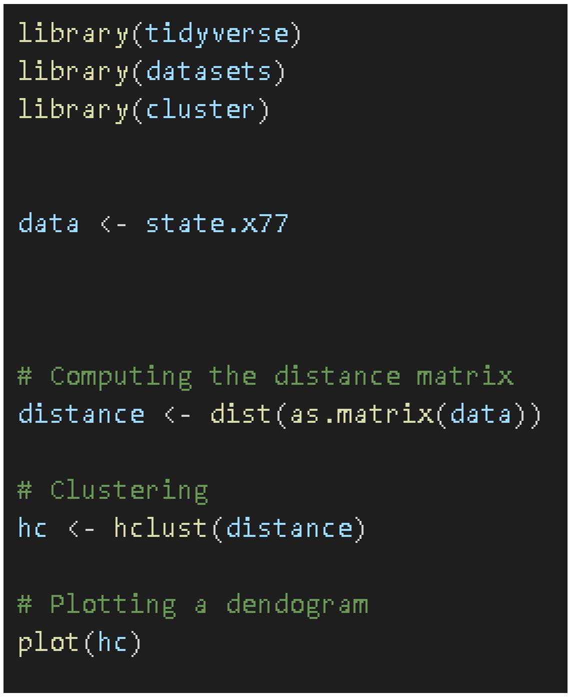
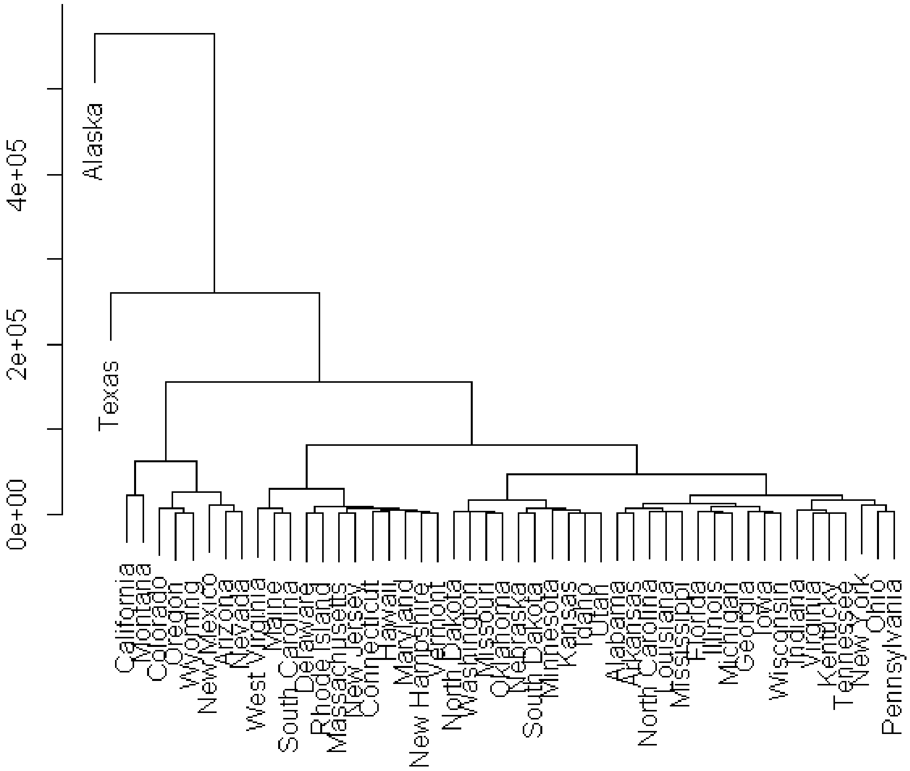
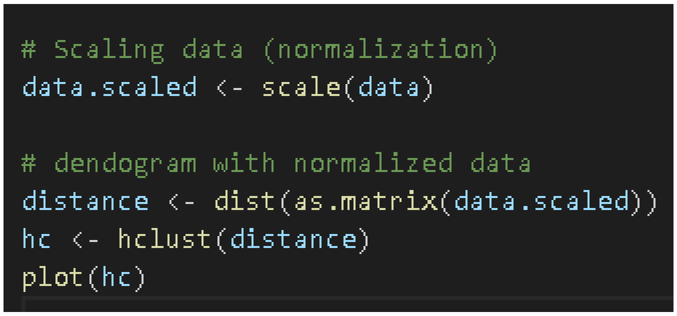
10 PROVE: CLUSTERING THE STATES

