Assignment_4

Olayinka Sikiru

2024-03-11

```
#load necessary libraries
library(caret)
library(tidyverse)
library(ggplot2)
library(factoextra)
```

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

```
library(flexclust)
```

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

```
library(ISLR)
```

```
#Setting working directory
library(readr)
Pharm <- read_csv("C:/Users/DELL/OneDrive/Desktop/Fundamentals of Machine Learning/Dataset/Pharm
aceuticals.csv")
```

```
## Rows: 21 Columns: 14
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (5): Symbol, Name, Median_Recommendation, Location, Exchange
## dbl (9): Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Checking Summary and Statistics of Dataset
summary(Pharm)
```

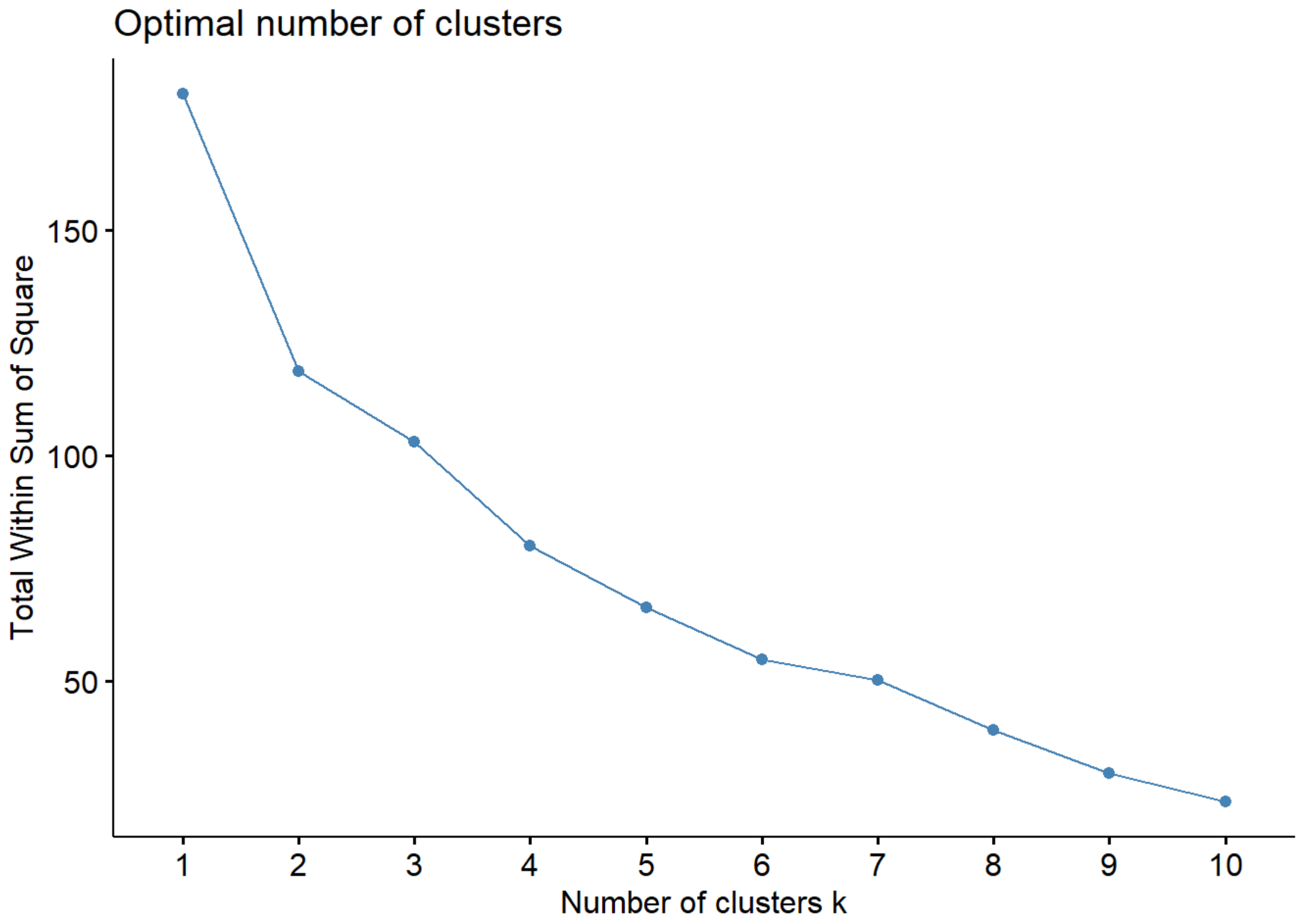
```
##     Symbol              Name             Market_Cap          Beta
##  Length:21          Length:21          Min.   :  0.41   Min.   :0.1800
##  Class :character   Class :character   1st Qu.:  6.30   1st Qu.:0.3500
##  Mode  :character   Mode  :character   Median : 48.19   Median :0.4600
##                                        Mean   : 57.65   Mean   :0.5257
##                                        3rd Qu.: 73.84   3rd Qu.:0.6500
##                                        Max.   :199.47   Max.   :1.1100
##     PE_Ratio          ROE             ROA        Asset_Turnover    Leverage
##  Min.   : 3.60   Min.   : 3.9   Min.   : 1.40   Min.   :0.3    Min.   :0.0000
##  1st Qu.:18.90   1st Qu.:14.9   1st Qu.: 5.70   1st Qu.:0.6    1st Qu.:0.1600
##  Median :21.50   Median :22.6   Median :11.20   Median :0.6    Median :0.3400
##  Mean   :25.46   Mean   :25.8   Mean   :10.51   Mean   :0.7    Mean   :0.5857
##  3rd Qu.:27.90   3rd Qu.:31.0   3rd Qu.:15.00   3rd Qu.:0.9    3rd Qu.:0.6000
##  Max.   :82.50   Max.   :62.9   Max.   :20.30   Max.   :1.1    Max.   :3.5100
##    Rev_Growth    Net_Profit_Margin Median_Recommendation   Location
##  Min.   :-3.17   Min.   : 2.6      Length:21             Length:21
##  1st Qu.: 6.38   1st Qu.:11.2      Class :character      Class :character
##  Median : 9.37   Median :16.1      Mode  :character      Mode  :character
##  Mean   :13.37   Mean   :15.7
##  3rd Qu.:21.87   3rd Qu.:21.1
##  Max.   :34.21   Max.   :25.5
##    Exchange
##  Length:21
##  Class :character
##  Mode  :character
##
##
##
```

