State-by-State Crime Analysis Using Hierarchical Clustering

Sejal Kankriya

2023-02-20

USArrests Dataset and Hierarchical Clustering

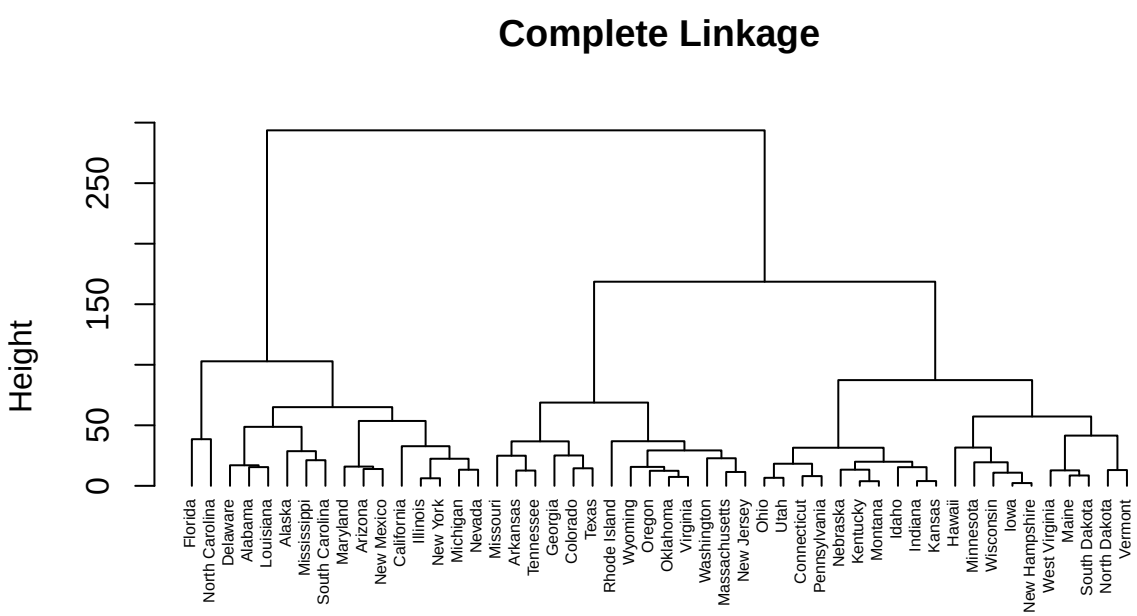
Consider the "USArrests" data. It is a built-in dataset you may directly get in RStudio. Performing hierarchical clustering on the observations (states)

```
head(USArrests)
```

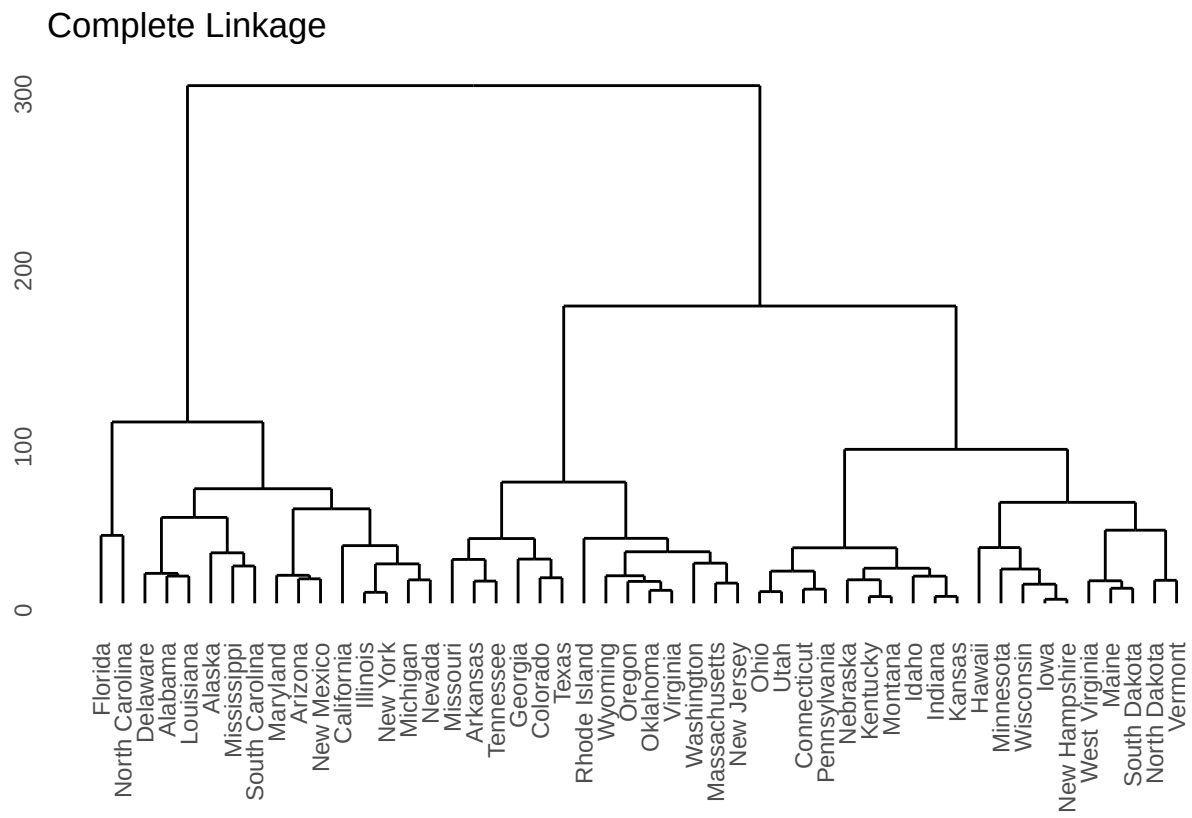
```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

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

