Chase Moreno

@2078503

MATH 4323

Homework 6

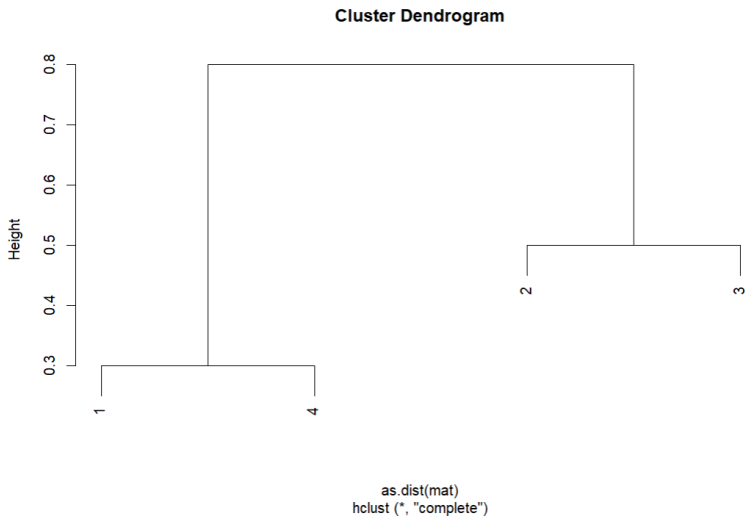
1. Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

$$\begin{pmatrix} & 0.4 & 0.75 & 0.3 \\ 0.4 & & 0.5 & 0.8 \\ 0.75 & 0.5 & & 0.45 \\ 0.3 & 0.8 & 0.45 & \end{pmatrix}$$

For instance, the dissimilarity between the first and second observations is 0.4, and the dissimilarity between the second and fourth observations is 0.8.

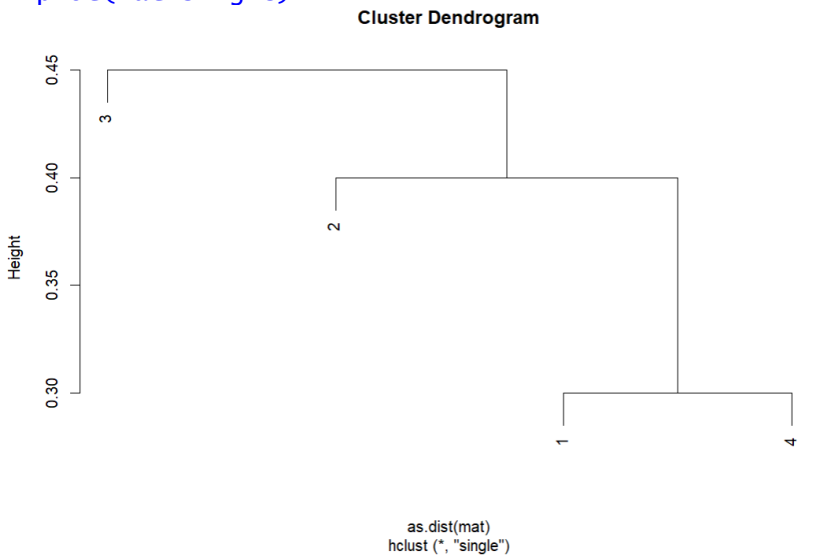
(a) On the basis of this dissimilarity matrix, sketch the dendrogram (by hand is fine) that results from hierarchically clustering these four observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.

```
> mat <- as.dist(matrix(c(0, .4, .75, .3,
+                         .4, 0, .5, .8,
+                         .75, .5, 0, .45,
+                         .3, .8, .45, 0), ncol = 4))
> mat
     1    2    3
2 0.40
3 0.75 0.50
4 0.30 0.80 0.45
> mat.complete = hclust(as.dist(mat),method = "complete")
> plot(mat.complete)
```



(b) Repeat (a), this time using single linkage clustering.

```
> mat.single = hclust(as.dist(mat),method = "single")
> plot(mat.single)
```



(c) Suppose that we cut the dendrogram obtained in (a) such that two clusters result. Which observations are in each cluster?

