Pharmaceuticals

Abhinav Reddy

2024-03-17

This below R code is importing the required libraries.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## Warning: package 'forcats' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.3
```

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

```
library(ISLR)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

This below R code is reading the CSV file named "Pharmaceuticals.csv" and removing the rows with missing values, and the cleaned data is stored in a new data frame named "CD"

```
inp_data <- read.csv("C:/Users/Abhinav Reddy/Desktop/FML/Assignment 4/Pharmaceuticals.csv")
CD<- na.omit(inp_data)
```

Question (a): Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on?

Answer:

Numerical variables(1 to 9) are taken to cluster 21 firms.

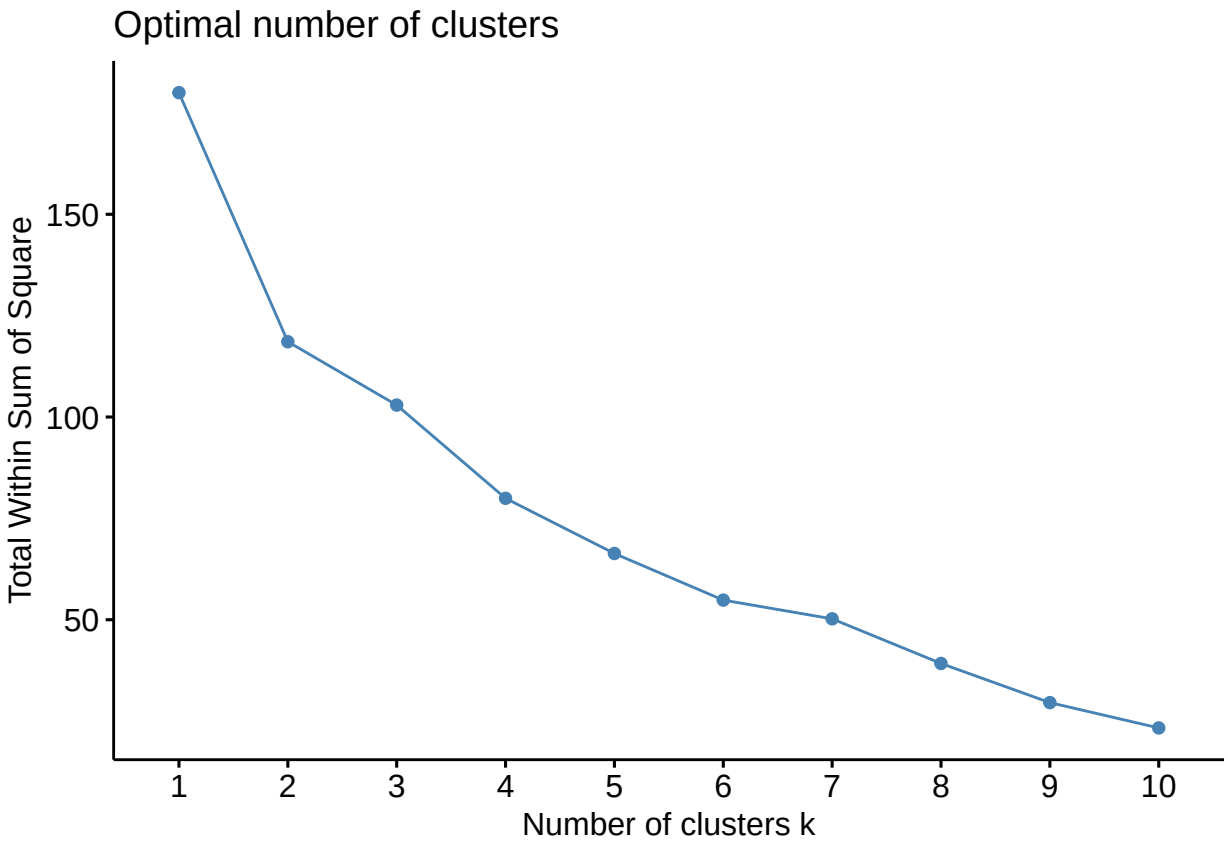
```
row.names(CD)<- CD[,1]
CD1<- CD%>% select('Market_Cap', 'Beta', 'PE_Ratio', 'ROE', 'ROA', 'Asset_Turnover', 'Leverage', 'Rev_G
head(CD1)
```

```
##     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
```

