# Q43

Rahul Atre

2023-11-29

## Q43 [Clustering]

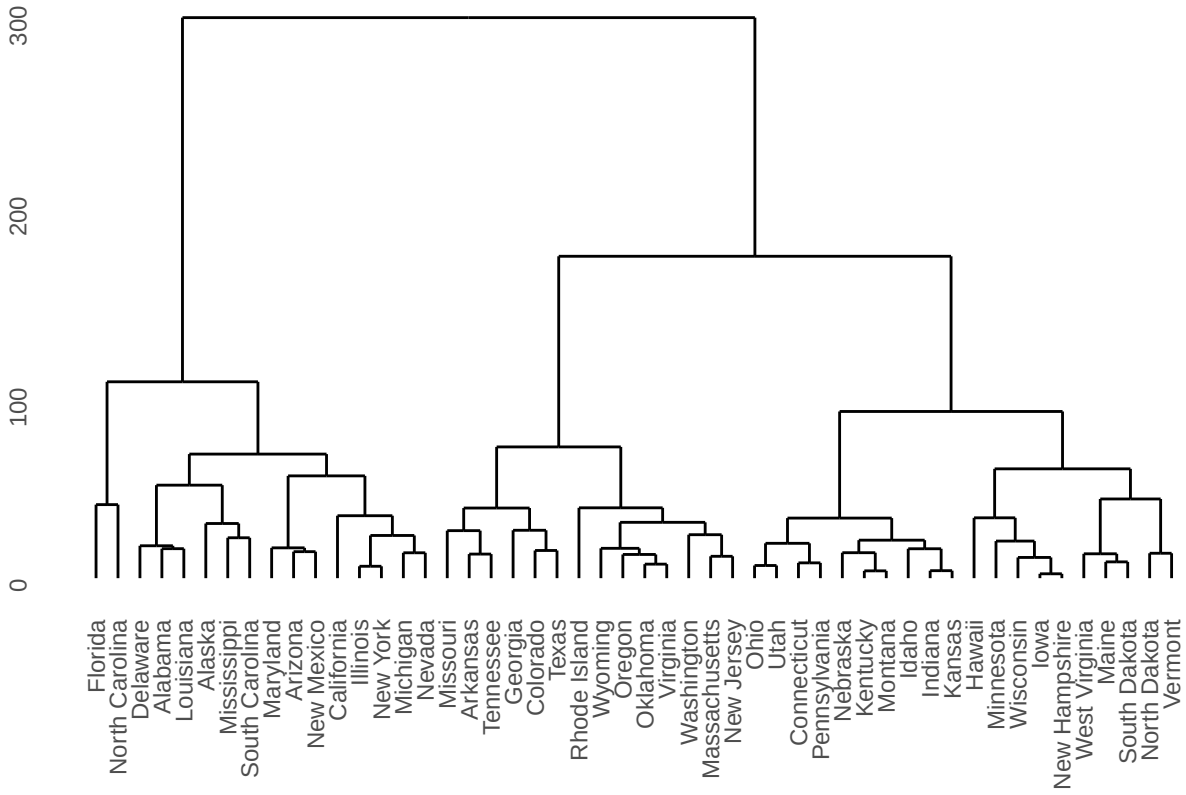
Consider the USArrests data. We perform hierarchical clustering on the states.

1. Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states (don't scale the data at first).

```
# Calculate the euclidean distance matrix
dist_matrix <- dist(USArrests)

# Perform hierarchical clustering with complete linkage
arrest_hier_cluster <- hclust(dist_matrix, method = "complete")

# Plot the dendrogram
ggdendrogram(arrest_hier_cluster)
```




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

```
# Cut the dendrogram into 3 clusters and assign to cluster_a
split_cluster <- cutree(arrest_hier_cluster, k = 3)
split_cluster
```

```
##        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
```

```
# Create a table of split_cluster
cluster_table <- table(split_cluster)
cluster_table
```

```
## split_cluster
##  1  2  3
## 16 14 20
```

Above we have generated all the states that belong to the three distinct clusters that were cut from the dendrogram. The table gives us the number of states that are contained within each cluster.

