Assignment 6 By Team 2

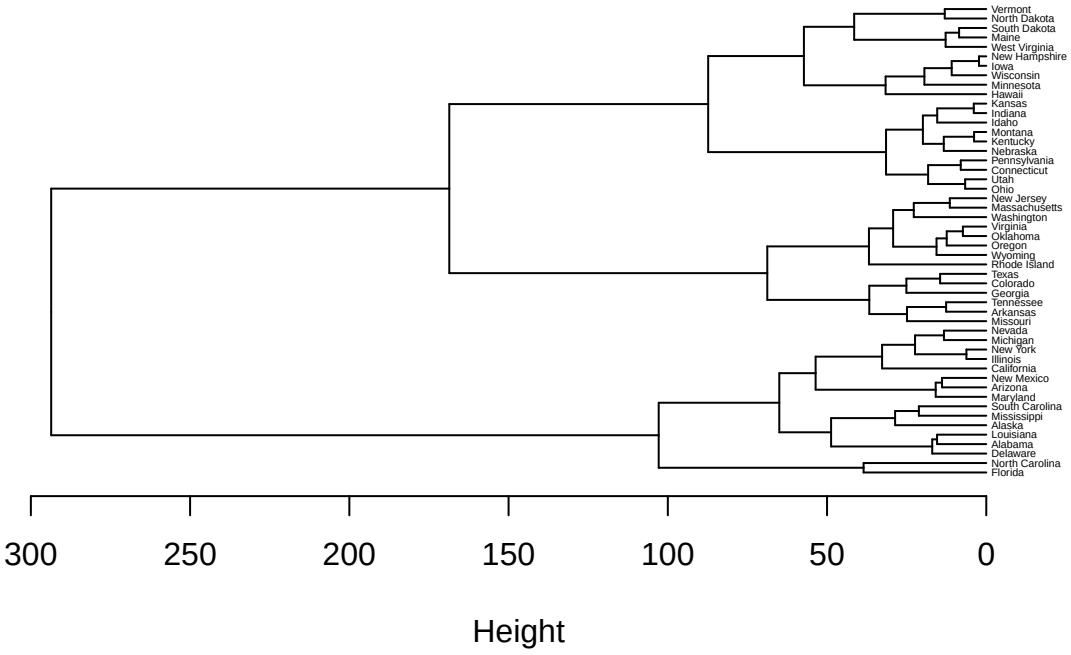
Shun-Lung Chang, Deepika Ganesan, Deepankar Upadhyay

1. First, perform hierarchical clustering on the states.

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

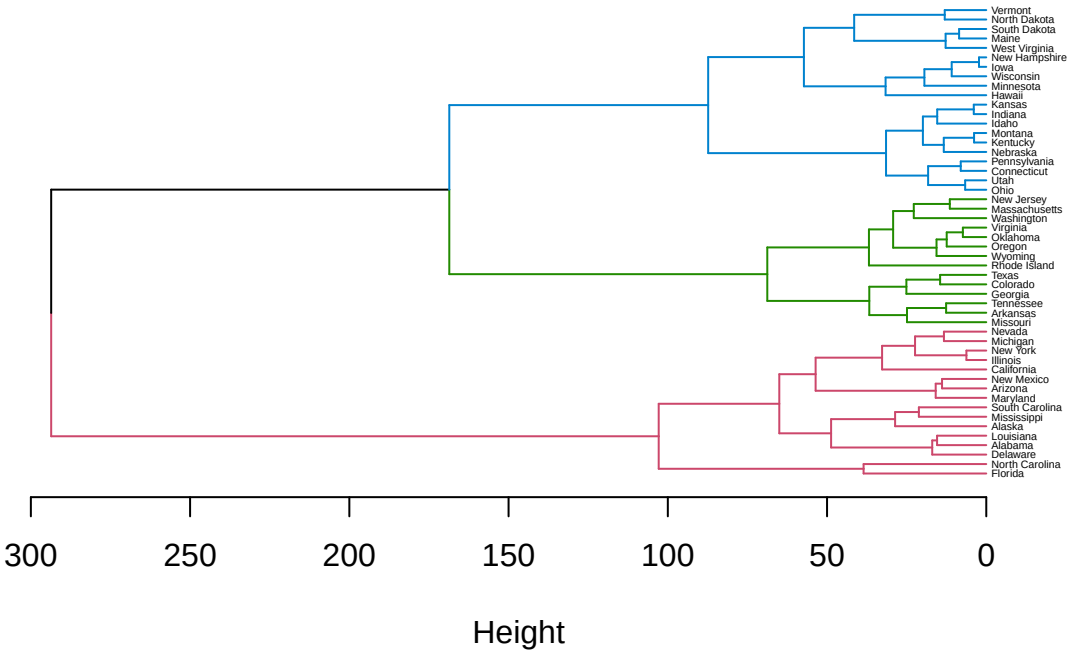
The `dist()` function is used to compute distance matrix. A distance matrix is a measure of similarity between two cases, based on set of numeric value. The default distance measure is the **Euclidean Distance**. The results are shown below.

```
hc <- USArrests %>%
        dist() %>%
        hclust()
dend <- hc %>% as.dendrogram()
dend %>%
    set("labels_cex", 0.3) %>%
    plot(horiz = TRUE, xlab = "Height")
```



