FML-Assignment4_Clustering

Anusha Banda

2023-11-12

Summary - The script loads necessary libraries and imports pharmaceutical data from a CSV file. It then ensures data integrity by removing missing values and performs k-means clustering on numerical variables (columns 1-9), scaling them for normalization. The optimal number of clusters (k=5) is determined using the Elbow Method and Gap Statistic. The k-means algorithm forms clusters, and mean values of variables within each cluster are visualized. Cluster interpretation reveals distinctive characteristics, guiding recommendations based on median recommendations (variables 10-12). Clusters 1, 2, 4, and 5 are generally recommended to be held. Each cluster is named for better understanding, and a cluster plot visually represents data point distribution across clusters.

Load the required packages, please see below,

```r
library(factoextra) # clustering algorithms & visualization
```

```
## Warning: package 'factoextra' was built under R version 4.3.2

## Loading required package: ggplot2

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

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

```
## Loading required package: lattice
```

We should import the data and i am using readr library to read my CSV file, please see below,

```r
library(readr)

pharma_data <- read_csv("C:\\Users\\banda\\Downloads\\Pharmaceuticals.csv", show_col_types = FALSE)

pharma_data
```

```
## # A tibble: 21 x 14
##    Symbol Name     Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##    <chr>  <chr>         <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
##  1 ABT    Abbott ~      68.4   0.32     24.7  26.4  11.8            0.7     0.42
##  2 AGN    Allerga~       7.58  0.41     82.5  12.9   5.5            0.9     0.6
##  3 AHM    Amersha~       6.3   0.46     20.7  14.9   7.8            0.9     0.27
##  4 AZN    AstraZe~      67.6   0.52     21.5  27.4  15.4            0.9     0
##  5 AVE    Aventis       47.2   0.32     20.1  21.8   7.5            0.6     0.34
##  6 BAY    Bayer AG      16.9   1.11     27.9   3.9   1.4            0.6     0
##  7 BMY    Bristol~      51.3   0.5      13.9  34.8  15.1            0.9     0.57
##  8 CHTT   Chattem~       0.41  0.85     26    24.1   4.3            0.6     3.51
##  9 ELN    Elan Co~       0.78  1.08      3.6  15.1   5.1            0.3     1.07
## 10 LLY    Eli Lil~      73.8   0.18     27.9  31    13.5            0.6     0.53
## # i 11 more rows
## # i 5 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>
```

1. Cluster the 21 firms using only the numerical variables (1–9). Justify the various decisions made during the cluster analysis, such as variable weights, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

Before initiating the clustering process, exclude any missing data and normalize variables to ensure they are comparable.

```
Pharma<- na.omit(pharma_data)
Pharma
```

```
## # A tibble: 21 x 14
##    Symbol Name     Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
##    <chr>  <chr>         <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>
##  1 ABT    Abbott ~      68.4   0.32     24.7  26.4  11.8            0.7     0.42
##  2 AGN    Allerga~       7.58  0.41     82.5  12.9   5.5            0.9     0.6
##  3 AHM    Amersha~       6.3   0.46     20.7  14.9   7.8            0.9     0.27
##  4 AZN    AstraZe~      67.6   0.52     21.5  27.4  15.4            0.9     0
##  5 AVE    Aventis       47.2   0.32     20.1  21.8   7.5            0.6     0.34
##  6 BAY    Bayer AG      16.9   1.11     27.9   3.9   1.4            0.6     0
##  7 BMY    Bristol~      51.3   0.5      13.9  34.8  15.1            0.9     0.57
##  8 CHTT   Chattem~       0.41  0.85     26    24.1   4.3            0.6     3.51
##  9 ELN    Elan Co~       0.78  1.08      3.6  15.1   5.1            0.3     1.07
## 10 LLY    Eli Lil~      73.8   0.18     27.9  31    13.5            0.6     0.53
## # i 11 more rows
## # i 5 more variables: Rev_Growth <dbl>, Net_Profit_Margin <dbl>,
## #   Median_Recommendation <chr>, Location <chr>, Exchange <chr>
```

Consider only the numerical variables (1-9) for clustering the 21 firms.

