# DSI_06- HW9 pg 550

Julia Gallucci

2023-03-13

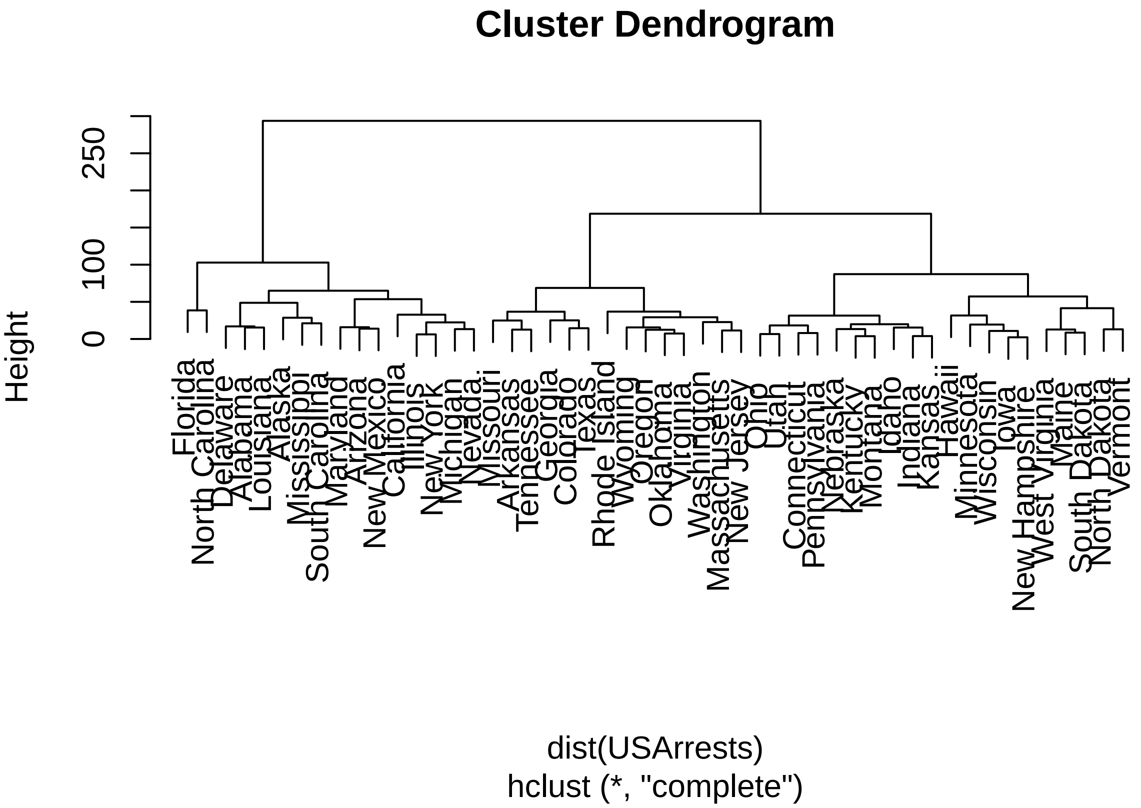
## 9. Consider the USArrests data. We will now perform hierarchical clustering on the states.

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

```
#install.packages("ISLR2")  Install package if you haven't already
library(ISLR2) #load library
attach(USArrests) #attach dataset
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

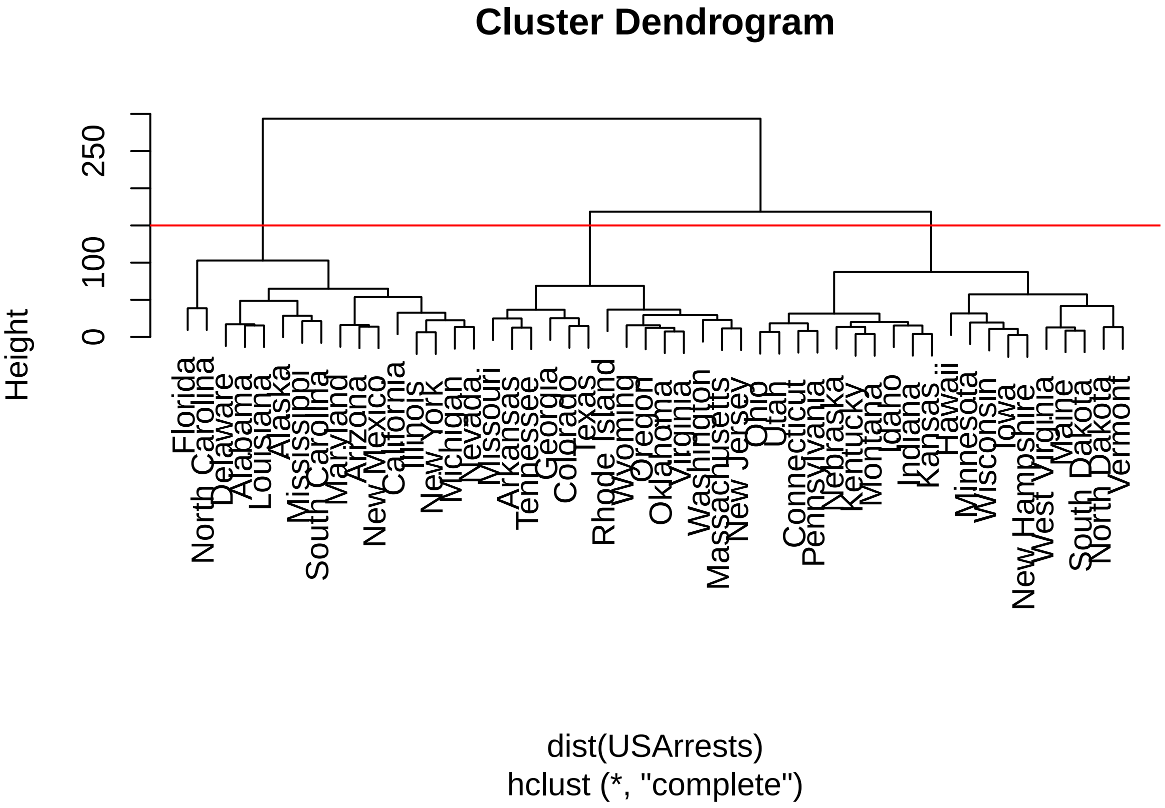
```
US.cluster = hclust(dist(USArrests), method="complete") #fit a hierarchical cluster using complete link
plot(US.cluster) #plot clustering
```



## (b)

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