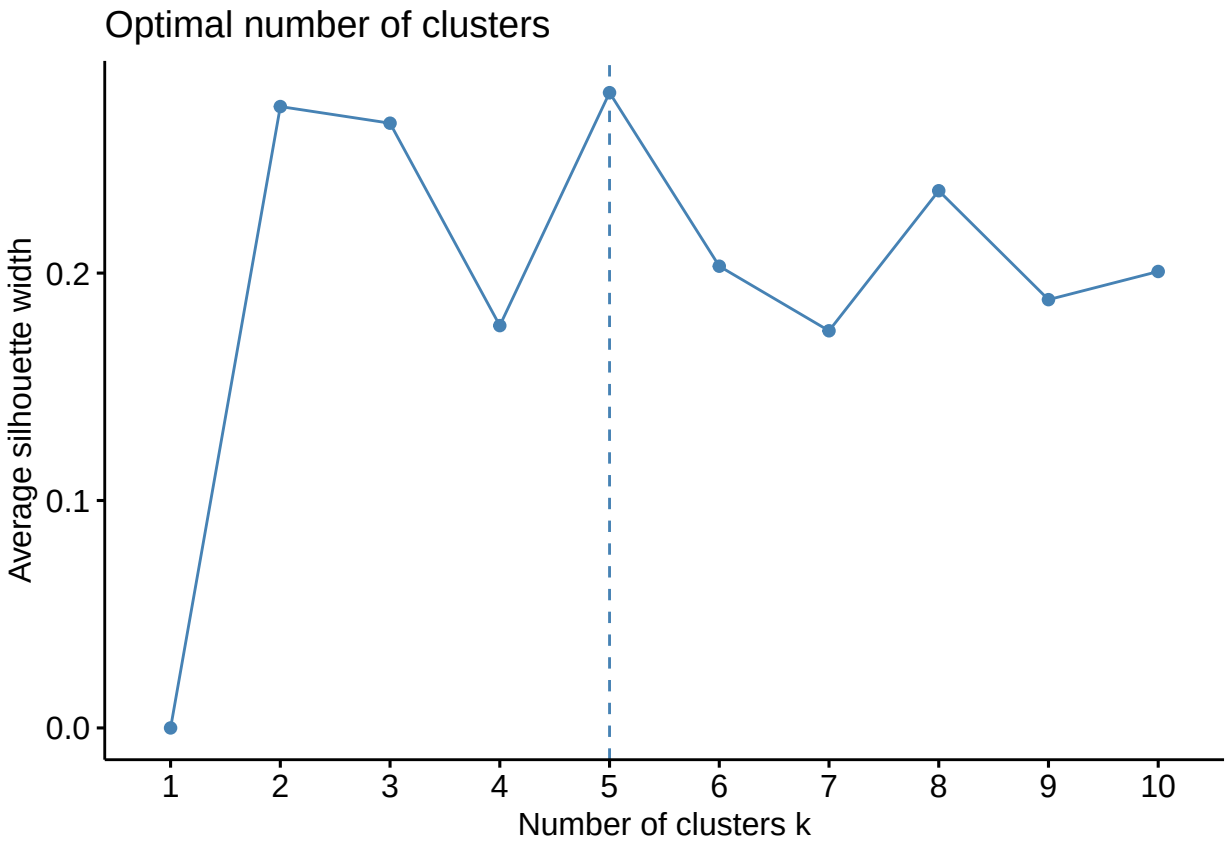
```
sc_data<-scale(CD1)
head(sc_data)
```

```
##     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
```

```
fviz_nbclust(sc_data, kmeans, method = "wss")
```




```
fviz_nbclust(sc_data, kmeans, method = "silhouette")
```



From above graph based on the silhouette method, selecting $\mathbf{k = 5}$ maximizes average silhouette width, indicating that the data points are effectively clustered into the distinct groups.

```
set.seed(123)
k_n5<- kmeans(sc_data,centers=5,nstart = 25)
k_n5
```