Solution 1

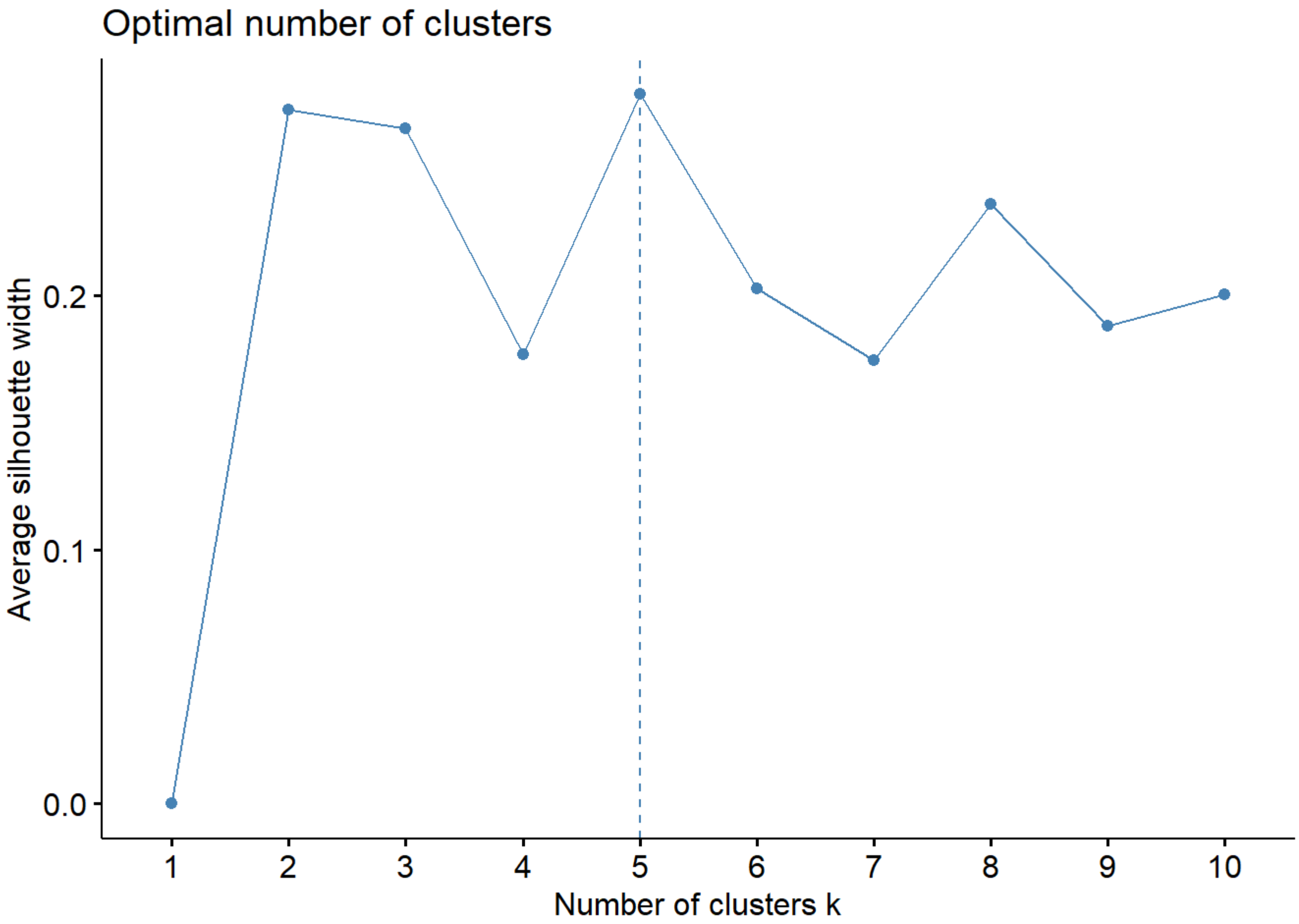
```
#Removing non-numerical variables
pharm.Pharm = Pharm[,c(3:11)]

#Normalizing the data
norm=preProcess(pharm.Pharm,method = c("center","scale"))
pharm.norm.Pharm = predict(norm,pharm.Pharm)
```

