Week 8 - AYUPod - Clustering

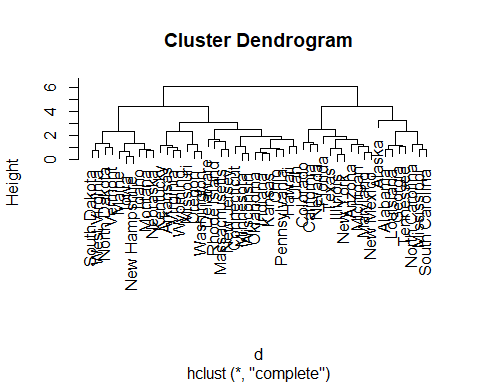
Table of Contents



(Source: kaggle.com)

## Hierarchical Clustering

library(cluster)  
library(factoextra)  
library(tidyverse)  
df <- USArrests  
df <- na.omit(df)  
df <- scale(df)  
# Dissimilarity matrix  
d <- dist(df, method = "euclidean")  
  
# Hierarchical clustering using Complete Linkage  
# Method could be single, average...  
  
hi\_clustering <- hclust(d, method = "complete" )  
  
# Plot the obtained dendrogram  
plot(hi\_clustering)

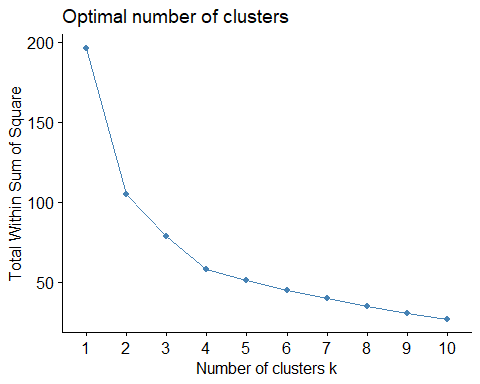


* Decide the number of clusters

We will use the elbow method to decide the number of cluster we should partition the data into. We will plot the total sum squares within clusters to determine how spread out the clusters are within themselves. If a cluster contains one point then the sum square within this cluster is zero. The more points a cluster has, the more likely it has larger sum squares. Thus, at the first step of hierarchical clustering where each point is a cluster, the total sum squares should be zero. At each step the total sum squares will be reduced.

We look for the elbow point of the graph, to identify the number of cluster. We can argue that the elbow point of this graph is at the number of cluster being 3. Thus we decide that the number of clusters for the data is 3.

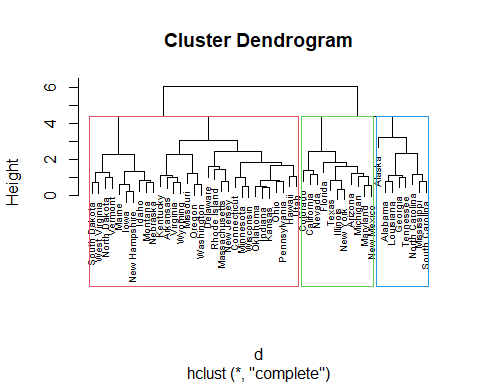
fviz\_nbclust(df, FUN = hcut, method = "wss")

 - Assign clusters to the observations

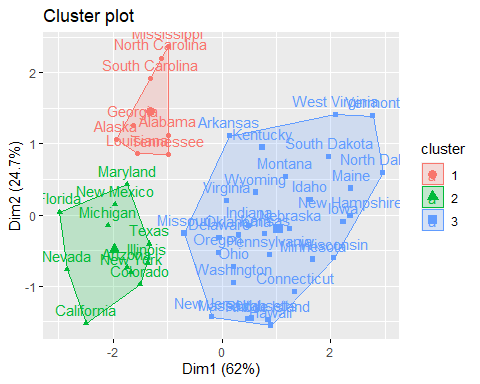
# Cut tree into 4 groups  
sub\_grp <- cutree(hi\_clustering, k = 3)  
USArrests = USArrests %>%  
 mutate(cluster = sub\_grp)

* Visualize the clusters

plot(hi\_clustering, cex = 0.6)  
rect.hclust(hi\_clustering, k = 3, border = 2:5)



fviz\_cluster(list(data = df, cluster = sub\_grp))



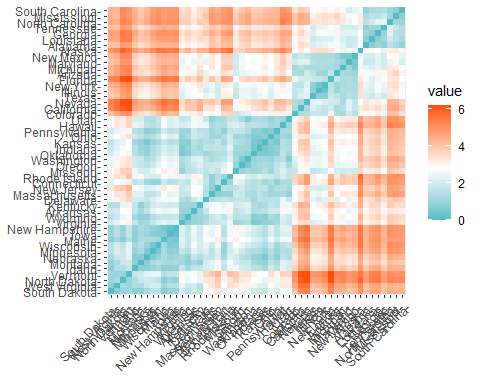
### Question

Working with the [Maill Customers](Mall_Customers.csv) dataset.