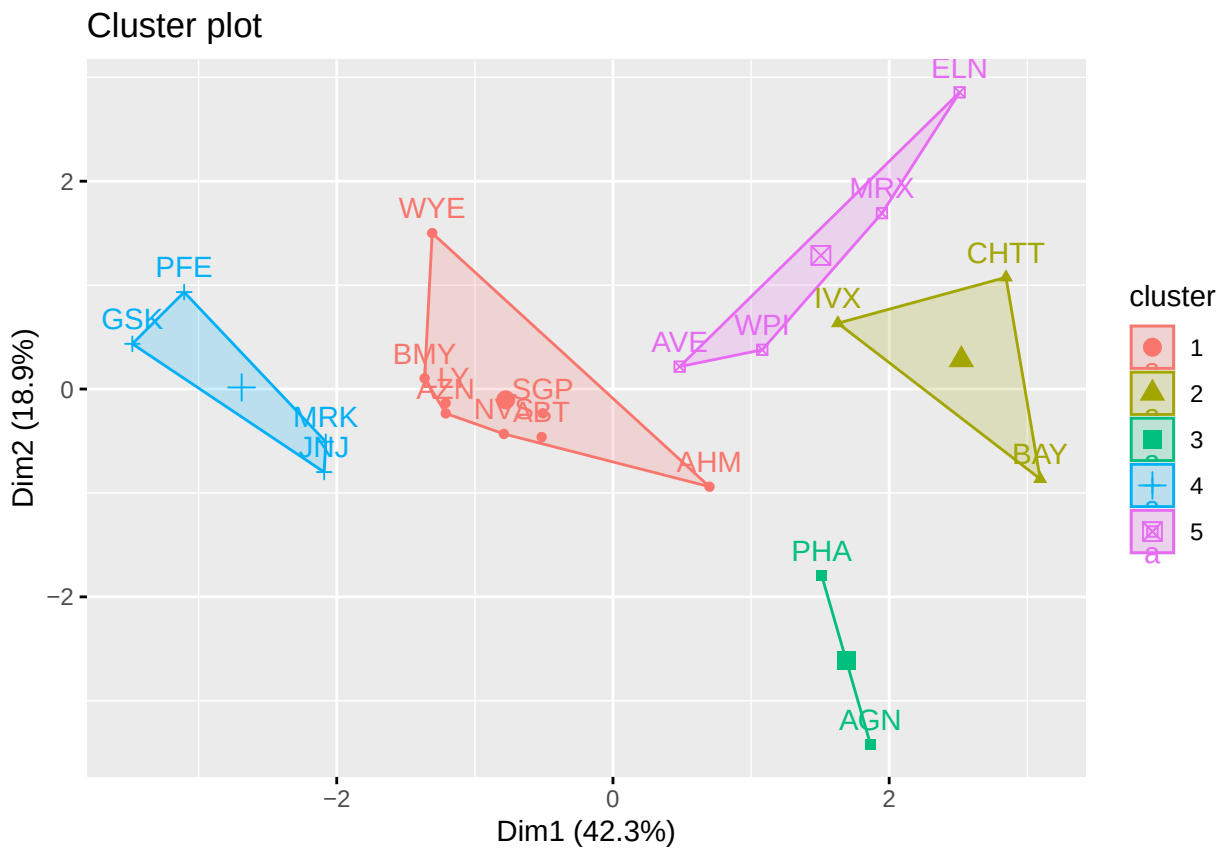
```
## K-means clustering with 5 clusters of sizes 8, 3, 2, 4, 4
##
## Cluster means:
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 2 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -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"
```

