#install package cluster.datasets

library(cluster.datasets)

library(tidyverse)

library(gridExtra)

data(all.mammals.milk.1956)

head(all.mammals.milk.1956)

plot1 <- all.mammals.milk.1956 %>%

ggplot(aes(x = "all mammals", y = water)) +

geom\_jitter(width = .025, height = 0, size = 2, alpha = .5, color = "blue") +

labs(x = "", y="percentage of water")

plot2 <- all.mammals.milk.1956 %>%

ggplot(aes(x = "all mammals", y = protein)) +

geom\_jitter(width = .02, height = 0, size = 2, alpha = .6, color = "orange") +

labs(x = "", y="percentage of protein")

plot3 <- all.mammals.milk.1956 %>%

ggplot(aes(x = "all mammals", y = fat)) +

geom\_jitter(width = .02, height = 0, size = 2, alpha = .6, color = "green") +

labs(x = "", y="percentage of fat")

plot4 <- all.mammals.milk.1956 %>%

ggplot(aes(x = "all mammals", y = lactose)) +

geom\_jitter(width = .02, height = 0, size = 2, alpha = .6, color = "red") +

labs(x = "", y="percentage of lactose")

plot5 <- all.mammals.milk.1956 %>%

ggplot(aes(x = "all mammals", y = ash)) +

geom\_jitter(width = .02, height = 0, size = 2, alpha = .6, color = "violet") +

labs(x = "", y="percentage of ash")

grid.arrange(plot1, plot2, plot3, plot4, plot5)

# As the initial centroids are defined randomly,

# we define a seed for purposes of reprodutability

set.seed(123)

# Let's remove the column with the mammals' names, so it won't be used in the clustering

input <- all.mammals.milk.1956[,2:6]

# The nstart parameter indicates that we want the algorithm to be executed 20 times.

# This number is not the number of iterations, it is like calling the function 20 times and then

# the execution with lower variance within the groups will be selected as the final result.

kmeans(input, centers = 3, nstart = 20)

#' Plots a chart showing the sum of squares within a group for each execution of the kmeans algorithm.

#' In each execution the number of the initial groups increases by one up to the maximum number of centers passed as argument.

#'

#' data is the The dataframe to perform the kmeans

#' nc is the The maximum number of initial centers

#'

wssplot <- function(data, nc=15, seed=123){

wss <- (nrow(data)-1)\*sum(apply(data,2,var))

for (i in 2:nc){

set.seed(seed)

wss[i] <- sum(kmeans(data, centers=i)$withinss)}

plot(1:nc, wss, type="b", xlab="Number of groups",

ylab="Sum of squares within a group")}

wssplot(input, nc = 20)

#Analysis of graph of above graph

#By Analysing the chart from right to left, we can see that when the

#number of groups (K) reduces from 4 to 3 there is a big increase in

#the sum of squares, bigger than any other previous increase. That means

#that when it passes from 4 to 3 groups there is a reduction in the

#clustering compactness (compactness, means the similarity within a group).

#So,choose K = 4 and run the K-means again.

set.seed(123)

clustering <- kmeans(input, centers = 4, nstart = 20)

clustering

#clustering validation with silhouette width

library(cluster)

library(factoextra)

sil <- silhouette(clustering$cluster, dist(input))

fviz\_silhouette(sil)

#Clustering Interpretation with interactive plot

library(GGally)

library(plotly)

all.mammals.milk.1956$cluster <- as.factor(clustering$cluster)

p <- ggparcoord(data = all.mammals.milk.1956, columns = c(2:6), groupColumn = "cluster", scale = "std") + labs(x = "milk constituent", y = "value (in standard-deviation units)", title = "Clustering")

ggplotly(p)