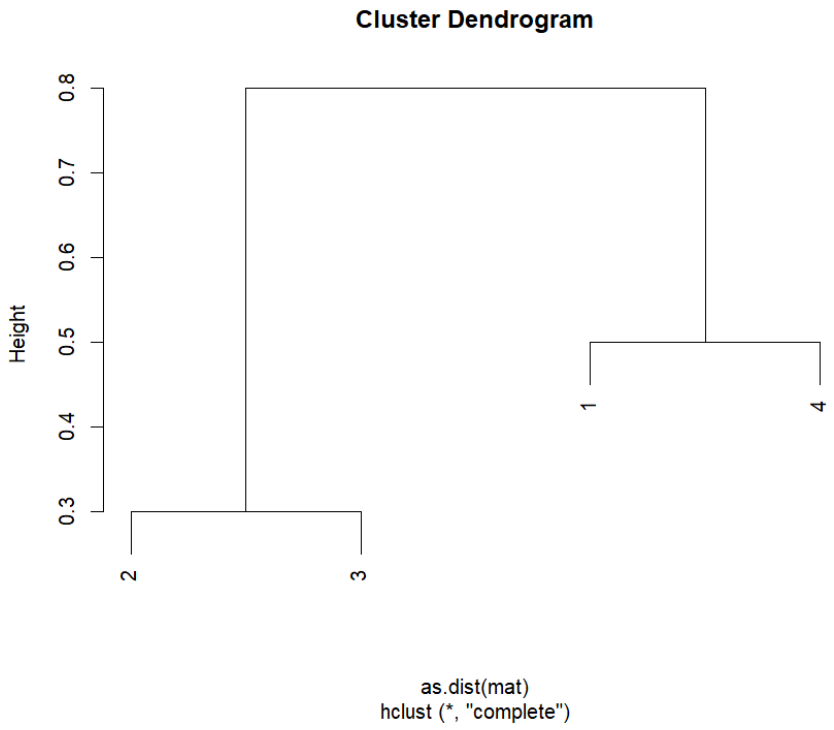
Observation 1 and 4 in one cluster. Observation 2 and 3 in another cluster.

(d) Suppose that we cut the dendrogram obtained in (b) such that two clusters result. Which observations are in each cluster?

Observation 1, 4, and 3 in one cluster. Observation 3 in another cluster.

(e) It is mentioned in the chapter that at each fusion in the dendrogram, the position of the two clusters being fused can be swapped without changing the meaning of the dendrogram. Draw a dendrogram that is equivalent to the dendrogram in (a), for which two or more of the leaves are repositioned, but for which the meaning of the dendrogram is the same.

```
> mat.complete = hclust(as.dist(mat),method = "complete")
> plot(mat.complete, labels = c(2, 1, 4, 3))
```



2. Here we work with the example introduced in lecture slides, where n = 9 observations are described by p = 2 predictors with the following dissimilarity matrix:

> round(dist(x),2)
1 2 3 4 5 6 7 8
2 0.66
3 1.08 1.70
4 0.97 1.52 1.15
5 2.02 1.48 2.80 2.98
6 0.36 0.96 1.00 0.61 2.38
7 2.24 1.74 2.93 3.20 0.32 2.59
8 1.46 1.04 2.16 2.43 0.65 1.82 0.79
9 1.96 2.01 1.91 2.77 1.84 2.25 1.75 1.37

We've already shown that clusters {5} and {7} will be merged first due to least dissimilarity (dist({5}, {7}) = 0.32). Afterwards, we calculated the distances from new cluster {5, 7} to all the other n−2 = 7 "clusters" ({1}, {2}, {3}, {4}, {6}, {8}, {9}) by using complete linkage. Here, proceed to calculate those distances by using

(a) Single linkage.

dist({1}, {5, 7}) = min(dist({1}, {5}), dist({1}, {7})) = min(2.02, 2.24) = 2.02

dist({2}, {5, 7}) = min(dist({2}, {5}), dist({2}, {7})) = min(1.48, 1.74) = 1.48

dist({3}, {5, 7}) = min(dist({3}, {5}), dist({3}, {7})) = min(2.80, 2.93) = 2.80

dist({4}, {5, 7}) = min(dist({4}, {5}), dist({4}, {7})) = min(2.98, 3.20) = 2.98

dist({6}, {5, 7}) = min(dist({6}, {5}), dist({6}, {7})) = min(2.38, 2.59) = 2.38

dist({8}, {5, 7}) = min(dist({8}, {5}), dist({8}, {7})) = min(0.65, 0.79) = 0.65

dist({9}, {5, 7}) = min(dist({9}, {5}), dist({9}, {7})) = min(1.84, 1.75) = 1.75

(b) Average linkage.

