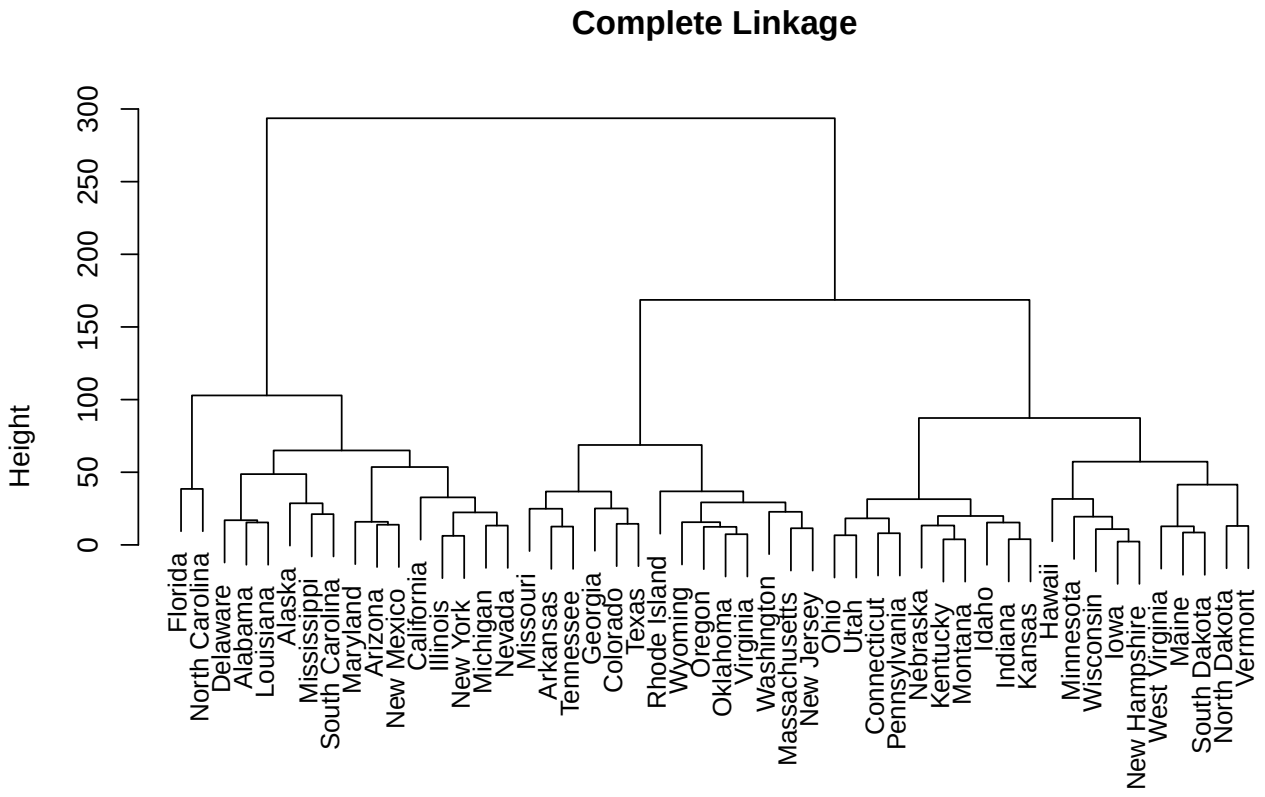
# STATS 415 - Homework 10 - Clustering

*Marian L. Schmidt*

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

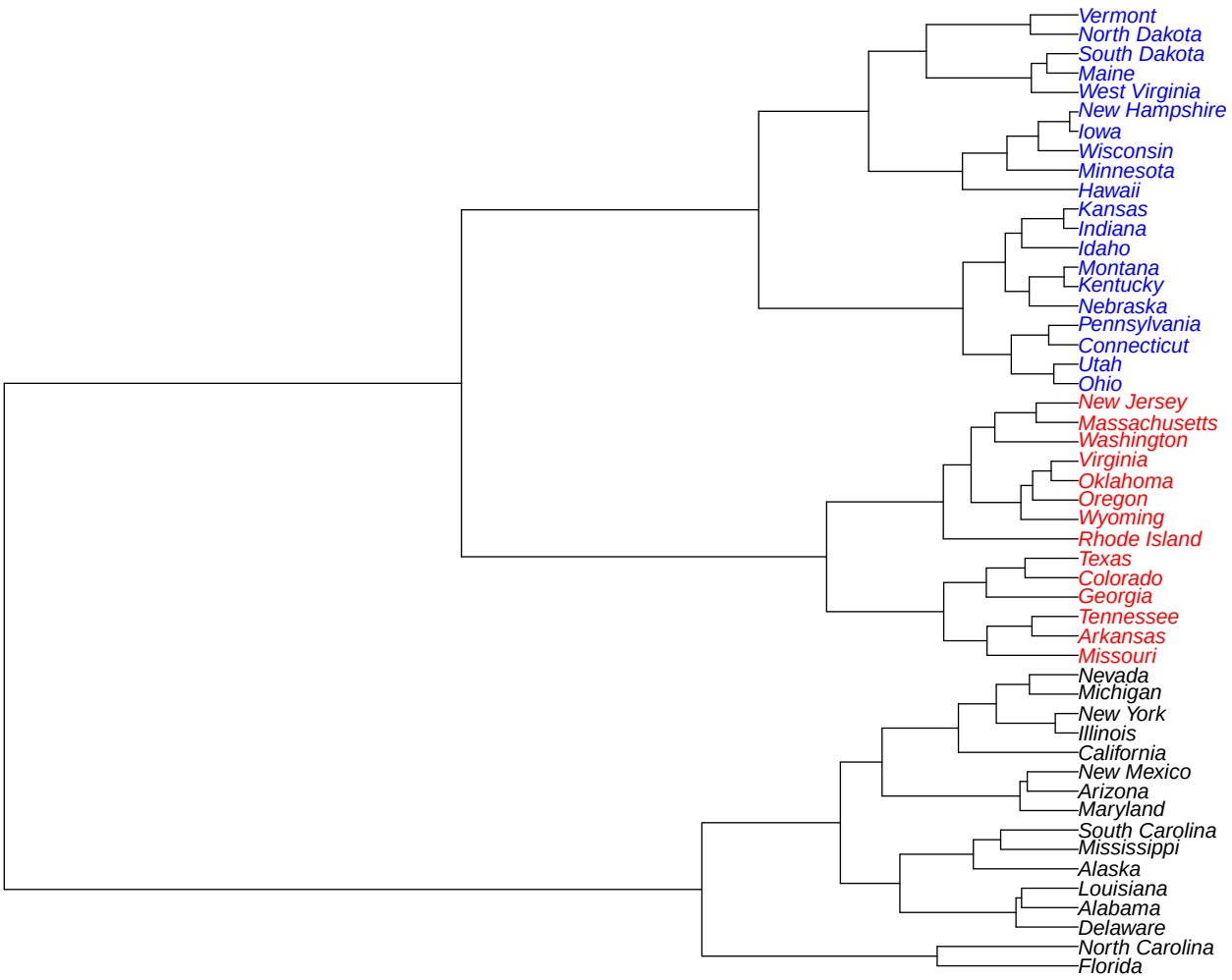
**(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.**

```
hc_complete <- hclust(dist(USArrests, method = "euclidean"), method="complete")
plot(hc_complete,main="Complete Linkage", xlab="", sub="", cex=.9)
```




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

```
clusters <- cutree(hc_complete, 3)
clust_1 <- clusters[clusters == 1]; # pull out the names of the states
clust_2 <- clusters[clusters == 2]; clust_3 <- clusters[clusters == 3]
mypal = c("black", "red", "blue")
plot(as.phylo(hc_complete), tip.color = mypal[cutree(hc_complete, 3)], main = "Complete Linkage")
```




