Clustering Assignment 4

Gaurav Kudeshia

2023-11-10

Assignment Overview:

In that assignment, I employed three clustering methods (K-means, DBSCAN, and Hierarchical clustering) to analyze financial data from 21 pharmaceutical firms. My goal was to comprehend the structure of the pharmaceutical industry using basic financial measures. The evaluation of clustering methods served to formulate and address business problems, demonstrating the aptitude of machine learning in handling relevant data for modeling frameworks. The analysis shed light on effective approaches for understanding industry dynamics and financial trends.

CONCLUSION

Investment Guidance for Pharmaceutical Industry

Objective: Identify the most promising clusters of pharmaceutical companies for investment.

- Clustering Methods:
- K-means: Produces the best and most cohesive clusters compared to DBSCAN and HIERARCHICAL methods.
- Elbow Method: Suggests 4 clusters, with 42.8% variance between clusters.
- Silhouette Method: Recommends multiple clusters, with a higher 63.5% variance between clusters.
- Optimal Clusters for Investment:

Cluster 4:

Favorable market capitalization, ROE, ROA, asset turnover, revenue growth, and net profit margin. Negative or lower beta, P/E ratio, and leverage.

Cluster 3:

Balanced market capitalization, ROE, ROA, and asset turnover. Positive revenue growth and net profit margin. Negative beta and lower P/E ratio.

Variable Analysis:

Market Capitalization: Larger values may indicate stability or growth potential.

Return on Equity (ROE): Positive values indicate profitability from shareholders’ equity.

Price/Earnings Ratio (P/E Ratio): Lower values might suggest a relatively cheaper stock.

Return on Assets (ROA): Positive values imply efficient asset utilization.

Net Profit Margin: Positive values indicate a favorable percentage of revenue translating into profit.

Asset Turnover: Positive values indicate effective asset use for revenue generation.

Estimated Revenue Growth: Positive values suggest potential revenue increase.

Beta: Negative values can provide risk diversification.

Leverage: Lower values suggest lower reliance on debt financing.

- Cluster Evaluation:

Silhouette method provides better separation between clusters, as evidenced by the higher 63.5% variance between clusters. Clustering Results:

Cluster 1: Buy Cluster

Highest median for the “Hold” recommendation. Companies from Switzerland and the United States. Listed on the NYSE. Cluster 2: Sceptical Cluster

Even distribution across AMEX, NASDAQ, and NYSE. Distinct “Hold” and “Moderate Buy” medians. Companies from the United States and Germany. Cluster 3: Moderate Buy Cluster

Listed on the NYSE. Separate counts for the United States, Ireland, and France. Equal “Moderate Buy” and “Moderate Sell” medians. Cluster 4: Hold Cluster

Distributed throughout the United States and the United Kingdom. Listed shares with the same “Hold” and “Moderate Buy” medians. Cluster 5: High Hold Cluster

Only on the NYSE. Equally distributed in the US and Canada. Medians for “Hold” and “Moderate Buy.” Pattern in Media Recommendation Variable:

“Hold” recommendation applies to Clusters 1 and 2. “Moderate Buy” recommendation for Clusters 3, 4, and 5.

SUMMARY

Data Preprocessing and Exploration

• Data Import: The provided dataset was imported into the R environment for analysis.

• Variable Type Identification: The data types of the variables were examined to identify numerical and categorical variables.

• Data Scaling: Numerical variables were scaled to ensure consistency in the data representation.

K-means Clustering Analysis

• Initial Clustering: K-means clustering was performed using the cluster package with k=3, resulting in three clusters of sizes 7, 7, and 7.

• Elbow Method: The elbow method was applied to determine the optimal value of k, suggesting k=4 as the most suitable number of clusters.

• Silhouette Method: The silhouette method was employed to further evaluate the optimal k value, indicating k=5 as the optimal choice.

• Final Clustering: K-means clustering with k=5 was implemented, resulting in five well-defined clusters of sizes 7, 4, 4, 3, and 3.

Cluster Properties Analysis

• Cluster Characteristics: Cluster properties were analyzed, including output centers, the number of companies in each cluster, and the cluster assignment of the 13th observation.

• Visualization: The clustering results were visualized using the fviz_cluster function to gain insights into the relationships between clusters.

DBSCAN Clustering Analysis

• DBSCAN Application: The DBSCAN cluster algorithm was applied with random values for eps=35 and min. points=1.

• Cluster Identification: Six clusters and zero noise points were identified using Euclidean distances as the similarity measure.

• Visualization and Interpretation: The DBSCAN clustering results were plotted and cluster details were printed using the fviz_cluster function.

Hierarchical Clustering Analysis

• Dendrogram Generation: Hierarchical clustering was performed, generating a dendrogram to illustrate the closeness of the 21 companies based on the variables.

• Heatmap Visualization: A heatmap was plotted for hierarchical clustering, providing a visual representation of the hierarchical relationships between the companies.

Further Analysis and Interpretation

• Numerical Variable Analysis: Numerical variables (10 to 12) that were not used in forming clusters were analyzed to interpret patterns and insights.

• Industry Insights: A structured analysis of the pharmaceutical industry based on financial metrics was conducted, revealing meaningful patterns and insights.

• Cluster Naming: Clusters were named appropriately using variables in the dataset, with names yet to be determined.