K-means

```
# Visualizing optimal number of clusters using the k-means algorithm and "wss" method
fviz_nbclust(pharm.norm.Pharm,kmeans,method = "wss")
```



```
# Visualizing optimal number of clusters using the k-means algorithm and "silhouette" method
fviz_nbclust(pharm.norm.Pharm,kmeans,method="silhouette")
```



k-means clustering

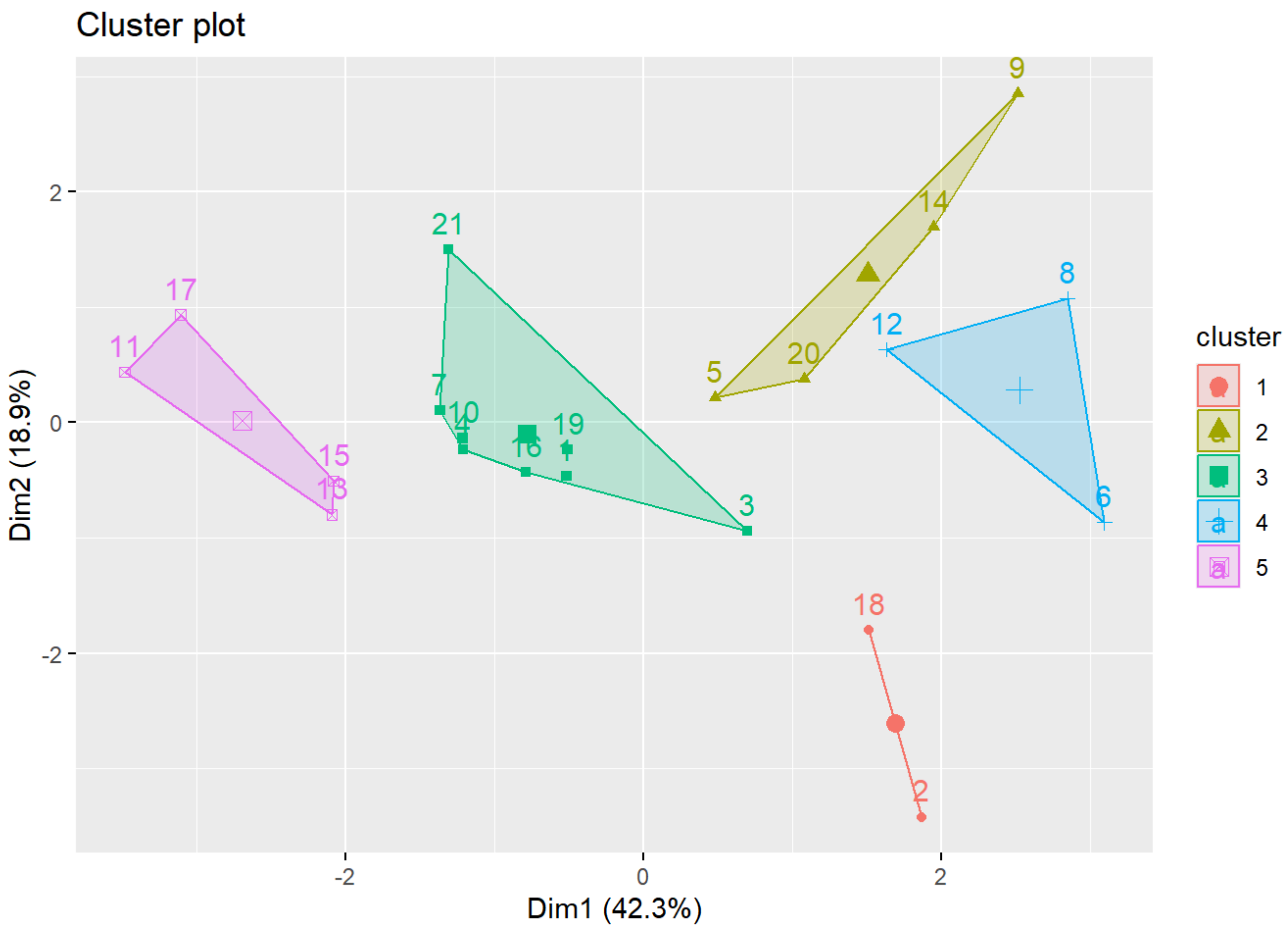
```
#Using k-means clustering on the normalized dataset with 5 clusters and 10 random starts
k = kmeans(pharm.norm.Pharm,centers=5,nstart = 10)

#Displaying the cluster centers
k$centers
```

```
##    Market_Cap       Beta    PE_Ratio        ROE        ROA Asset_Turnover
## 1 -0.43925134 -0.4701800  2.70002464 -0.8349525 -0.9234951      0.2306328
## 2 -0.76022489  0.2796041 -0.47742380 -0.7438022 -0.8107428     -1.2684804
## 3 -0.03142211 -0.4360989 -0.31724852  0.1950459  0.4083915      0.1729746
## 4 -0.87051511  1.3409869 -0.05284434 -0.6184015 -1.1928478     -0.4612656
## 5  1.69558112 -0.1780563 -0.19845823  1.2349879  1.3503431      1.1531640
##      Leverage Rev_Growth Net_Profit_Margin
## 1 -0.14170336 -0.1168459      -1.416514761
## 2  0.06308085  1.5180158      -0.006893899
## 3 -0.27449312 -0.7041516       0.556954446
## 4  1.36644699 -0.6912914      -1.320000179
## 5 -0.46807818  0.4671788       0.591242521
```

```
#Displaying the Clusters size
k$size
```

```
## [1] 2 4 8 3 4
```