dist({1}, {5, 7}) = [dist({1}, {5}) + dist({1}, {7})] / 2 = [2.02 + 2.24] / 2 = 2.13

dist({2}, {5, 7}) = [dist({2}, {5}) + dist({2}, {7})] / 2 = [1.48 + 1.74] / 2 = 1.61

dist({3}, {5, 7}) = [dist({3}, {5}) + dist({3}, {7})] / 2 = [2.80 + 2.93] / 2 = 2.865

dist({4}, {5, 7}) = [dist({4}, {5}) + dist({4}, {7})] / 2 = [2.98 + 3.20] / 2 = 3.09

dist({6}, {5, 7}) = [dist({6}, {5}) + dist({6}, {7})] / 2 = [2.38 + 2.59] / 2 = 2.485

dist({8}, {5, 7}) = [dist({8}, {5}) + dist({8}, {7})] / 2 = [0.65 + 0.79] / 2 = 0.72

dist({9}, {5, 7}) = [dist({9}, {5}) + dist({9}, {7})] / 2 = [1.84 + 1.75] / 2 = 1.795

See the slide #9 from ”Unsupervised Learning. Hierarchical Clustering” lecture, for examples of calculations in case of complete linkage.

3. Suppose that for a particular data set, we perform hierarchical clustering using single linkage and using complete linkage. We obtain two dendrograms.

(a) At a certain point on the single linkage dendrogram, the clusters {1, 2, 3} and {4, 5} fuse. On the complete linkage dendrogram, the clusters {1, 2, 3} and {4, 5} also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

Not enough information to tell because I need more additional information. Also, the height of the fusion point depends on the dissimilarity between the clusters that are being fused. So, it is hard to tell which fusion will occur higher on the tree or fuse at the same height.

(b) At a certain point on the single linkage dendrogram, the clusters {5} and {6} fuse. On the complete linkage dendrogram, the clusters {5} and {6} also fuse at a certain point. Which fusion will occur higher on the tree, or will they fuse at the same height, or is there not enough information to tell?

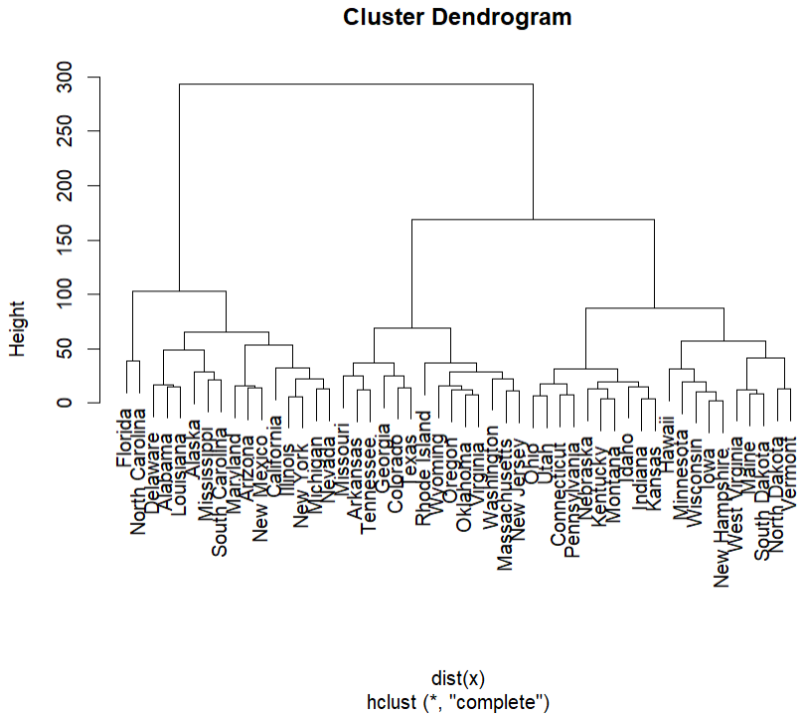
They will fuse at the same height.

