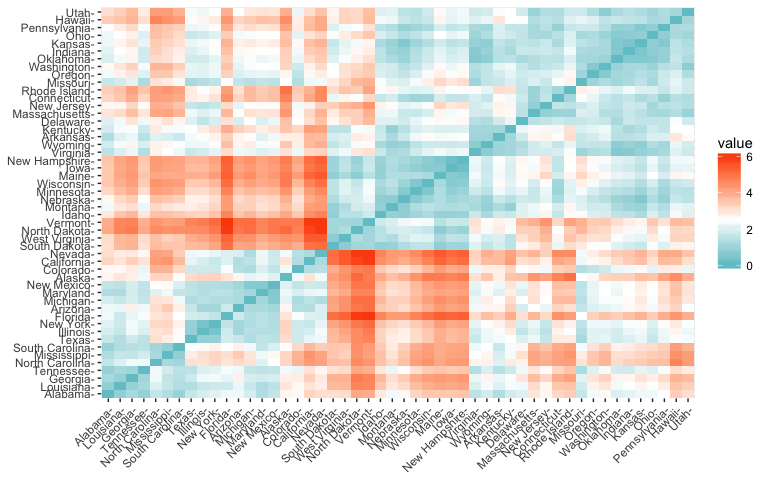
1. get\_dist() – to obtain distance matrix
2. fviz\_dist() – to visualize distance matrix
3. kmeans() – to get k-means distance
4. fviz\_cluster() – to visualize clusters
5. fviz\_nbclust(df, kmeans, method = "wss") - elbow method
6. fviz\_nbclust(df, kmeans, method = "silhouette") - average silhouette method
7. fviz\_gap\_stat(gap\_stat) - gap statistic method

distance <- get\_dist(df)

fviz\_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))



k2 <- kmeans(df, centers = 2, nstart = 25)

str(k2)

## List of 9

## $ cluster : Named int [1:50] 1 1 1 2 1 1 2 2 1 1 ...

## ..- attr(\*, "names")= chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...

## $ centers : num [1:2, 1:4] 1.005 -0.67 1.014 -0.676 0.198 ...

## ..- attr(\*, "dimnames")=List of 2

## .. ..$ : chr [1:2] "1" "2"

## .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"

## $ totss : num 196

## $ withinss : num [1:2] 46.7 56.1

## $ tot.withinss: num 103

## $ betweenss : num 93.1

## $ size : int [1:2] 20 30

## $ iter : int 1

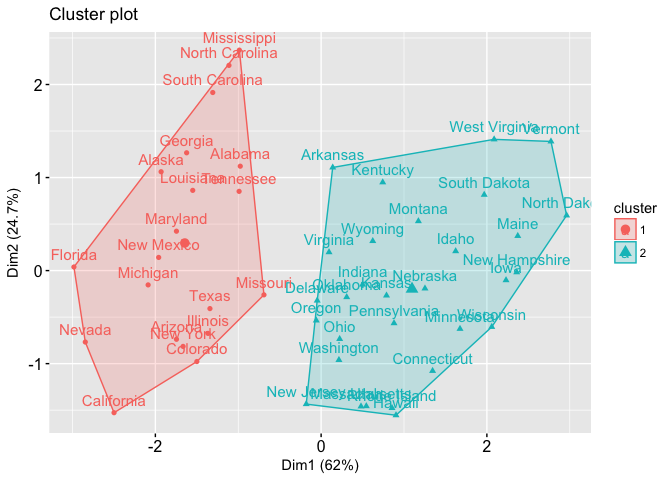
## $ ifault : int 0

## - attr(\*, "class")= chr "kmeans"

The output of kmeans is a list with several bits of information. The most important being:

* cluster: A vector of integers (from 1:k) indicating the cluster to which each point is allocated.
* centers: A matrix of cluster centers.
* totss: The total sum of squares.
* withinss: Vector of within-cluster sum of squares, one component per cluster.
* tot.withinss: Total within-cluster sum of squares, i.e. sum(withinss).
* betweenss: The between-cluster sum of squares, i.e. $totss-tot.withinss$.
* size: The number of points in each cluster.

fviz\_cluster(k2, data = df)



df %>%

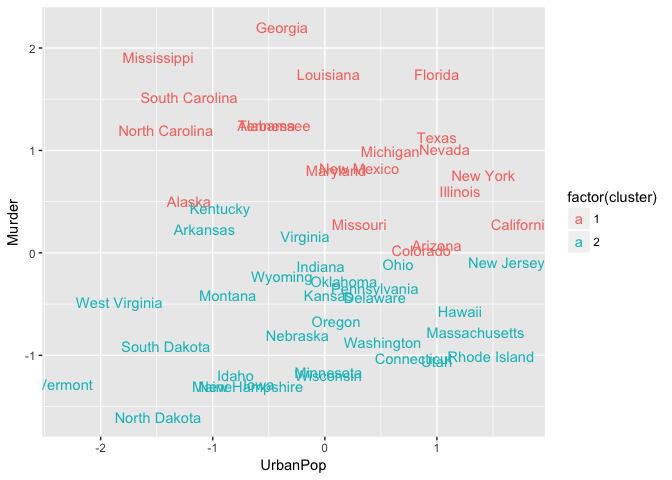
as\_tibble() %>%

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)

