Q43-WilliamKennedy-300015367

William Kennedy

2023-12-02

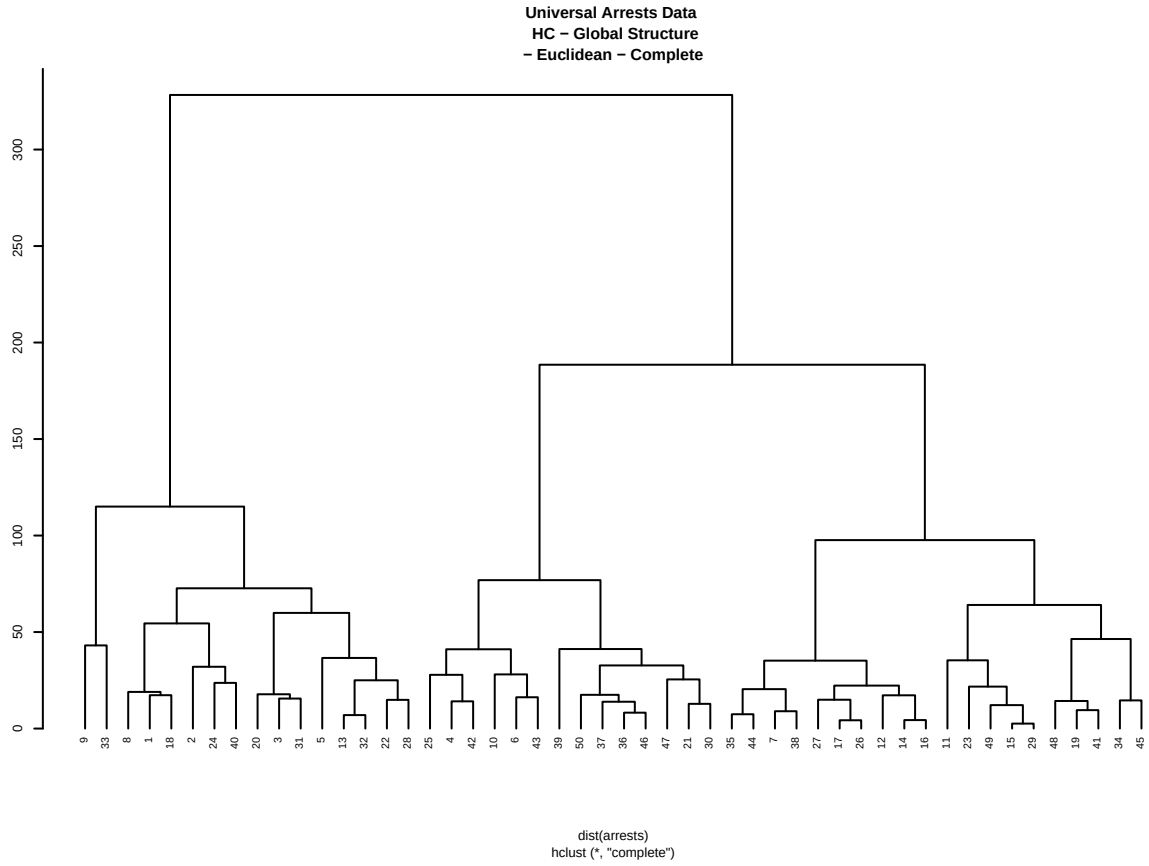
```
arrests = read.csv("USArrests.csv")

arrests.noX = arrests[,c(-1)]
```

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

```
par(cex=0.45)
hclust.arrests = hclust(dist(arrests))
```

