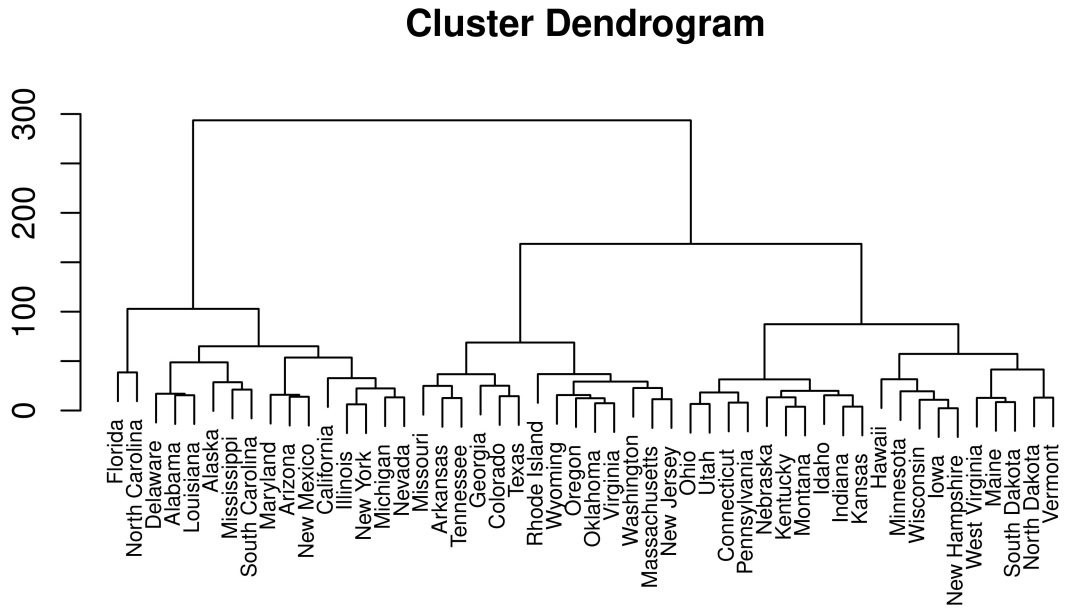
HW7_Q9

9a. Consider the USArrests data. We will now perform hierarchical clustering on the states.

```
data = USArrests
set.seed(1)

cluster = hclust(dist(data, method = "euclidean"), method = "complete")
plot(cluster, cex = 0.65, xlab = "", ylab = "", sub="")
```



9b. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

```
clusters3 = cutree(cluster, 3)
clusters3
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
```

```
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

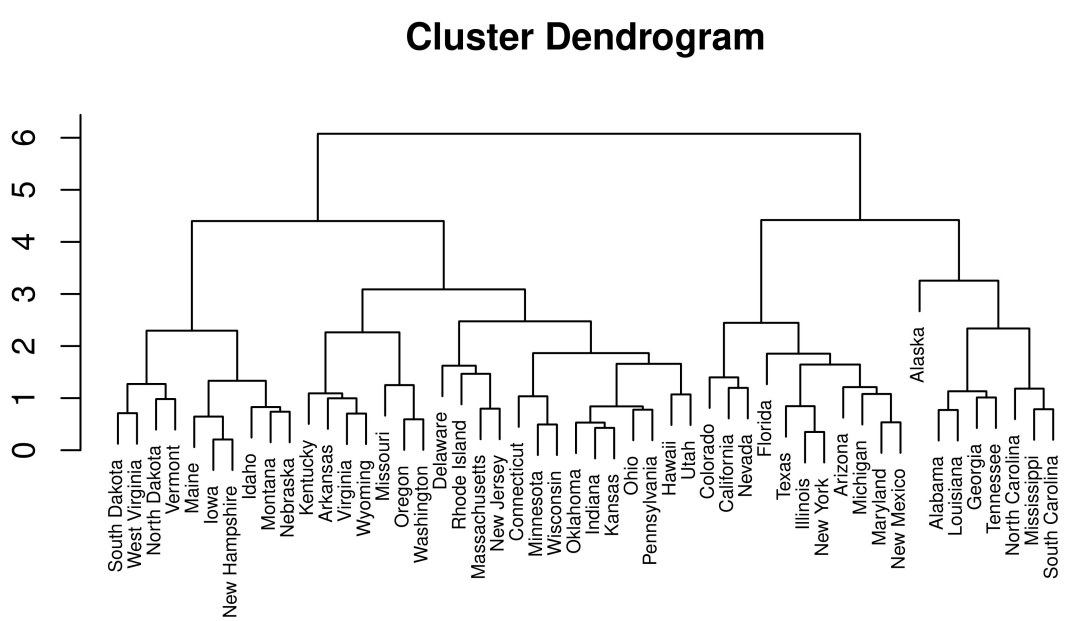
```
table(clusters3)
```

```
## clusters3
##  1  2  3
## 16 14 20
```

9c. Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

```
scaled = scale(USArrests)
set.seed(1)

cluster_scaled = hclust(dist(scaled, method = "euclidean"), method = "complete")
plot(cluster_scaled, cex = 0.6, xlab = "", ylab = "", sub="")
```




```
clusters3_scaled = cutree(cluster_scaled, 3)
clusters3_scaled
```

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

```
table(clusters3_scaled)
```

```
## clusters3_scaled
##  1  2  3
##  8 11 31
```

9d. 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.

Scaling the variables so that the standard deviation = 1, caused a shift in which states were assigned to which cluster. Previously, the states were separated 16, 14, 20 in clusters 1, 2, 3 respectively, whereas now it is at 8, 11, 31. For this dataset, there was increase in dissimilarity.