***Above we see the three clusters include the following states:***

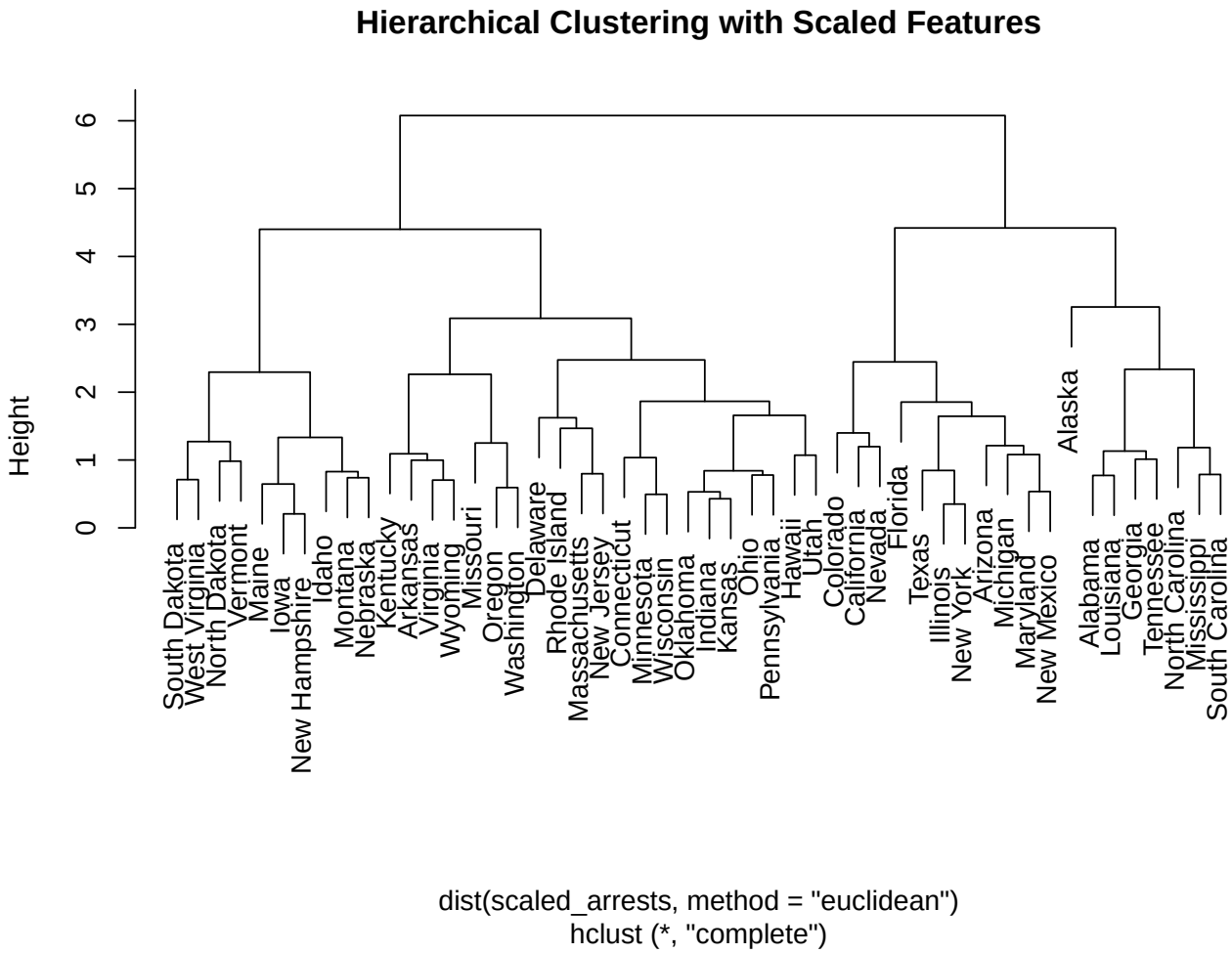
- ***First cluster:*** *Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, South Carolina*

- ***Second cluster:*** *Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming*

- ***Third cluster:*** *Connecticut, Hawaii, Idaho, Indiana, Iowa, Kansas, Kentucky, Maine, Minnesota, Montana, Nebraska, New Hampshire, North Dakota, Ohio, Pennsylvania, South Dakota, Utah, Vermont, West Virginia, Wisconsin*