Loaded datasets from the necessary libraries

```r
library('tidyverse')
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ─
─
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ─
─
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conf
licts to become errors
```

```
library('factoextra')
```

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

```
library('ISLR')
library('cluster')
library('factoextra')
library('fpc')
library('ggplot2')
library('gridExtra')
```

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

Loaded required data set

```
pharma.data.actual <- read.csv("/Users/gauravkudeshia/Downloads/Pharmaceuticals.cs
v")
```

Eliminated unnecessary variables

```
pharma.data.numeric <- pharma.data.actual[ ,-c(1,2,12,13,14)]
pharma.data.numeric
```

Scaled variables for facilitating direct comparison

```
pharma.data.scaled <- scale(pharma.data.numeric)
pharma.data.scaled
```

To ensure reproducibility, used a seed set function in R for consistent random number generation, took k = 3 and conducted clustering on the data using k means

```
set.seed(2)
k <- 3
mod.kmeans <- kmeans(pharma.data.scaled, k)
mod.kmeans
```

```
## K-means clustering with 3 clusters of sizes 7, 7, 7
##
## Cluster means:
##   Market_Cap        Beta   PE_Ratio        ROE         ROA Asset_Turnover
## 1 -0.1799275 -0.81238208 -0.2271431 -0.3387161 -0.04563784     -0.1976853
## 2  0.9547543 -0.06120687 -0.3576482  1.0818081  1.10336187      0.8566361
## 3 -0.7748268  0.87358895  0.5847912 -0.7430920 -1.05772403     -0.6589509
##     Leverage  Rev_Growth Net_Profit_Margin
## 1 -0.4168821 -0.14141325         0.1923035
## 2 -0.2797499 -0.01818848         0.7082574
## 3  0.6966320  0.15960172        -0.9005609
##
## Clustering vector:
##  [1] 1 3 1 2 1 3 2 3 3 1 2 3 2 3 2 1 2 3 1 1 2
##
## Within cluster sum of squares by cluster:
## [1] 19.81945 25.26414 57.86025
##  (between_SS / total_SS =  42.8 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

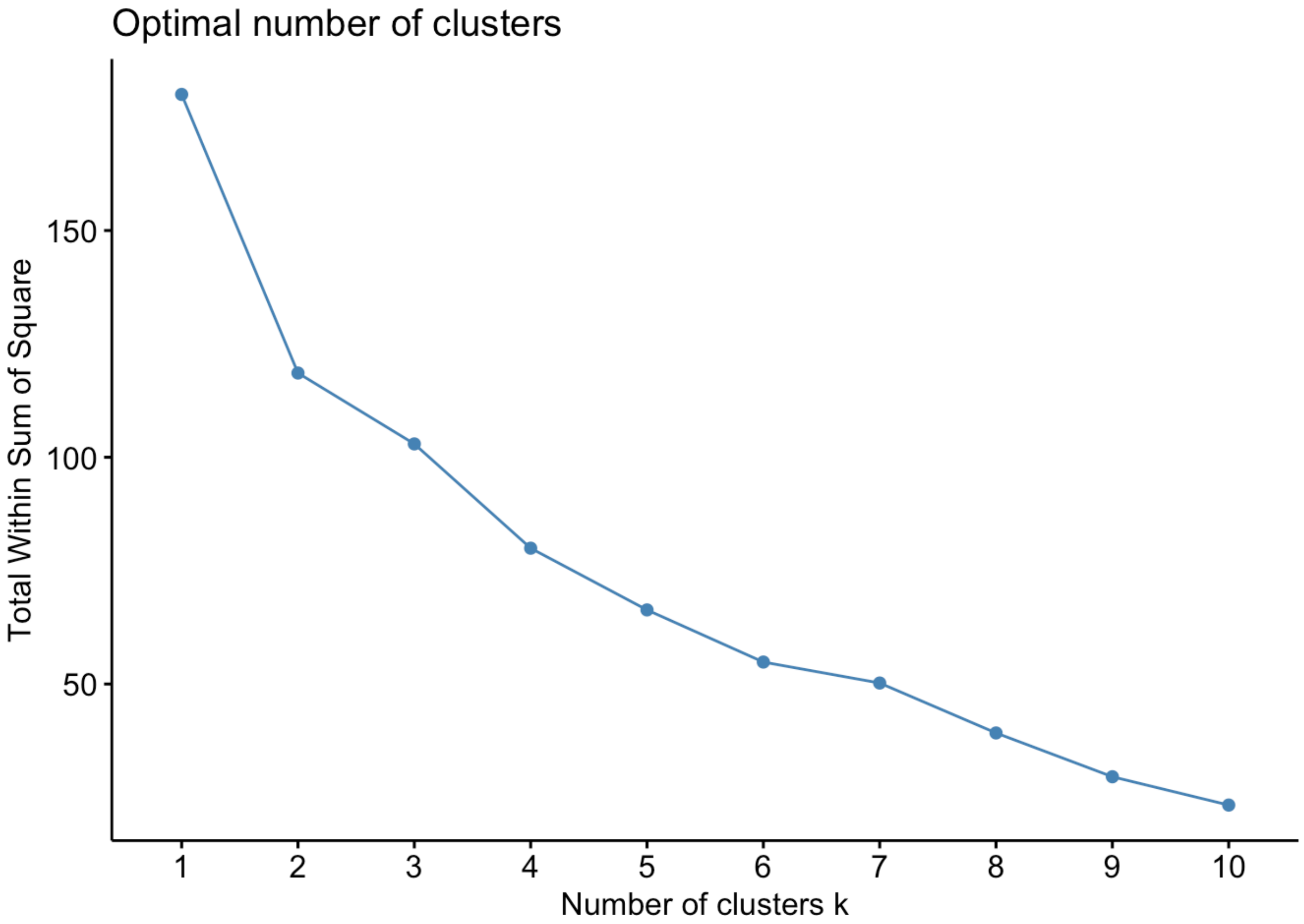
This code extracts cluster assignments from a k-means model, calculates mean values for each cluster, and prints a summary of the means

```
cluster_assignments <- mod.kmeans$cluster
cluster_means <- aggregate(pharma.data.scaled, by = list(cluster_assignments), FUN
= mean)
head(cluster_means)
```

