Clustering Analysis

January 14, 2021

# Importing data

data <- read\_xlsx("COVID-19 and economic data in Nigeria vs3.xlsx", sheet = 1, skip = 1) %>%  
 select(-c(2, 3, 4, 7, 14, 17:21, 25)) %>%  
 janitor::clean\_names()

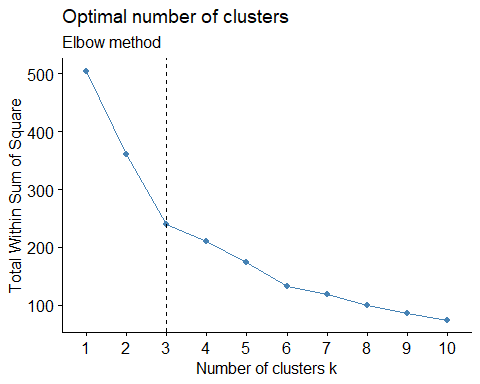
The columns used for this clustering are:

states, population\_female, population\_male, no\_of\_cases\_lab\_confirmed, no\_of\_cases\_on\_admission, no\_discharged, no\_of\_deaths, discharge\_rate, fatality\_rate, x2020\_initial\_budget\_bn\_presented, x2020\_revised\_budget\_bn\_due\_to\_covid\_19, primary\_health\_care, secondary\_health\_care, tertiary, covid\_19\_lab

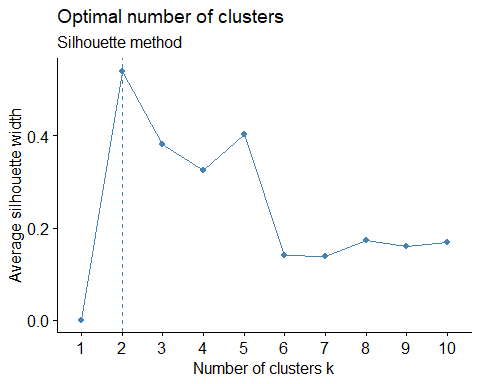
data <- data %>%  
 mutate\_at(5, ~ as.numeric(.)) %>%  
 tidyimpute::impute\_mean() %>%  
 column\_to\_rownames(var = "states")  
  
data <- scale(data)

# Method of choosing K

# Elbow method  
fviz\_nbclust(data, kmeans, method = "wss") +  
 geom\_vline(xintercept = 3, linetype = 2) + # add line for better visualization  
 labs(subtitle = "Elbow method") # add subtitle



# Silhouette method  
  
fviz\_nbclust(data, kmeans, method = "silhouette") +  
 labs(subtitle = "Silhouette method")



# k-means clustering

km.res <- eclust(data, "kmeans", k = 3, nstart = 50, graph = FALSE)  
  
# k-means group number of each observation  
  
  
km.res

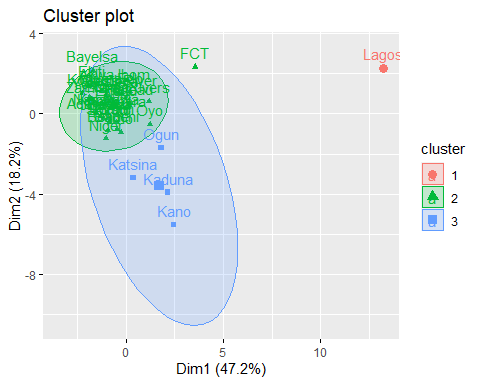
K-means clustering with 3 clusters of sizes 1, 32, 4  
  
Cluster means:  
 population\_female population\_male no\_of\_cases\_lab\_confirmed  
1 1.8026773 1.9106505 5.4201493  
2 -0.3185461 -0.3178441 -0.1818750  
3 2.0976997 2.0650901 0.0999625  
 no\_of\_cases\_on\_admission no\_discharged no\_of\_deaths discharge\_rate  
1 3.4008920 5.54795351 4.8669132 0.500151526  
2 -0.1221707 -0.18465808 -0.1740525 -0.016733878  
3 0.1271428 0.09027624 0.1756921 0.008833142  
 fatality\_rate x2020\_initial\_budget\_bn\_presented  
1 -0.4995605 4.6890142  
2 0.0615585 -0.1515702  
3 -0.3675779 0.0403079  
 x2020\_revised\_budget\_bn\_due\_to\_covid\_19 primary\_health\_care  
1 5.1890696 0.6268082  
2 -0.1972279 -0.1772102  
3 0.2805554 1.2609796  
 secondary\_health\_care tertiary covid\_19\_lab  
1 4.3763529 0.005545864 3.4353638  
2 -0.1006345 -0.279777282 -0.1973013  
3 -0.2890126 2.236831789 0.7195694  
  
