Assignment 4 Kmeans

Franklin Ngochi

2023-03-19

#Loading required libraries

```
library(conflicted)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.0     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.1     v tibble    3.1.8
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
```

```
library(factoextra)
```

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

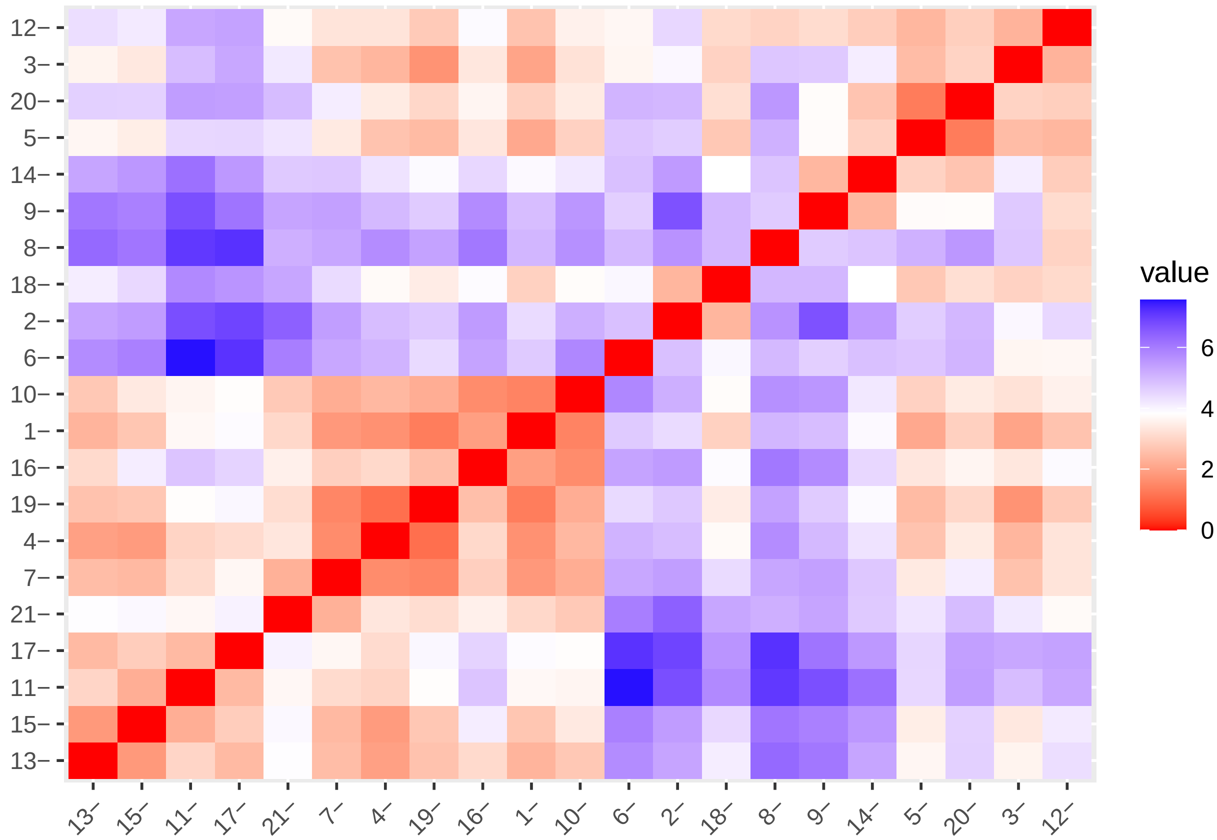
#Importing Dataframe

```
Pharm<-read.csv(file="C:\\Users\\ngoch\\Downloads\\Pharmaceuticals.csv", header=TRUE, sep=",")
colnames(Pharm)
```

```
##  [1] "Symbol"                "Name"                  "Market_Cap"
##  [4] "Beta"                  "PE_Ratio"              "ROE"
##  [7] "ROA"                   "Asset_Turnover"        "Leverage"
## [10] "Rev_Growth"            "Net_Profit_Margin"     "Median_Recommendation"
## [13] "Location"              "Exchange"
```