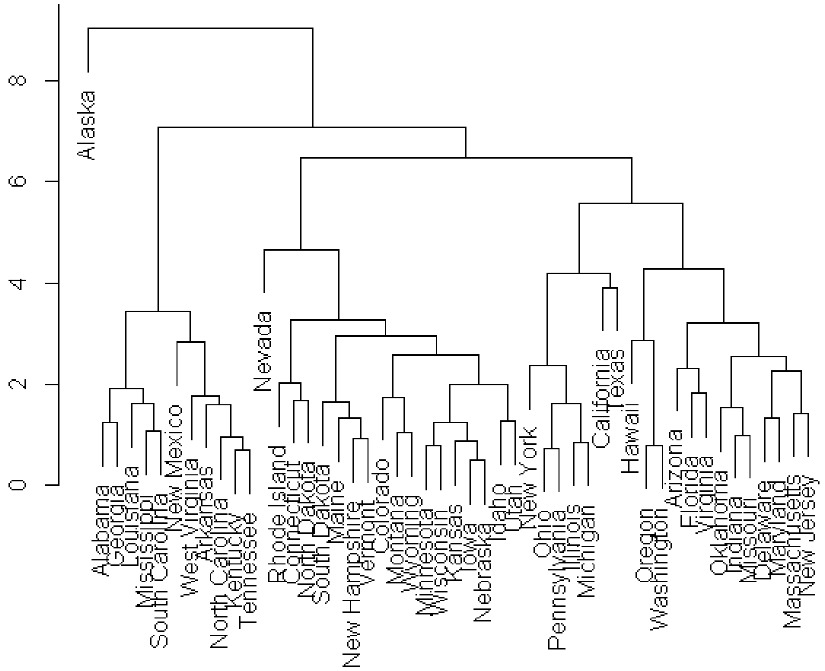


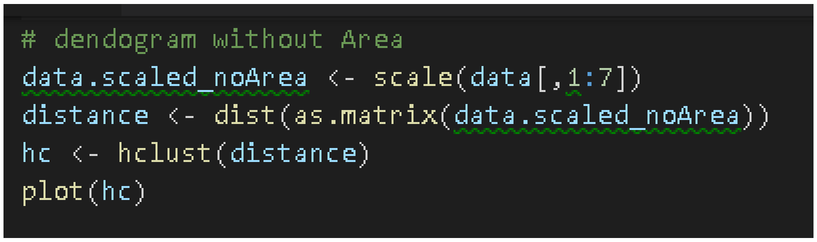
```
library(tidyverse)
library(datasets)
library(cluster)


data <- state.x77



# Computing the distance matrix
distance <- dist(as.matrix(data))

# Clustering
hc <- hclust(distance)

# Plotting a dendogram
plot(hc)
```



