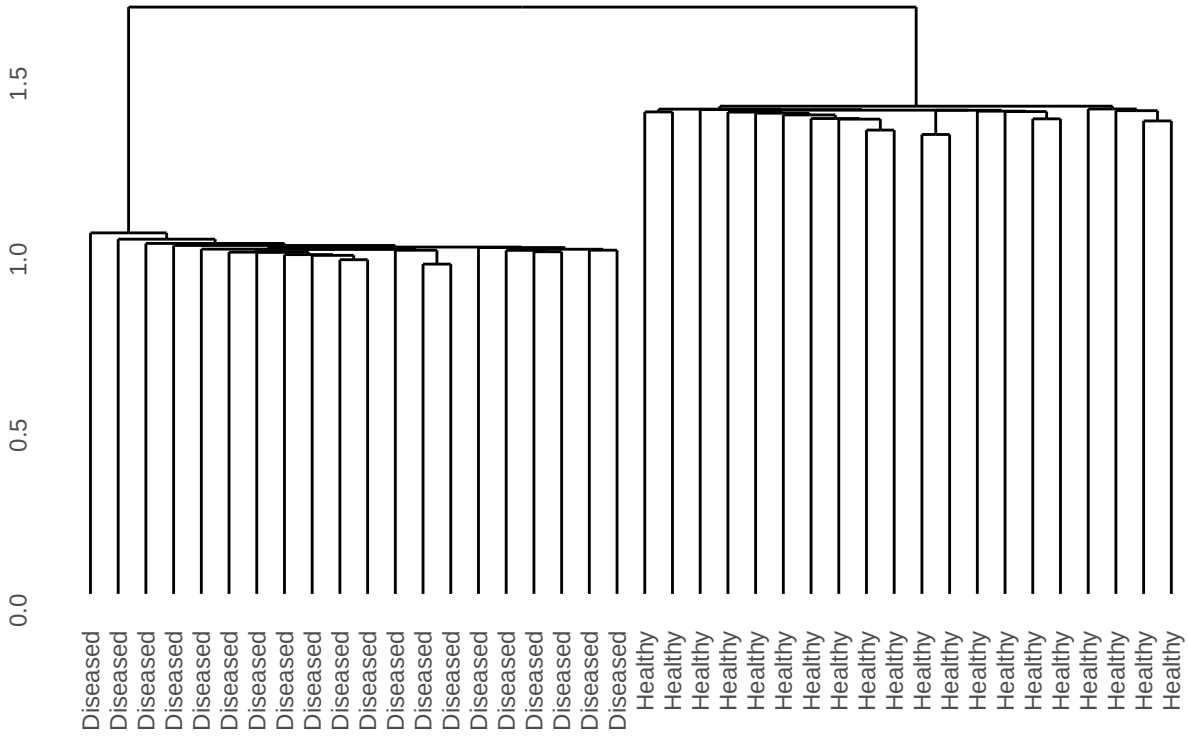

Complete Linkage (Euclidean distance)
Height
0.8
1.2
1.6
2.0
Diseased
Healthy
dd.dist
hclust (*, "complete")