Applied.

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

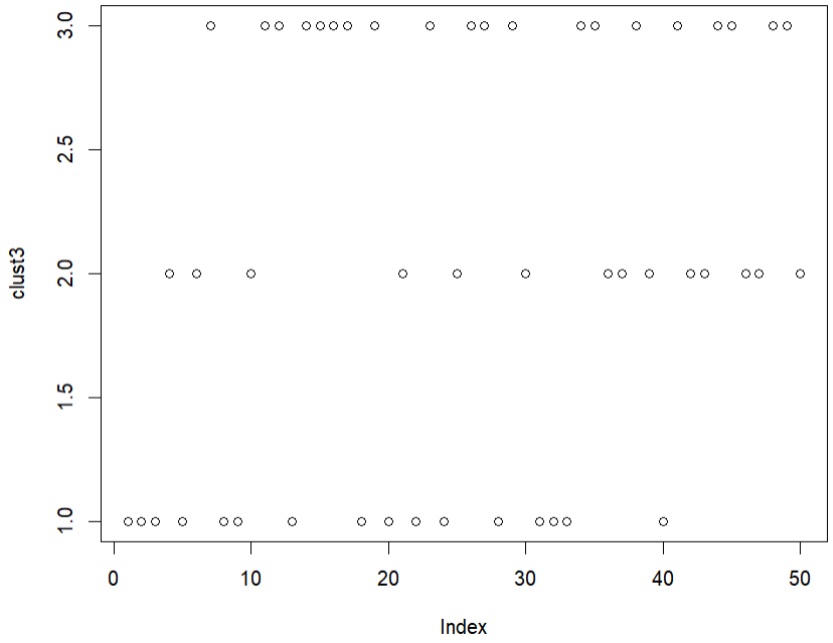
(a) Using hierarchical clustering (Eucledian distance as the dissimilarity measure) with complete linkage, cluster the states. Provide the dendrogram.

```
> x <- USArrests
> comp = hclust(dist(x), method = "complete")
> plot(comp)
```



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

```
> clust3 <- cutree(comp, 3)
> clust3
       Alabama         Alaska        Arizona       Arkansas     California       Colorado
             1              1              1              2              1              2
   Connecticut       Delaware        Florida        Georgia         Hawaii          Idaho
             3              1              1              2              3              3
      Illinois        Indiana           Iowa         Kansas       Kentucky      Louisiana
             1              3              3              3              3              1
         Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi
             3              1              2              1              3              1
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey
             2              3              3              1              3              2
    New Mexico       New York North Carolina   North Dakota           Ohio       Oklahoma
             1              1              1              3              3              2
        Oregon   Pennsylvania   Rhode Island South Carolina   South Dakota      Tennessee
             2              3              2              1              3              2
         Texas           Utah        Vermont       Virginia     Washington  West Virginia
             2              3              3              2              2              3
     Wisconsin        Wyoming
             3              2
> plot(clust3)
```