Clustering vector:  
 Abia Adamawa Akwa Ibom Anambra Bauchi Bayelsa   
 2 2 2 2 2 2   
 Benue Borno Cross River Delta Ebonyi Edo   
 2 2 2 2 2 2   
 Ekiti Enugu Gombe Imo Jigawa Kaduna   
 2 2 2 2 2 3   
 Kano Katsina Kebbi Kogi Kwara Lagos   
 3 3 2 2 2 1   
 Nasarawa Niger Ogun Ondo Osun Oyo   
 2 2 3 2 2 2   
 Plateau Rivers Sokoto Taraba Yobe Zamfara   
 2 2 2 2 2 2   
 FCT   
 2   
  
Within cluster sum of squares by cluster:  
[1] 0.00000 212.78593 26.31504  
 (between\_SS / total\_SS = 52.6 %)  
  
Available components:  
  
 [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
 [6] "betweenss" "size" "iter" "ifault" "silinfo"   
[11] "nbclust" "data"

km.res$centers

population\_female population\_male no\_of\_cases\_lab\_confirmed  
1 1.8026773 1.9106505 5.4201493  
2 -0.3185461 -0.3178441 -0.1818750  
3 2.0976997 2.0650901 0.0999625  
 no\_of\_cases\_on\_admission no\_discharged no\_of\_deaths discharge\_rate  
1 3.4008920 5.54795351 4.8669132 0.500151526  
2 -0.1221707 -0.18465808 -0.1740525 -0.016733878  
3 0.1271428 0.09027624 0.1756921 0.008833142  
 fatality\_rate x2020\_initial\_budget\_bn\_presented  
1 -0.4995605 4.6890142  
2 0.0615585 -0.1515702  
3 -0.3675779 0.0403079  
 x2020\_revised\_budget\_bn\_due\_to\_covid\_19 primary\_health\_care  
1 5.1890696 0.6268082  
2 -0.1972279 -0.1772102  
3 0.2805554 1.2609796  
 secondary\_health\_care tertiary covid\_19\_lab  
1 4.3763529 0.005545864 3.4353638  
2 -0.1006345 -0.279777282 -0.1973013  
3 -0.2890126 2.236831789 0.7195694

# Visualize k-means clusters

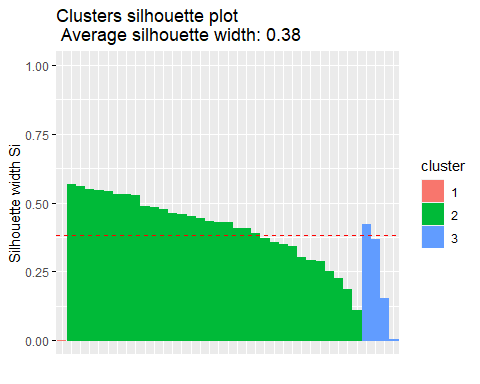
fviz\_cluster(km.res, ellipse.type = "norm", ellipse.level = 0.95)



# Visualize the silhouette of clusters

fviz\_silhouette(km.res)

cluster size ave.sil.width  
1 1 1 0.00  
2 2 32 0.41  
3 3 4 0.24



# Hierarchical clustering

res.hc <- eclust(data, "hclust",  
 k = 3,  
 method = "ward.D2", graph = FALSE  
)  
  
res.hc

Call:  
stats::hclust(d = x, method = hc\_method)  
  
