Assignment 6 By Team 2

Shun-Lung Chang, Deepika Ganesan, Deepankar Upadhyay

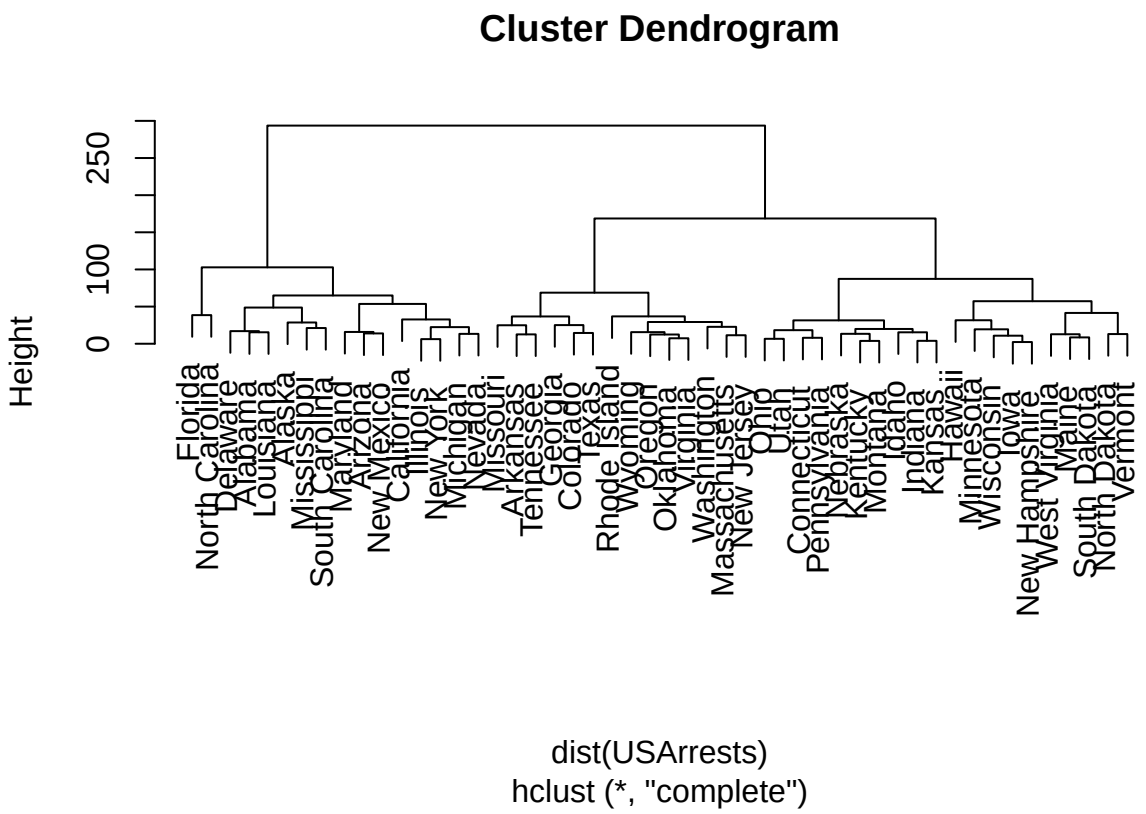
1. First, perform hierarchical clustering on the states.