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

The plot belows has shown that three clusters in different colors (blue, green, and red). We also show the various states that belong to each cluster.



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

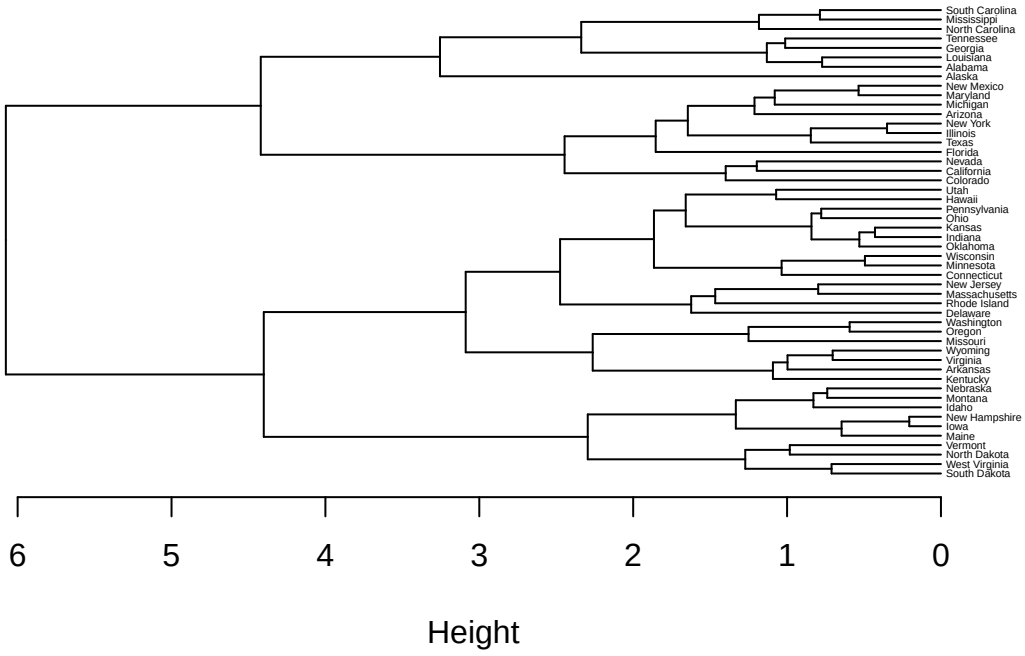
```
 [1] "Alabama"        "Alaska"         "Arizona"        "California"
 [5] "Delaware"       "Florida"        "Illinois"       "Louisiana"
 [9] "Maryland"       "Michigan"       "Mississippi"    "Nevada"
[13] "New Mexico"     "New York"       "North Carolina" "South Carolina"
```