Cluster method : ward.D2   
Distance : euclidean   
Number of objects: 37

res.hc$cluster

Abia Adamawa Akwa Ibom Anambra Bauchi Bayelsa   
 1 1 1 1 1 1   
 Benue Borno Cross River Delta Ebonyi Edo   
 1 1 1 1 1 1   
 Ekiti Enugu Gombe Imo Jigawa Kaduna   
 1 1 1 1 1 2   
 Kano Katsina Kebbi Kogi Kwara Lagos   
 2 1 1 1 1 3   
 Nasarawa Niger Ogun Ondo Osun Oyo   
 1 1 1 1 1 1   
 Plateau Rivers Sokoto Taraba Yobe Zamfara   
 1 1 1 1 1 1   
 FCT   
 1

res.hc$size

[1] 34 2 1

res.hc$height

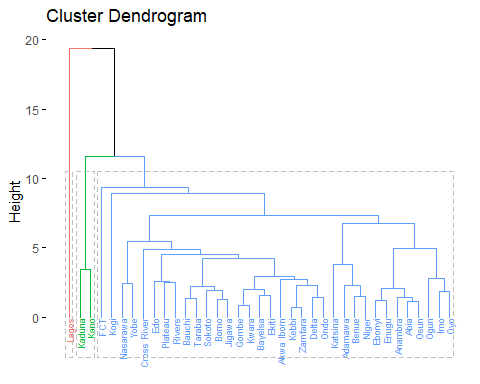
[1] 0.6989078 0.8576963 1.1495256 1.1689516 1.2414005 1.3238035  
 [7] 1.3958933 1.4331866 1.4774845 1.5463315 1.8313158 1.9487046  
[13] 1.9735127 2.0627740 2.2103994 2.3070884 2.3090155 2.4005970  
[19] 2.4673650 2.5307109 2.7337310 2.7662067 2.9258495 3.4490733  
[25] 3.7957496 4.2508480 4.4744893 4.8562227 4.9476561 5.4513167  
[31] 6.7350829 7.3480336 8.8814011 9.3730961 11.6129345 19.3987821

res.hc$silinfo

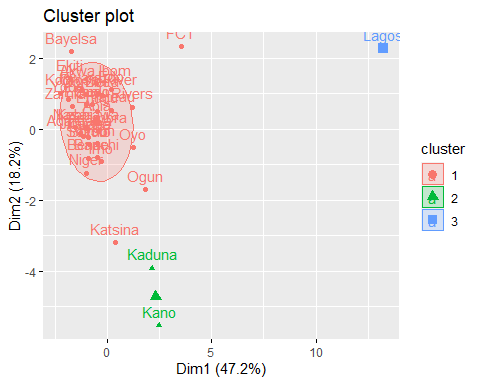
$widths  
 cluster neighbor sil\_width  
Kwara 1 2 0.61888861  
Zamfara 1 2 0.61134562  
Gombe 1 2 0.61076365  
Ebonyi 1 2 0.59987164  
Ondo 1 2 0.59976642  
Kebbi 1 2 0.59885691  
Ekiti 1 2 0.59400632  
Taraba 1 2 0.57705054  
Bayelsa 1 2 0.57114779  
Jigawa 1 2 0.54942870  
Enugu 1 2 0.54872312  
Yobe 1 2 0.54123138  
Delta 1 2 0.53709669  
Abia 1 2 0.53485078  
Akwa Ibom 1 2 0.53037460  
Osun 1 2 0.52743246  
Borno 1 2 0.52527268  
Plateau 1 2 0.50779386  
Bauchi 1 2 0.50378545  
Anambra 1 2 0.49548011  
Sokoto 1 2 0.48500965  
Benue 1 2 0.48251078  
Adamawa 1 2 0.47856268  
Edo 1 2 0.47322943  
Imo 1 2 0.44257205  
Cross River 1 2 0.43687862  
Niger 1 2 0.41814190  
Rivers 1 2 0.39544087  
Nasarawa 1 2 0.37991059  
Oyo 1 2 0.36796029  
Kogi 1 2 0.28893753  
FCT 1 2 0.15077313  
Ogun 1 2 0.06599418  
Katsina 1 2 -0.11027082  
Kano 2 1 0.52002618  
Kaduna 2 1 0.43979744  
Lagos 3 2 0.00000000  
  
$clus.avg.widths  
[1] 0.4687888 0.4799118 0.0000000  
  
$avg.width  
[1] 0.45672

# Dendrogram

fviz\_dend(res.hc, rect = TRUE, show\_labels = TRUE, cex = 0.5)