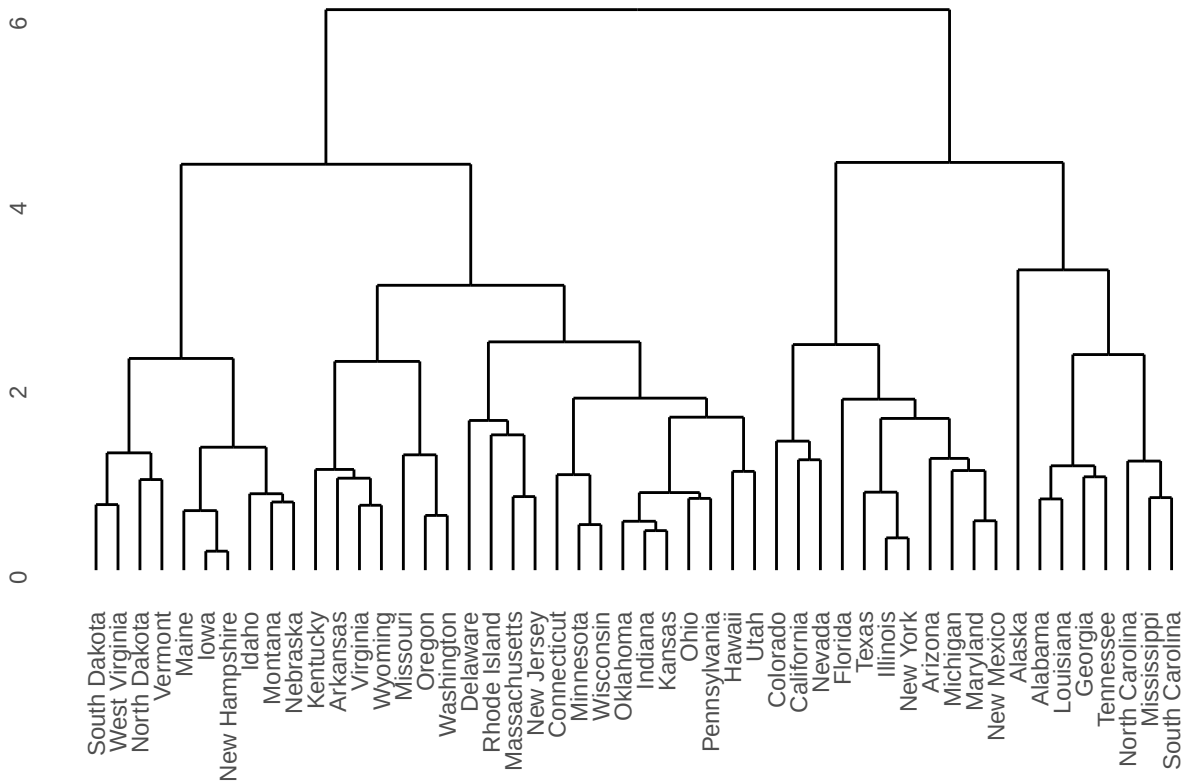
3. Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.

```
# Scale the data
scaled_data <- scale(USArrests)

# Calculate the distance matrix
dist_matrix <- dist(scaled_data)

# Perform hierarchical clustering with complete linkage
scale_cluster <- hclust(dist_matrix, method = "complete")

# Plot the dendrogram
ggdendrogram(scale_cluster)
```




4. 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.

In my opinion, I believe that the variables should be scaled before the inter-observation dissimilarities are computed due to the data being in different measurement of units (across variables). It is generally good practice to scale the data as we are unsure of the relationship between each of the variables, as we could risk giving parameters with larger magnitudes more importance. Overall, in this particular case, scaling the observations increases the dissimilarities, ensuring all variables have the same range and importance.