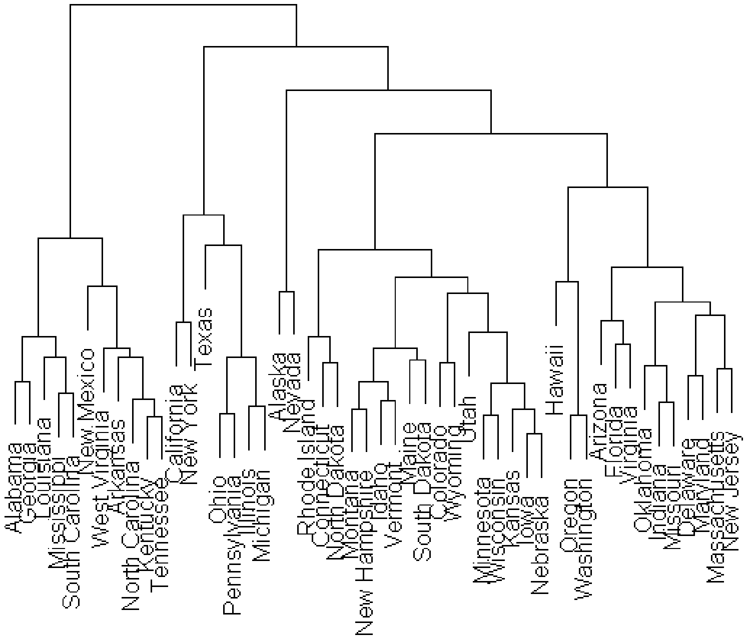


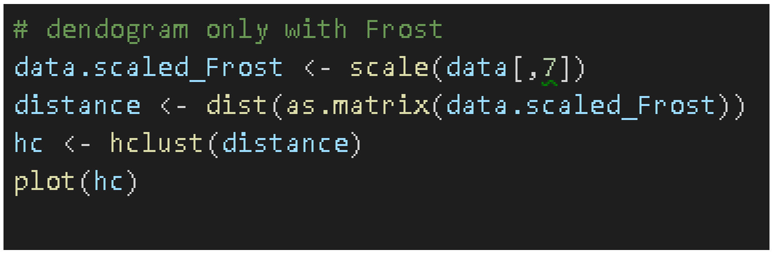
```
# Scaling data (normalization)
data.scaled <- scale(data)

# dendogram with normalized data
distance <- dist(as.matrix(data.scaled))
hc <- hclust(distance)
plot(hc)
```



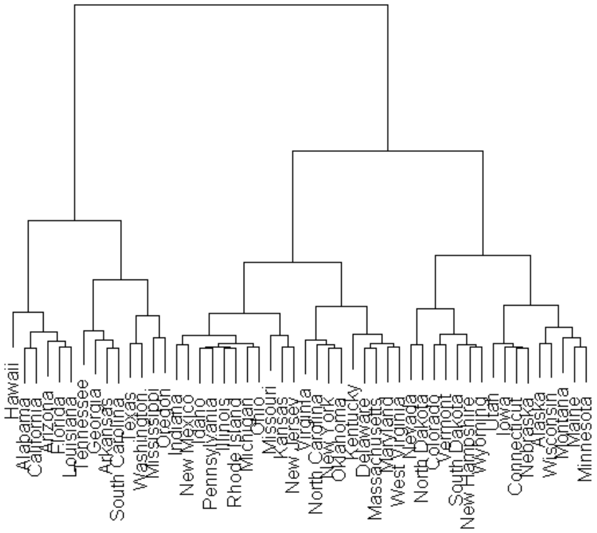


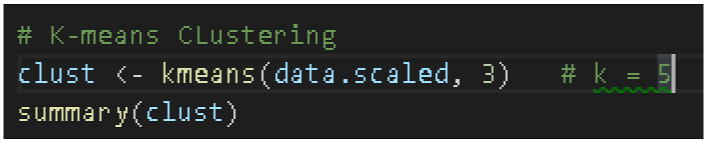
```
# dendogram without Area
data.scaled_noArea <- scale(data[,1:7])
distance <- dist(as.matrix(data.scaled_noArea))
hc <- hclust(distance)
plot(hc)
```



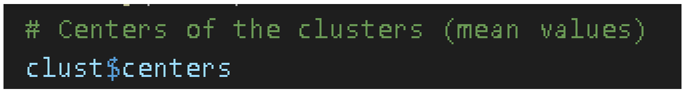


```
# dendogram only with Frost
data.scaled_Frost <- scale(data[,7])
distance <- dist(as.matrix(data.scaled_Frost))
hc <- hclust(distance)
plot(hc)
```