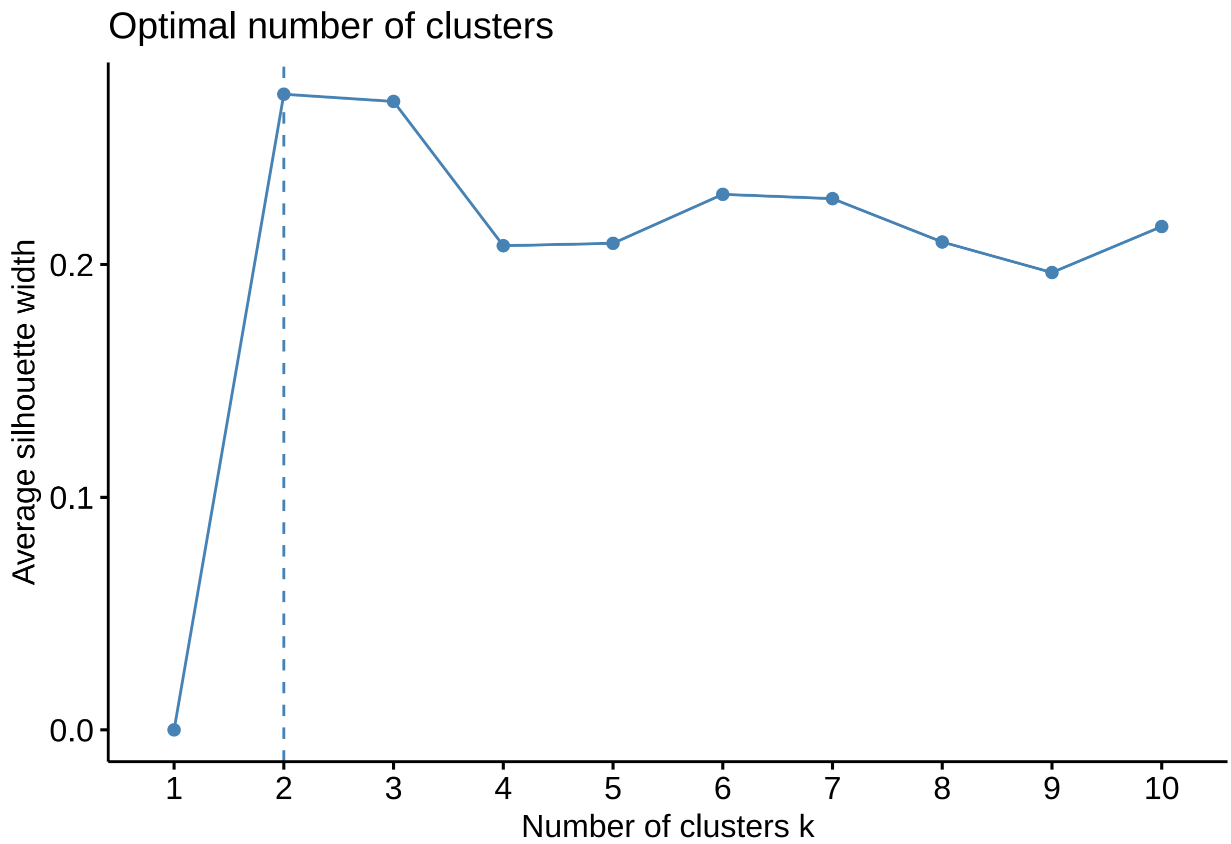
#scaling the dataframe

```
ScaledPharm<-scale(Pharm[, 3:11])
distance<-get_dist(ScaledPharm)
fviz_dist(distance)
```



#Deriving the optimal number of clusters for Kmeans clustering

```
set.seed(456)
fviz_nbclust(ScaledPharm, FUNcluster = hcut, method="silhouette")
```




```r
#the optimal number of clusters is 2 because it corresponds to the highest silhouette width
```

#Performing Kmeans clustering for k=2 (a)

```r
#Distance measure chosen for Kmeans clustering is Euclidean because the data is numerical and has been
k2<-kmeans(ScaledPharm, centers=2, nstart=25)#Kmeans clustering
k2
```