```
##   Group.1 Market_Cap        Beta   PE_Ratio        ROE         ROA
## 1       1 -0.1799275 -0.81238208 -0.2271431 -0.3387161 -0.04563784
## 2       2  0.9547543 -0.06120687 -0.3576482  1.0818081  1.10336187
## 3       3 -0.7748268  0.87358895  0.5847912 -0.7430920 -1.05772403
##   Asset_Turnover   Leverage  Rev_Growth Net_Profit_Margin
## 1     -0.1976853 -0.4168821 -0.14141325         0.1923035
## 2      0.8566361 -0.2797499 -0.01818848         0.7082574
## 3     -0.6589509  0.6966320  0.15960172        -0.9005609
```

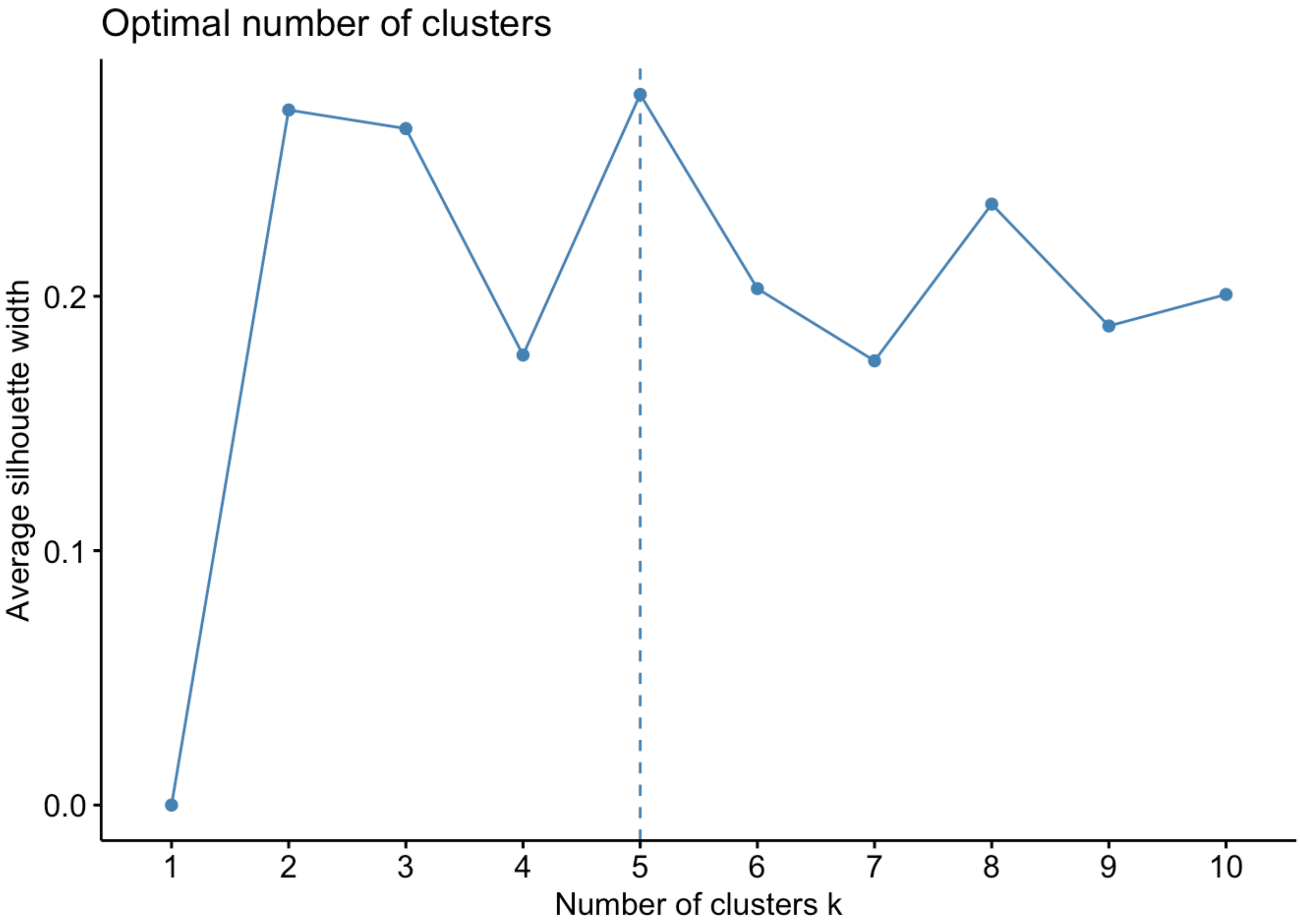
Used “elbow chart” to determine k

```
fviz_nbclust(pharma.data.scaled, kmeans, method = "wss")
```



We can determine k to be 4 **Silhouette method**

```
fviz_nbclust(pharma.data.scaled, kmeans, method = "silhouette")
```



Based on the Silhouette method results obtained above, a value of “k” equal to 5 was determined. Subsequently, this value was applied, and the dataset was clustered accordingly

```
set.seed(2)
k <- 5
mod.kmeans_sil <- kmeans(pharma.data.scaled, k)
mod.kmeans_sil
```