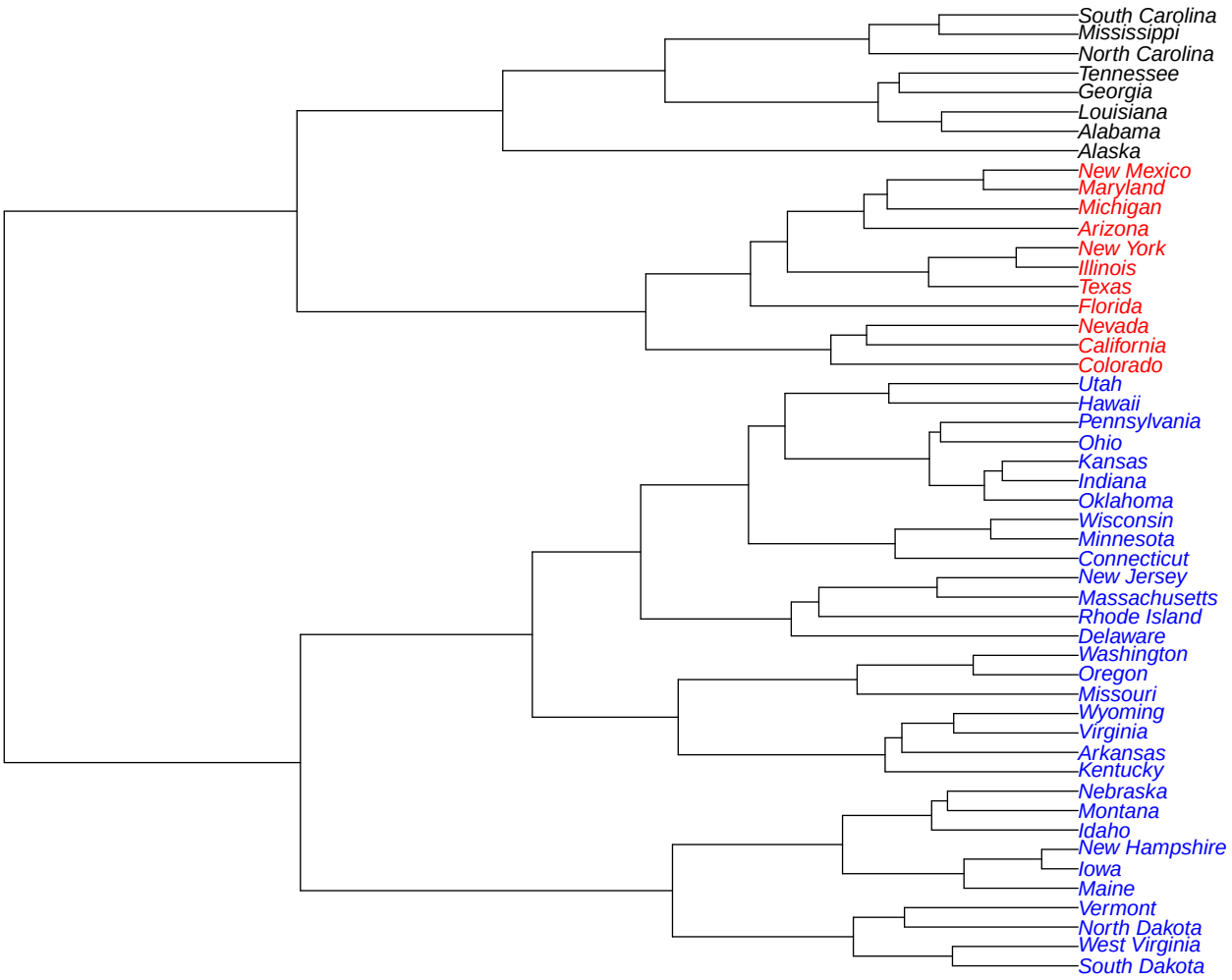
**(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. Now which states belong to which clusters?**

```
scaled_arrests <- scale(USArrests)
hc_scaled <- hclust(dist(scaled_arrests, method = "euclidean"), method="complete")
plot(hc_scaled, main="Hierarchical Clustering with Scaled Features")
```




```
plot(as.phylo(hc_scaled), tip.color = mypal[cutree(hc_scaled, 3)], main = "Complete Linkage")
```

