K-Means for clustering

Ram

11/6/2021

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

```
## 'data.frame':    21 obs. of  14 variables:
##  $ Symbol               : chr  "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr  "Abbott Laboratories" "Allergan, Inc."
"Amersham plc" "AstraZeneca PLC" ...
##  $ Market_Cap           : num  68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num  0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08
0.18 ...
##  $ PE_Ratio             : num  24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6
27.9 ...
##  $ ROE                  : num  26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1
31 ...
##  $ ROA                  : num  11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5
...
##  $ Asset_Turnover       : num  0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num  0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53
...
##  $ Rev_Growth           : num  7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num  16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3
23.4 ...
##  $ Median_Recommendation: chr  "Moderate Buy" "Moderate Buy" "Strong Buy"
"Moderate Sell" ...
##  $ Location             : chr  "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr  "NYSE" "NYSE" "NYSE" "NYSE" ...
```

Load all required libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa
```

```
library(cluster)
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

To remove any missing value that might be present in the data

```
Pharmadata <- na.omit(Pharmadata)
```

Collecting numerical variables from column 1 to 9 to cluster 21 firms

```
row.names(Pharmadata)<- Pharmadata[,1]
P1<- Pharmadata[, 3:11]
head(P1)
```

```
##     Market_Cap Beta PE_Ratio  ROE  ROA Asset_Turnover Leverage Rev_Growth
## ABT      68.44 0.32     24.7 26.4 11.8            0.7     0.42       7.54
## AGN       7.58 0.41     82.5 12.9  5.5            0.9     0.60       9.16
## AHM       6.30 0.46     20.7 14.9  7.8            0.9     0.27       7.05
## AZN      67.63 0.52     21.5 27.4 15.4            0.9     0.00      15.00
## AVE      47.16 0.32     20.1 21.8  7.5            0.6     0.34      26.81
## BAY      16.90 1.11     27.9  3.9  1.4            0.6     0.00      -3.17
##     Net_Profit_Margin
## ABT              16.1
## AGN               5.5
## AHM              11.2
## AZN              18.0
## AVE              12.9
## BAY               2.6
```

Scaling the data using Scale function

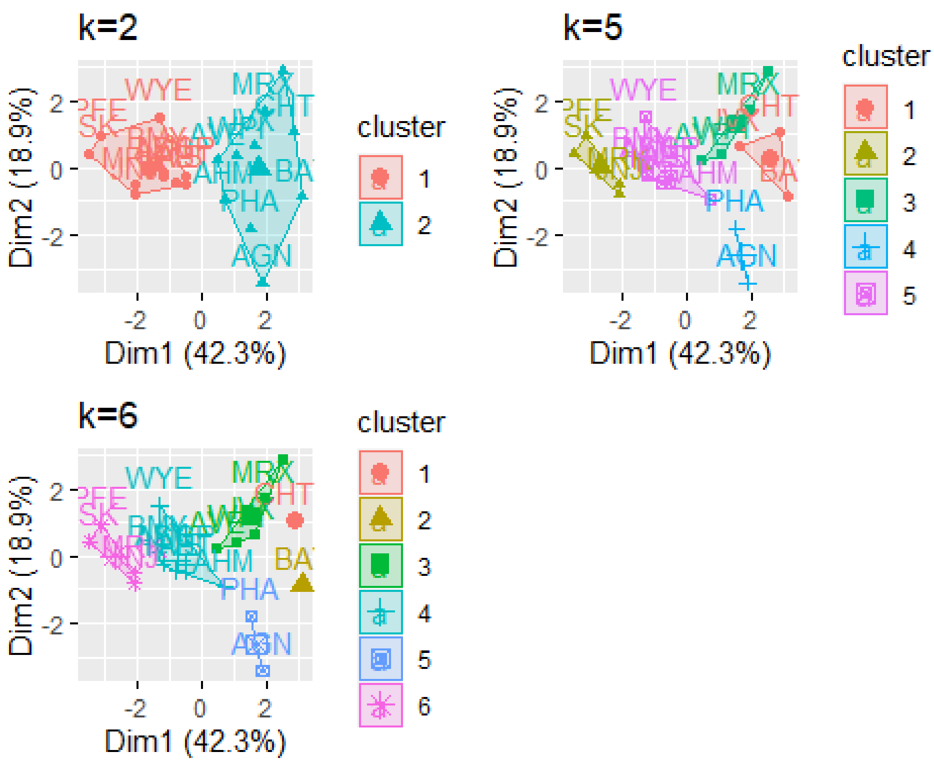
```
dataframe<- scale(P1)
head(dataframe)
```

```
##     Market_Cap        Beta    PE_Ratio         ROE        ROA
Asset_Turnover
## ABT  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121
0.0000000
```

```
## AGN -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871
0.9225312
## AHM -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700
0.9225312
## AZN  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259
0.9225312
## AVE -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -
0.4612656
## BAY -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -
0.4612656
##       Leverage Rev_Growth Net_Profit_Margin
## ABT -0.2120979 -0.5277675        0.06168225
## AGN  0.0182843 -0.3811391       -1.55366706
## AHM -0.4040831 -0.5721181       -0.68503583
## AZN -0.7496565  0.1474473        0.35122600
## AVE -0.3144900  1.2163867       -0.42597037
## BAY -0.7496565 -1.4971443       -1.99560225
```