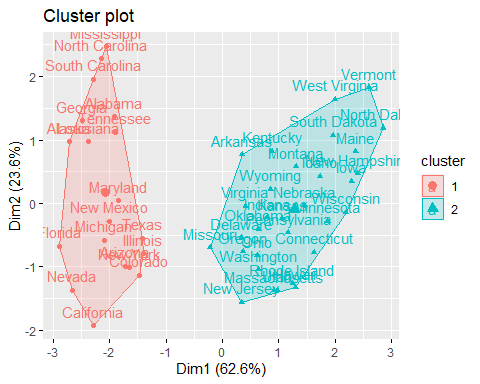
* Plot the total sum squares within clusters and use the elbow method to decide the number of clusters.
* Visualize the clusters with the selected number of clusters.

## K-means Clustering

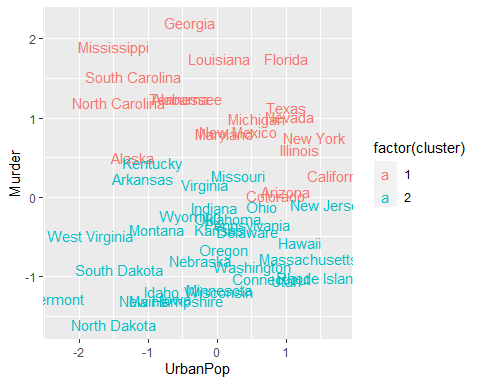
library(tidyverse) # data manipulation  
library(cluster) # clustering algorithms  
library(factoextra) # clustering algorithms & visualization  
df <- USArrests  
  
df <- na.omit(df)  
  
df <- scale(df)  
distance <- get\_dist(df)  
fviz\_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))



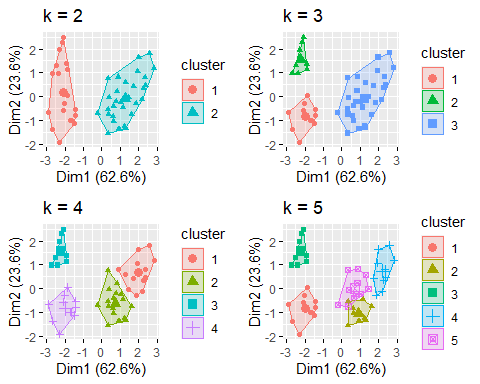
k2 <- kmeans(df, centers = 2, nstart = 25)  
fviz\_cluster(k2, data = df)



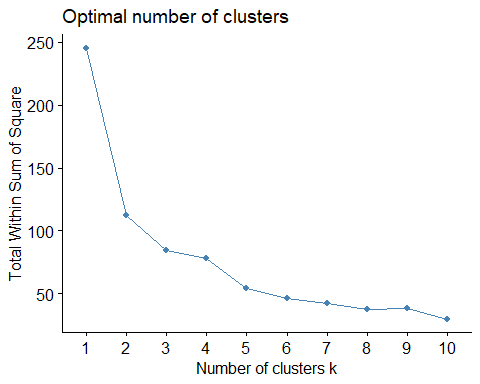
df %>%  
 as\_tibble() %>%  
 mutate(cluster = k2$cluster,  
 state = row.names(USArrests)) %>%  
 ggplot(aes(UrbanPop, Murder, color = factor(cluster), label = state)) +  
 geom\_text()



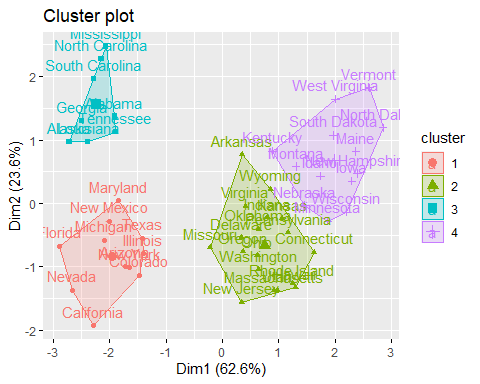
k3 <- kmeans(df, centers = 3, nstart = 25)  
k4 <- kmeans(df, centers = 4, nstart = 25)  
k5 <- kmeans(df, centers = 5, nstart = 25)  
  
# plots to compare  
p1 <- fviz\_cluster(k2, geom = "point", data = df) + ggtitle("k = 2")  
p2 <- fviz\_cluster(k3, geom = "point", data = df) + ggtitle("k = 3")  
p3 <- fviz\_cluster(k4, geom = "point", data = df) + ggtitle("k = 4")  
p4 <- fviz\_cluster(k5, geom = "point", data = df) + ggtitle("k = 5")  
  
library(gridExtra)  
grid.arrange(p1, p2, p3, p4, nrow = 2)



set.seed(123)  
  
fviz\_nbclust(df, kmeans, method = "wss")



final <- kmeans(df, 4, nstart = 25)  
  
fviz\_cluster(final, data = df)



### Question:

Working with the [Maill Customers](Mall_Customers.csv) dataset.

* Plot the total sum squares within clusters and use the elbow method to decide the number of clusters.
* Visualize the clusters with the selected number of clusters.