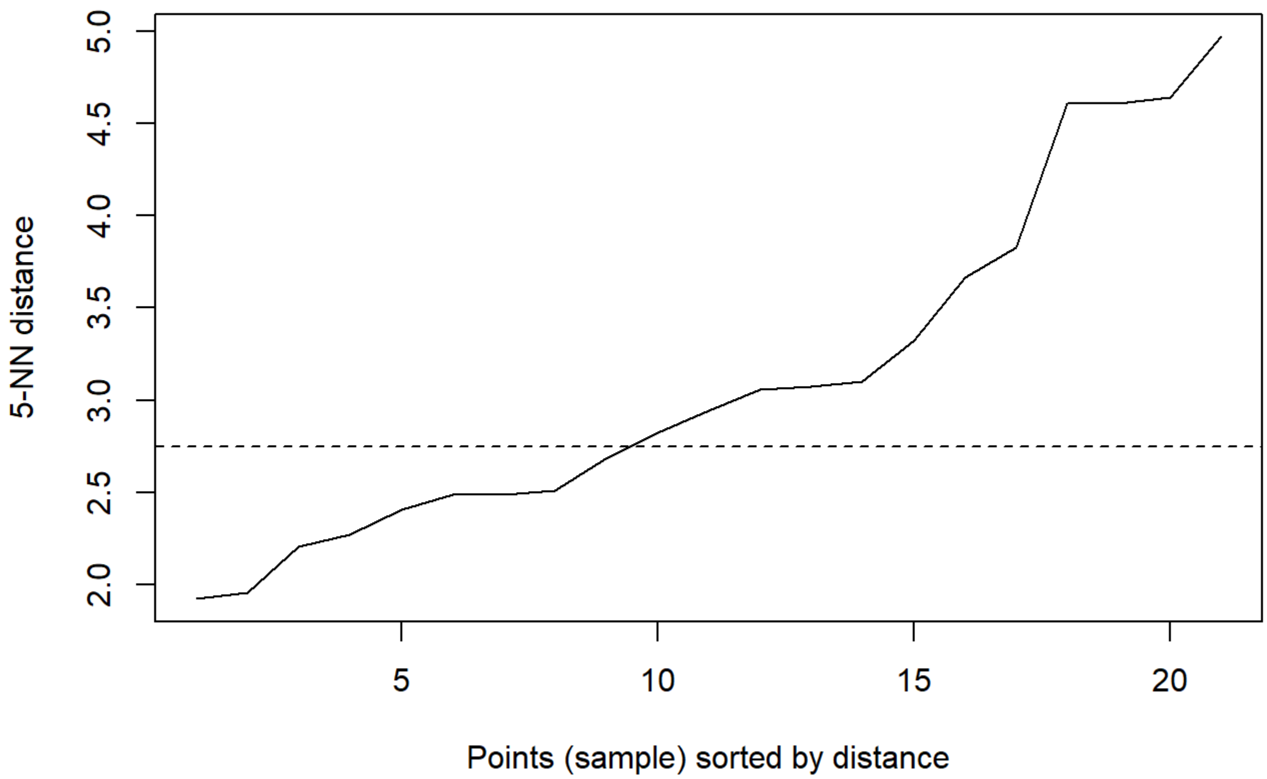
```
#Visualing the Cluster
fviz_cluster(k,data = pharm.norm.Pharm)
```



Interpretation - The k-means plot provides an optimal representation of clusters, grouping together points that are close to each other into the same cluster. It offers a clear visualization of the dataset's structure, making it easier to analyze the inherent patterns within the data.