**Complete Linkage**

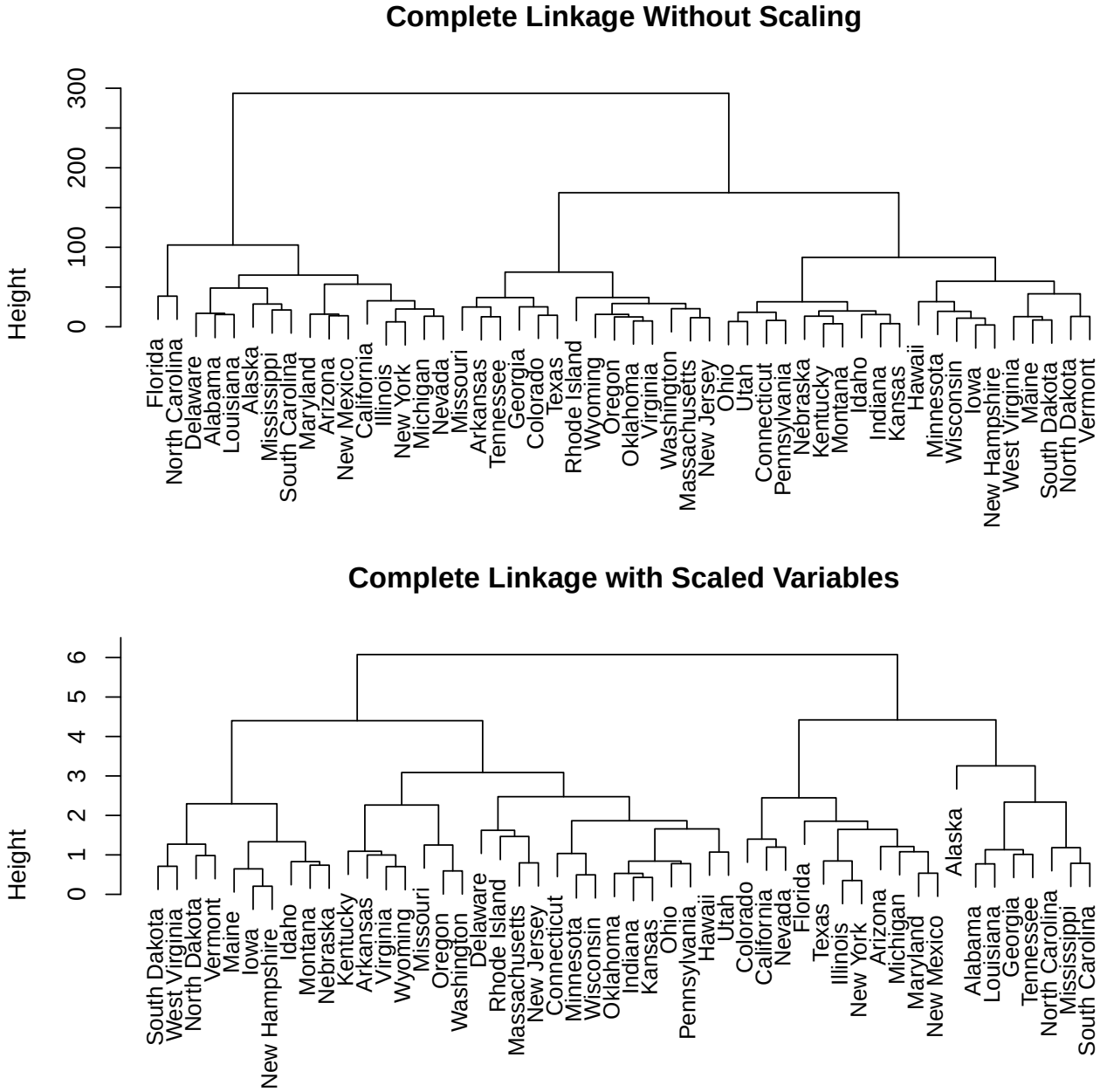



```
scaled_clusters <- cutree(hc_scaled, 3)
clust_1 <- scaled_clusters[scaled_clusters == 1]; # pull out the names of the states
clust_2 <- scaled_clusters[scaled_clusters == 2]; clust_3 <- scaled_clusters[scaled_clusters == 3]
```

***With scaling the variables to have a standard deviation of one, we now see that the three clusters include the following states:***

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

**(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.**




*Scaling the variables impacts the clusters that are obtained, the branch lengths, and the height of the tree. For example, without scaling, Michigan clusters with Nevada while with scaling Michigan clusters nearby Arizona. In addition, the height of the un-scaled tree is 300 while the height of the scaled tree is 6. Without scaling, we cut the tree at a height of ~150 whereas we cut the scaled tree at a height of ~4 to obtain 3 clusters. In addition, the branch for Alaska (and many other states) is shorter in the scaled tree.*

*In this scenario, scaling is more appropriate because* `Murder`*,* `Assault`*, and* `Rape` *all have unites of per 100,000 people while* `UrbanPop` *is the percentage of the state population that lives in urban areas. Therefore, it is imporant to scale so that the units of* `UrbanPop` *has an equal contribution to the hierarchical clustering algorithm as the other variables.*