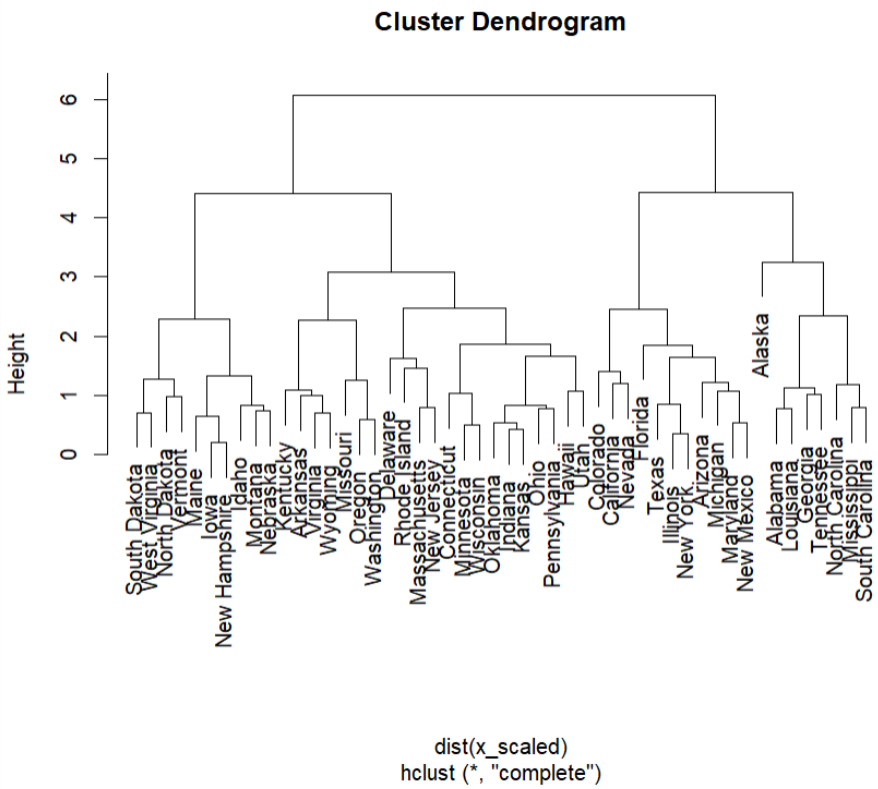
States in cluster 1 are Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, and South Carolina.

States in cluster 2 are Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, and Wyoming.

States in cluster 3 are Connecticut, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, Ohio, Pennsylvania, South Dakota, Utah, Vermont, West Virginia, and Wisconsin.

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

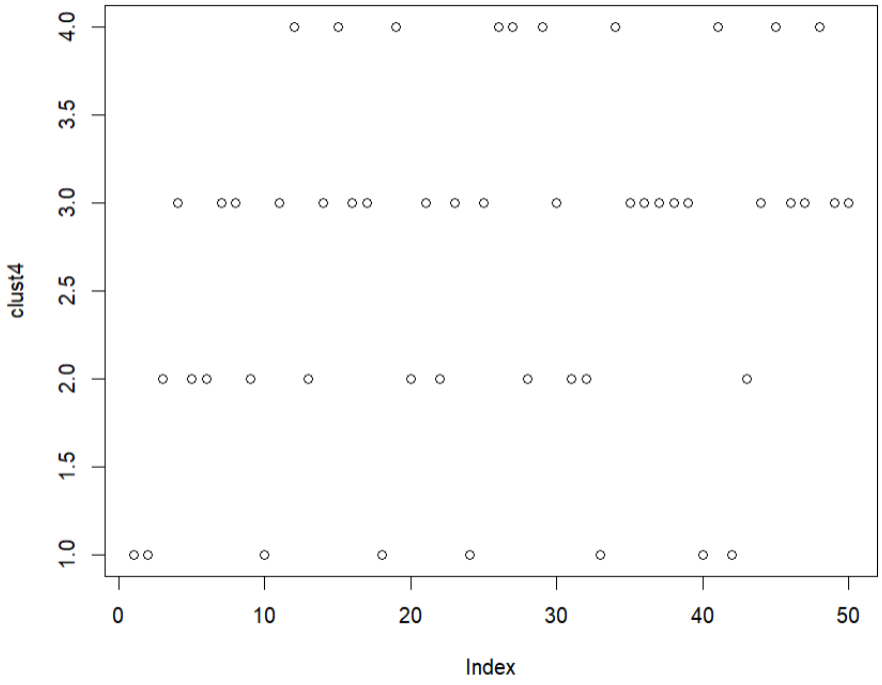
```
> x_scaled<- scale(x)
> comp_scaled <- hclust(dist(x_scaled), method = "complete")
> plot(comp_scaled)
```



(d) Judging by the dendrogram, what appears to be a good number K of natural clusters? Cut the dendrogram at a height corresponding to that number K.