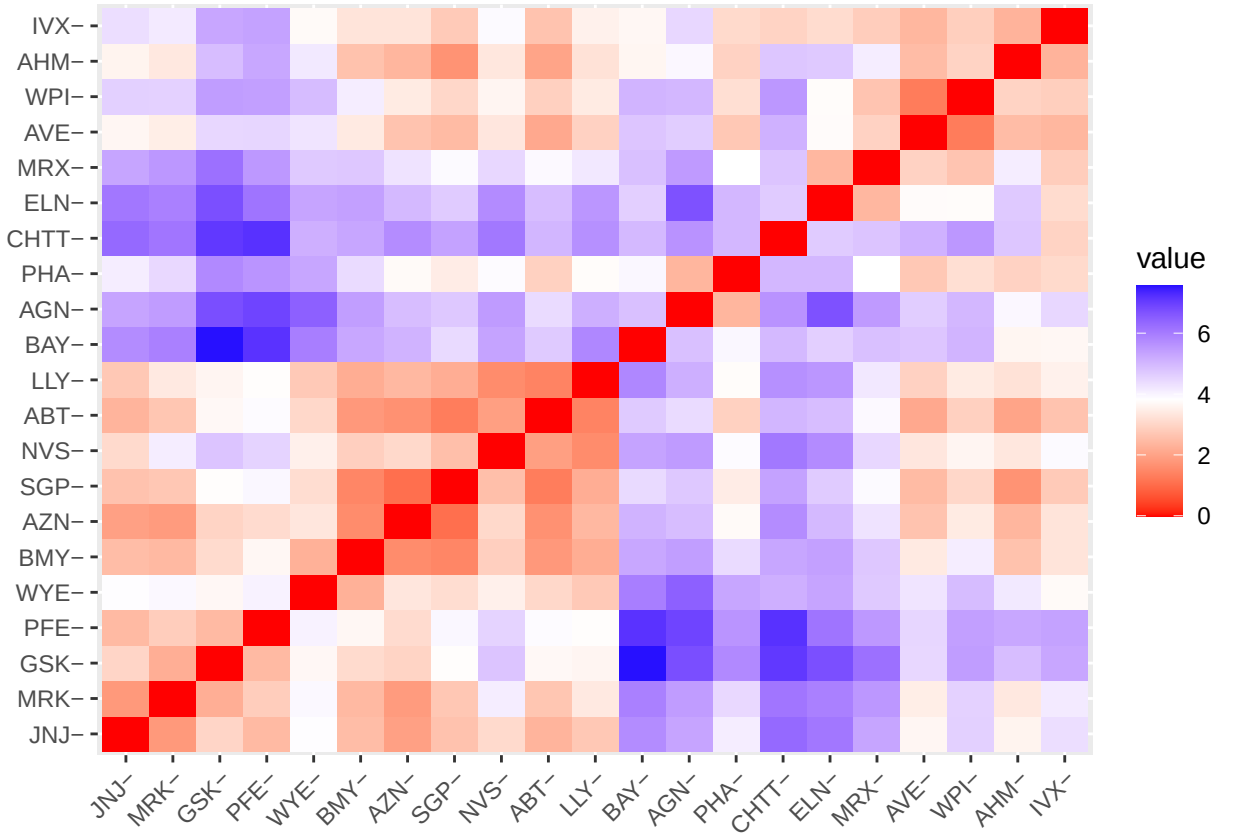
```
k_n5$centers
```

```
##    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
```

```
fviz_cluster(k_n5,data = sc_data)
```




```
D<- dist(sc_data, method = "euclidean")
fviz_dist(D)
```




```
Fit_<-kmeans(sc_data,5)
aggregate(sc_data,by=list(Fit_$cluster),FUN=mean)
```