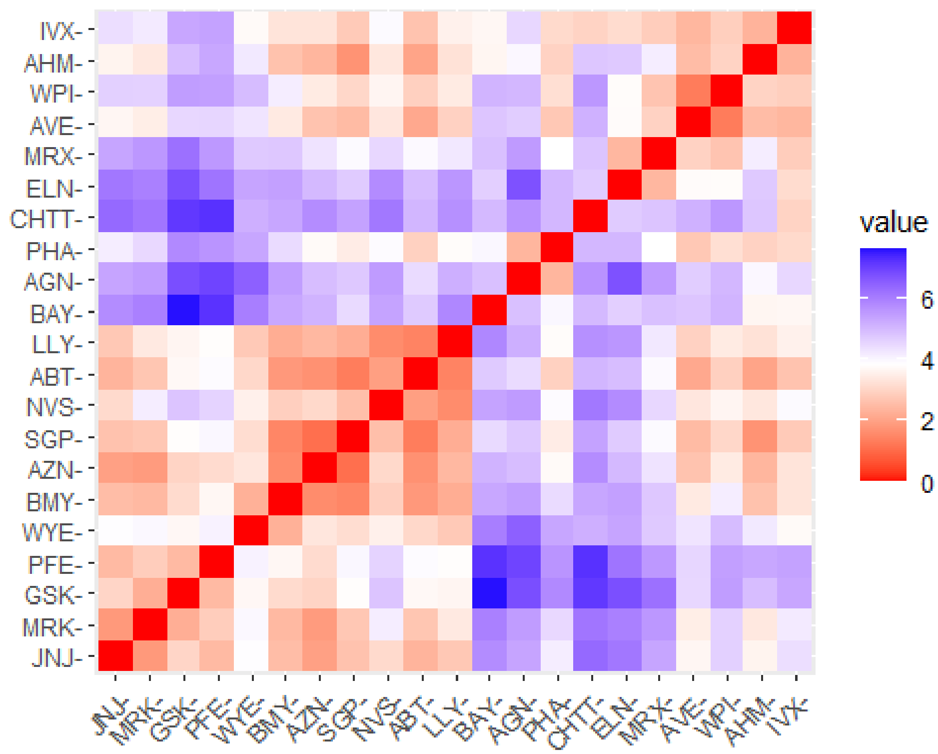
Computing K-means clustering in R for different centers Using multiple values of K and examine the differences in results

```
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)
```



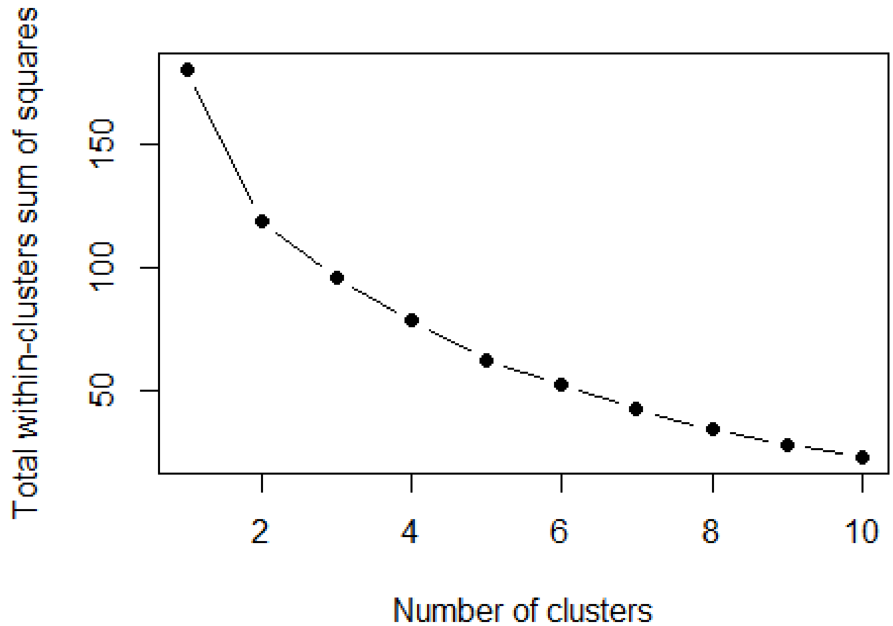
Determining optimal clusters using Elbow method

```
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.