```
row.names(Pharma) <- Pharma$Symbol
```

```
## Warning: Setting row names on a tibble is deprecated.
```

```
Pharma_1 <- Pharma[, 3:11]

head(Pharma_1)
```

```
## # A tibble: 6 x 9
##   Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Growth
##        <dbl> <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>      <dbl>
## 1      68.4   0.32     24.7  26.4  11.8            0.7     0.42       7.54
## 2       7.58  0.41     82.5  12.9   5.5            0.9     0.6        9.16
## 3       6.3   0.46     20.7  14.9   7.8            0.9     0.27       7.05
## 4      67.6   0.52     21.5  27.4  15.4            0.9     0         15
## 5      47.2   0.32     20.1  21.8   7.5            0.6     0.34      26.8
## 6      16.9   1.11     27.9   3.9   1.4            0.6     0         -3.17
## # i 1 more variable: Net_Profit_Margin <dbl>
```

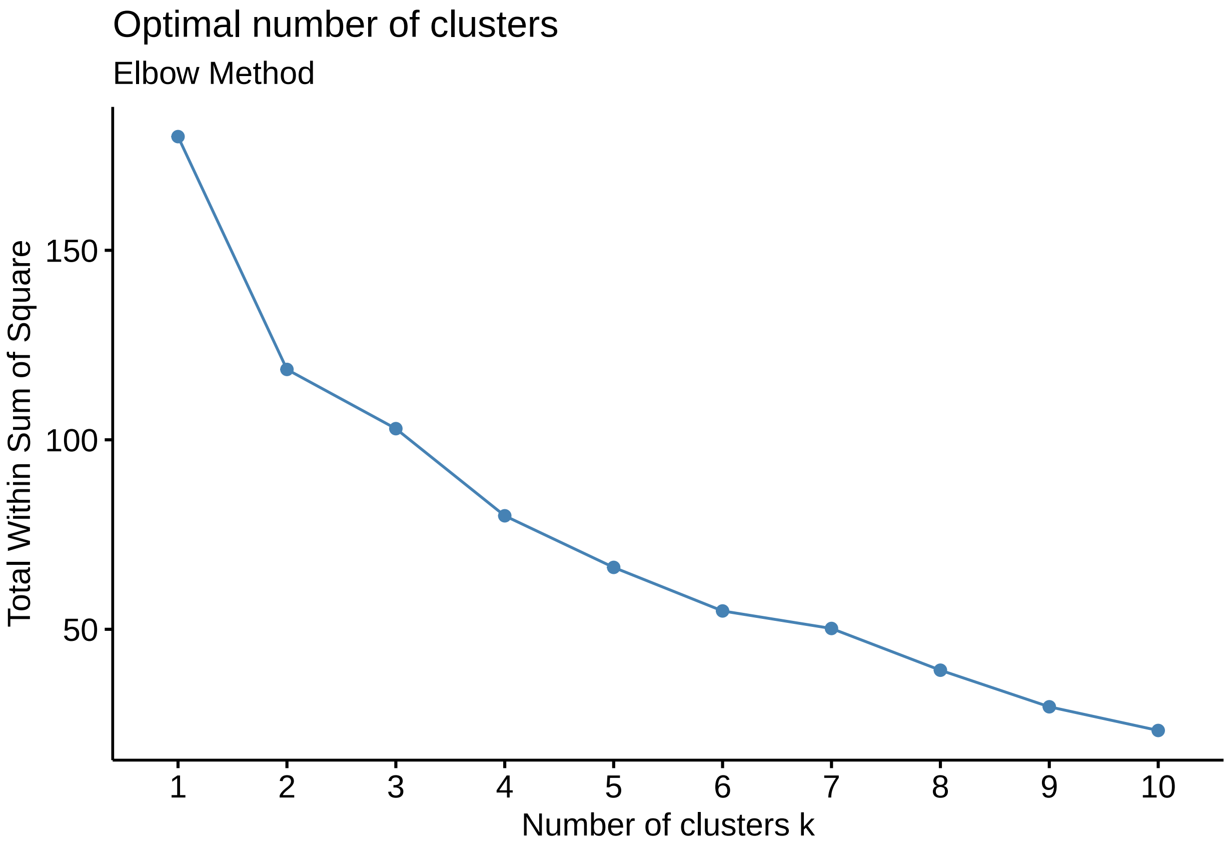
Normalize the quantitative variables of the dataframe

