Assignment 6 By Team 2

Shun-Lung Chang, Deepika Ganesan, Deepankar Upadhyay

1. First, perform hierarchical clustering on the states.

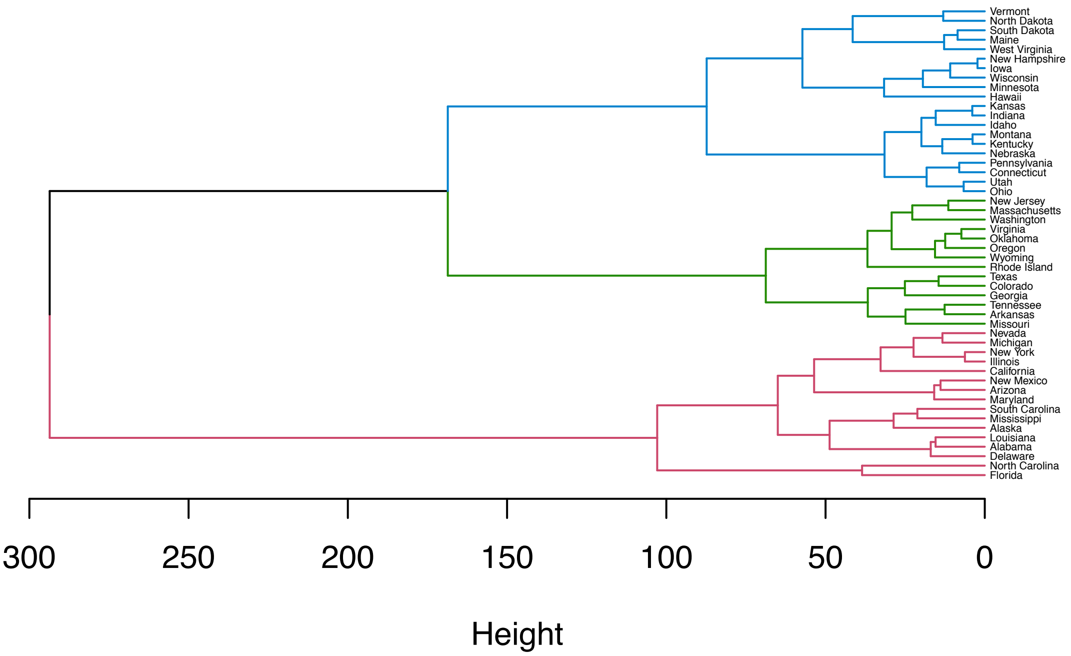
a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states

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

```
Call:
hclust(d = .)

Cluster method   : complete
Distance         : euclidean
Number of objects: 50
```

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



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"
```

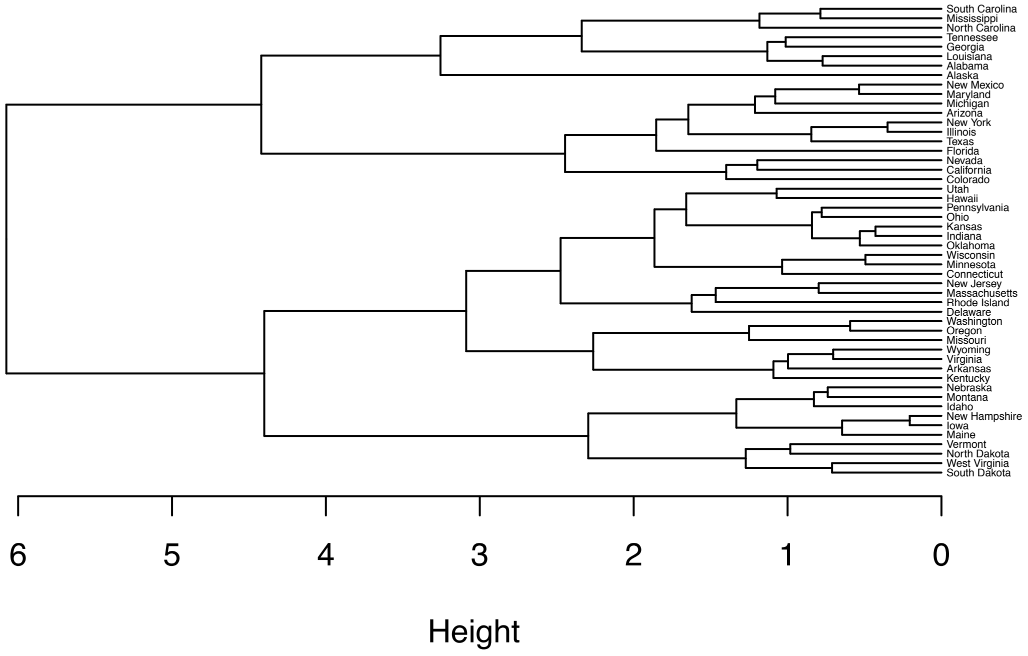
3. Group Red

```
 [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

```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")
```

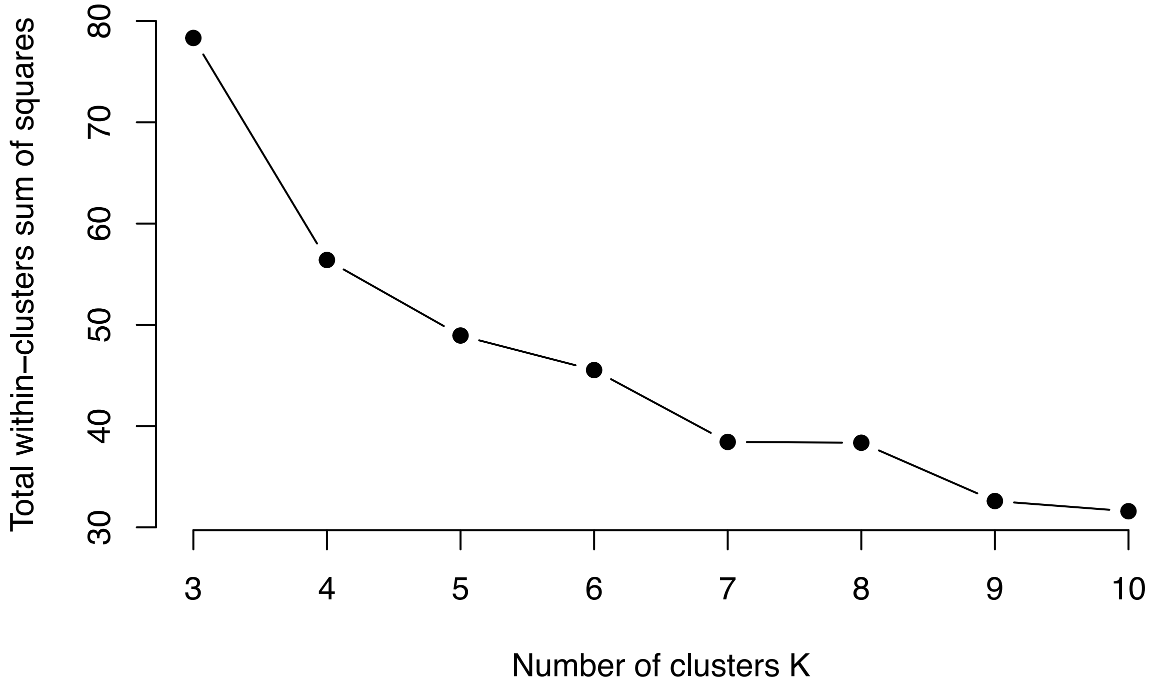


d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer

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

```
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
Number of clusters K