# Assignment 6 By Team 2

*Shun-Lung Chang, Deepika Ganesan, Deepankar Upadhyay*

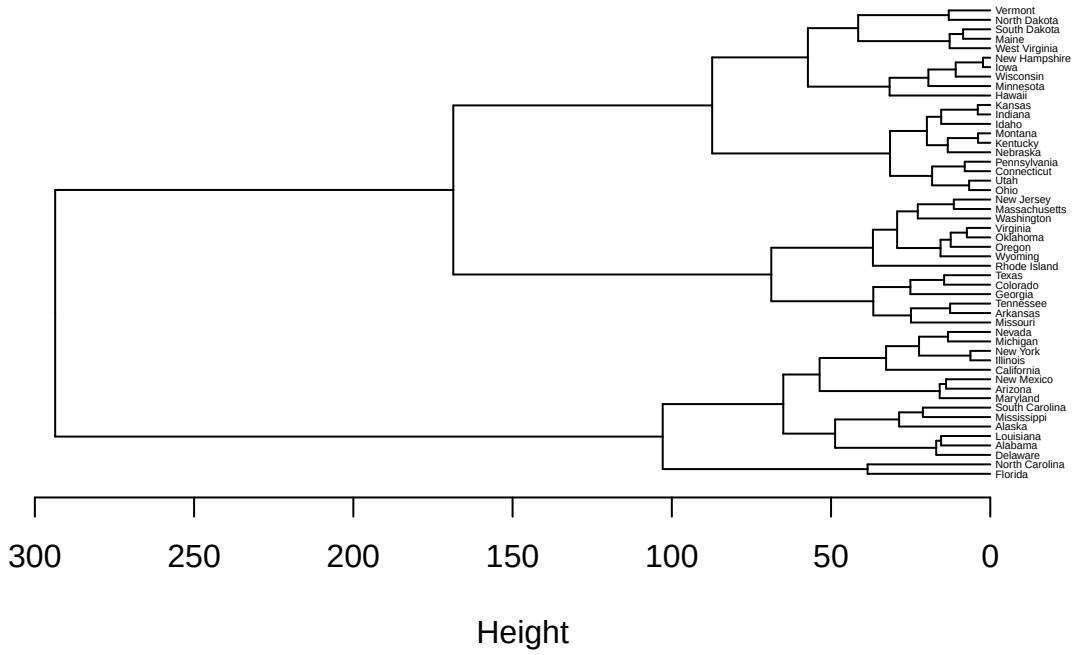
The analysis was construted in R, and the code can be found *here*.