I'm thinking K = 4 appears to be a good number K of natural clusters.

```
> clust4 <- cutree(comp_scaled, 4)
> clust4
       Alabama         Alaska        Arizona       Arkansas     California       Colorado
             1              1              2              3              2              2
   Connecticut       Delaware        Florida        Georgia         Hawaii          Idaho
             3              3              2              1              3              4
      Illinois        Indiana           Iowa         Kansas       Kentucky      Louisiana
             2              3              4              3              3              1
         Maine       Maryland  Massachusetts       Michigan      Minnesota    Mississippi
             4              2              3              2              3              1
      Missouri        Montana       Nebraska         Nevada  New Hampshire     New Jersey
             3              4              4              2              4              3
    New Mexico       New York North Carolina   North Dakota           Ohio       Oklahoma
             2              2              1              4              3              3
        Oregon   Pennsylvania   Rhode Island South Carolina   South Dakota      Tennessee
             3              3              3              1              4              1
         Texas           Utah        Vermont       Virginia     Washington  West Virginia
             2              3              4              3              3              4
     Wisconsin        Wyoming
             3              3
> plot(clust4)
```



(e) Which states belong to which clusters from (d)? Provide the cluster assignment output. Describe which aspects unite the states within each of the clusters. E.g. ”Cluster 1 contains states with low urban populations and low counts of murder, rape & assault.”

I provided the cluster assignment output in part d.

States in cluster 1 are Alabama, Alaska, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee.

States in cluster 2 are Arizona, California, Colorado, Florida, Illinois, Michigan, Nevada, New Mexico, New York, and Texas.

States in cluster 3 are Arkansas, Connecticut, Delaware, Hawaii, Indiana, Kansas, Kentucky, Massachusetts, Minnesota, Missouri, New Jersey, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Utah, Virginia, Washington, Wisconsin, and Wyoming.

States in cluster 4 are Idaho, Iowa, Maine, Montana, Nebraska, New Hampshire, North Dakota, South Dakota, Vermont, and West Virginia.

```
> x[1]
               Murder
Alabama          13.2
Alaska           10.0
Arizona           8.1
Arkansas          8.8
California        9.0
Colorado          7.9
Connecticut       3.3
Delaware          5.9
Florida          15.4
Georgia          17.4
Hawaii            5.3
Idaho             2.6
Illinois         10.4
Indiana           7.2
Iowa              2.2
Kansas            6.0
Kentucky          9.7
Louisiana        15.4
Maine             2.1
Maryland         11.3
```

Massachusetts	4.4
Michigan	12.1
Minnesota	2.7
Mississippi	16.1
Missouri	9.0
Montana	6.0
Nebraska	4.3
Nevada	12.2
New Hampshire	2.1
New Jersey	7.4
New Mexico	11.4
New York	11.1
North Carolina	13.0
North Dakota	0.8
Ohio	7.3
Oklahoma	6.6
Oregon	4.9
Pennsylvania	6.3
Rhode Island	3.4
South Carolina	14.4
South Dakota	3.8
Tennessee	13.2
Texas	12.7
Utah	3.2
Vermont	2.2
Virginia	8.5
Washington	4.0
West Virginia	5.7
Wisconsin	2.6
Wyoming	6.8

```
> x[2]
```

	Assault
Alabama	236
Alaska	263
Arizona	294
Arkansas	190
California	276
Colorado	204
Connecticut	110
Delaware	238
Florida	335
Georgia	211
Hawaii	46
Idaho	120
Illinois	249
Indiana	113
Iowa	56
Kansas	115
Kentucky	109
Louisiana	249
Maine	83
Maryland	300
Massachusetts	149
Michigan	255
Minnesota	72
Mississippi	259
Missouri	178
Montana	109
Nebraska	102
Nevada	252
New Hampshire	57
New Jersey	159
New Mexico	285
New York	254
North Carolina	337
North Dakota	45
Ohio	120
Oklahoma	151
Oregon	159
Pennsylvania	106
Rhode Island	174
South Carolina	279
South Dakota	86
Tennessee	188
Texas	201
Utah	120
Vermont	48
Virginia	156
Washington	145
West Virginia	81
Wisconsin	53
Wyoming	161

```
> x[3]
```

