HW 6

Elizabeth Hutton

October 23, 2018

1 10.9 Applied

a.
```
library(ISLR)
attach(USArrests)

# dendogram w/ complete linkage, euclidean distance
hc.complete = hclust(dist(USArrests), method = "complete")
plot(hc.complete)
```

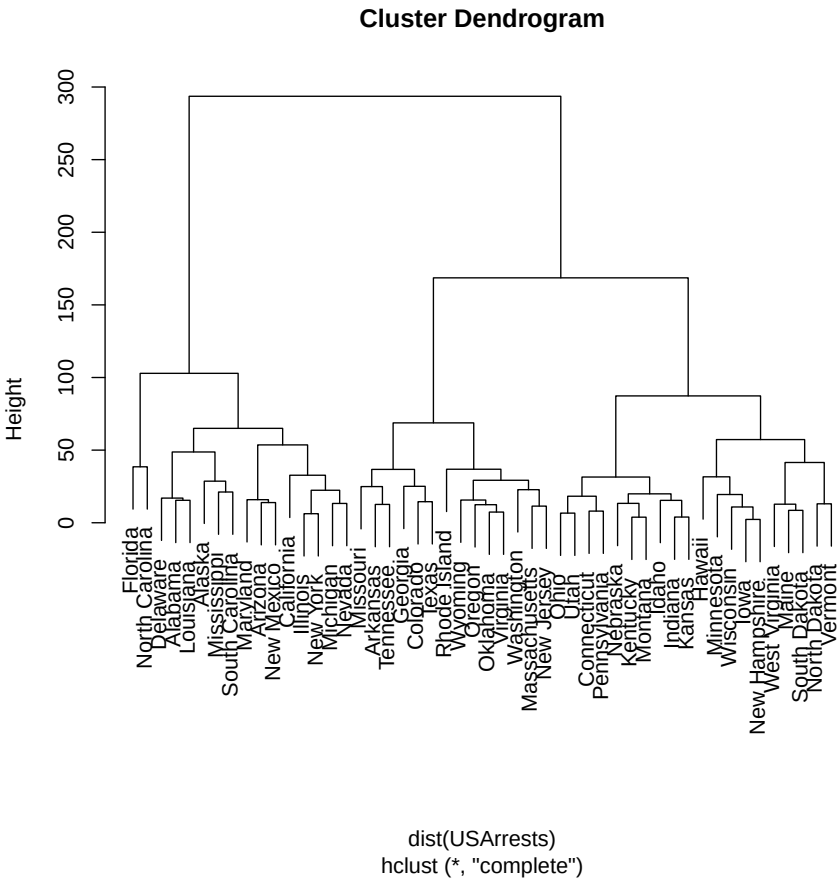

Figure 1: Complete Dendogram

b.
```
# cut tree to form 3 clusters
cut = cutree(hc.complete, 3)

# cluster 1
cut[cut == 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
```

```
# cluster 2
cut[cut == 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
```

```
# cluster 3
cut[cut == 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
```

c.
```
# repeat w/ scaled variables
scaled = scale(USArrests)
hc.scaled = hclust(dist(scaled), method = "complete")
plot(hc.scaled)

# cut tree to form 3 clusters
```

```r
cut = cutree(hc.scaled, 3)

# cluster 1
cut[cut == 1]
```

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

```r
# cluster 2
cut[cut == 2]
```

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

```r
# cluster 3
cut[cut == 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
```