## 1. First, perform hierarchical clustering on the states.

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

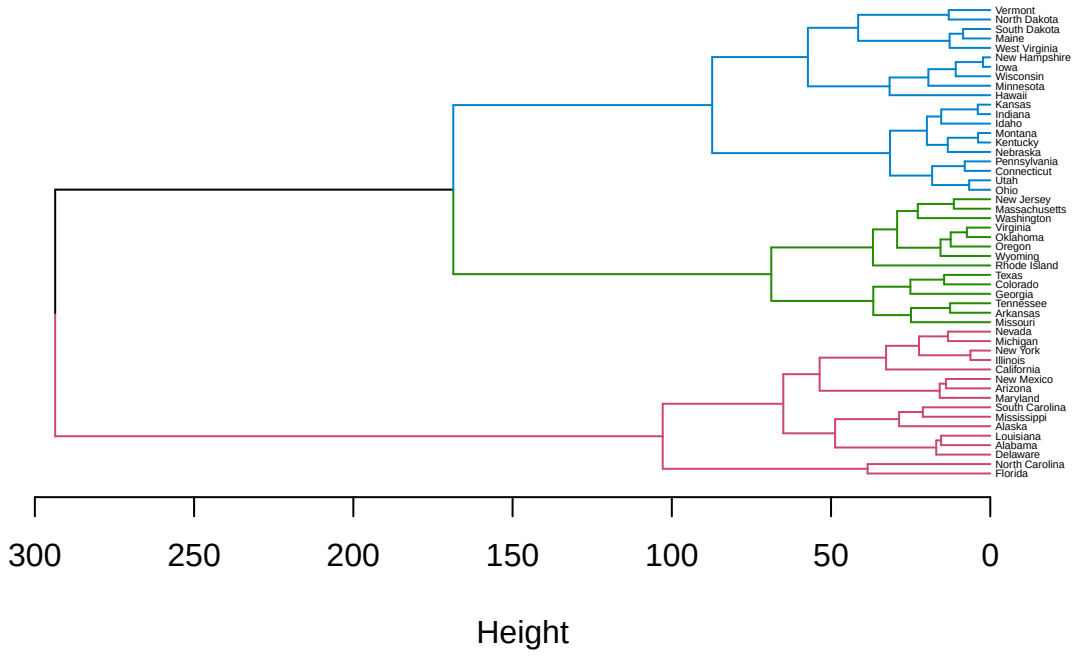
To perform hierarchical clustering (`hclust()` function), a distance matrix, which measures similarity between elements of a set, is required. In R, `dist()` function (default measure is **Euclidean Distance**) can assist us in deriving the matrix. After a distance matrix is computed, the `hclust()` function can then be applied to the matrix, and the result is shown as the dendrogram below.

```
hc <- USArrests %>%
        dist() %>%
        hclust()
```




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

The three distinct clusters can be acquired by cutting the dendrogram at the height of 170 as the plot below. The three cluster are labeled in different color (blue, green, and red) and states in different clusters are also listed.




1. Group Blue

```
 [1] "Connecticut"   "Hawaii"        "Idaho"          "Indiana"
 [5] "Iowa"          "Kansas"        "Kentucky"       "Maine"
 [9] "Minnesota"     "Montana"       "Nebraska"       "New Hampshire"
[13] "North Dakota"  "Ohio"          "Pennsylvania"   "South Dakota"
[17] "Utah"          "Vermont"       "West Virginia"  "Wisconsin"
```

2. Group Green

```
 [1] "Arkansas"      "Colorado"      "Georgia"       "Massachusetts"
 [5] "Missouri"      "New Jersey"    "Oklahoma"      "Oregon"
 [9] "Rhode Island"  "Tennessee"     "Texas"         "Virginia"
[13] "Washington"    "Wyoming"
```

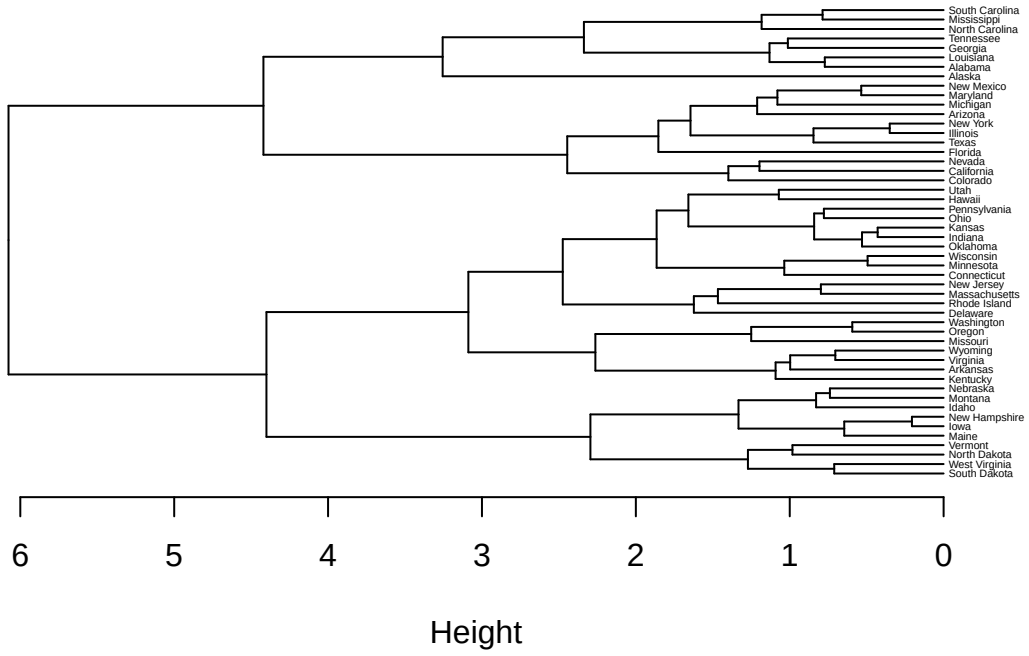
3. Group Red

```
 [1] "Alabama"        "Alaska"         "Arizona"        "California"
 [5] "Delaware"       "Florida"        "Illinois"       "Louisiana"
 [9] "Maryland"       "Michigan"       "Mississippi"    "Nevada"
[13] "New Mexico"     "New York"       "North Carolina" "South Carolina"
```