```
Pharma_2<-scale(Pharma_1)
head(Pharma_2)
```

```
##       Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
## [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
## [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
##        Leverage Rev_Growth Net_Profit_Margin
## [1,] -0.2120979 -0.5277675        0.06168225
## [2,]  0.0182843 -0.3811391       -1.55366706
## [3,] -0.4040831 -0.5721181       -0.68503583
## [4,] -0.7496565  0.1474473        0.35122600
## [5,] -0.3144900  1.2163867       -0.42597037
## [6,] -0.7496565 -1.4971443       -1.99560225
```

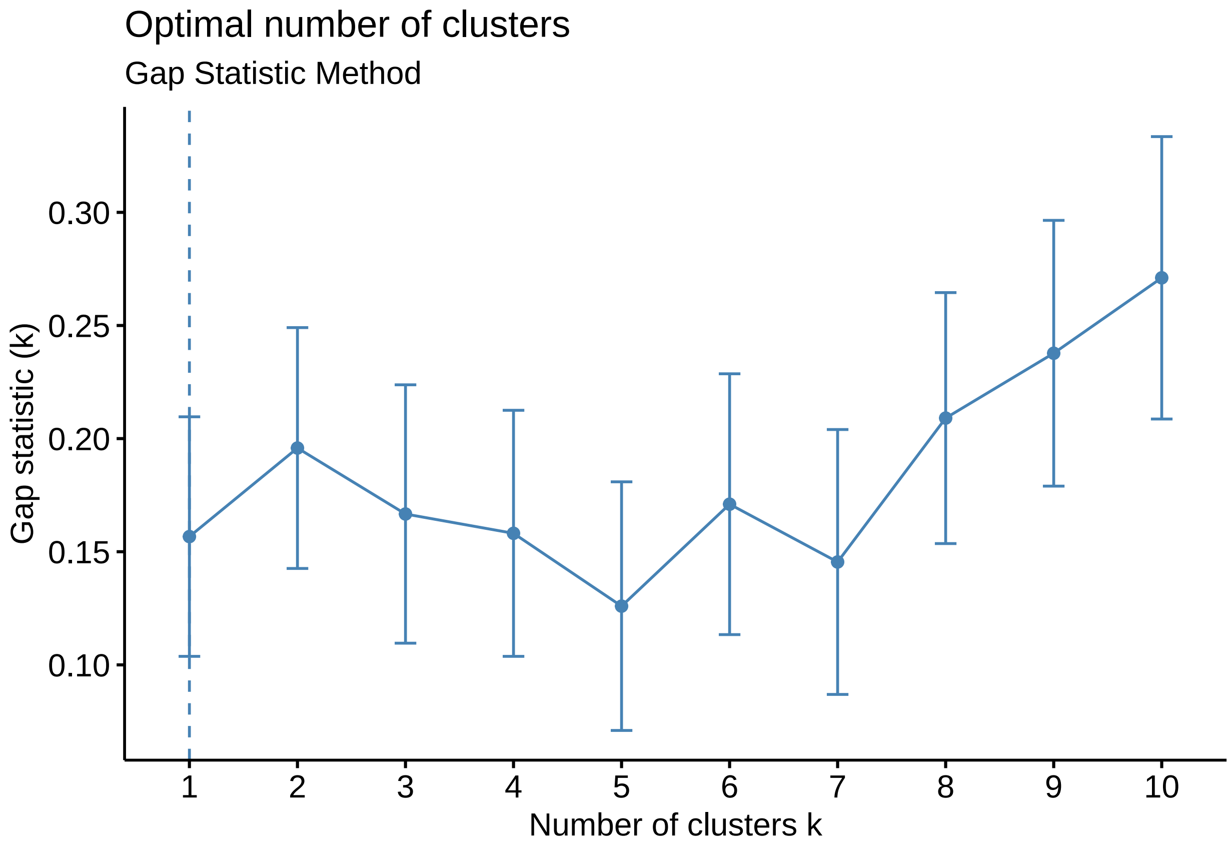
Apply the Elbow Method to identify the most appropriate number of clusters for the cluster analysis.