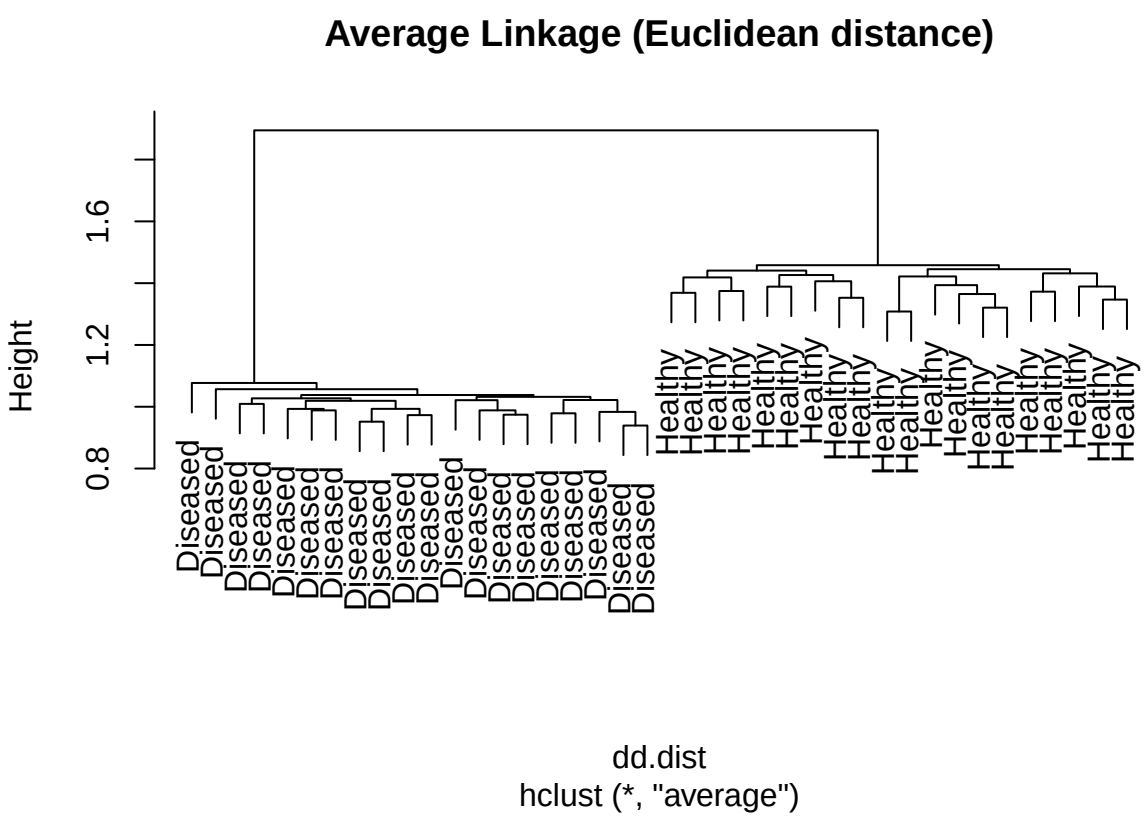


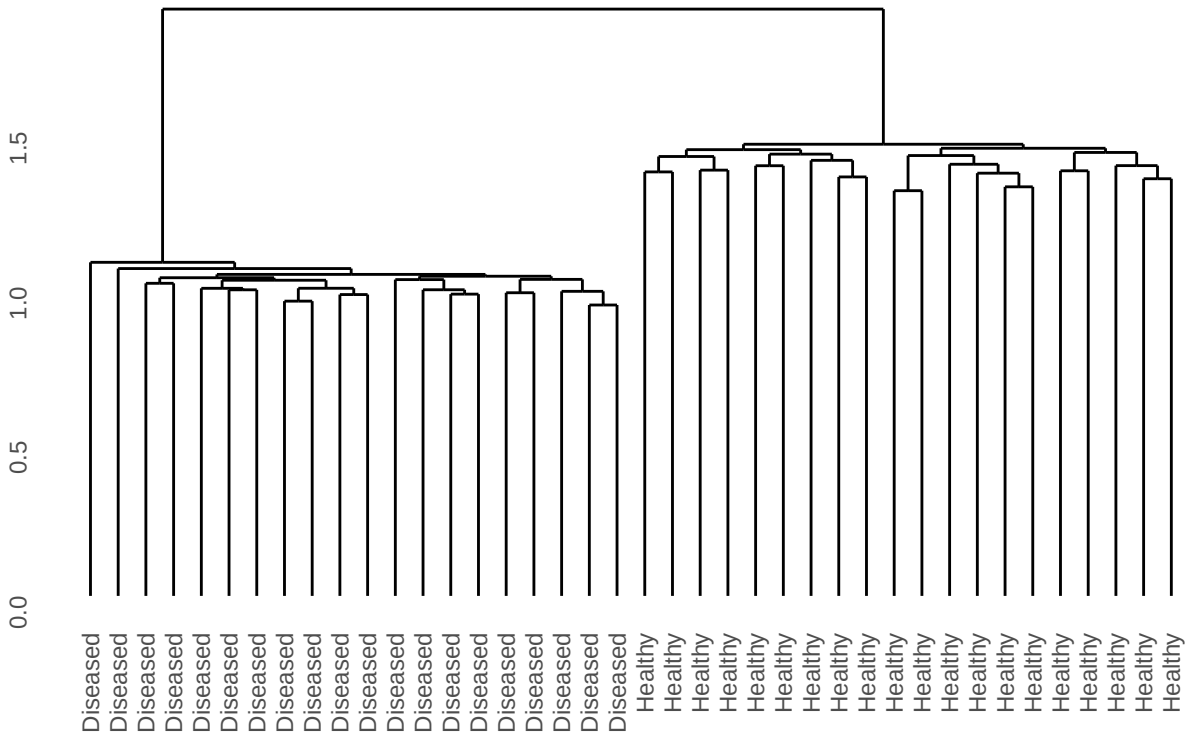
Does the linkage change the groups?