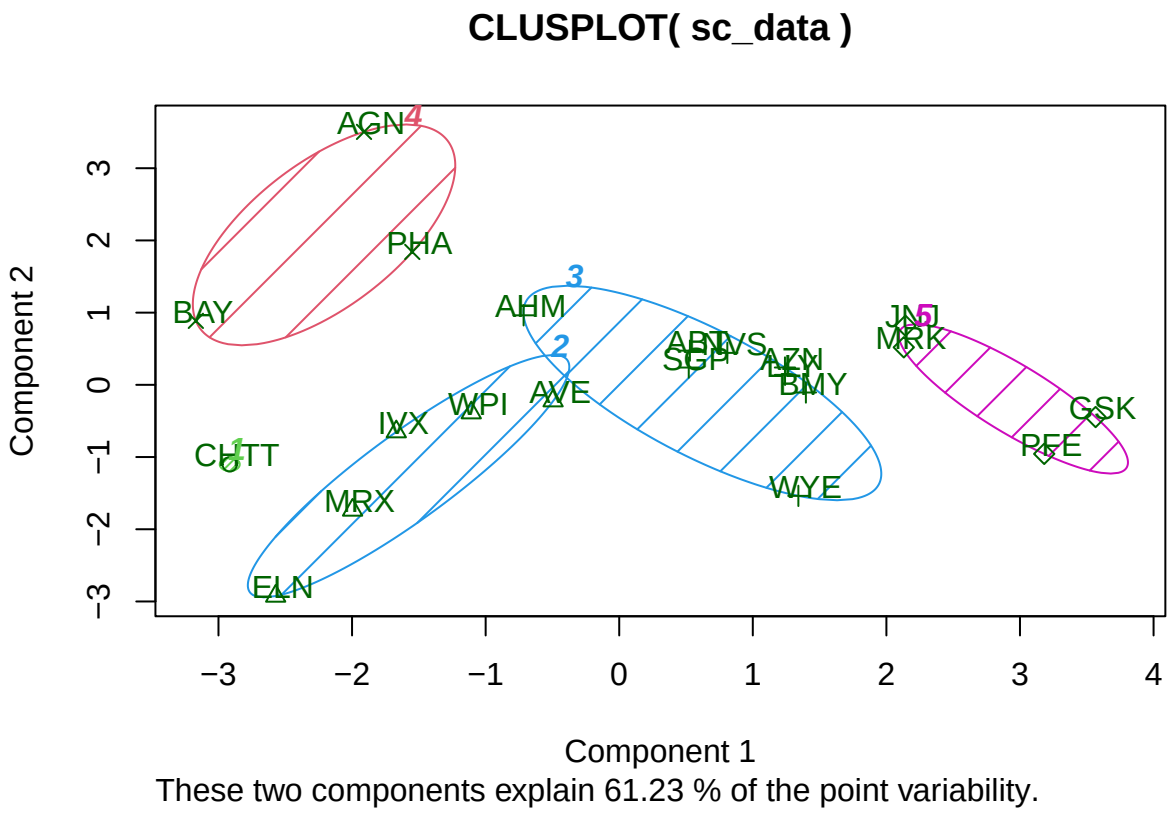
```
##   Group.1  Market_Cap       Beta    PE_Ratio        ROE        ROA
## 1       1 -0.97676686  1.2630872  0.03299122 -0.1123792 -1.1677918
## 2       2 -0.79605926  0.3205014 -0.45014035 -0.6533148 -0.7881923
## 3       3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915
## 4       4 -0.52462814  0.4451409  1.84984387 -1.0404550 -1.1865838
## 5       5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1  -4.612656e-01  3.7427970 -0.6327607        -1.2488842
## 2  -1.107037e+00  0.2717048  1.2256188        -0.1486179
## 3   1.729746e-01 -0.2744931 -0.7041516         0.5569544
## 4   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 5   1.153164e+00 -0.4680782  0.4671788         0.5912425
```

```
CD3<-data.frame(sc_data,Fit_$cluster)
head(CD3)
```

```
##     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 Fit_.cluster
## ABT -0.2120979 -0.5277675        0.06168225            3
## AGN  0.0182843 -0.3811391       -1.55366706            4
## AHM -0.4040831 -0.5721181       -0.68503583            3
## AZN -0.7496565  0.1474473        0.35122600            3
## AVE -0.3144900  1.2163867       -0.42597037            2
## BAY -0.7496565 -1.4971443       -1.99560225            4
```

```
library(cluster)
clusplot(sc_data,Fit_$cluster,color = TRUE,shade = TRUE,labels = 2,lines = 0)
```



Justification

-Variable Selection: numerical variables are used for clustering, ensuring compatibility with k-means algorithm.

-Scaling: Data is scaled to prevent variables with larger scales from dominating the clustering process.

-weights for different variables: Equal weights are assigned to all numerical variables.

-Determining the Number of Clusters: Optimal number of clusters (k=5) is chosen based on silhouette method, indicating well-separated clusters.

-Clustering Algorithm: K-means clustering. Visualization: Cluster is visualized using fviz_cluster and clusplot.

Question (b): Interpret the clusters with respect to the numerical variables used in forming the clusters?