```
fviz_nbclust(Pharma_2, kmeans, method = "wss") + labs(subtitle = "Elbow Method")
```



Utilize the Gap Statistic Method to determine the suitable number of clusters for the analysis.

```
fviz_nbclust(Pharma_2, kmeans, method = "gap_stat") + labs(subtitle = "Gap Statistic Method")
```

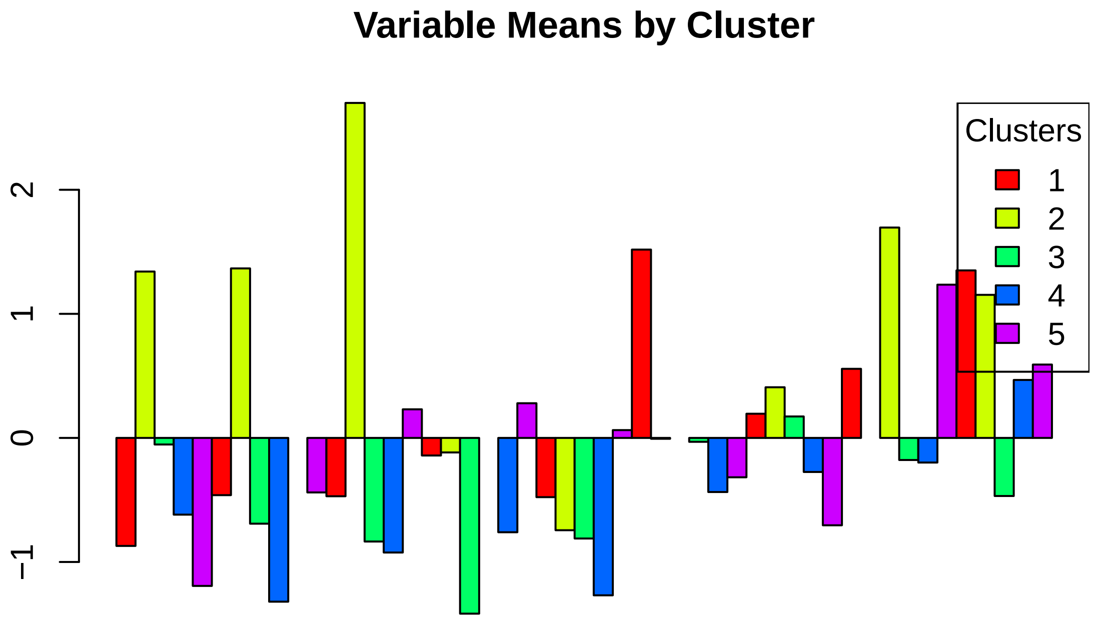


Based on the plots above, we can discern 5 clusters, effectively illustrating the variations present in the data.

```
set.seed(64060)
k5<- kmeans(Pharma_2,centers=5,nstart = 25)
```

In the visualization process, we showcase the mean values of each variable within individual clusters.