```
## K-means clustering with 5 clusters of sizes 7, 4, 4, 3, 3
## 
## Cluster means:
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.08926902 -0.4618336 -0.3208615  0.3260892  0.5396003   6.589509e-02
## 2  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431   1.153164e+00
## 3 -0.96247577  1.1949250 -0.3639982 -0.5200697 -0.9610792  -1.153164e+00
## 4 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022  -1.537552e-01
## 5 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838  -3.330669e-16
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.2559803 -0.7230135         0.7343816
## 2 -0.4680782  0.4671788         0.5912425
## 3  1.4773718  0.7120120        -0.3688236
## 4 -0.4040831  0.6917224        -0.4005718
## 5 -0.3443544 -0.5769454        -1.6095439
## 
## Clustering vector:
##  [1] 1 5 4 1 4 5 1 3 3 1 2 3 2 3 2 1 2 5 1 4 1
## 
## Within cluster sum of squares by cluster:
## [1] 16.655937  9.284424 19.219788  5.511294 14.938904
##  (between_SS / total_SS =  63.5 %)
## 
## Available components:
## 
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Computed the average value of the numerical variables within each cluster and subsequently displayed the results

```
cluster.silhouette.assignment <- mod.kmeans_sil$cluster
cluster.means.silhouette <- aggregate(pharma.data.scaled, by = list(cluster.silhou
ette.assignment), FUN = mean)
head(cluster.means.silhouette)
```

```
##   Group.1  Market_Cap       Beta   PE_Ratio        ROE        ROA
## 1       1  0.08926902 -0.4618336 -0.3208615  0.3260892  0.5396003
## 2       2  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431
## 3       3 -0.96247577  1.1949250 -0.3639982 -0.5200697 -0.9610792
## 4       4 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022
## 5       5 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838
##   Asset_Turnover   Leverage Rev_Growth Net_Profit_Margin
## 1   6.589509e-02 -0.2559803 -0.7230135         0.7343816
## 2   1.153164e+00 -0.4680782  0.4671788         0.5912425
## 3  -1.153164e+00  1.4773718  0.7120120        -0.3688236
## 4  -1.537552e-01 -0.4040831  0.6917224        -0.4005718
## 5  -3.330669e-16 -0.3443544 -0.5769454        -1.6095439
```

```
# Visualize the output

mod.kmeans_sil$centers # output the centers
```

```
##    Market_Cap       Beta   PE_Ratio        ROE        ROA Asset_Turnover
## 1  0.08926902 -0.4618336 -0.3208615  0.3260892  0.5396003   6.589509e-02
## 2  1.69558112 -0.1780563 -0.1984582  1.2349879  1.3503431   1.153164e+00
## 3 -0.96247577  1.1949250 -0.3639982 -0.5200697 -0.9610792  -1.153164e+00
## 4 -0.66114002 -0.7233539 -0.3512251 -0.6736441 -0.5915022  -1.537552e-01
## 5 -0.52462814  0.4451409  1.8498439 -1.0404550 -1.1865838  -3.330669e-16
##     Leverage Rev_Growth Net_Profit_Margin
## 1 -0.2559803 -0.7230135         0.7343816
## 2 -0.4680782  0.4671788         0.5912425
## 3  1.4773718  0.7120120        -0.3688236
## 4 -0.4040831  0.6917224        -0.4005718
## 5 -0.3443544 -0.5769454        -1.6095439
```

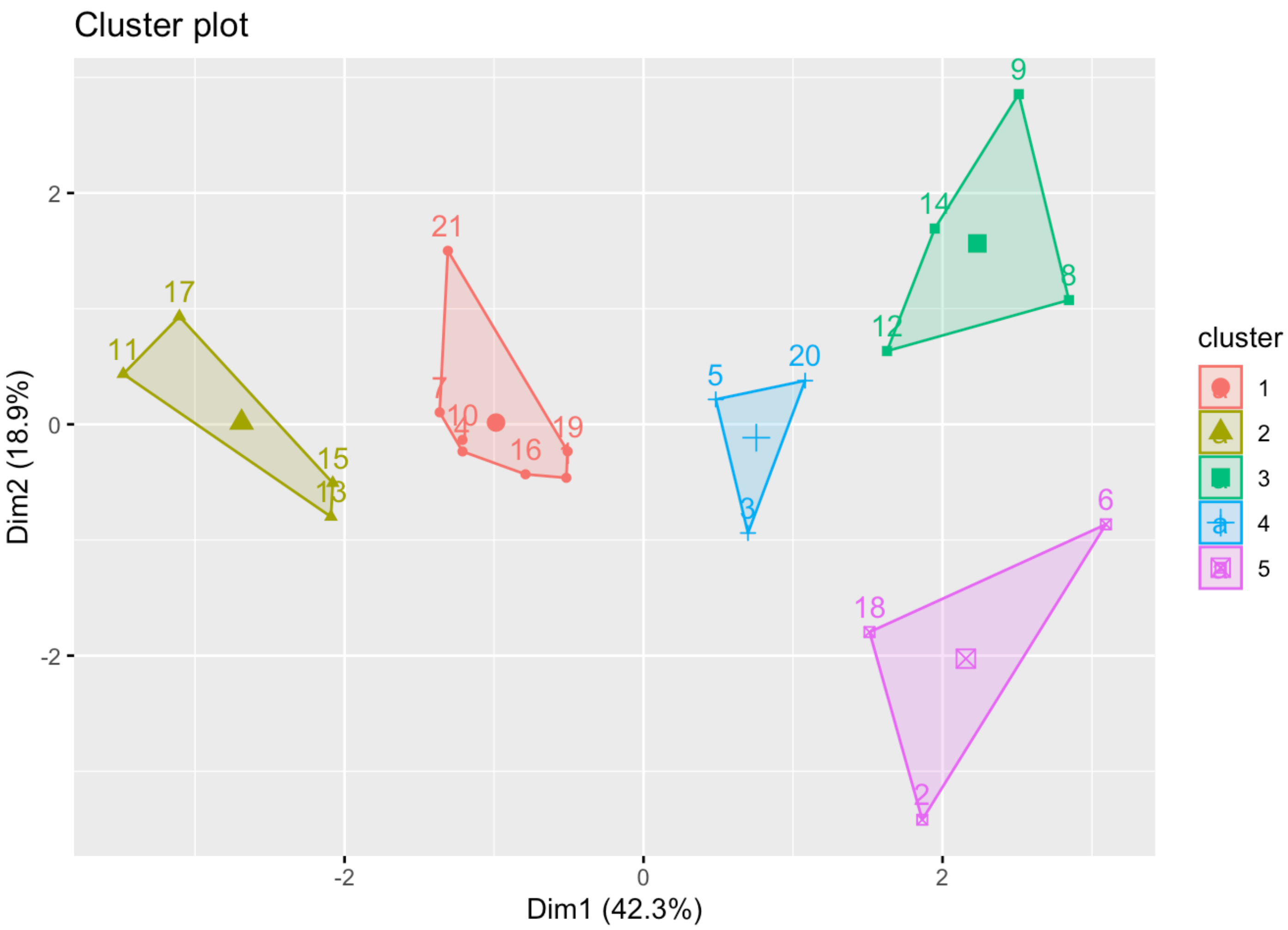
```
mod.kmeans_sil$size # Number of companies in each cluster
```

```
## [1] 7 4 4 3 3
```

```
mod.kmeans_sil$cluster[13] # Identify the cluster of the 13th observation as an ex
ample
```

```
## [1] 2
```

```
fviz_cluster(mod.kmeans_sil, data = pharma.data.scaled) # Visualize the output
```