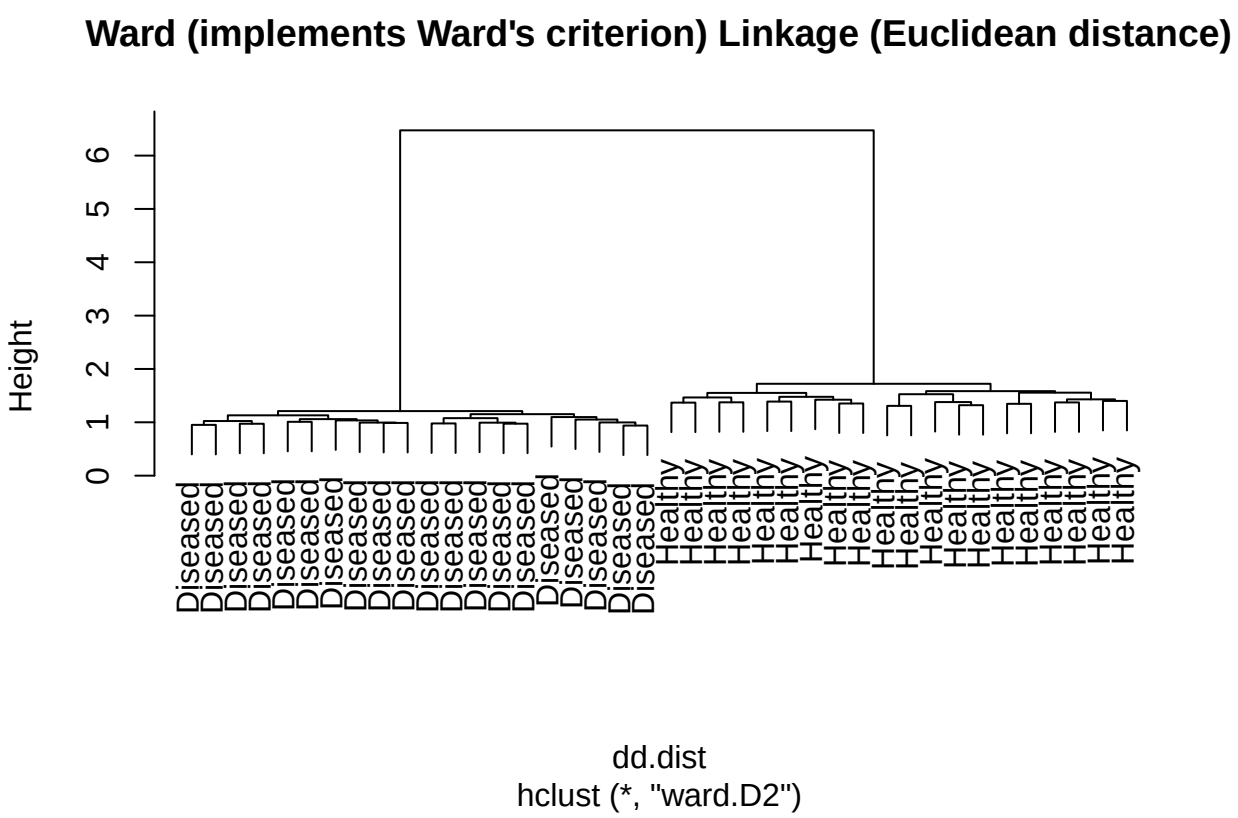

Single Linkage (Euclidean distance)
Height
0.8
1.0
1.2
1.4
1.6
Diseased
Healthy
dd.dist
hclust (*, "single")