	UrbanPop
Alabama	58
Alaska	48
Arizona	80
Arkansas	50
California	91
Colorado	78
Connecticut	77
Delaware	72
Florida	80
Georgia	60
Hawaii	83
Idaho	54
Illinois	83
Indiana	65
Iowa	57
Kansas	66
Kentucky	52
Louisiana	66
Maine	51
Maryland	67
Massachusetts	85
Michigan	74
Minnesota	66
Mississippi	44
Missouri	70
Montana	53
Nebraska	62
Nevada	81
New Hampshire	56
New Jersey	89
New Mexico	70
New York	86
North Carolina	45
North Dakota	44
Ohio	75
Oklahoma	68

```
Oregon               67
Pennsylvania         72
Rhode Island         87
South Carolina       48
South Dakota         45
Tennessee            59
Texas                80
Utah                 80
Vermont              32
Virginia             63
Washington           73
West Virginia        39
Wisconsin            66
Wyoming              60
```

```
> x[4]
```

```
               Rape
Alabama        21.2
Alaska         44.5
Arizona        31.0
Arkansas       19.5
California     40.6
Colorado       38.7
Connecticut    11.1
Delaware       15.8
Florida        31.9
Georgia        25.8
Hawaii         20.2
Idaho          14.2
Illinois       24.0
Indiana        21.0
Iowa           11.3
Kansas         18.0
Kentucky       16.3
Louisiana      22.2
Maine           7.8
Maryland       27.8
Massachusetts  16.3
Michigan       35.1
Minnesota      14.9
Mississippi    17.1
Missouri       28.2
Montana        16.4
Nebraska       16.5
Nevada         46.0
New Hampshire   9.5
New Jersey     18.8
New Mexico     32.1
New York       26.1
North Carolina 16.1
North Dakota    7.3
Ohio           21.4
Oklahoma       20.0
Oregon         29.3
Pennsylvania   14.9
Rhode Island    8.3
South Carolina 22.5
South Dakota   12.8
Tennessee      26.9
Texas          25.5
Utah           22.9
Vermont        11.2
Virginia       20.7
Washington     26.2
West Virginia   9.3
Wisconsin      10.8
Wyoming        15.6
```

Cluster 1 and 2 contains states with a high murder rate and high assault.

Cluster 2 and 3 contains states with a high urban pop.

Cluster 4 contains states with a low urban pop and low rape rate.

(f) In your opinion, is there a reason to scale those variables before the inter-observation dissimilarities are computed? Why?

Yes, there is a reason to scale those variables before the inter-observation dissimilarities are computed if they are in a different measure unit.

5. 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 1000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

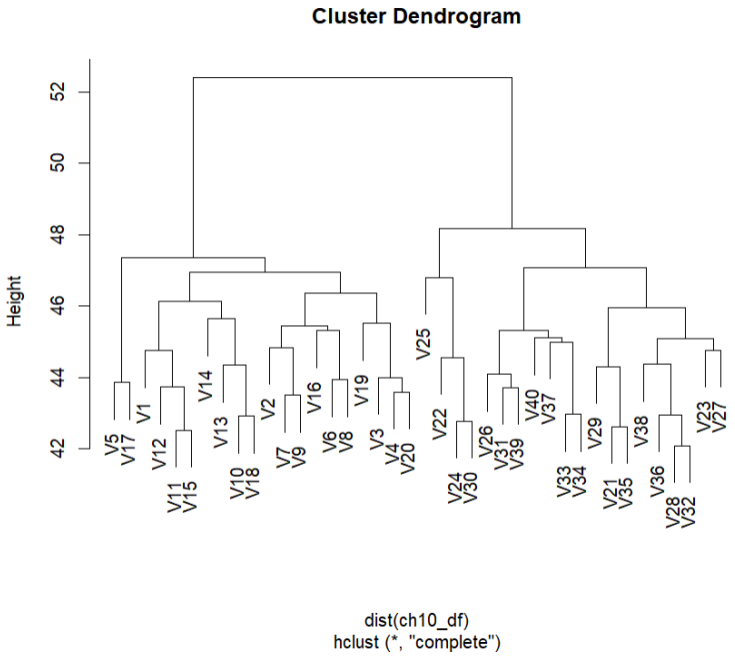
(a) Load in the data using read.csv(). You will need to select header = F.

```
> Ch10Ex11 <- read.csv("C:/Users/chase/Downloads/Ch10Ex11.csv", h
eader=FALSE)
>   View(Ch10Ex11)
```