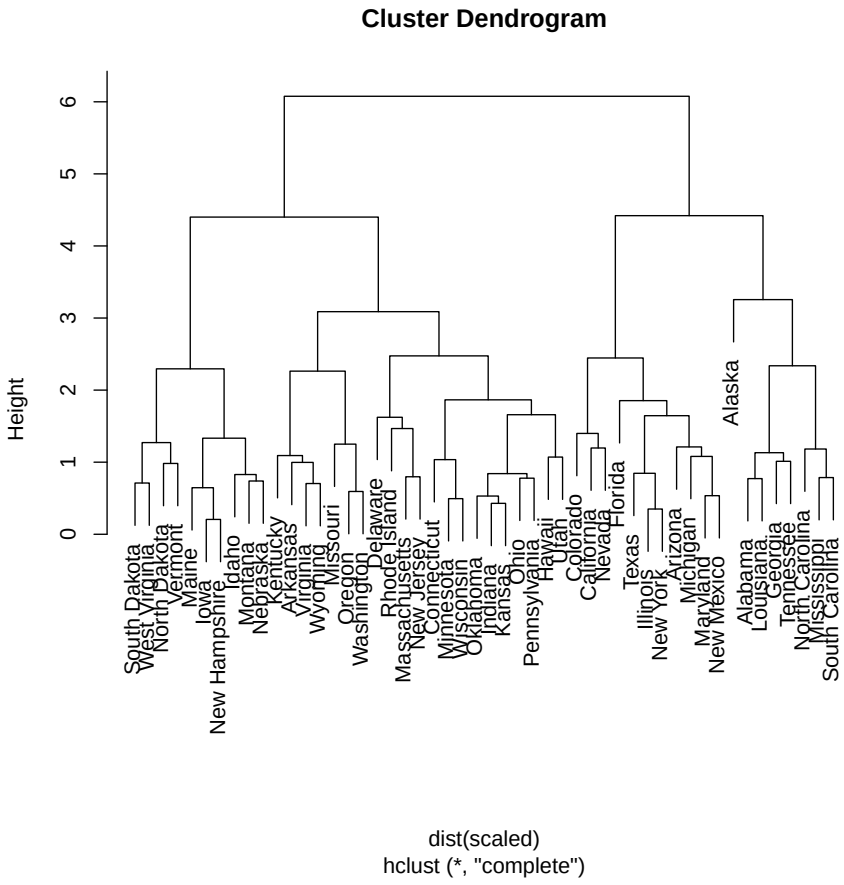

Figure 2: Scaled Dendogram

d. Scaling the variables before making the tree results in a less even distribution of the states over the 3 clusters. There are far more states in cluster 3 than in the other two. However, because the variables all have different units and scales, pre-scaling the data is appropriate.

2 10.9 Applied

a.
```r
data = read.csv("Ch10Ex11.csv", header = FALSE)
```

b.
```r
# complete dendogram
dm = as.dist(1 - cor(data))
cor.complete = hclust(dm, method = "complete")
plot(cor.complete)
```

```r
# single dendogram
cor.single = hclust(dm, method = "single")
plot(cor.single)
```

```r
# single dendogram
cor.avg = hclust(dm, method = "average")
plot(cor.avg)
```