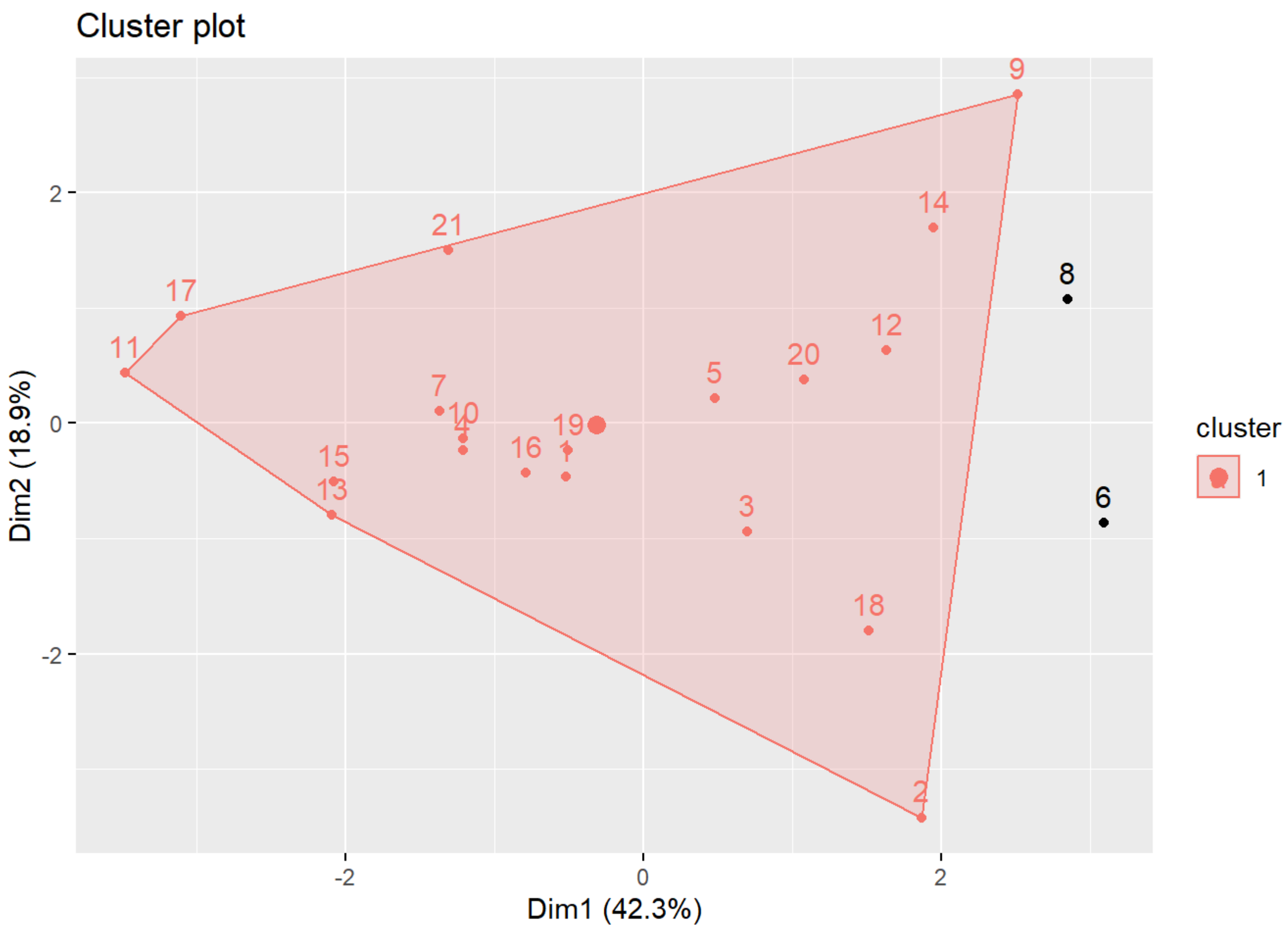
DBSCAN Clustering

```
set.seed(123)

#Generating kNN distance plot for the normalized dataset using k=5
dbscan::kNNdistplot(pharm.norm.Pharm,k=5)


abline(h = 2.75, lty = "dashed")
```



Therefore, the optimal value for epsilon, which is 2.75 (knee-point), indicates that we can build a DBSCAN model using this value.

```
db= dbscan::dbscan(pharm.norm.Pharm,eps=2.75,minPts = 2)
fviz_cluster(db,pharm.norm.Pharm)
```