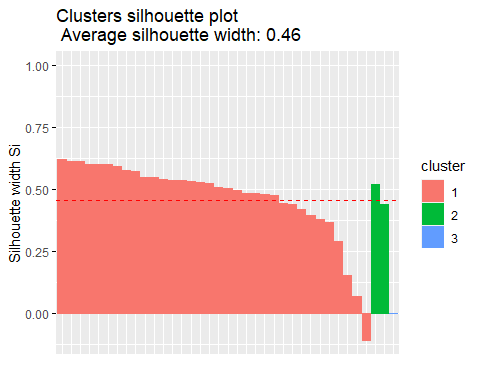


fviz\_cluster(res.hc, ellipse.type = "norm", ellipse.level = 0.68)

 # Visualize the silhouette of clusters

fviz\_silhouette(res.hc)

cluster size ave.sil.width  
1 1 34 0.47  
2 2 2 0.48  
3 3 1 0.00



# Combining hierarchical clustering and k-means

res.hk <- hkmeans(data, 3)  
  
names(res.hk)

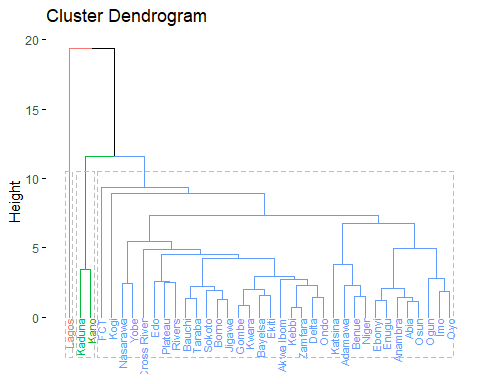
[1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
 [6] "betweenss" "size" "iter" "ifault" "data"   
[11] "hclust"

res.hk

Hierarchical K-means clustering with 3 clusters of sizes 32, 4, 1  
  
Cluster means:  
 population\_female population\_male no\_of\_cases\_lab\_confirmed  
1 -0.3185461 -0.3178441 -0.1818750  
2 2.0976997 2.0650901 0.0999625  
3 1.8026773 1.9106505 5.4201493  
 no\_of\_cases\_on\_admission no\_discharged no\_of\_deaths discharge\_rate  
1 -0.1221707 -0.18465808 -0.1740525 -0.016733878  
2 0.1271428 0.09027624 0.1756921 0.008833142  
3 3.4008920 5.54795351 4.8669132 0.500151526  
 fatality\_rate x2020\_initial\_budget\_bn\_presented  
1 0.0615585 -0.1515702  
2 -0.3675779 0.0403079  
3 -0.4995605 4.6890142  
 x2020\_revised\_budget\_bn\_due\_to\_covid\_19 primary\_health\_care  
1 -0.1972279 -0.1772102  
2 0.2805554 1.2609796  
3 5.1890696 0.6268082  
 secondary\_health\_care tertiary covid\_19\_lab  
1 -0.1006345 -0.279777282 -0.1973013  
2 -0.2890126 2.236831789 0.7195694  
3 4.3763529 0.005545864 3.4353638  
  
Clustering vector:  
 Abia Adamawa Akwa Ibom Anambra Bauchi Bayelsa   
 1 1 1 1 1 1   
 Benue Borno Cross River Delta Ebonyi Edo   
 1 1 1 1 1 1   
 Ekiti Enugu Gombe Imo Jigawa Kaduna   
 1 1 1 1 1 2   
 Kano Katsina Kebbi Kogi Kwara Lagos   
 2 2 1 1 1 3   
 Nasarawa Niger Ogun Ondo Osun Oyo   
 1 1 2 1 1 1   
 Plateau Rivers Sokoto Taraba Yobe Zamfara   
 1 1 1 1 1 1   
 FCT   
 1   
  
Within cluster sum of squares by cluster:  
[1] 212.78593 26.31504 0.00000  
 (between\_SS / total\_SS = 52.6 %)  
  
Available components:  
  
 [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
 [6] "betweenss" "size" "iter" "ifault" "data"   
[11] "hclust"

# Visualize the tree

fviz\_dend(res.hk, cex = 0.6, rect = TRUE)



# Visualize the hkmeans final clusters

fviz\_cluster(res.hk, ellipse.type = "norm", ellipse.level = 0.68)