```
set.seed(8)

data("USArrests")

# Load the data
data <- USArrests

# Omiting the missing values if any
data <- na.omit(data)

# Computing the distance matrix using euclidean method
distance_matrix <- dist(data, method = "euclidean")

# Performing hierarchical clustering
hc.complete <- hclust(distance_matrix, method = "complete")

# Plot the dendrogram
plot(hc.complete, main = "Complete Linkage",
    xlab = "", sub = "", cex = .5, hang = -1)
```




```
# ggplot
ggdendrogram(hc.complete, segements=TRUE, labels=TRUE, leaf_labels = TRUE, rotate=FALSE, theme_dendro =
 labs(title='Complete Linkage')
```



Cut the dendrogram at a height that results in three distinct clusters and interpreting the clusters

```
# Determining the cut tree
clusters <- cutree(hc.complete, 3)
clusters
```

```
##        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
```

```r
table(clusters)
```

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

```r
cat("\nStates belonging to Cluster 1\n")
```

```
##
## States belonging to Cluster 1
```

```r
subset(row.names(USArrests), clusters == 1)
```

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

```r
cat("\nStates belonging to Cluster 2\n")
```

```
##
## States belonging to Cluster 2
```

```r
subset(row.names(USArrests), clusters == 2)
```

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

```r
cat("\nStates belonging to Cluster 3\n")
```

```
##
## States belonging to Cluster 3
```

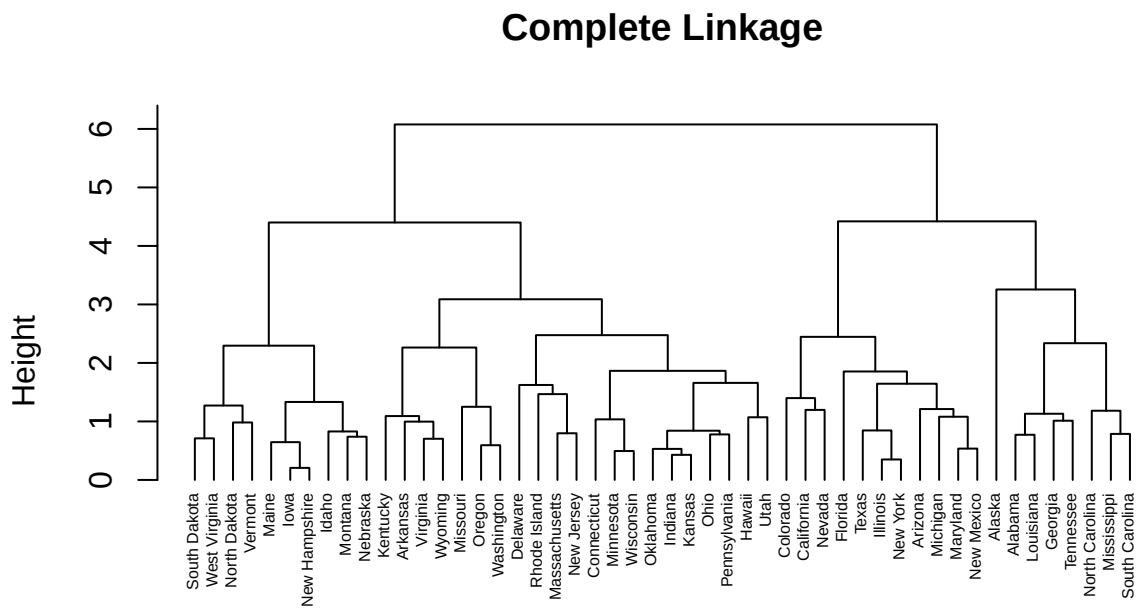
```r
subset(row.names(USArrests), clusters == 3)
```

