---
title: "K-Means for clustering"
author: "Ram"
date: "11/6/2021"
output: word_document

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
setwd("C:/Users/ramne/Desktop/ML Assignment/K-Means")
Pharmadata<- read.csv("Pharmaceuticals.csv", header = TRUE)
str(Pharmadata)
```

Load all required libraries

```{r}
library(tidyverse)
library(factoextra)
library(cluster)
library(ggplot2)
library(gridExtra)
```

To remove any missing value that might be present in the data
```{r}
Pharmadata <- na.omit(Pharmadata)
```
Collecting numerical variables from column 1 to 9 to cluster 21 firms
```{r}
row.names(Pharmadata)<- Pharmadata[,1]
P1<- Pharmadata[, 3:11]
head(P1)
```
Scaling the data using Scale function
```{r}
dataframe<- scale(P1)
head(dataframe)
```
Computing K-means clustering in R for different centers
Using multiple values of K and examine the differences in results
```{r}
kmeans <- kmeans(dataframe, centers = 2, nstart = 30)
kmeans1<- kmeans(dataframe, centers = 5, nstart = 30)
kmeans2<- kmeans(dataframe, centers = 6, nstart = 30)
Plot1<-fviz_cluster(kmeans, data = dataframe)+ggtitle("k=2")
plot2<-fviz_cluster(kmeans1, data = dataframe)+ggtitle("k=5")
plot3<-fviz_cluster(kmeans2, data = dataframe)+ggtitle("k=6")
grid.arrange(Plot1,plot2,plot3, nrow = 2)
```

Determining optimal clusters using Elbow method
```{r}
distance<- dist(dataframe, method = "euclidean")# for calculating
distance matrix between rows of a data matrix.
```

```
fviz_dist(distance)# Visualizing a distance matrix
```
For each k, calculate the total within-cluster sum of square (wss)
tot.withinss is total within-cluster sum of squares
Compute and plot wss for k = 1 to k = 10
extract wss for 2-15 clusters
The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters k =5.
```{r}
set.seed(123)
wss<- function(k){
  kmeans(dataframe, k, nstart =10)$tot.withinss
}
k.values<- 1:10
wss_clusters<- map_dbl(k.values, wss)
plot(k.values, wss_clusters,
 type="b", pch = 16, frame = TRUE,
xlab="Number of clusters",
ylab="Total within-clusters sum of squares")
```
Final analysis and Extracting results using 5 clusters and Visualize the results
```{r}
set.seed(123)
final<- kmeans(dataframe, 5, nstart = 25)
print(final)
fviz_cluster(final, data = dataframe)
P1%>%
  mutate(Cluster = final$cluster) %>%
  group_by(Cluster)%>% summarise_all("mean")
clusplot(dataframe,final$cluster, color = TRUE, labels = 2,lines = 0)
```
b) Interpret the clusters with respect to the numerical variables used in forming the clusters
Cluster 1 - AHM,SGP,WYE,BMY,AZN, ABT, NVS, LLY
Cluster 2 - BAY, CHTT, IVX
Cluster 3 - AGN, PHA
Cluster 4 - JNJ, MRK, PFE,GSK
Cluster 5 - WPI, MRX,ELN,AVE
```{r}
ClusterForm<- Pharmadata[,c(12,13,14)]%>% mutate(clusters = final$cluster)%>% arrange(clusters, ascending = TRUE)
ClusterForm
```
c)Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)
```{r, message=FALSE, warning=FALSE, fig.width=10}
p1<-ggplot(ClusterForm, mapping = aes(factor(clusters), fill=Median_Recommendation))+geom_bar(position = 'dodge')+labs(x ='Number of clusters')
p2<- ggplot(ClusterForm, mapping = aes(factor(clusters),fill = Location))+geom_bar(position = 'dodge')+labs(x ='Number of clusters')
p3<- ggplot(ClusterForm, mapping = aes(factor(clusters),fill = Exchange))+geom_bar(position = 'dodge')+labs(x ='Number of clusters')
grid.arrange(p1,p2,p3)
```
As per graph, Cluster 1 Suggests to Hold to Moderate Sell
Cluster 2 Suggests to Hold

Cluster 3 Suggests to Hold to Moderate Buy
Cluster 4 suggests to Hold to Moderate Buy
Cluster 5 suggests to Moderate Buy to Moderate Sell

d)Provide an appropriate name for each cluster using any or all of the variables in the dataset.
Cluster1-Sell Cluster
Cluster2-Hold Cluster
Cluster3-Buy Cluster
Cluster4-High Buy Cluster
Cluster5-Buy-Sell Cluster