DBSCAN

**Performed Clustering on the same data set

```
set.seed(12)
DBSCAN.model <- dbscan::dbscan(pharma.data.numeric, eps = 35, MinPts = 1) #perform
clustering
```

```
## Warning in dbscan::dbscan(pharma.data.numeric, eps = 35, MinPts = 1):
## converting argument MinPts (fpc) to minPts (dbscan)!
```

```
print(DBSCAN.model) #print cluster details
```

```
## DBSCAN clustering for 21 objects.
## Parameters: eps = 35, minPts = 1
## Using euclidean distances and borderpoints = TRUE
## The clustering contains 6 cluster(s) and 0 noise points.
##
##  1  2  3  4  5  6
## 15  1  2  1  1  1
##
## Available fields: cluster, eps, minPts, dist, borderPoints
```

I utilized the “fpc” package to evaluate the cluster and displayed the generated model. In the outcome, the seed reveals the core points, and when added together, they total “21”

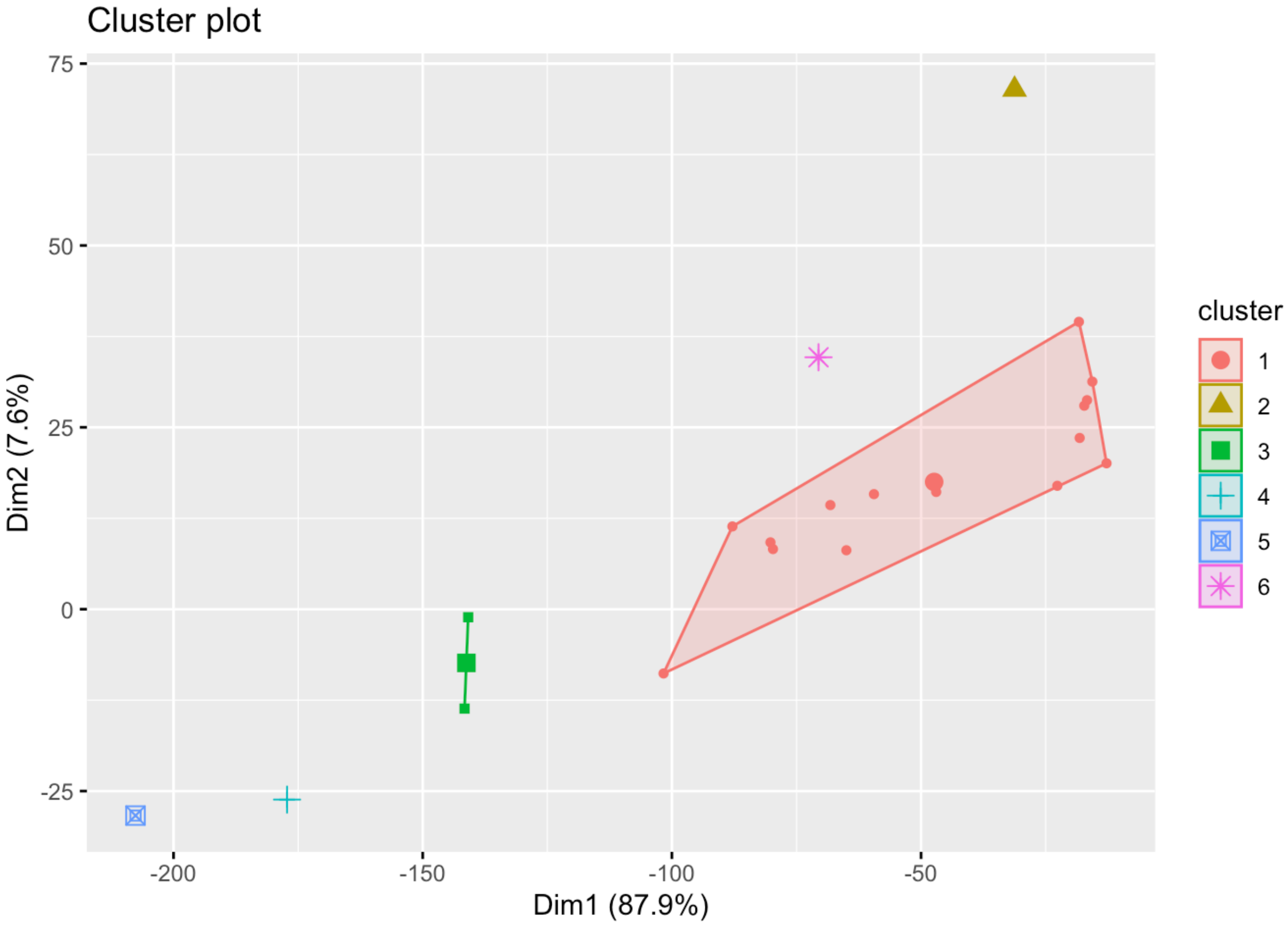
```
set.seed(123)
DBSCAN.model <- fpc::dbscan(pharma.data.numeric, eps = 35, MinPts = 1) # DBSCAN using fpc package

print(DBSCAN.model) # show clusters' details
```

```
## dbscan Pts=21 MinPts=1 eps=35
##        1 2 3 4 5 6
## seed  15 1 2 1 1 1
## total 15 1 2 1 1 1
```