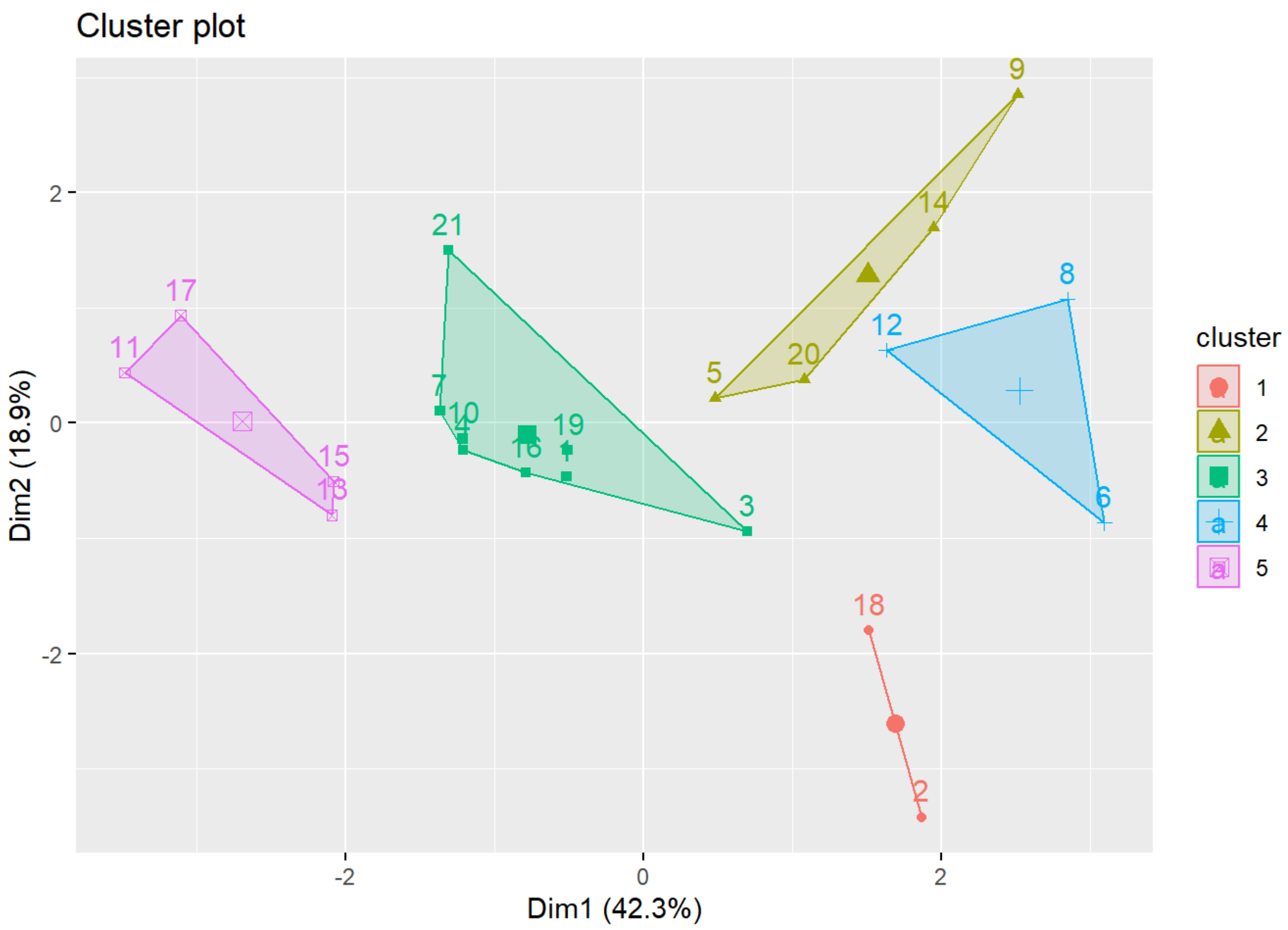


Interpretation - The DBSCAN method appears inappropriate for clustering the provided data as it predominantly groups all data points into a single cluster. Moreover, selecting a smaller value for epsilon results in most points being classified as outliers. Therefore, DBSCAN is not an optimal method for this dataset.

Justification - The rationale behind our choices stems from the absence of explicit weightage for variables, leading us to normalize the data and accord equal significance to all factors. Through rigorous experimentation with various clustering methods, the k-means approach emerged as the most effective for our dataset, providing a clearer clustering depiction wherein closely positioned data points are grouped together within clusters. To determine the appropriate number of clusters, we utilized techniques such as the elbow method and distribution plots to identify optimal values for ‘k’ and ‘eps’, ensuring a comprehensive analysis of the data.

Solution 2

```
fviz_cluster(k,data = pharm.norm.Pharm)
```



In terms of the numerical values used in clustering, the data points within the same k-means clusters exhibit proximity to each other, contrasting with those in distinct groups. Observing these clusters allows for a clearer interpretation of the data.

Solution 3

Cluster 1

```
pharm.norm.Pharm[c(6,8,12),]
```

```
## # A tibble: 3 × 9
##   Market_Cap  Beta PE_Ratio    ROE    ROA Asset_Turnover Leverage Rev_Growth
##        <dbl> <dbl>    <dbl>  <dbl>  <dbl>          <dbl>    <dbl>      <dbl>
## 1     -0.695 2.28    0.149  -1.45  -1.71          -0.461   -0.750    -1.50
## 2     -0.977 1.26    0.0330 -0.112 -1.17          -0.461    3.74     -0.633
## 3     -0.939 0.484  -0.341  -0.291 -0.698         -0.461    1.11      0.0560
## # ℹ 1 more variable: Net_Profit_Margin <dbl>
```

This particular cluster displays high Beta values alongside an average PE-ratio, while all other variables fall below average levels. However, the variables leverage and Rev_growth exhibit a mixture of values within this cluster.

Cluster 2

```
pharm.norm.Pharm[c(2,18),]
```

```
## # A tibble: 2 × 9
##   Market_Cap   Beta PE_Ratio    ROE    ROA Asset_Turnover Leverage Rev_Growth
##        <dbl>  <dbl>    <dbl>  <dbl>  <dbl>          <dbl>    <dbl>      <dbl>
## 1    -0.854  -0.451     3.50 -0.855 -0.942          0.923   0.0183     -0.381
## 2    -0.0241 -0.490     1.90 -0.815 -0.905         -0.461  -0.302       0.147
## # ℹ 1 more variable: Net_Profit_Margin <dbl>
```

Within this cluster, the PE ratio stands out as notably high, while the remaining variables fall below average levels. Nonetheless, the cluster also exhibits mixed values for leverage and rev_growth.

Cluster 3

```
pharm.norm.Pharm[c(1,3,4,7,10,16,19,21),]
```

```
## # A tibble: 8 × 9
##   Market_Cap    Beta PE_Ratio     ROE    ROA Asset_Turnover Leverage Rev_Growth
##        <dbl>   <dbl>    <dbl>   <dbl>  <dbl>          <dbl>    <dbl>      <dbl>
## 1      0.184 -0.801   -0.0467  0.0401  0.242          0      -0.212      -0.528
## 2     -0.876 -0.256   -0.292  -0.722  -0.510          0.923  -0.404      -0.572
## 3      0.170 -0.0223  -0.243   0.106   0.918          0.923  -0.750       0.147
## 4     -0.108 -0.100   -0.709   0.597   0.862          0.923  -0.0201     -0.966
## 5      0.276 -1.35     0.149   0.345   0.561         -0.461  -0.0713     -0.648
## 6      0.665 -1.31    -0.237  -0.523   0.129         -0.923  -0.673      -1.45
## 7     -0.402 -0.0612  -0.402  -0.212   0.523          0.461  -0.750      -0.435
## 8     -0.161  0.406   -0.758   1.93    0.542         -0.461   0.684      -1.18
## # ℹ 1 more variable: Net_Profit_Margin <dbl>
```

This specific cluster displays notably high net profit values, although the remaining variables exhibit mixed values.