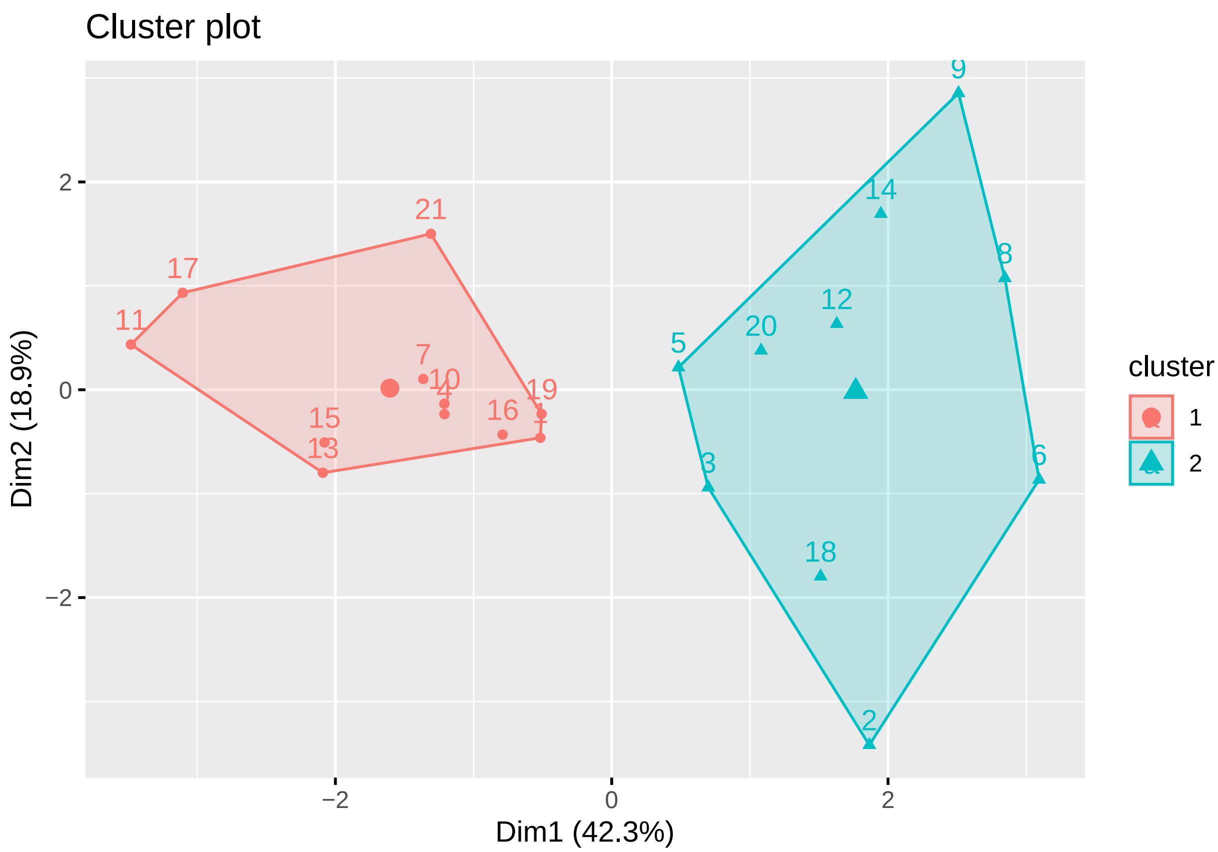
```
## K-means clustering with 2 clusters of sizes 11, 10
##
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.6733825 -0.3586419 -0.2763512  0.6565978  0.8344159      0.4612656
## 2 -0.7407208  0.3945061  0.3039863 -0.7222576 -0.9178575     -0.5073922
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.3331068 -0.2902163         0.6823310
## 2  0.3664175  0.3192379        -0.7505641
##
## Clustering vector:
##  [1] 1 2 2 1 2 2 1 2 2 1 1 2 1 2 1 1 1 2 1 2 1
##
## Within cluster sum of squares by cluster:
## [1] 43.30886 75.26049
##  (between_SS / total_SS =  34.1 %)
##
```

```
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
k2$size#Size of each cluster
```

```
## [1] 11 10
```

```
fviz_cluster(k2, data=ScaledPharm)#Visualize clusters
```



#Interprete the clusters with respect to numerical variables (b)

#Cluster 1 represents companies having positive Market_Cap, ROE, ROA, Asset_Turnover and NetProfitMargin, with negative Beta, PE-Ratio, Leverage and Revenue Growth, while Cluster 2 represents the reverse.

#Ten and eleven companies are in Cluster 1 and 2 respectively

#The within cluster sum of square errors are 43.30886 and 75.26049 respectively for cluster 1 and 2. This suggests that the distribution in Cluster 1 is more compact than in Cluster 2.