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

Standardised data can be obtained through `scale()` function. The dendrogram indicates the hierarchical tree of the standardised dataset.

```
scaled_USArrests <- scale(USArrests)
hc_scaled <- hclust(dist(scaled_USArrests))
```



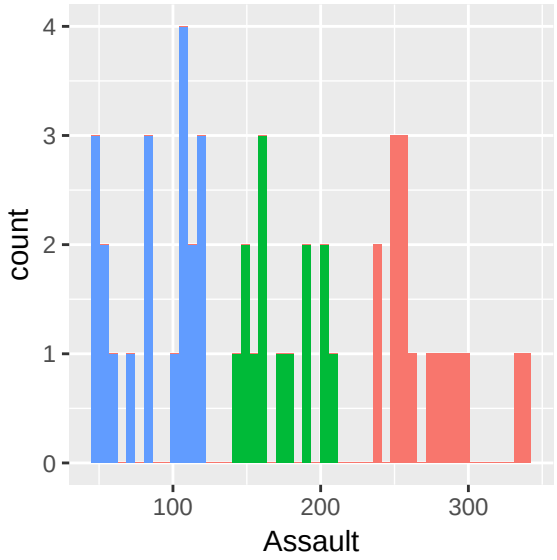

**d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer**

The heights in the two dendrograms differ significantly. The difference could result from the `Assault` variable, since its standard deviation is considerably higher than that of the other three variables.

```
apply(USArrests, 2, sd)
```

```
   Murder   Assault  UrbanPop      Rape
 4.355510 83.337661 14.474763  9.366385
```

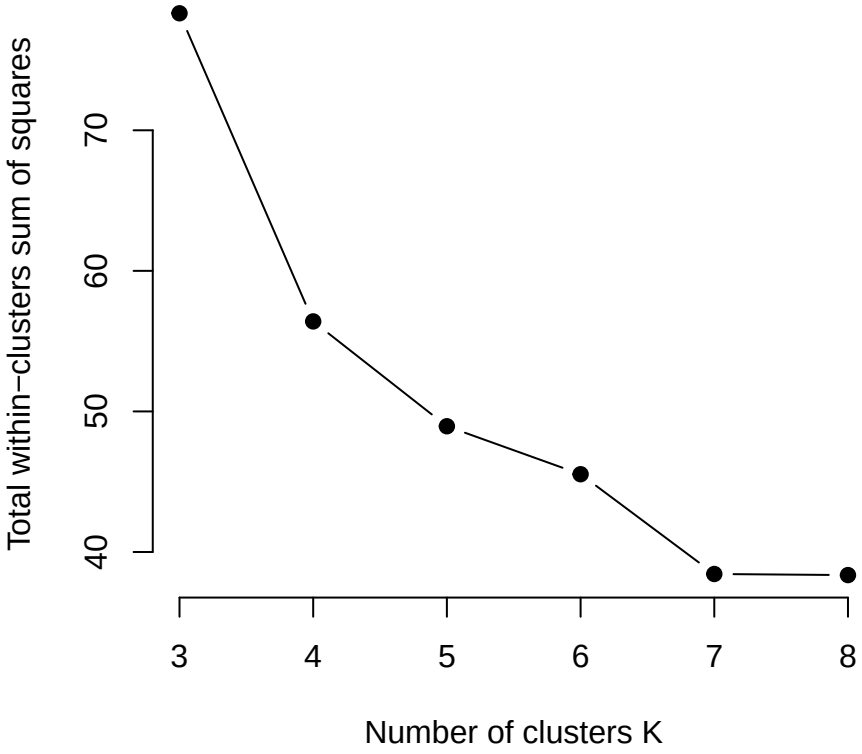
The histogram also reveals that the `Assault` variable dominates the clusters derived from the original dataset.