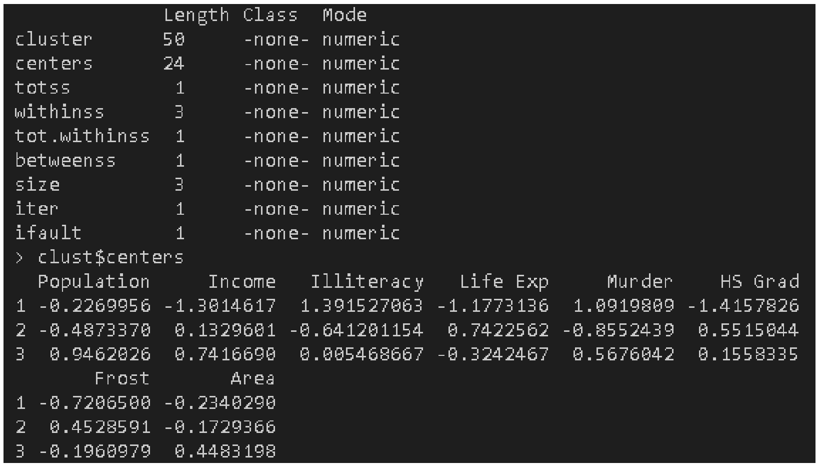




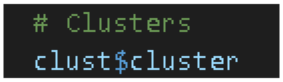
```
# K-means CLustering
clust <- kmeans(data.scaled, 3)    # k = 5
summary(clust)
```



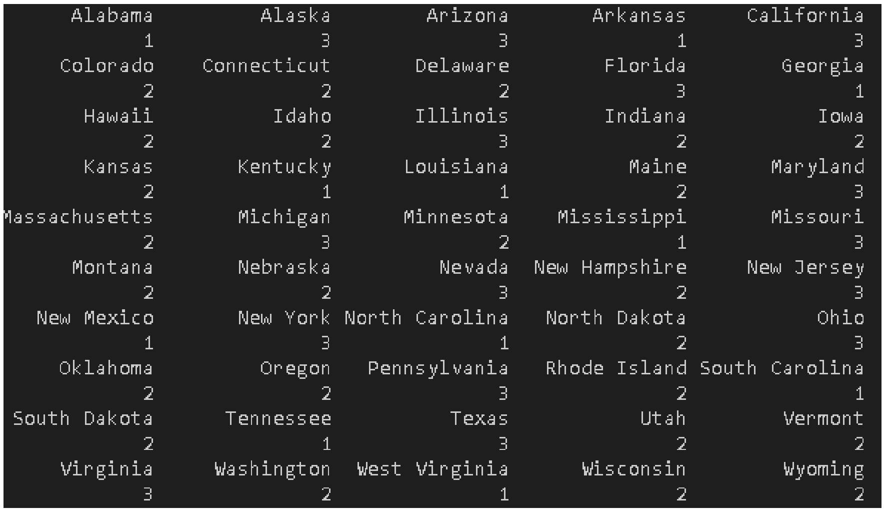
```
# Centers of the clusters (mean values)
clust$centers
```



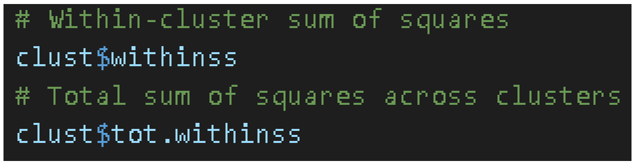
```
             Length Class  Mode
cluster      50     -none- numeric
centers      24     -none- numeric
totss         1     -none- numeric
withinss      3     -none- numeric
tot.withinss  1     -none- numeric
betweenss     1     -none- numeric
size          3     -none- numeric
iter          1     -none- numeric
ifault        1     -none- numeric
> clust$centers
  Population     Income  Illiteracy   Life Exp     Murder    HS Grad
1 -0.2269956 -1.3014617 1.391527063 -1.1773136  1.0919809 -1.4157826
2 -0.4873370  0.1329601 -0.641201154  0.7422562 -0.8552439  0.5515044
3  0.9462026  0.7416690 0.005468667 -0.3242467  0.5676042  0.1558335
       Frost       Area
1 -0.7206500 -0.2340290
2  0.4528591 -0.1729366
3 -0.1960979  0.4483198
```