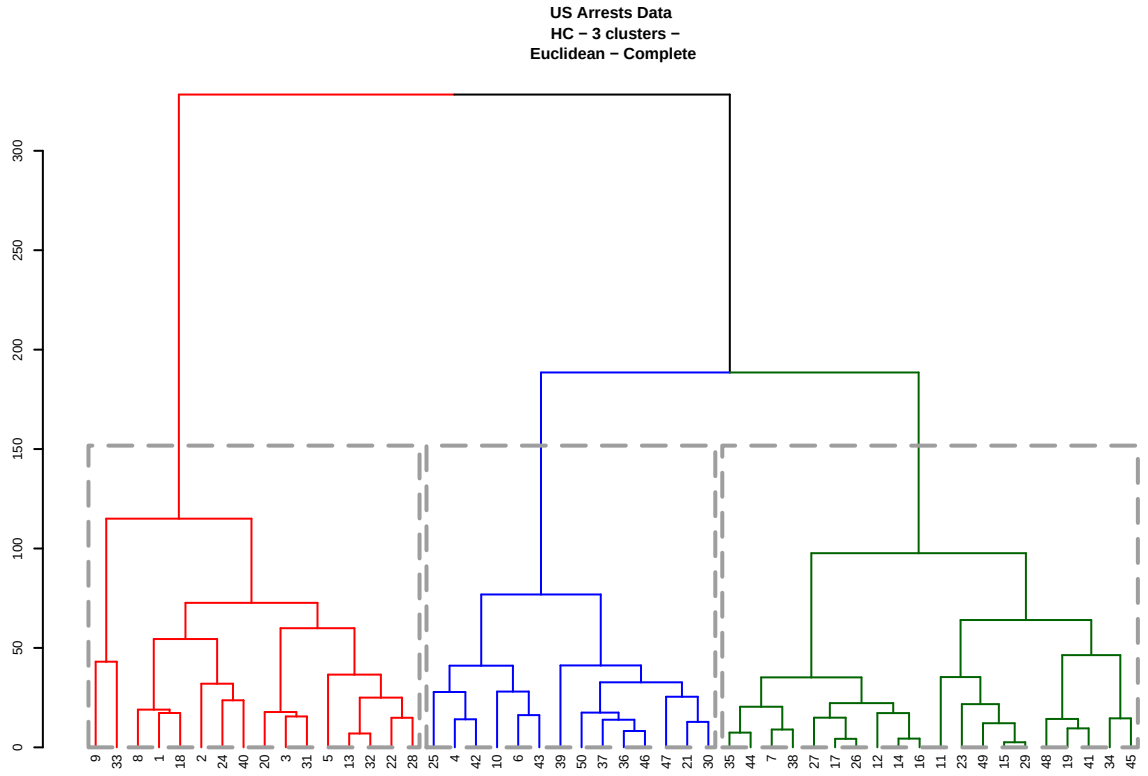
```
## Warning in dist(arrests): NAs introduced by coercion
```

```
plot(hclust.arrests, hang = -1, cex=0.7,
main = "Universal Arrests Data \n HC - Global Structure
- Euclidean - Complete", ylab="")
```



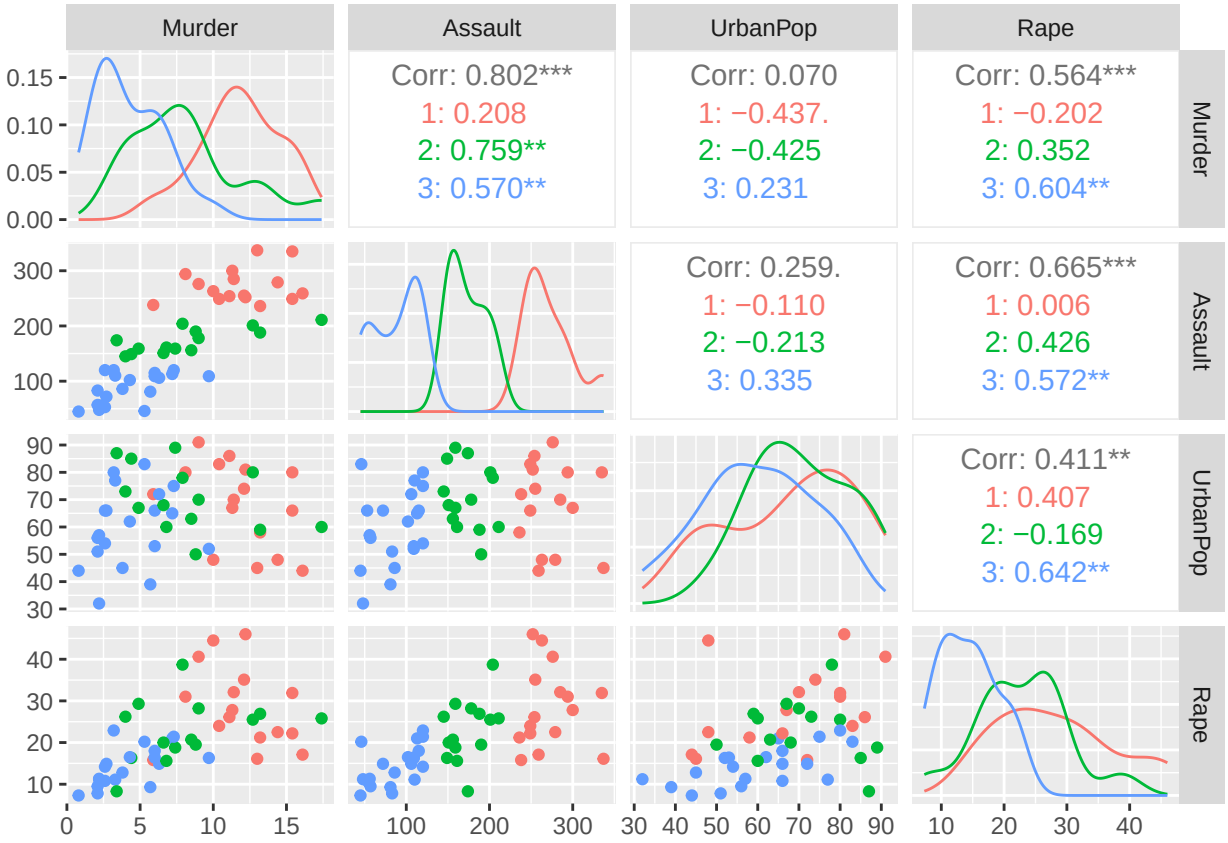
2. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

```
library(ggplot2)
my_dens = function(data, mapping, ..., low = "#132B43",
  high = "#56B1F7") {
  ggplot(data = data, mapping=mapping) +
  geom_density(..., alpha=0.7)
}


# k=3, complete, Euclidean
par(cex=0.45)
hclust.arrests |> as.dendrogram() |>
dendextend::set("branches_k_color",
value = c("red", "blue", "darkgreen"), k = 3) |>
plot(main = "US Arrests Data \n HC - 3 clusters -
Euclidean - Complete")
hclust.arrests |> as.dendrogram() |>
dendextend::rect.dendrogram(k=3, border = 8, lty = 5,
lwd = 2, lower_rect=0)
```




```
GGally::ggpairs(arrests.noX,
aes(color=as.factor(cutree(hclust.arrests,
k = 3))), diag=list(continuous=my_dens))
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```