(b) Apply hierarchical clustering to the samples (using Eucledian distance), and plot the dendrogram, for the following linkages:

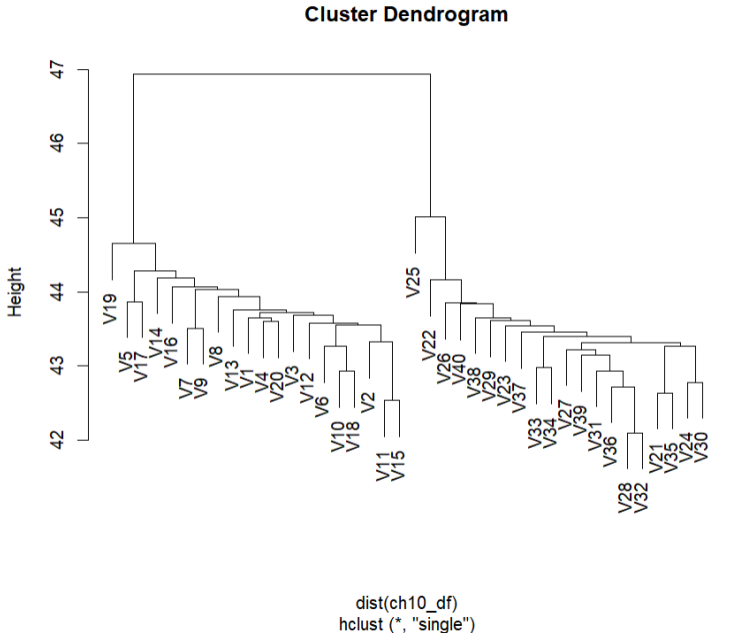
i. Complete.

```
> ch10_df <- t(Ch10Ex11)
> comp = hclust(dist(ch10_df), method = "complete")
> plot(comp)
```



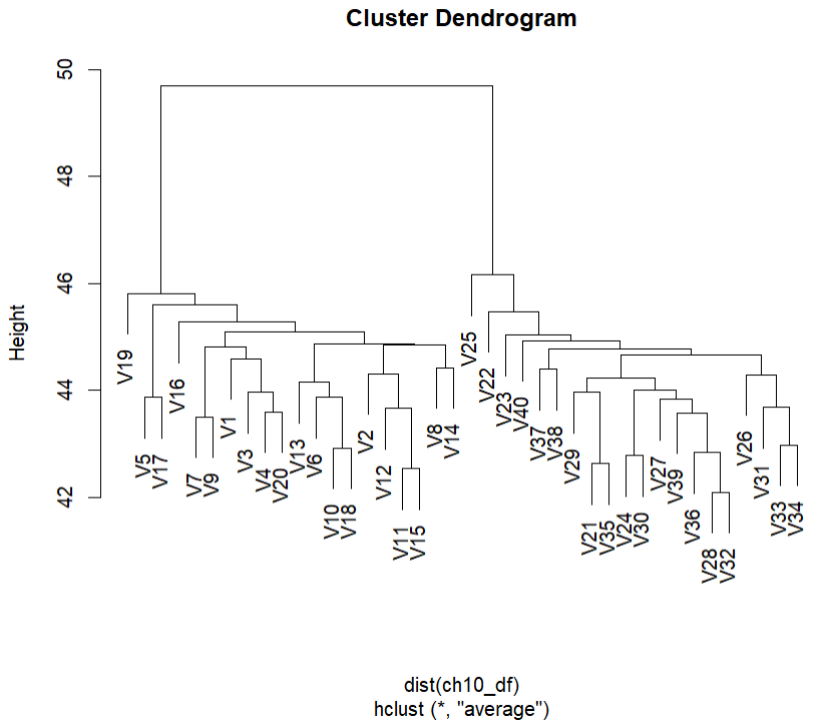
ii. Single.

```
> sing = hclust(dist(ch10_df), method = "single")
> plot(sing)
```



iii. Average

```
> avg <- hclust(dist(ch10_df), method = "average")
> plot(avg)
```



Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?

Yes, the genes do separate the samples into the two groups. No, my results do not depend on the type of linkage that I used because all are separated into two groups.