Plotted

```
fviz_cluster(DBSCAN.model, pharma.data.numeric,   stand = FALSE, frame = FALSE, geom = "point")
```

```
## Warning: argument frame is deprecated; please use ellipse instead.
```



Hierarchical

Compute Euclidean distance

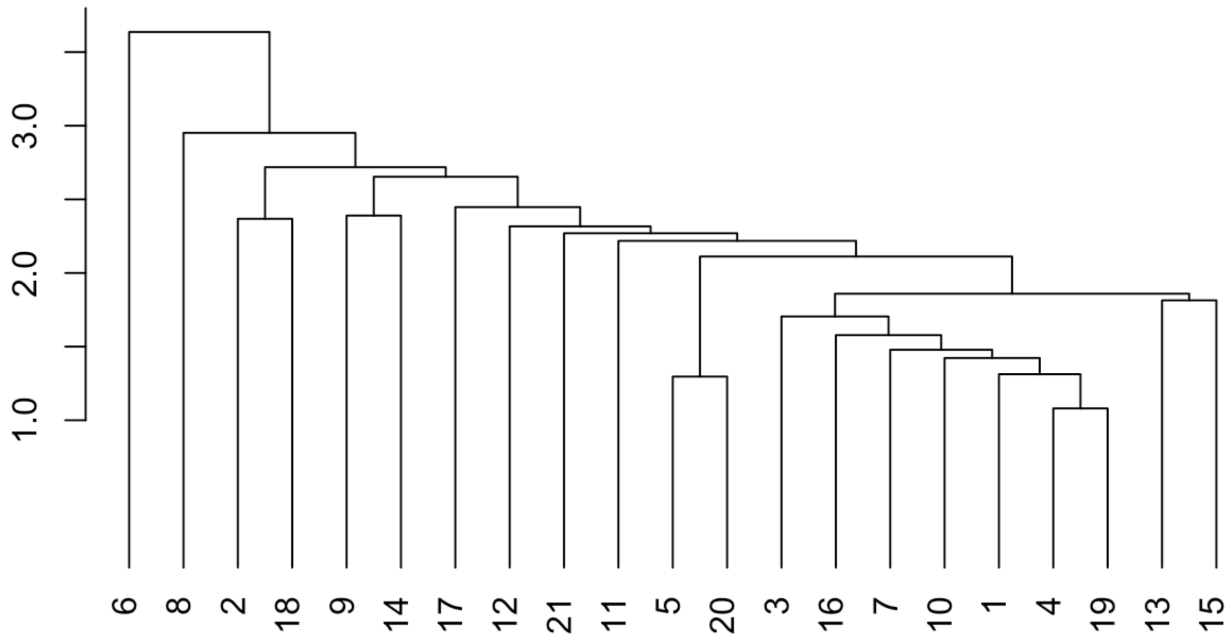
```
d <- dist(pharma.data.numeric, method = "euclidean")
d.norm <- dist(pharma.data.numeric[,c(5,6)], method = "euclidean")
```

Normalized the data and computed normalized distance based on variables ROA and Asset_Turnover

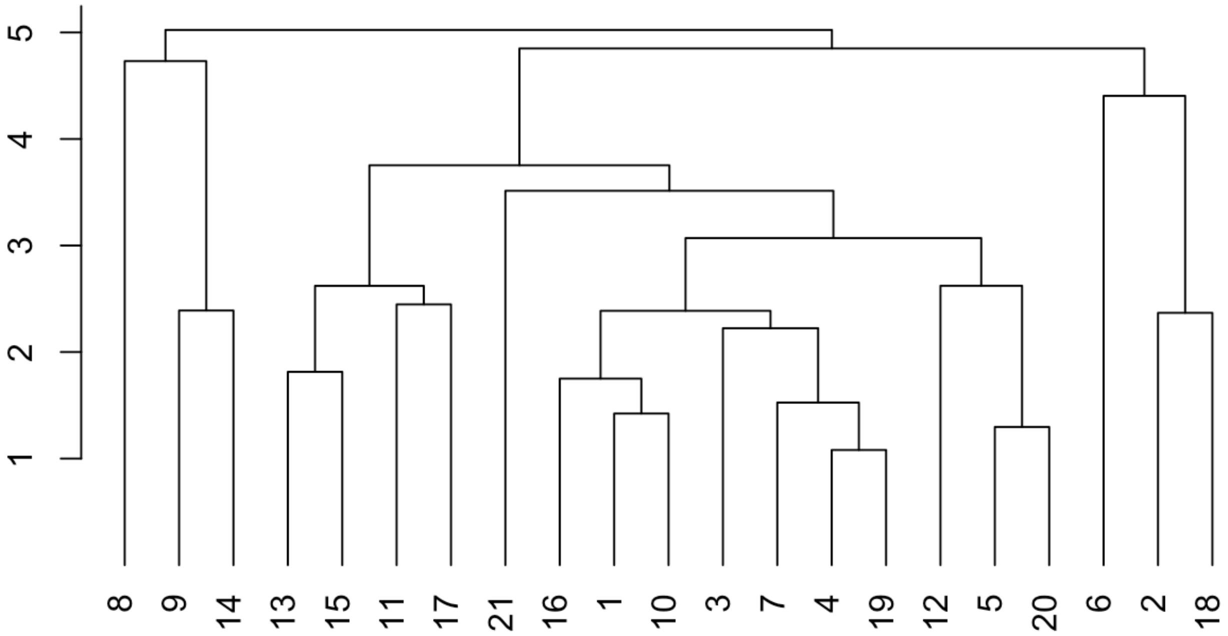
```
pharma.data.numeric.norm <- sapply(pharma.data.numeric, scale)
row.names(pharma.data.numeric.norm) <- row.names(pharma.data.numeric)
d.norm <- dist(pharma.data.numeric.norm[,c(5,6)], method = "euclidean")
pharma.data.numeric.norm
```