```
table(cutree(hclust.arrests, k = 3))
```

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

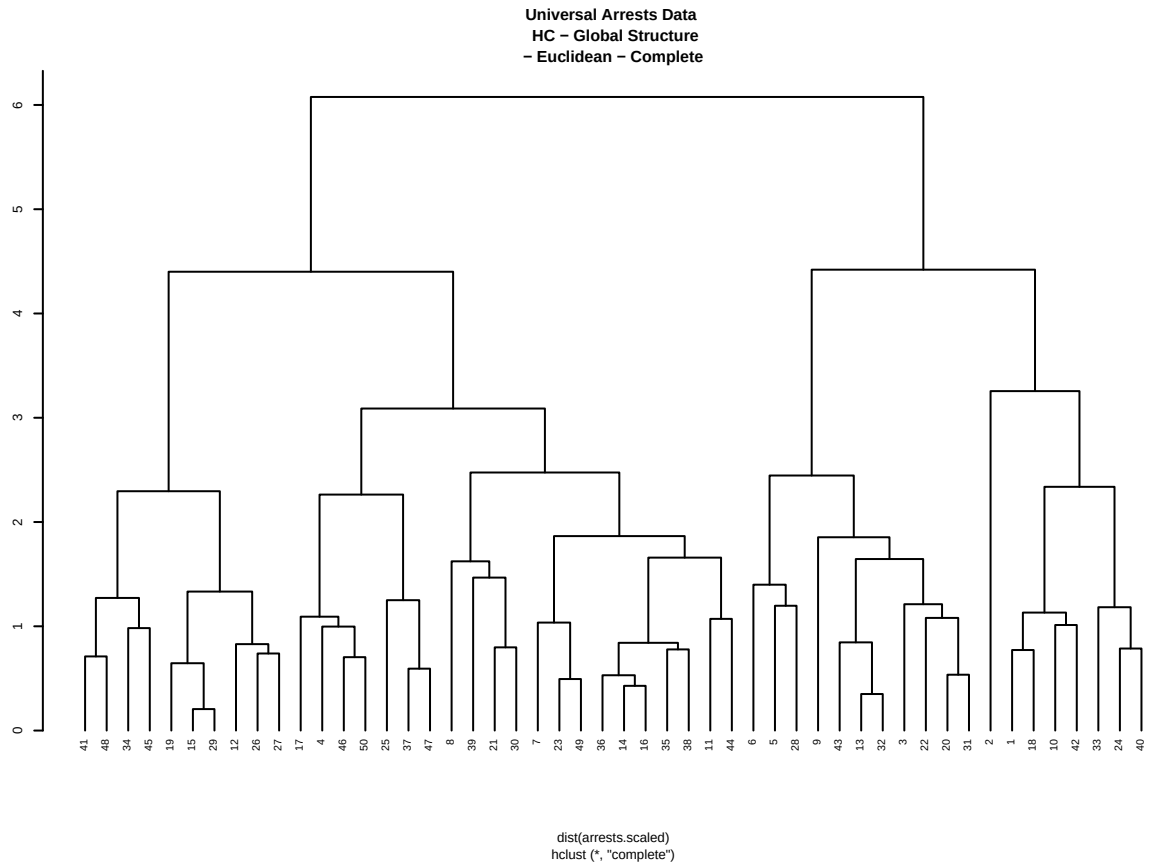
In the first (red) cluster the observations that belong are $C1 = 9, 33, 8, 1, 18, 2, 24, 40, 20, 3, 31, 5, 13, 32, 22, 28$

In the second (blue) cluster the observations that belong are $C2 = 25, 4, 42, 10, 6, 43, 39, 50, 37, 46, 47, 21, 30$

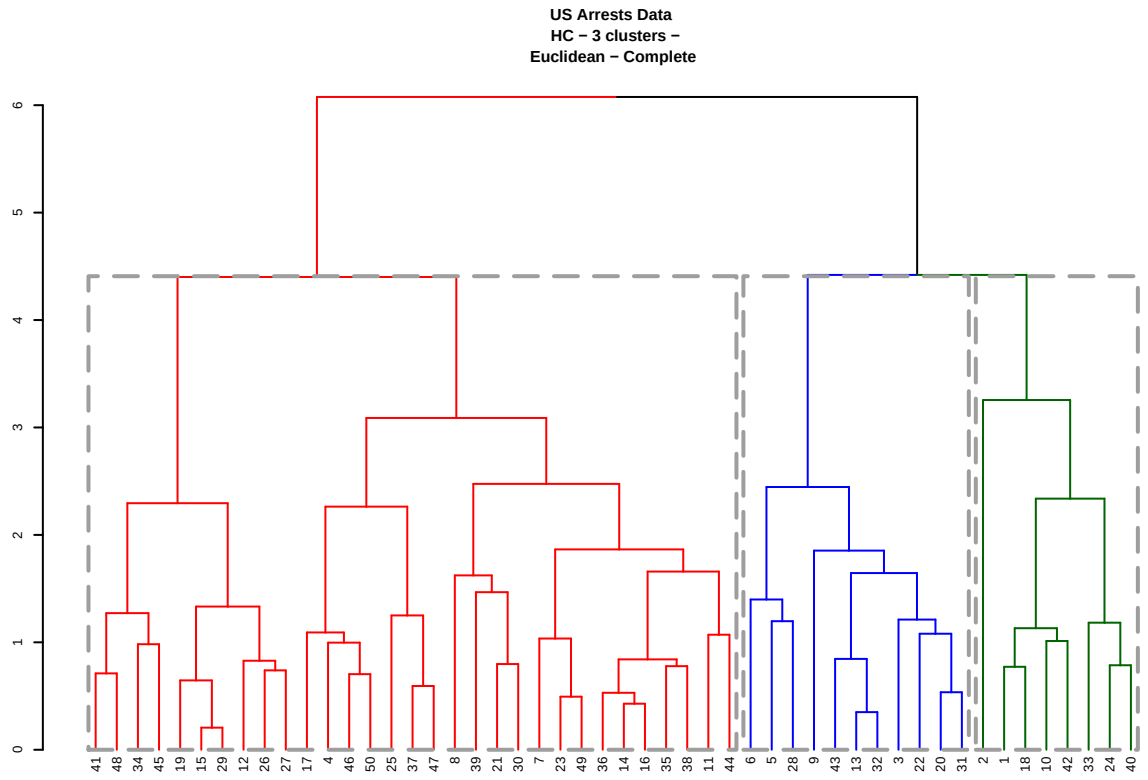
In the third (gray) cluster the observations that belong are $C3 = 35, 44, 7, 38, 27, 17, 26, 12, 14, 16, 11, 23, 49, 15, 29, 48, 19, 41, 34$

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

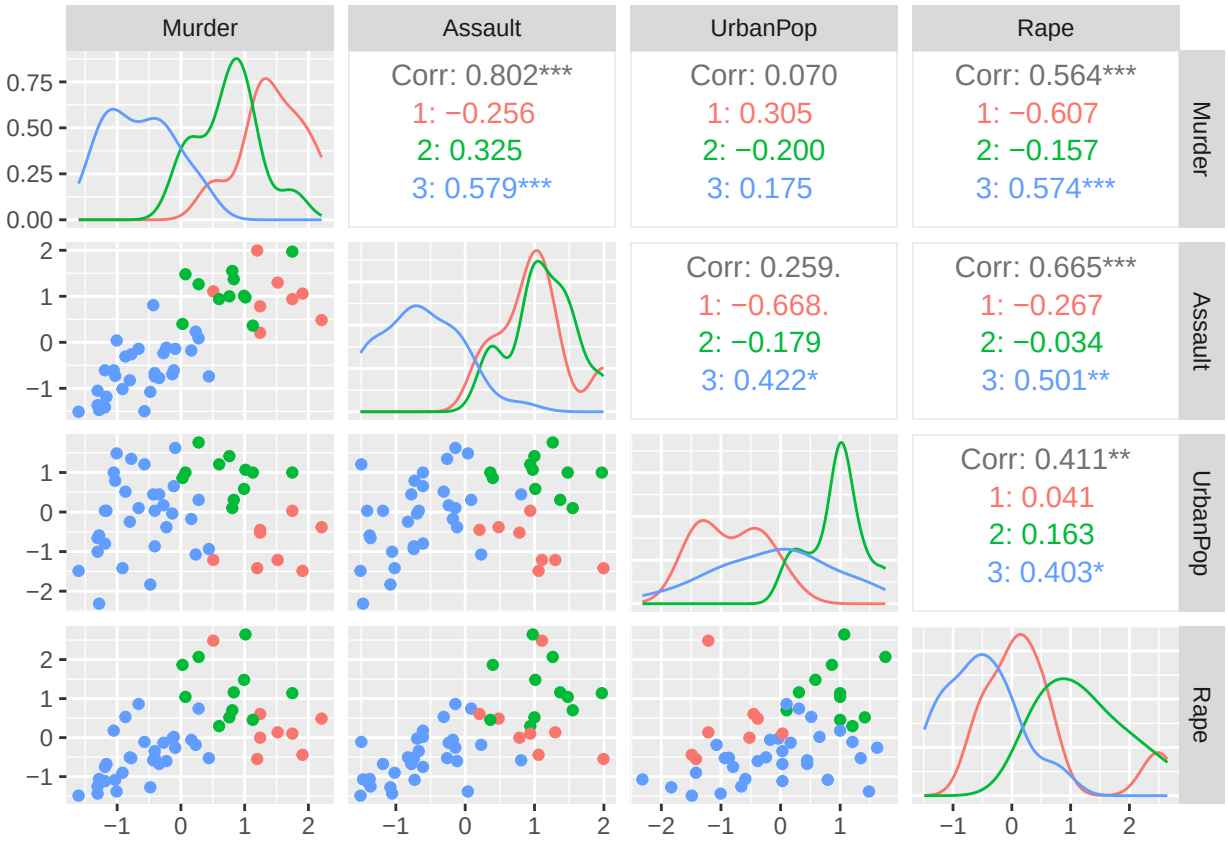
```
arrests.scaled = data.frame(scale(arrests.noX))
par(cex=0.45)
hclust.arrests.2 = hclust(dist(arrests.scaled))
plot(hclust.arrests.2, hang = -1, cex=0.7,
main = "Universal Arrests Data \n HC - Global Structure
- Euclidean - Complete", ylab="")
```




```
# k=3, complete, Euclidean
par(cex=0.45)
hclust.arrests.2 |> as.dendrogram() |>
dendextend::set("branches_k_color",
value = c("red", "blue", "darkgreen"), k = 3) |>
plot(main = "US Arrests Data \n HC - 3 clusters -
Euclidean - Complete")
hclust.arrests.2 |> as.dendrogram() |>
dendextend::rect.dendrogram(k=3, border = 8, lty = 5,
lwd = 2, lower_rect=0)
```




```
GGally::ggpairs(arrests.scaled,
aes(color=as.factor(cutree(hclust.arrests.2,
k = 3))), diag=list(continuous=my_dens))
```




```
table(cutree(hclust.arrests.2, k = 3))
```

```
##
##  1  2  3
##  8 11 31
```

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.

Scaling the variables changes the whole global clustering structure, scaling this dataset before performing hierarchial clustering creates a lower level in the tree with k=3.