In light of the case above, the dataset should be standardised if it contains variables with high variance.

## 2. Perform k-means clustering, selecting a suitable range for k. Compare the results with the ones from question 1

In this study, the cluster values from 3 to 8 are chosen, and we picked the desired number through the elbow method. As can be seen from the elbow diagram, there exists solely slight drop in the within-cluster sum of squares from 7 to 8. Therefore, we selected 7 as the number of centers.




The following tables show the clusters derived from the two approaches. Though the order is different, a number of clusters are consistent, such as cluster with "Alabama", with "Alaska" and with "Arizona". But the members in other clusters differ given that they are grouped by two methods.

**States in each cluster (Hierachical clustering)**

```
Group 1

[1] "Alabama"        "Georgia"        "Louisiana"      "Mississippi"
[5] "North Carolina" "South Carolina" "Tennessee"

Group 2

[1] "Alaska"

Group 3

 [1] "Arizona"    "California" "Colorado"   "Florida"    "Illinois"
 [6] "Maryland"   "Michigan"   "Nevada"     "New Mexico" "New York"
[11] "Texas"

Group 4

[1] "Arkansas"   "Kentucky"   "Missouri"   "Oregon"     "Virginia"
[6] "Washington" "Wyoming"

Group 5
```

```
 [1] "Connecticut"  "Hawaii"       "Indiana"      "Kansas"
 [5] "Minnesota"    "Ohio"         "Oklahoma"     "Pennsylvania"
 [9] "Utah"         "Wisconsin"
```

```
Group 6
```

```
[1] "Delaware"      "Massachusetts" "New Jersey"    "Rhode Island"
```

```
Group 7
```

```
 [1] "Idaho"         "Iowa"          "Maine"         "Montana"
 [5] "Nebraska"      "New Hampshire" "North Dakota"  "South Dakota"
 [9] "Vermont"       "West Virginia"
```

**States in each cluster (Kmeans clustering)**

```
Group 1
```

```
[1] "Connecticut"   "Hawaii"        "Massachusetts" "New Jersey"
[5] "Ohio"          "Pennsylvania"  "Rhode Island"  "Utah"
[9] "Washington"
```

```
Group 2
```

```
[1] "Arizona"    "Florida"    "Illinois"   "Maryland"   "Michigan"
[6] "New Mexico" "New York"   "Texas"
```

```
Group 3
```

```
[1] "Alabama"        "Georgia"        "Louisiana"      "Mississippi"
[5] "North Carolina" "South Carolina" "Tennessee"
```

```
Group 4
```

```
[1] "Idaho"         "Iowa"          "Minnesota"     "Montana"
[5] "Nebraska"      "New Hampshire" "Wisconsin"
```

```
Group 5
```

```
[1] "Alaska"     "California" "Colorado"   "Nevada"
```

```
Group 6
```

```
[1] "Maine"         "North Dakota"  "South Dakota"  "Vermont"
[5] "West Virginia"
```

```
Group 7
```

```
 [1] "Arkansas" "Delaware" "Indiana"  "Kansas"   "Kentucky" "Missouri"
 [7] "Oklahoma" "Oregon"   "Virginia" "Wyoming"
```