```
cluster_means <- aggregate(Pharma_2, by = list(k5$cluster), FUN = mean)
barplot(t(cluster_means[, -1]), beside = TRUE, col = rainbow(5), main = "Variable Means by Cluster")
legend("topright", legend = 1:5, title = "Clusters", fill = rainbow(5))
```



Applied K-Means Cluster Analysis for segmenting the data into 5 clusters.

```
fit<-kmeans(Pharma_2,5)
fit
```

```
## K-means clustering with 5 clusters of sizes 4, 3, 4, 3, 7
##
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431   1.153164e+00
## 2 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022  -1.537552e-01
## 3 -0.96247577  1.1949250 -0.3639982 -0.5200697 -0.9610792  -1.153164e+00
## 4 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838   1.480297e-16
## 5  0.08926902 -0.4618336 -0.3208615  0.3260892  0.5396003   6.589509e-02
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.4680782  0.4671788         0.5912425
## 2 -0.4040831  0.6917224        -0.4005718
## 3  1.4773718  0.7120120        -0.3688236
## 4 -0.3443544 -0.5769454        -1.6095439
## 5 -0.2559803 -0.7230135         0.7343816
##
## Clustering vector:
##  [1] 5 4 2 5 2 4 5 3 3 5 1 3 1 3 1 5 1 4 5 2 5
##
## Within cluster sum of squares by cluster:
## [1]  9.284424  5.511294 19.219788 14.938904 16.655937
##  (between_SS / total_SS =  63.5 %)
```

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

Calculated the mean for each cluster across all quantitative variables.

```
aggregate(Pharma_2,by=list(fit$cluster),FUN=mean)
```

```
##   Group.1  Market_Cap       Beta   PE_Ratio        ROE        ROA
## 1       1  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431
## 2       2 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022
## 3       3 -0.96247577  1.1949250 -0.3639982 -0.5200697 -0.9610792
## 4       4 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838
## 5       5  0.08926902 -0.4618336 -0.3208615  0.3260892  0.5396003
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1   1.153164e+00 -0.4680782  0.4671788         0.5912425
## 2  -1.537552e-01 -0.4040831  0.6917224        -0.4005718
## 3  -1.153164e+00  1.4773718  0.7120120        -0.3688236
## 4   1.480297e-16 -0.3443544 -0.5769454        -1.6095439
## 5   6.589509e-02 -0.2559803 -0.7230135         0.7343816
```