Answer:

Interpreting clusters using mean values of quantitative variables shows the cluster characteristics and differences among the clusters.

```
aggregate(sc_data,by=list(Fit_$cluster),FUN=mean)
```

```
##   Group.1  Market_Cap       Beta    PE_Ratio        ROE        ROA
## 1       1 -0.97676686  1.2630872  0.03299122 -0.1123792 -1.1677918
## 2       2 -0.79605926  0.3205014 -0.45014035 -0.6533148 -0.7881923
## 3       3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915
## 4       4 -0.52462814  0.4451409  1.84984387 -1.0404550 -1.1865838
## 5       5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1  -4.612656e-01  3.7427970 -0.6327607        -1.2488842
## 2  -1.107037e+00  0.2717048  1.2256188        -0.1486179
## 3   1.729746e-01 -0.2744931 -0.7041516         0.5569544
## 4   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 5   1.153164e+00 -0.4680782  0.4671788         0.5912425
```

Interpreting:

- **Cluster 1**:
  - Highest: Beta (Mean: 1.263)
  - Lowest: Asset_Turnover (Mean: -0.461)
  - Moderate: Market_Cap (Mean: -0.977), PE_Ratio (Mean: 0.033), ROE (Mean: -0.112), ROA (Mean: -1.168), Leverage (Mean: 3.743), Rev_Growth (Mean: -0.633), Net_Profit_Margin (Mean: -1.249)
- **Cluster 2**:
  - Highest: Rev_Growth (Mean: 1.226)
  - Lowest: Asset_Turnover (Mean: -1.107)
  - Moderate: Market_Cap (Mean: -0.796), Beta (Mean: 0.321), PE_Ratio (Mean: -0.450), ROE (Mean: -0.653), ROA (Mean: -0.788), Leverage (Mean: 0.272), Net_Profit_Margin (Mean: -0.149)
- **Cluster 3**:
  - Highest: Asset_Turnover (Mean: 0.173), Net_Profit_Margin (Mean: 0.557)
  - Lowest: Leverage (Mean: -0.274), Rev_Growth (Mean: -0.704)
  - Moderate: Market_Cap (Mean: -0.031), Beta (Mean: -0.436), PE_Ratio (Mean: -0.317), ROE (Mean: 0.195), ROA (Mean: 0.408)
- **Cluster 4**:
  - Highest: PE_Ratio (Mean: 1.850)
  - Lowest: Asset_Turnover (Mean: 0), Net_Profit_Margin (Mean: -1.610)
  - Moderate: Market_Cap (Mean: -0.525), Beta (Mean: 0.445), ROE (Mean: -1.040), ROA (Mean: -1.187), Leverage (Mean: -0.344), Rev_Growth (Mean: -0.577)
- **Cluster 5**:
  - Highest: Net_Profit_Margin (Mean: 0.591)
  - Lowest: Leverage (Mean: -0.468)

– Moderate: Market_Cap (Mean: 1.696), Beta (Mean: -0.178), PE_Ratio (Mean: -0.198), ROE (Mean: 1.235), ROA (Mean: 1.350), Asset_Turnover (Mean: 1.153), Rev_Growth (Mean: 0.467)

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

Answer:

```
Pharma_Pattern <- inp_data %>% select(c(12,13,14)) %>% mutate(Cluster = k_n5$cluster)
print(Pharma_Pattern)
```

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

There is a pattern in the clusters with respect to the Median_Recommendation variable.

- Cluster 1: Mix of recommendations like Moderate Buy, Strong Buy, Moderate Sell, and Hold, with companies from various locations and exchanges.
- Cluster 2: Mostly filled with Hold recommendations, mainly listed on NYSE and NASDAQ.
- Cluster 3: Dominated by Hold recommendations, primarily listed on NYSE.
- Cluster 4: Mainly consists of Hold and Moderate Buy recommendations, listed predominantly on NYSE.
- Cluster 5: Mix of recommendations like Moderate Buy, Moderate Sell, and Hold, with companies from various locations and exchanges.

Question (d): Provide an appropriate name for each cluster using any or all of the variables in the dataset?