This code performs hierarchical clustering on the normalized pharmaceutical data and plots the resulting dendrograms using the single and average linkage methods

```
d.norm <- dist(pharma.data.numeric.norm, method = "euclidean")
hc1 <- hclust(d.norm, method = "single")
plot(hc1, hang = -1, ann = FALSE)
```



```
hc2 <- hclust(d.norm, method = "average")
plot(hc2, hang = -1, ann = FALSE)
```



This code assigns each pharmaceutical company to one of six clusters using the single and average linkage hierarchical clustering results

```
memb <- cutree(hc1, k = 6)
memb
```

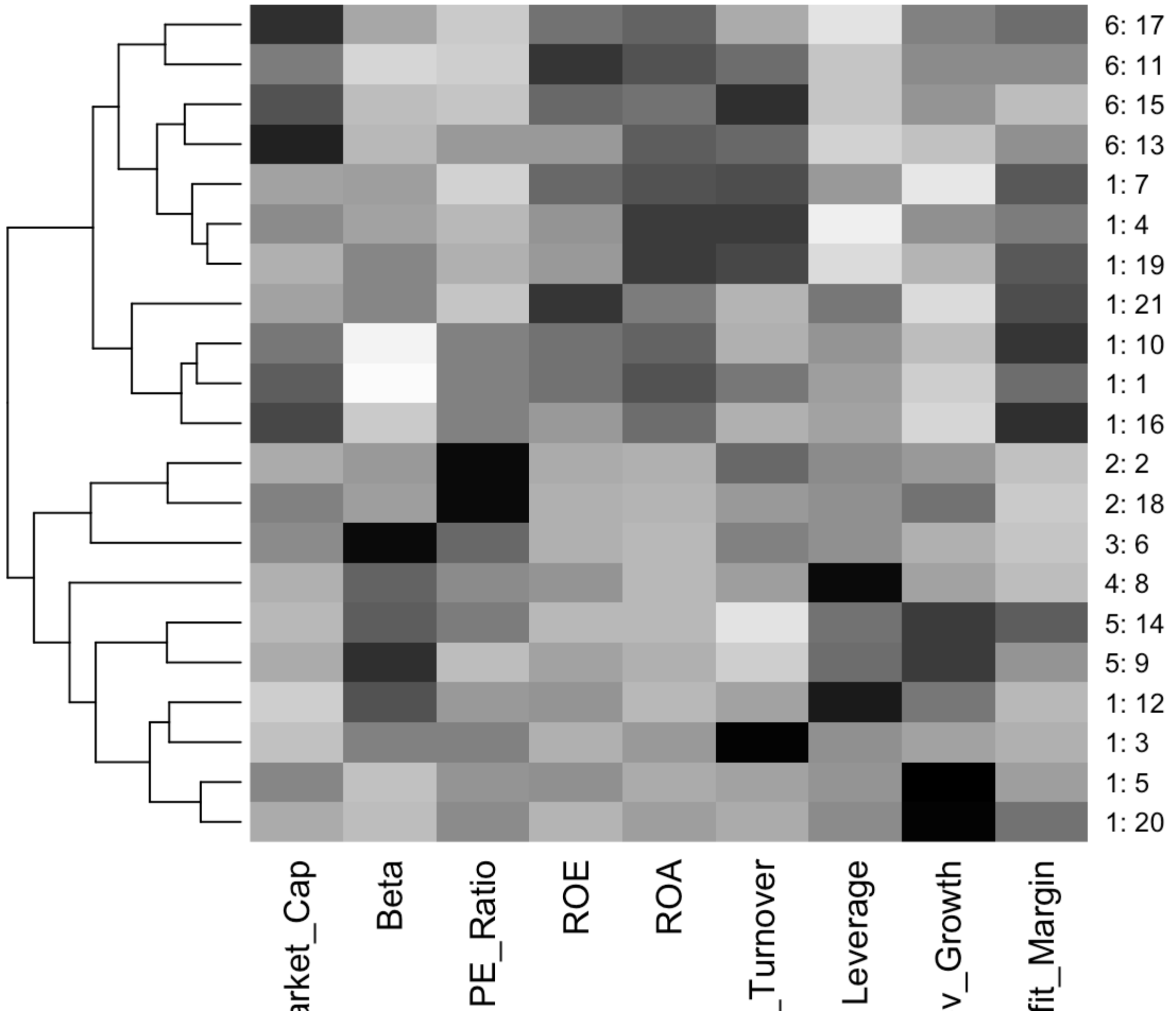
```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 
##  1  2  1  1  1  3  1  4  5  1  1  1  1  5  1  1  6  2  1  1  1
```

```
memb <- cutree(hc2, k = 6)
memb
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 
##  1  2  1  1  1  3  1  4  5  1  6  1  6  5  6  1  6  2  1  1  1
```