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

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

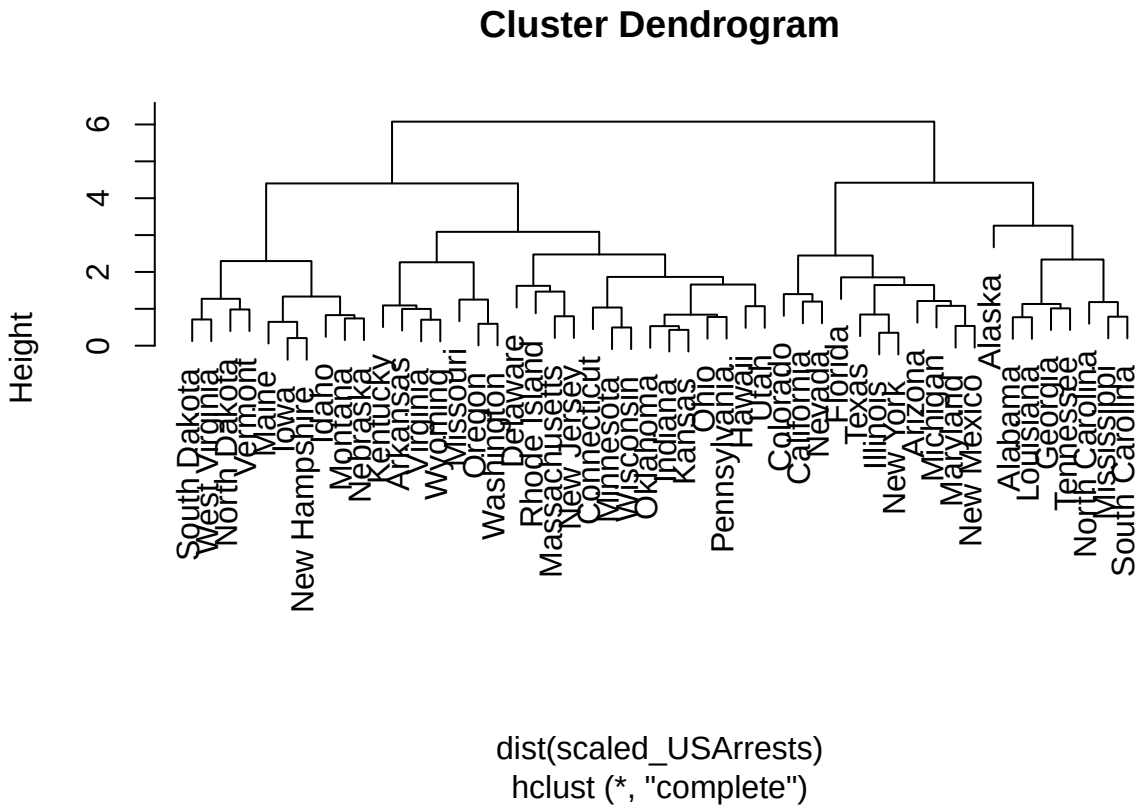
```
##
## Call:
## hclust(d = dist(USArrests))
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 50
```

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



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

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

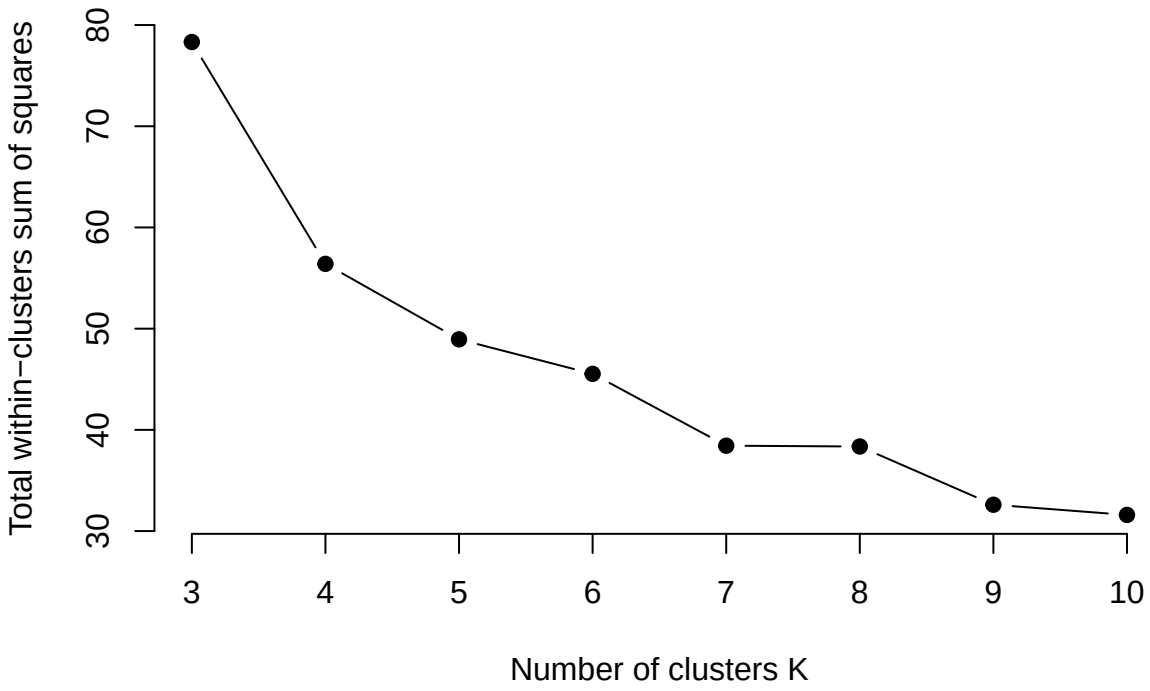


d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer

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
30
40
50
60
70
80
3
4
5
6
7
8
9
10
Number of clusters K