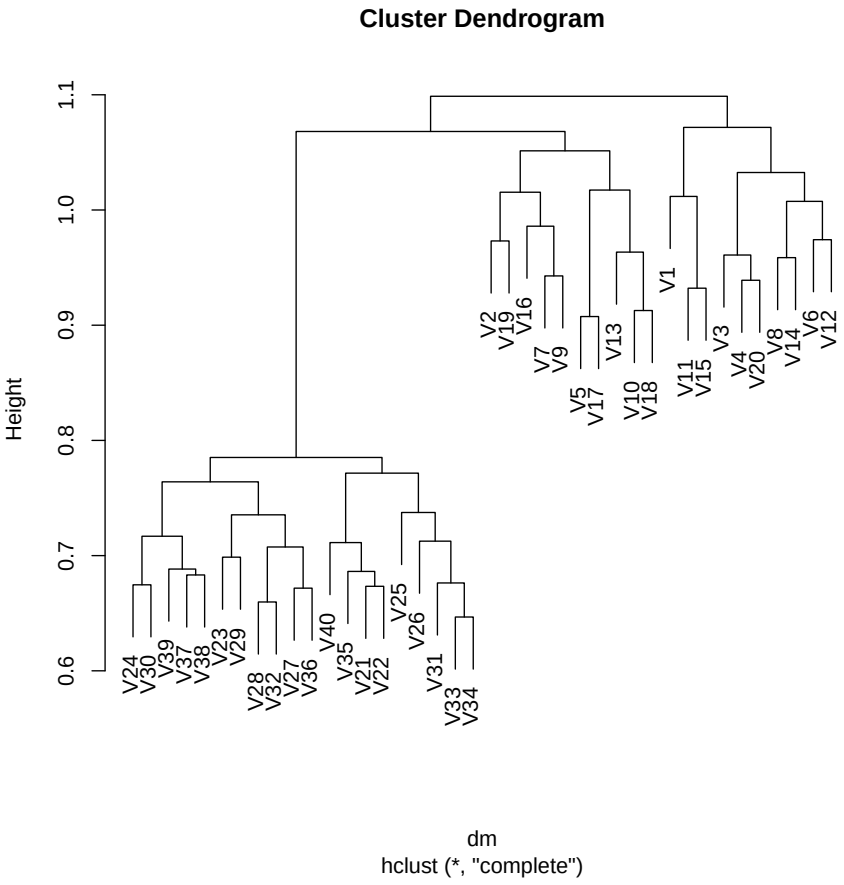

Figure 3: Complete Dendogram

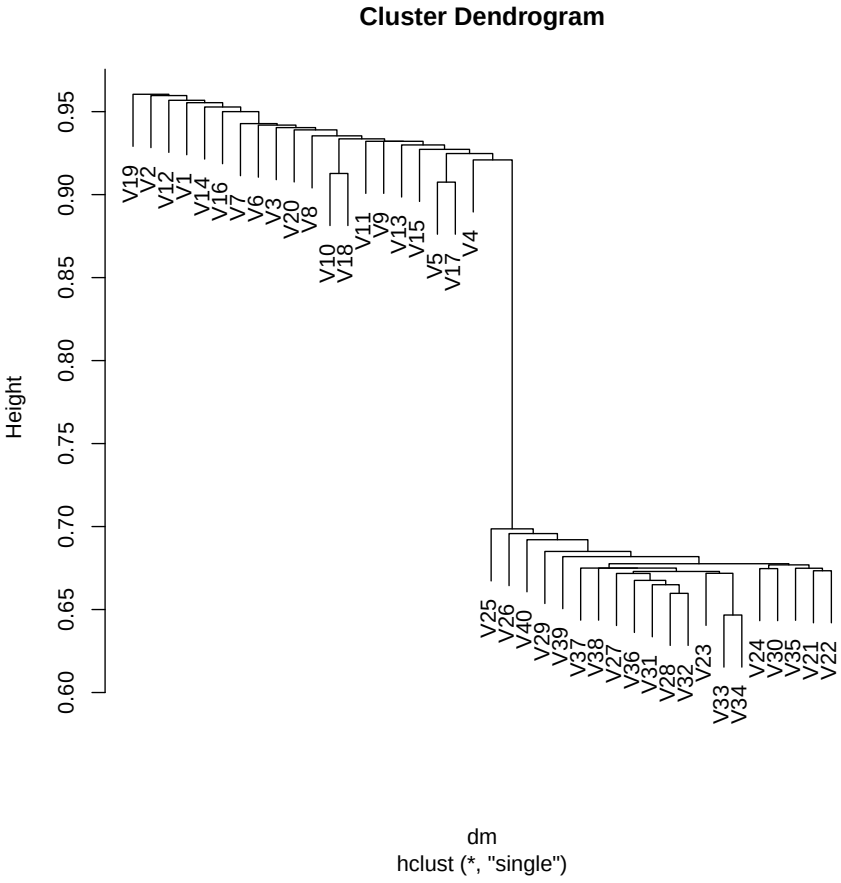

Figure 4: Single Dendogram

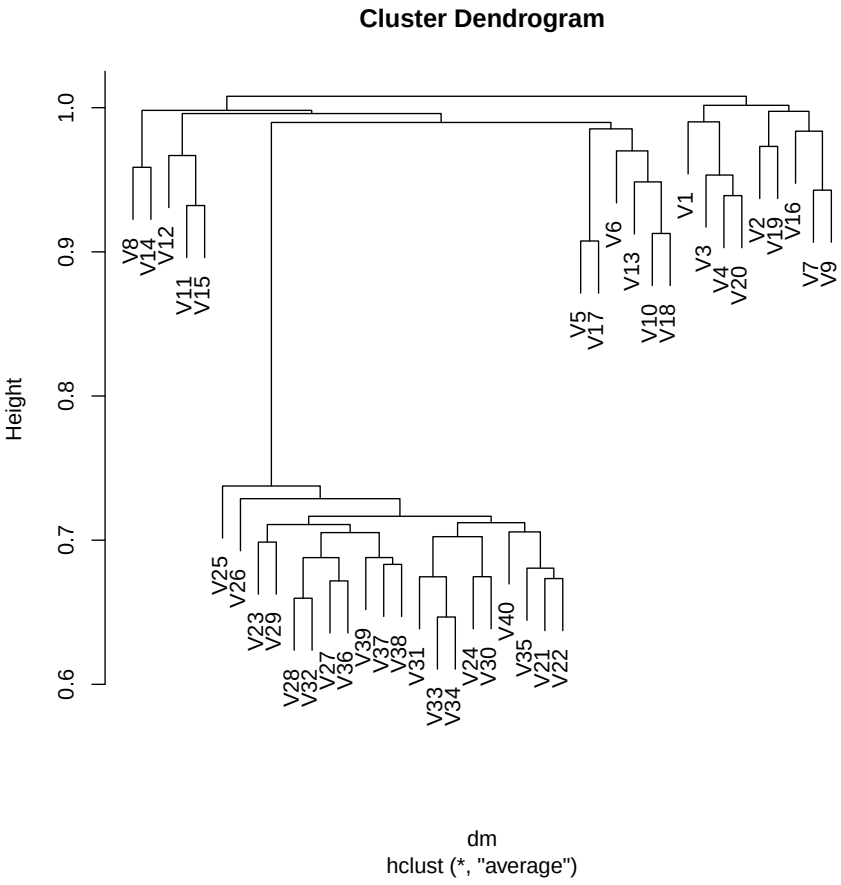

Figure 5: Average Dendogram

When using complete or single linkage, the dendogram cleanly separates the data into its two natural clusters. Using average linkage does not.

c. You could perform PCA to determine which genes are the most different between the two clusters, diseased vs. healthy. The genes with the largest absolute value of loading will be more different across clusters. For example, the 10 most different genes are listed below:

```
# PCA to find most different genes
pr.out = prcomp(t(data))
load = apply(pr.out$rotation, 1, sum)
ordered = order(abs(load), decreasing = TRUE)

# the most different genes
ordered[1:10]

##  [1] 865  68 911 428 624  11 524 803 980 822
```