Answer: Appropriate name for each cluster:

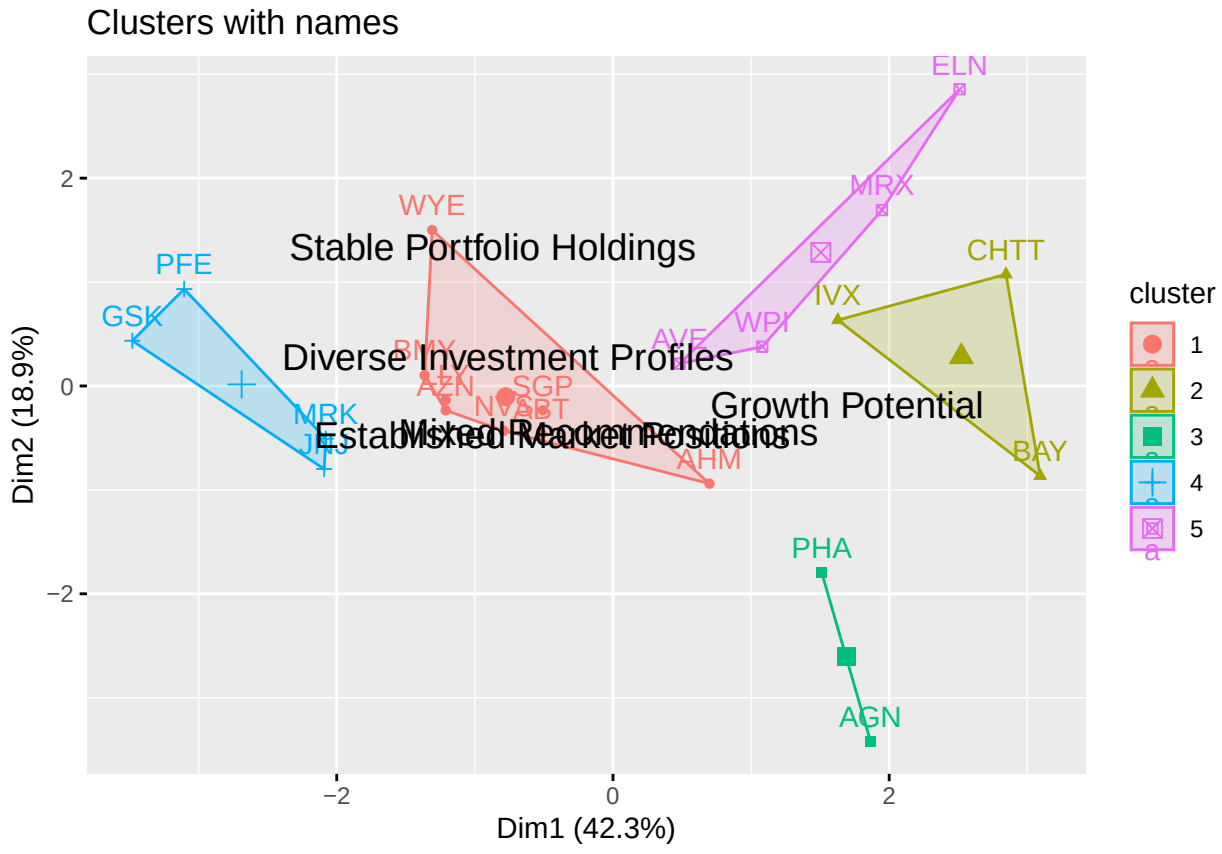
- **Cluster 1:** Mixed Recommendations
- **Cluster 2:** Stable Portfolio Holdings
- **Cluster 3:** Established Market Positions
- **Cluster 4:** Growth Potential
- **Cluster 5:** Diverse Investment Profiles

```
library(ggplot2)


cn <- c("Mixed Recommendations",
                  "Stable Portfolio Holdings",
                  "Established Market Positions",
                  "Growth Potential",
                  "Diverse Investment Profiles")

p <- fviz_cluster(k_n5, data = sc_data,
                  main = "Clusters with names") +
  annotate("text", x = k_n5$centers[,1],
           y = k_n5$centers[,2], label = cn,
           color = "black", size = 5)

print(p)
```



Thank You!!!