Cluster 4

```
pharm.norm.Pharm[c(11,17,15,13),]
```

```
## # A tibble: 4 × 9
##   Market_Cap   Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage Rev_Growth
##        <dbl>  <dbl>    <dbl> <dbl> <dbl>          <dbl>    <dbl>      <dbl>
## 1       1.10 -0.684   -0.457 2.46  1.84           1.38    -0.314      0.769
## 2       2.42  0.484   -0.114 1.31  1.63           0.461   -0.545      1.10
## 3       1.28 -0.256   -0.402 0.981 0.843          1.85    -0.391      0.360
## 4       1.98 -0.256    0.180 0.186 1.09           0.923   -0.622     -0.362
## # ℹ 1 more variable: Net_Profit_Margin <dbl>
```

All cluster displays high values. Besides, Beta, PE ratio and Leverage

Cluster 5

```
pharm.norm.Pharm[c(5,9,14,20),]
```

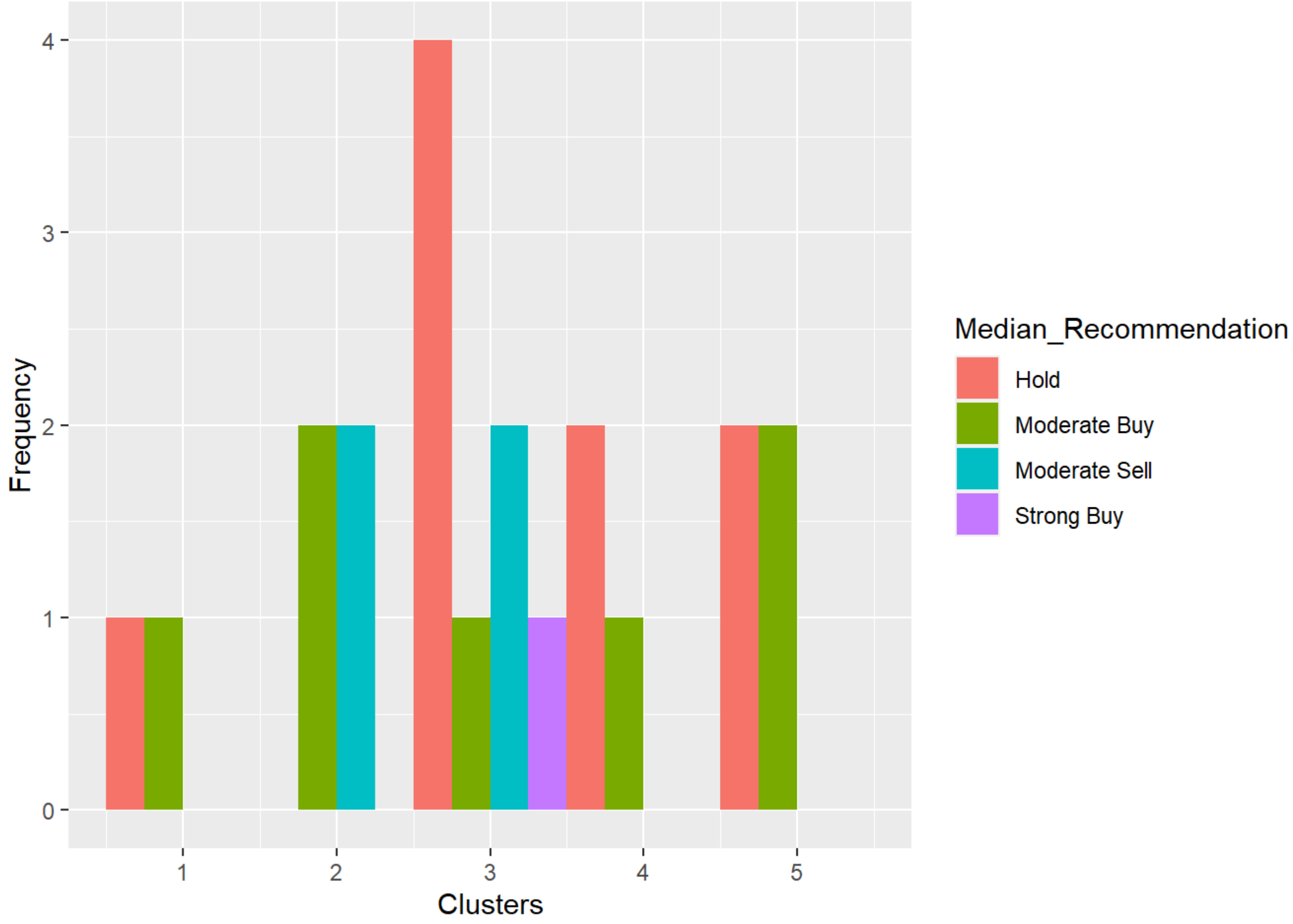
```
## # A tibble: 4 × 9
##   Market_Cap   Beta PE_Ratio    ROE    ROA Asset_Turnover Leverage Rev_Growth
##        <dbl>  <dbl>    <dbl>  <dbl>  <dbl>          <dbl>    <dbl>      <dbl>
## 1     -0.179 -0.801   -0.329 -0.265 -0.566         -0.461   -0.314       1.22
## 2     -0.970  2.16    -1.34  -0.709 -1.02          -1.85     0.620       1.89
## 3     -0.963  0.874    0.192 -0.968 -0.961         -1.85     0.441       1.54
## 4     -0.928 -1.11    -0.433 -1.03  -0.698         -0.923   -0.494       1.43
## # ℹ 1 more variable: Net_Profit_Margin <dbl>
```