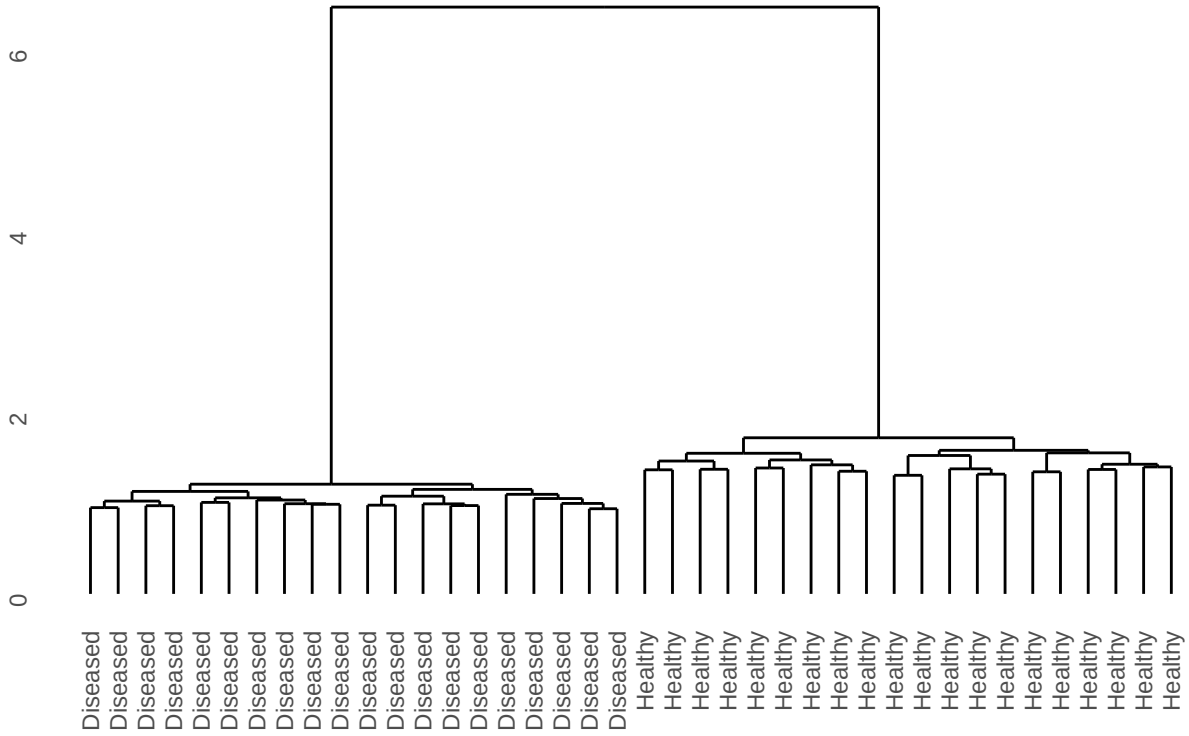
Single linkage in ggplot2




Average Linkage (Euclidean distance)
Height
0.8
1.2
1.6
Diseased
Healthy
dd.dist
hclust (*, "average")


Average linkage in ggplot2
0.0
0.5
1.0
1.5
Diseased
Healthy


Ward (implements Ward's criterion) Linkage (Euclidean distance)
Height
0
1
2
3
4
5
6
Diseased
Healthy
dd.dist
hclust (*, "ward.D2")



The ggplot dendrogram pulls all labels down to the x-axis which is frustrating. However, no matter what the linkage used is, the samples cluster into 2 very distinct groups of Diseased and Healthy patients.

(c) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question and apply it here.

To determine the difference between the 2 different groups 'healthy/diseased', discover.nci.nih.gov mentions that t-tests can be used for gene expression. Here, we aren't looking at *how* significant the genes are so correction of the p-value (with an FDR) shouldn't matter, the only thing we are interested in are *which genes differ the most across the two groups*. To do this, a t-test is calculated for each of the 'genes' and then sorted based on the resulting p-vlaue. This sorted vector is than used. If we were given the raw read counts rather than something that looks like fpkm or something like that, we could use DESeq2 to also calculate which genes are most significantly different between the 2 different treatments (DESeq2 doesn't accept adjusted values because it calculates something similar within the object itself based on the total library size, correcting for large differences between samples).

```
## Genes that are the most different between the two groups:

##  [1] 502 600 589 590 565 584 593 554 538 528
```

The top 10 different genes by t-test significance. Here are the gene names listed above, and then the mean/standard deviation for each subset by the healthy and diseased groups.

```
##    Gene Healthy_Mean Diseased_Mean Healthy_SD Diseased_SD
## 1   502  -0.22692579      1.752314  0.8086420   0.6350892
## 2   600  -0.63335914      1.582609  0.8854333   0.7827552
## 3   589   0.14451706      1.899517  0.6371375   0.7078835
## 4   590  -0.09871516      1.577093  0.6622189   0.7394857
## 5   565  -0.11694254      1.784578  0.8754414   0.7748617
## 6   584  -0.29152291      1.739882  0.9342506   0.9715113
## 7   593  -0.09412908      1.747109  0.9955027   0.7348276
## 8   554  -0.24370881      1.651003  0.9646793   0.8795647
## 9   538  -0.17716210      1.657133  1.0104167   0.7657701
## 10  528  -0.39622798      1.141856  0.7442581   0.7909581
```

Little addon to show that the correction of the p-value is not needed in order to find *which genes differ the most across the two groups* (for peace of mind).

```
headttest = head(order(ttest.pvalues, decreasing = false), 10)
#cat("most significant genes from unadjusted p-values:\n")
headttest
```

```
##  [1] 502 600 589 590 565 584 593 554 538 528
```

```
headttest.adj = head(order(p.adjust(ttest.pvalues, method="bonferroni"), decreasing = false), 10)
#cat("\nmost significant genes from adjusted p-values:\n")
headttest.adj
```

```
##  [1] 502 600 589 590 565 584 593 554 538 528
```

Wonderful.

References

- science.smith.edu
- stats.stackexchange.com
- stackoverflow.com
- discover.nci.nih.gov