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

After applying `scale()` function in R, we obtain a standardised representation of the data set. The results are shown below.

```r
scaled_USArrests <- scale(USArrests)
hc_scaled <- hclust(dist(scaled_USArrests))
dend_scaled <- hc_scaled %>% as.dendrogram()

dend_scaled %>%
    set("labels_cex", 0.3) %>%
    plot(horiz = TRUE, xlab = "Height")
```



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

Firstly we can see that the height in two dendrograms is significantly different. The difference could result from the `Assault` variable, since its variance is larger than that of the other three variables. Therefore, the result could dominated by this variable.

```
apply(USArrests, 2, var)
```

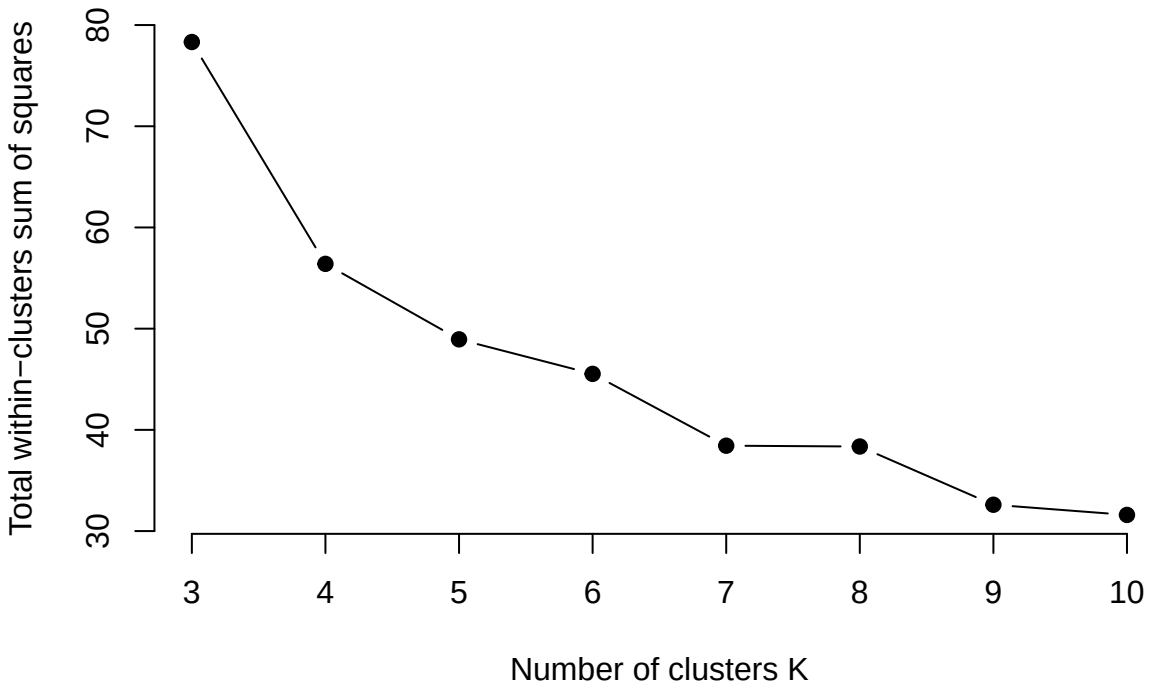
```
    Murder    Assault   UrbanPop       Rape
  18.97047 6945.16571  209.51878   87.72916
```

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

We choose the number of cluster from 3 to 10, and then see the elbow graph to see the change in total withinss between two cluster values (7 and 8) . . . .

```
set.seed(42)
k_max <- 10
wss <-
    sapply(3:k_max,  function(k) {kmeans(scaled_USArrests, k, iter.max = 50)$tot.withinss})

plot(3:k_max, wss,
     type = "b", pch = 19, frame = FALSE,
     xlab = "Number of clusters K",
     ylab = "Total within-clusters sum of squares")
```


Total within−clusters sum of squares
30
40
50
60
70
80
3
4
5
6
7
8
9
10
Number of clusters K