```
##  [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"
```

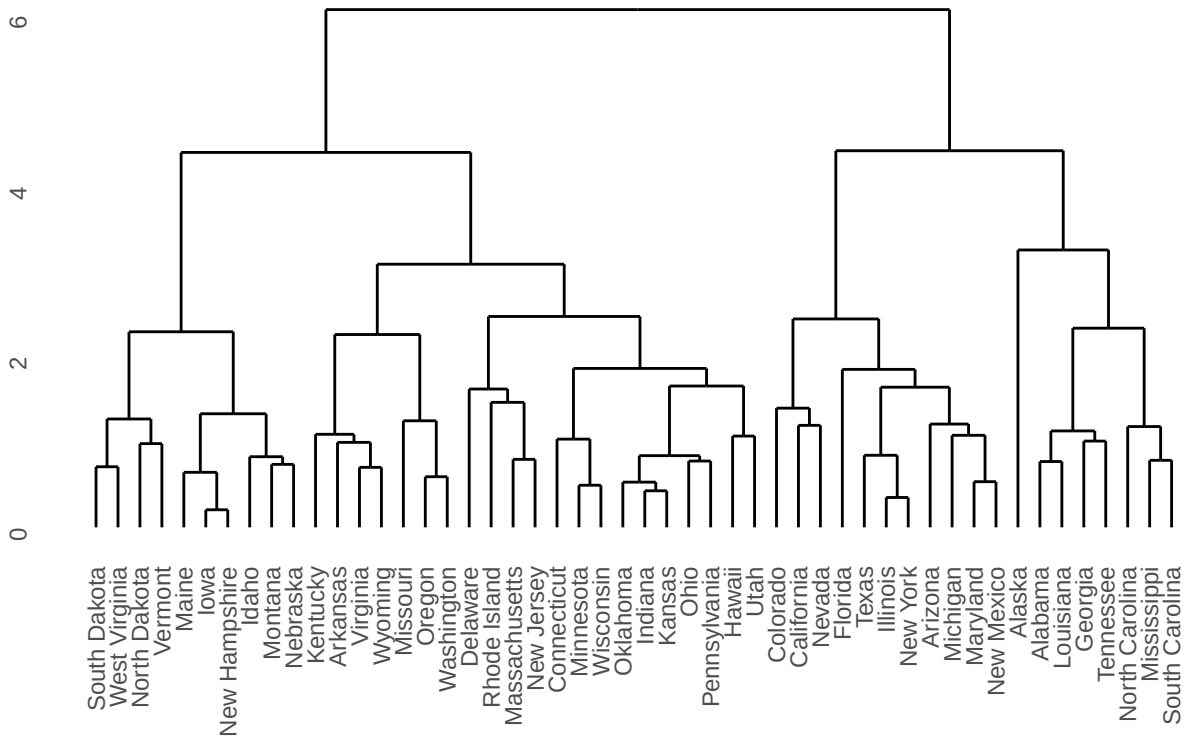
Cluster 1 contains the states with higher levels of violent crimes and arrests, such as California, New York, Florida, and Illinois. Cluster 2 includes states that have moderate levels of violent crimes and arrests, such as Arkansas, Georgia, and Tennessee. Finally, Cluster 3 consists of states with lower levels of violent crimes and arrests, such as Maine, Montana, and Vermont.

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

```r
set.seed(6)
data("USArrests")
# Load the data
data <- USArrests
# Omiting the missing values if any
data <- na.omit(data)
# Standardizing the data before clustering
data_scaled <- scale(data)
# Computing the distance matrix using euclidean method
distance_matrix <- dist(data_scaled, method = "euclidean")
# Performing hierarchical clustering
hc.complete_scaled <- hclust(distance_matrix, method = "complete")
# Plot the dendrogram
plot(hc.complete_scaled, main = "Complete Linkage",
    xlab = "", sub = "", cex = .5, hang = -1)
```




```r
# ggplot
ggdendrogram(hc.complete_scaled, segements=TRUE, labels=TRUE, leaf_labels = TRUE, rotate=FALSE, theme_d
 labs(title='Complete Linkage')
```




```
# Determining the cut tree
scaled_cutree <- cutree(hc.complete, 3)
scaled_cutree
```

```
##        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
```

```
table(scaled_cutree)
```

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

```
cat("\nStates belonging to Cluster 1\n")
```

```
##
## States belonging to Cluster 1
```

```
subset(row.names(USArrests), scaled_cutree == 1)
```

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

```
cat("\nStates belonging to Cluster 2\n")
```

```
##
## States belonging to Cluster 2
```

```
subset(row.names(USArrests), scaled_cutree == 2)
```

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

```
cat("\nStates belonging to Cluster 3\n")
```

```
##
## States belonging to Cluster 3
```

```
subset(row.names(USArrests), scaled_cutree == 3)
```

```
##  [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"
```

The 50 states in the USArrests dataset have been separated into three distinct clusters based on their similarities and differences, according to the output of the cutree() tool.

Alabama, Alaska, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, and Tennessee are all part of Cluster 1. When the clusters are interpreted, we can observe that Cluster 1 predominantly consists of states in the United States' Southeastern area, such as Alabama, Georgia, and Louisiana, which have

relatively high crime rates across all four categories examined in the dataset (assault, murder, rape, and robbery).

Arizona, California, Colorado, Florida, Illinois, Maryland, Michigan, Nevada, New Mexico, New York, and Texas are part of Cluster 2. These states have intermediate rates of violent crime, murder, rape and assault arrest rates.

Cluster 3 consists of the remaining 32 states, which had lower rates of violent crime and arrest rates for murder, rape, and assault when compared to the other states in the dataset.