This code creates a heatmap of the normalized pharmaceutical data, with the rows reordered according to the single linkage hierarchical clustering results and the columns colored in grayscale

```
row.names(pharma.data.numeric.norm) <- paste(memb, ": ", row.names(pharma.data.numeric), sep = "")
heatmap(as.matrix(pharma.data.numeric.norm), Colv = NA, hclustfun = hclust,
        col=rev(paste("gray",1:99,sep="")))
```



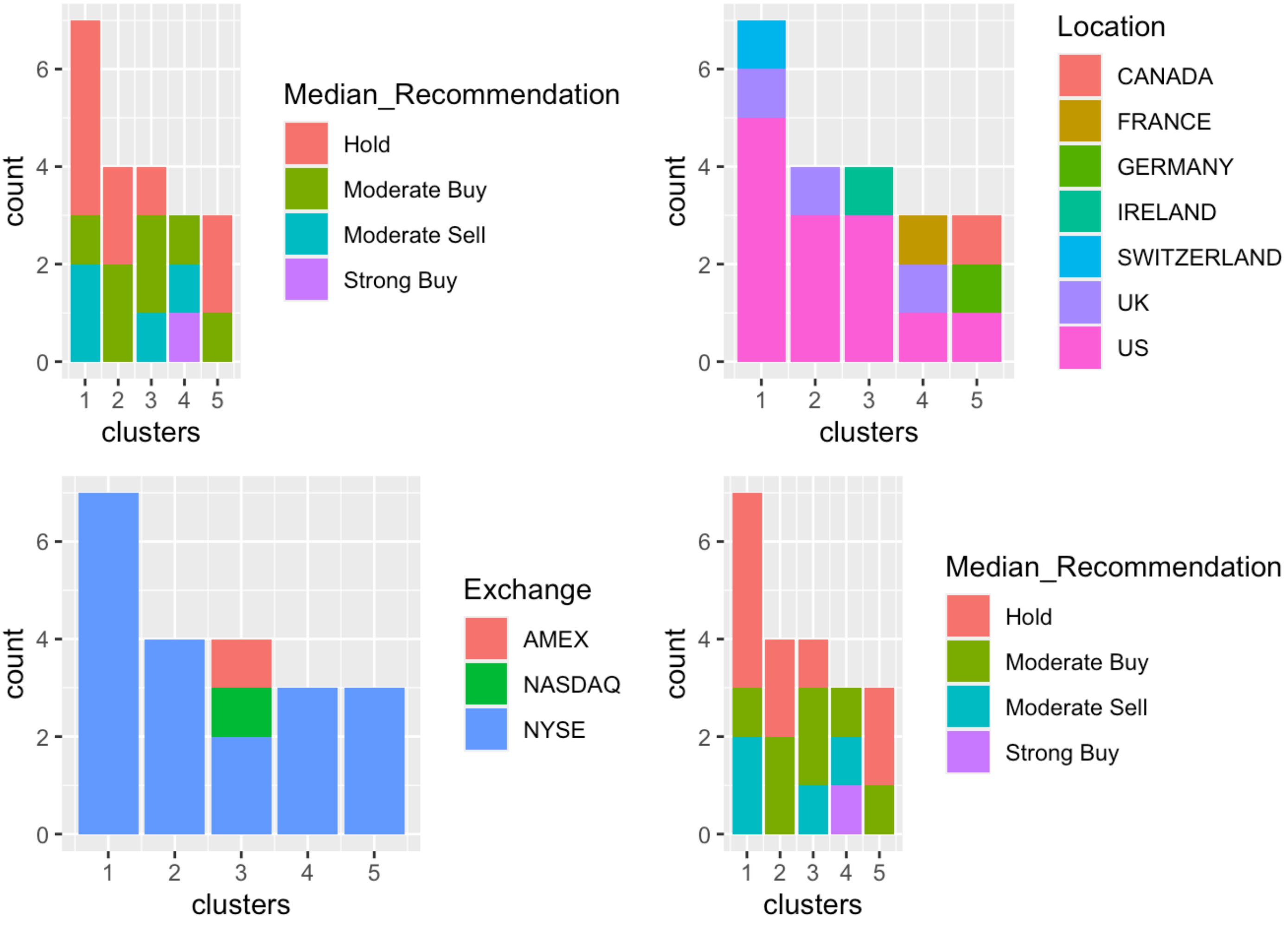
b) Interpret the clusters with respect to the numerical variables used in forming the clusters.

```
set.seed(3)
p <- pharma.data.actual[,c(12,13,14)]%>% mutate(clusters = mod.kmeans_sil$cluster)
```

(c)Is there a pattern in the clusters with respect to the numerical variables (10 to 12)?

Four stacked bar graphs visualize the distribution of pharmaceutical companies across clusters based on recommendation, location, exchange, and median recommendation.

```
p.1 <- ggplot(p, mapping = aes(clusters, fill = Median_Recommendation)) + geom_bar
(position = "stack")
p.2 <- ggplot(p, mapping = aes(clusters, fill = Location)) + geom_bar(position = "
stack")
p.3 <- ggplot(p, mapping = aes(clusters, fill = Exchange)) + geom_bar(position = "
stack")
p.4 <- ggplot(p, mapping = aes(clusters, fill = Median_Recommendation)) + geom_bar
(position = "stack")
grid.arrange(p.1, p.2, p.3, p.4, ncol = 2)
```



The code classifies the pharmaceutical companies into five clusters: Buy Cluster, Sceptical Cluster, Moderate Buy Cluster, Hold Cluster, and High Hold Cluster

(D)Provide an appropriate name for each cluster using any or all of the variables in the dataset.

Cluster 1 :- Buy Cluster Cluster 2 :- Sceptical Cluster Cluster 3 :- Moderate Buy Cluster Cluster 4 :- Hold Cluster Cluster 5 :- High Hold Cluster