```
Pharma_3<-data.frame(Pharma_2,fit$cluster)
Pharma_3
```

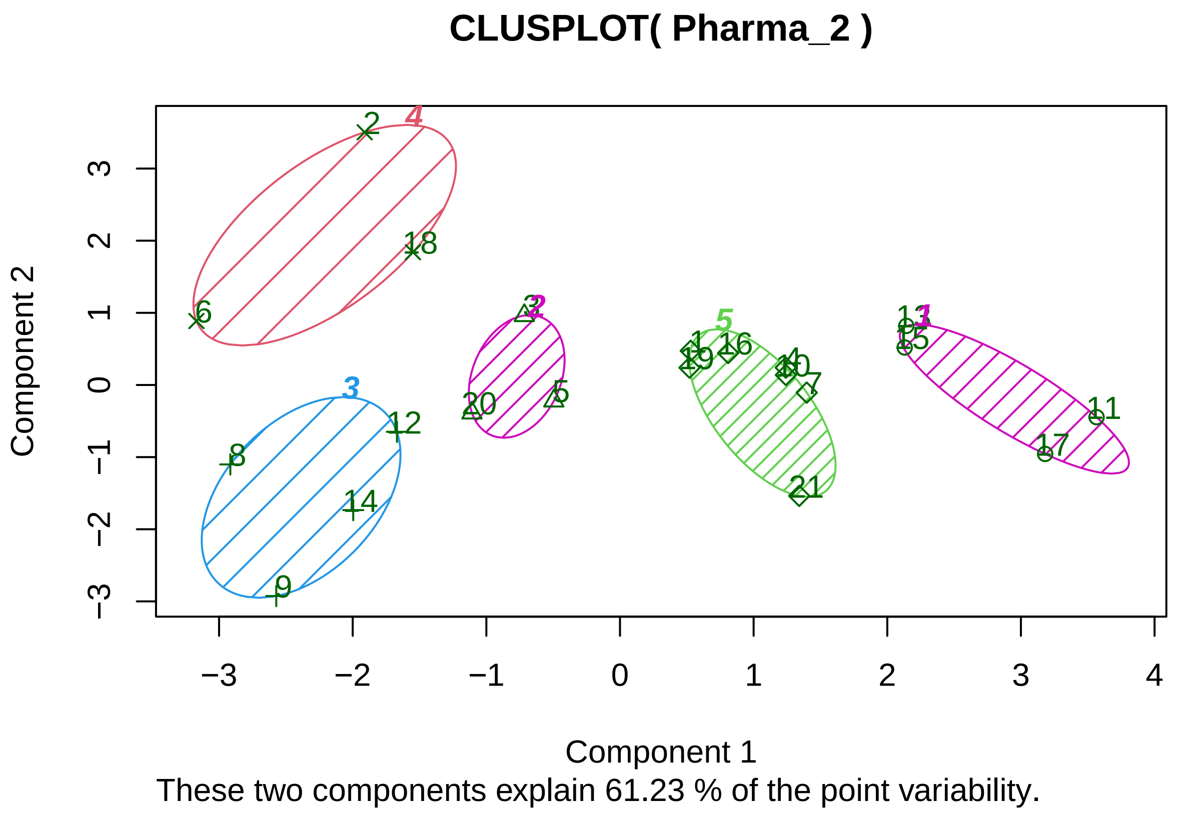
```
##    Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## 1   0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121      0.0000000
## 2  -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871      0.9225312
## 3  -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700      0.9225312
## 4   0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259      0.9225312
## 5  -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461     -0.4612656
## 6  -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612     -0.4612656
## 7  -0.1078688 -0.10015669 -0.70887325  0.59693581  0.8617498      0.9225312
## 8  -0.9767669  1.26308721  0.03299122 -0.11237924 -1.1677918     -0.4612656
## 9  -0.9704532  2.15893320 -1.34037772 -0.70899938 -1.0174553     -1.8450624
## 10  0.2762415 -1.34655112  0.14948233  0.34502953  0.5610770     -0.4612656
## 11  1.0999201 -0.68440408 -0.45749769  2.45971647  1.8389364      1.3837968
## 12 -0.9393967  0.48409069 -0.34100657 -0.29136529 -0.6979905     -0.4612656
## 13  1.9841758 -0.25595600  0.18013789  0.18593083  1.0872544      0.9225312
## 14 -0.9632863  0.87358895  0.19240011 -0.96753478 -0.9610792     -1.8450624
## 15  1.2782387 -0.25595600 -0.40231769  0.98142435  0.8429577      1.8450624
## 16  0.6654710 -1.30760129 -0.23677768 -0.52338423  0.1288598     -0.9225312
## 17  2.4199899  0.48409069 -0.11415545  1.31287998  1.6322239      0.4612656
## 18 -0.0240846 -0.48965495  1.90298017 -0.81506519 -0.9047030     -0.4612656
## 19 -0.4018812 -0.06120687 -0.40231769 -0.21181593  0.5234929      0.4612656
## 20 -0.9281345 -1.11285216 -0.43297324 -1.03382590 -0.6979905     -0.9225312
## 21 -0.1614497  0.40619104 -0.75792214  1.92938746  0.5422849     -0.4612656
##       Leverage  Rev_Growth Net_Profit_Margin fit.cluster
## 1  -0.21209793 -0.52776752        0.06168225           5
## 2   0.01828430 -0.38113909       -1.55366706           4
## 3  -0.40408312 -0.57211809       -0.68503583           2
## 4  -0.74965647  0.14744734         0.35122600          5
## 5  -0.31449003  1.21638667        -0.42597037          2
## 6  -0.74965647 -1.49714434        -1.99560225          4
## 7  -0.02011273 -0.96584257         0.74744375          5
## 8   3.74279705 -0.63276071        -1.24888417          3
## 9   0.61983791  1.88617085        -0.36501379          3
## 10 -0.07130879 -0.64814764         1.17413980          5
## 11 -0.31449003  0.76926048         0.82363947          1
## 12  1.10620040  0.05603085        -0.71551412          3
## 13 -0.62166634 -0.36213170         0.33598685          1
## 14  0.44065173  1.53860717         0.85411776          3
## 15 -0.39128411  0.36014907        -0.24310064          1
## 16 -0.67286239 -1.45369888         1.02174835          5
## 17 -0.54487226  1.10143723         1.44844440          1
## 18 -0.30169102  0.14744734        -1.27936246          4
## 19 -0.74965647 -0.43544591         0.29026942          5
## 20 -0.49367621  1.43089863        -0.09070919          2
## 21  0.68383297 -1.17763919         1.49416183          5
```

Here we can view the cluster plot, please see below,

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 4.3.2
```

```
clusplot(Pharma_2,fit$cluster,color = TRUE,shade = TRUE,labels = 2,lines = 0)
```



