```
library(readr)
library(ggplot2)
library(tidyverse)
library(MASS)
library(tibble)
library(caret)
library(ISLR)
library(Hmisc)
library(leaps)
library(NbClust)
library(factoextra)
summary(Pharmaceuticals)
colnames(Pharmaceuticals)
str(Pharmaceuticals)
Pharmaceutical_data <- Pharmaceuticals
Pharmaceutical_data <- read.csv("Pharmaceuticals.csv", header =TRUE)
#Set row names to the Name Column
row.names(Pharmaceutical_data) <- Pharmaceutical_data[,2]

#Remove all categorical variables
Pharmaceutical_data$Symbol<- NULL
Pharmaceutical_data$Name <- NULL
Pharmaceutical_data$Median_Recommendation<-NULL
Pharmaceutical_data$Location <-NULL
Pharmaceutical_data$Exchange <-NULL
str(Pharmaceutical_data)
# Normalizing the data
Pharmaceutical_data.norm <- sapply(Pharmaceutical_data, scale)
Pharmaceutical_data.norm

#to find the distance between the observations
distance <- dist(Pharmaceutical_data.norm, method = "euclidean")
fviz_dist(distance)
#hierarchical clustering
fit.average <- hclust(distance, method="average")
plot(fit.average, hang = -1, cex=0.8, main ="average link clustering")

nc <- NbClust(Pharmaceutical_data.norm, distance = "euclidean", min.nc = 2,
max.nc = 10, method = "average")
nc <- NbClust(Pharmaceutical_data.norm, distance = "euclidean", min.nc = 2,
max.nc = 10, method = "average")
# using the optimal number of clusters as 6
clusters <- cutree(fit.average, k=6)
table(clusters)
rect.hclust(fit.average, k=6, border="red" )
table(clusters)


# Perform k-means cluster analysis

k3 <- kmeans(Pharmaceutical_data.norm, centers = 3, nstart = 25)
k3$centers
k3$size
k3$cluster
```

```
fviz_cluster(k3, data = Pharmaceutical_data.norm )
#checking for any outliers
dist(k3$centers)
aggregate(Pharmaceutical_data.norm,by=list(k3$cluster),FUN=mean)
```