```
plot(US.cluster) + abline(h=150, col = "red") #visualize 3 clusters
```



```
## integer(0)
```

```
US.clusters_3 <- cutree(US.cluster, k = 3)
print(US.clusters_3)
```

```
##        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
```

```
print(table(US.clusters_3))
```

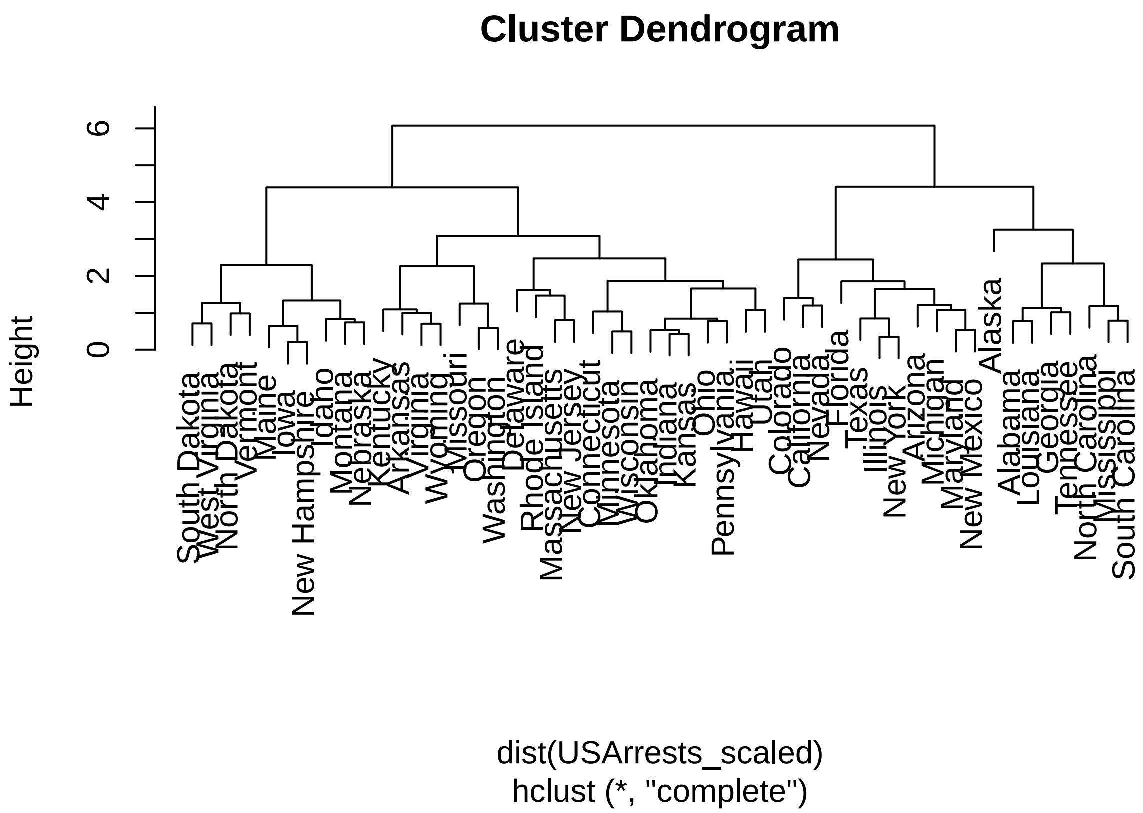
```
## US.clusters_3
##  1  2  3 
## 16 14 20
```

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

```
USArrests_scaled <- as.data.frame(scale(USArrests)) #scale all variables in USArrests dataframe
sd(USArrests_scaled$Murder) #confirm scaling worked!
```

```
## [1] 1
```

```
US.cluster_scaled = hclust(dist(USArrests_scaled), method="complete") #fit a hierarchical cluster using
plot(US.cluster_scaled) #plot clustering
```



## (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.

```
US.clusters_3_scaled <- cutree(US.cluster_scaled, k = 3)
print(US.clusters_3_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
```

```
print(table(US.clusters_3_scaled))
```

```
## US.clusters_3_scaled
##  1  2  3
##  8 11 31
```

```
table(US.clusters_3, US.clusters_3_scaled)
```

```
##               US.clusters_3_scaled
## US.clusters_3  1  2  3
##              1  6  9  1
##              2  2  2 10
##              3  0  0 20
```

```
same_membership <- (6 + 2 + 20) / 50
```

Scaling of the variables does indeed change the cluster membership of certain states! It appears as though only 56% of states were assigned to the same membership when comparing scaled and non-scaled data.Scaling maybe useful for this dataset, given UrbanPop is recorded as a different unit (Percent), compared to Murder, Assault and Rape which are reported as per 100,000. (hint: use ?USArrests)