2.Interpret the clusters in relation to the numerical variables that were used to form the clusters. By examining the mean values of all quantitative variables within each cluster.

Group 1 comprises JNJ, MRK, PFE, and GSK, exhibiting the maximum Market_cap, ROA, ROE, Asset_Turnover, and the minimum Beta and PE_Ratio.

Cluster 2 showcases the utmost Rev_Growth and the lowest PE_Ratio and Asset_Turnover.

In Cluster 3, there are elevated values for Beta and Leverage, coupled with lower Market_Cap, ROE, ROA, Leverage, Rev_Growth, and Net_Profit_Margin.

The PE_Ratio is notably high in Cluster 4, whereas Leverage and Asset_Turnover are minimal.

The fifth cluster encompasses AZN, ABT, NVS, BMY, WYE, SGP, LLY, and it stands out with the highest Net_Profit_Margin, along with lower leverage and Beta.

2(A)Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)
There is a pattern in the clusters for the Media recommendation variable.

In Cluster 1, there is an even distribution of "Hold" and "Moderate Buy" recommendations, attributable to its peak values in Market_Cap, ROE, ROA, and Asset_Turnover.

For Cluster 2, a "Hold" recommendation is advisable, given its minimal PE_Ratio and Asset_Turnover.

Cluster 3, characterized by the highest Beta and Leverage, predominantly receives a "Moderate Buy" recommendation.

A "Hold" recommendation is suggested for Cluster 4, which exhibits the highest PE_Ratio.

Cluster 5, featuring the highest Net_Profit_Margin, is generally advised to be held.

Regarding variables (10 to 12), discernible trends are evident among Clusters 1–3, where a substantial majority receives a "Moderate Buy" recommendation.

Clusters 1, 2, 4, and 5 are typically suggested to be held.

3 .Provide an appropriate name for each cluster using any or all of the variables in the dataset. Cluster-1 - Purchase (or) Maintain a Moderate Holding.

Cluster-2 - Minimal PE_Ratio, Asset_Turnover, or Sustain.

Cluster-3 - Elevated Beta, Purchase Cluster (or Leverage Cluster).

Cluster-4 - Marked by a substantial PE_Ratio (or considerable Holding).

Cluster-5 - Marked by a significant net profit margin (or substantial Holding).