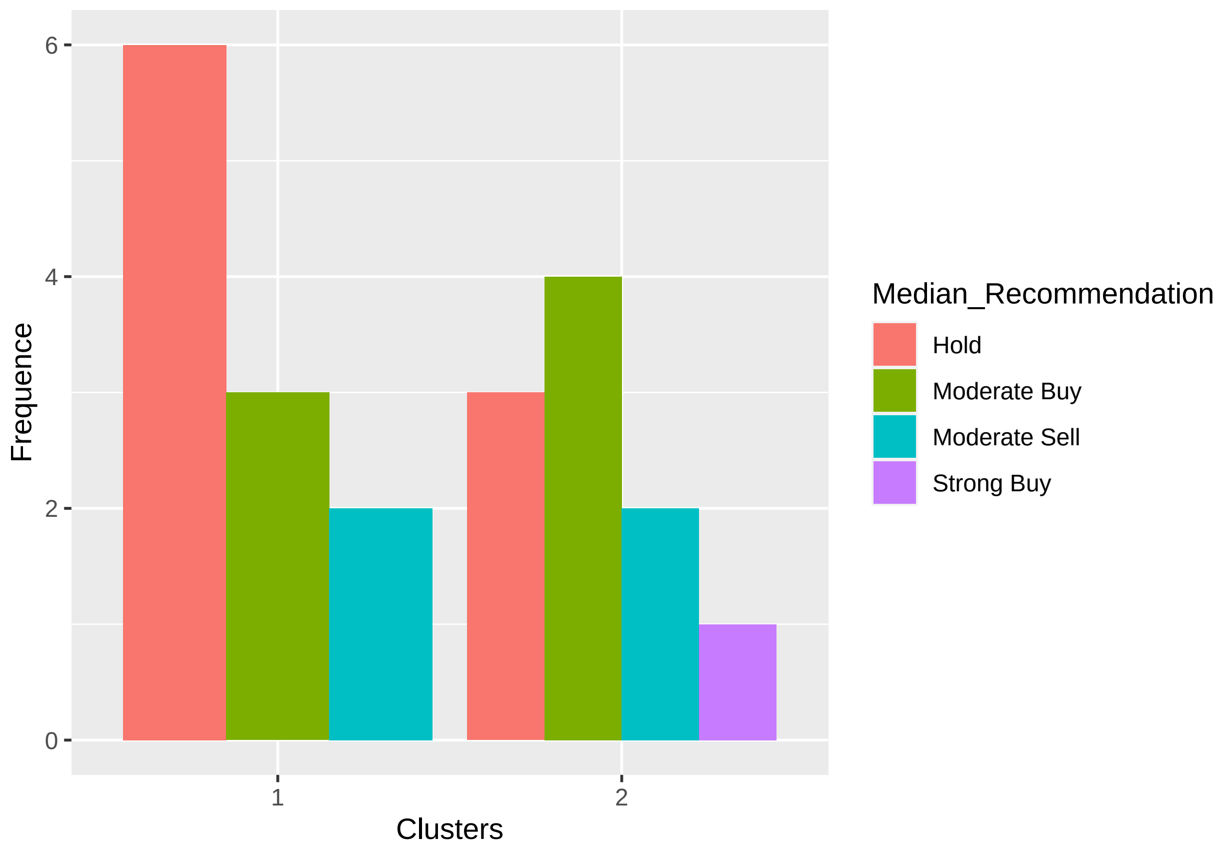
#The overall sum of square errors is 34.1 % which is relatively low and suggests an overall compact distribution.

#Pattern in the cluster (c)

```r
Pattern <- Pharm %>% select(c(12,13,14)) %>% mutate(Cluster = k2$cluster)
print(Pattern)
```

```
##    Median_Recommendation    Location Exchange Cluster
## 1           Moderate Buy          US     NYSE       1
## 2           Moderate Buy      CANADA     NYSE       2
## 3             Strong Buy          UK     NYSE       2
## 4          Moderate Sell          UK     NYSE       1
## 5           Moderate Buy      FRANCE     NYSE       2
## 6                   Hold     GERMANY     NYSE       2
## 7          Moderate Sell          US     NYSE       1
## 8           Moderate Buy          US   NASDAQ       2
## 9          Moderate Sell     IRELAND     NYSE       2
## 10                  Hold          US     NYSE       1
## 11                  Hold          UK     NYSE       1
## 12                  Hold          US     AMEX       2
## 13          Moderate Buy          US     NYSE       1
## 14          Moderate Buy          US     NYSE       2
## 15                  Hold          US     NYSE       1
## 16                  Hold SWITZERLAND     NYSE       1
## 17          Moderate Buy          US     NYSE       1
## 18                  Hold          US     NYSE       2
## 19                  Hold          US     NYSE       1
## 20         Moderate Sell          US     NYSE       2
## 21                  Hold          US     NYSE       1
```

```r
Median_Recommendation <- ggplot(Pattern, mapping = aes(factor(Cluster), fill=Median_Recommendation)) + g
Median_Recommendation
```




```
#There appears to be a pattern consisting of brokerages recommending a strong buy for companies in Clus
```

#Provide appropriate names for each cluster (d)

#Cluster1 could be named LOW RISK HIGH RETURN because although they have on average negative revenue growth, they show on average a positive market capitalization, ROE, ROA, Asset Turnover and Net Profit. They also have negative leverage, negative Beta and PE_Ratio

#Cluster2 on the contrary is HIGH RISK LOW RETURN because companies in this cluster portray opposite characteristics to cluster 1 companies. That is, higher Beta and PE ratio on average as well as higher leverage while running negative net profit margins, market capitalization, asset turnover etc.