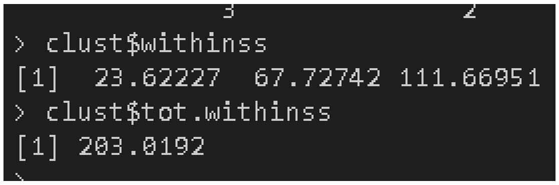
```
# Clusters
clust$cluster
```



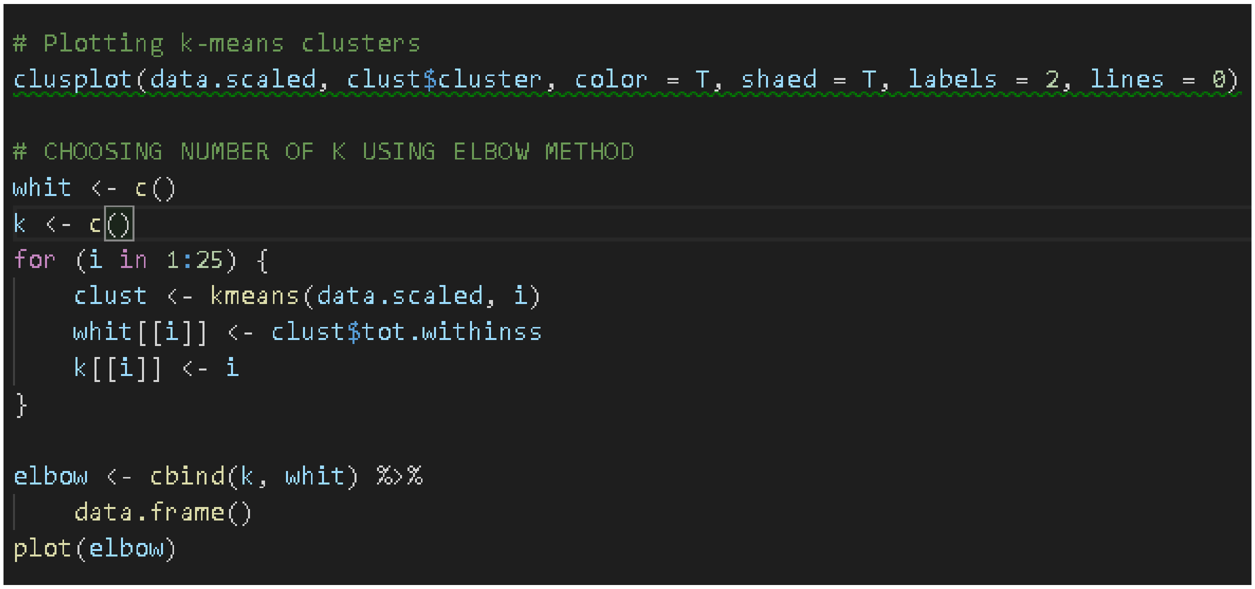
```
       Alabama         Alaska        Arizona       Arkansas     California 
             1              3              3              1              3 
      Colorado    Connecticut       Delaware        Florida        Georgia 
             2              2              2              3              1 
        Hawaii          Idaho       Illinois        Indiana           Iowa 
             2              2              3              2              2 
        Kansas       Kentucky      Louisiana          Maine       Maryland 
             2              1              1              2              3 
 Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
             2              3              2              1              3 
       Montana       Nebraska         Nevada  New Hampshire     New Jersey 
             2              2              3              2              3 
    New Mexico       New York North Carolina   North Dakota           Ohio 
             1              3              1              2              3 
      Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
             2              2              3              2              1 
  South Dakota      Tennessee          Texas           Utah        Vermont 
             2              1              3              2              2 
      Virginia     Washington  West Virginia      Wisconsin        Wyoming 
             3              2              1              2              2 
```



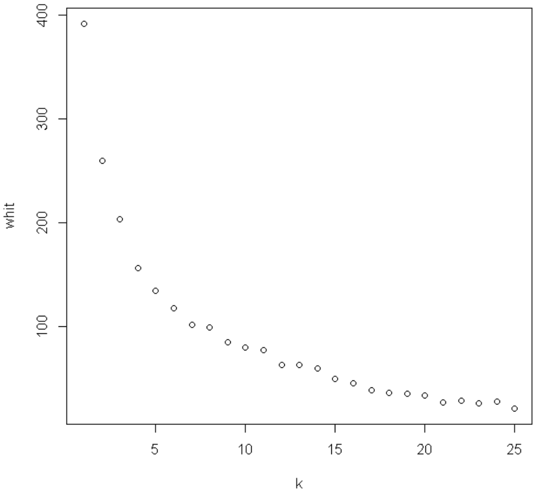
```
# Within-cluster sum of squares
clust$withinss
# Total sum of squares across clusters
clust$tot.withinss
```