```
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")
```



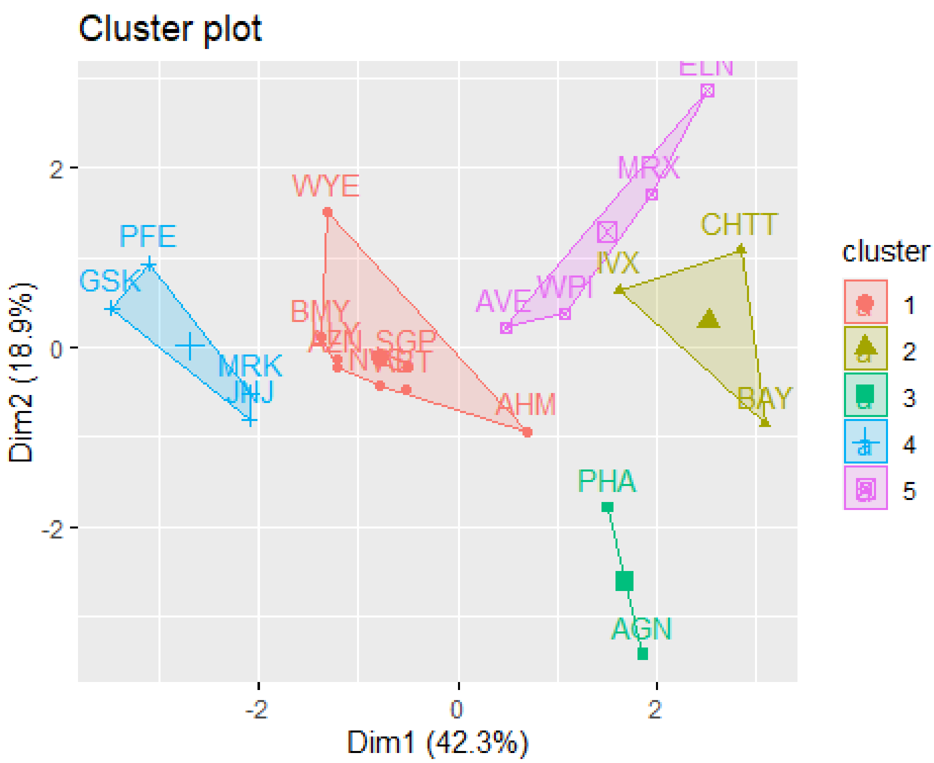
Final analysis and Extracting results using 5 clusters and Visualize the results

```
set.seed(123)
final<- kmeans(dataframe, 5, nstart = 25)
print(final)

## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 3 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 4  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
## 5 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.27449312 -0.7041516       0.556954446
## 2  1.36644699 -0.6912914      -1.320000179
## 3 -0.14170336 -0.1168459      -1.416514761
## 4 -0.46807818  0.4671788       0.591242521
## 5  0.06308085  1.5180158      -0.006893899
##
## Clustering vector:
##  ABT  AGN  AHM  AZN  AVE  BAY  BMY CHTT  ELN  LLY  GSK  IVX  JNJ  MRX  MRK
NVS
##    1    3    1    1    5    2    1    2    5    1    4    2    4    5    4
1
```