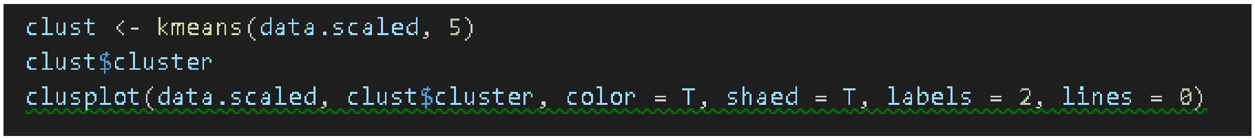


```
             3                  2
> clust$withinss
[1]  23.62227  67.72742 111.66951
> clust$tot.withinss
[1] 203.0192
>
```

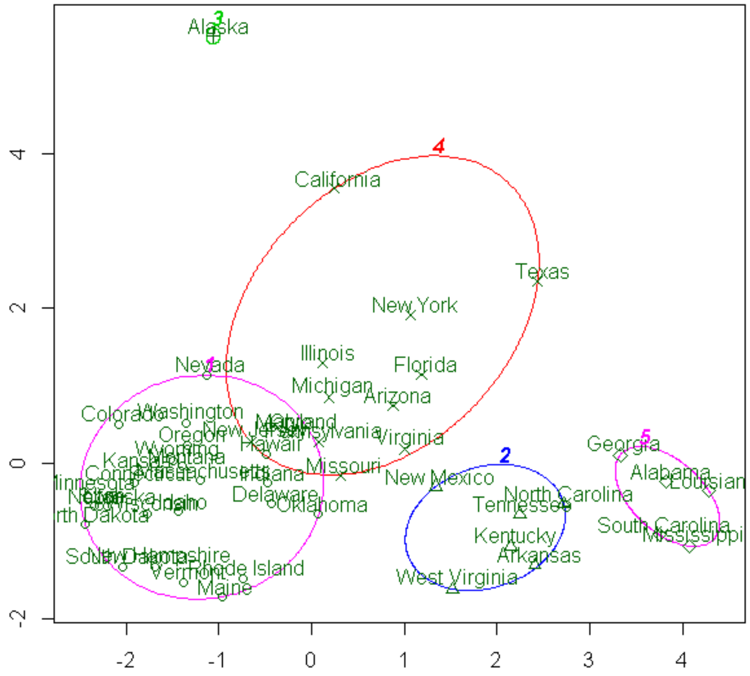


```
# Plotting k-means clusters
clusplot(data.scaled, clust$cluster, color = T, shaed = T, labels = 2, lines = 0)

# CHOOSING NUMBER OF K USING ELBOW METHOD
whit <- c()
k <- c()
for (i in 1:25) {
    clust <- kmeans(data.scaled, i)
    whit[[i]] <- clust$tot.withinss
    k[[i]] <- i
}

elbow <- cbind(k, whit) %>%
    data.frame()
plot(elbow)
```





```
clust <- kmeans(data.scaled, 5)
clust$cluster
clusplot(data.scaled, clust$cluster, color = T, shaed = T, labels = 2, lines = 0)
```


Alaska
California
Texas
New York
Illinois
Florida
Michigan
Arizona
Nevada
Colorado
Washington
Oregon
New Jersey
Maryland
Pennsylvania
Hawaii
Virginia
Wyoming
Kansas
Montana
Connecticut
Massachusetts
Indiana
Missouri
linnesota
Nebraska
Wisconsin
Utah
Idaho
Delaware
Oklahoma
rth Dakota
South Dakota
New Hampshire
Vermont
Rhode Island
Maine
New Mexico
North Carolina
Tennessee
Kentucky
Arkansas
West Virginia
Georgia
Alabama
Louisian
South Carolina
Mississippi