```
##   PFE   PHA   SGP   WPI   WYE
##     4     3     1     5     1
##
## Within cluster sum of squares by cluster:
## [1] 21.879320 15.595925  2.803505  9.284424 12.791257
##  (between_SS / total_SS =  65.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(final, data = dataframe)
```

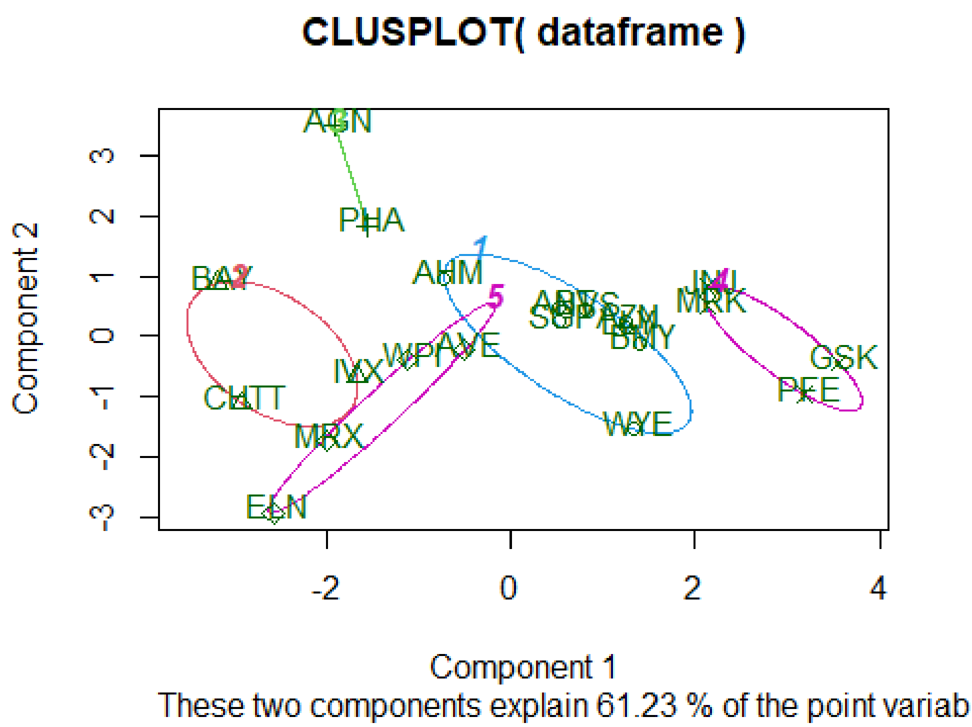



```
P1%>%
  mutate(Cluster = final$cluster) %>%
  group_by(Cluster)%>% summarise_all("mean")
```

```
## # A tibble: 5 x 10
##   Cluster Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##     <int>      <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1       1      55.8  0.414     20.3  28.7 12.7           0.738    0.371
## 2       2       6.64 0.87      24.6  16.5  4.17          0.6      1.65
## 3       3      31.9  0.405     69.5  13.2  5.6           0.75     0.475
## 4       4     157.   0.48      22.2  44.4 17.7           0.95     0.22
```

```
## 5        5       13.1  0.598      17.7  14.6  6.2            0.425    0.635
## # ... with 2 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>

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

```
ClusterForm<- Pharmadata[,c(12,13,14)]%>% mutate(clusters = final$cluster)%>%
arrange(clusters, ascending = TRUE)
ClusterForm

##       Median_Recommendation    Location Exchange clusters
## ABT            Moderate Buy          US     NYSE        1
## AHM              Strong Buy          UK     NYSE        1
## AZN           Moderate Sell          UK     NYSE        1
## BMY           Moderate Sell          US     NYSE        1
## LLY                    Hold          US     NYSE        1
## NVS                    Hold SWITZERLAND     NYSE        1
## SGP                    Hold          US     NYSE        1
## WYE                    Hold          US     NYSE        1
## BAY                    Hold     GERMANY     NYSE        2
## CHTT           Moderate Buy          US   NASDAQ        2
## IVX                    Hold          US     AMEX        2
## AGN            Moderate Buy      CANADA     NYSE        3
## PHA                    Hold          US     NYSE        3
## GSK                    Hold          UK     NYSE        4
```

```
## JNJ          Moderate Buy        US      NYSE        4
## MRK                  Hold        US      NYSE        4
## PFE          Moderate Buy        US      NYSE        4
## AVE          Moderate Buy    FRANCE      NYSE        5
## ELN         Moderate Sell   IRELAND      NYSE        5
## MRX          Moderate Buy        US      NYSE        5
## WPI         Moderate Sell        US      NYSE        5
```

c)Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

```
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