This particular cluster demonstrates low values across Marketcap, PE-ratio, ROE, ROA, and Net_Profit, while the remaining variables exhibit mixed values.

- Creating barchart to examine patterns in variables that were not utilized in the clustering process.

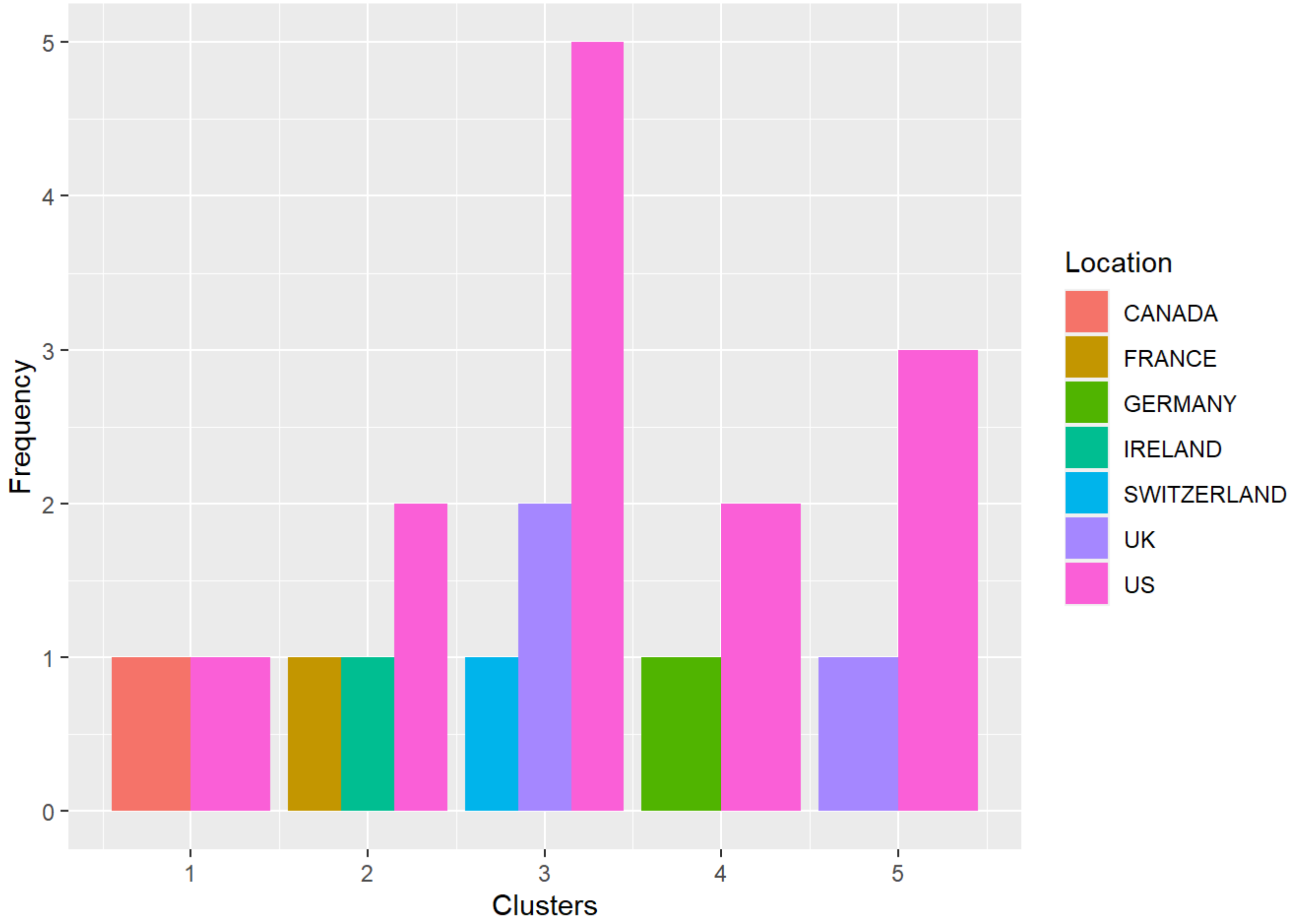
Evaluating the recommendations of different clusters.

```
Pharm.2 <- Pharm %>%
  select(c(1, 12, 13, 14)) %>%
  mutate(cluster = k$cluster)

# Creating a bar plot to visualize the frequency of median recommendations within clusters.
ggplot(Pharm.2, aes(x = cluster, fill = Median_Recommendation)) +
  geom_histogram(position = "dodge", binwidth = 1, aes(y = after_stat(count))) +
  labs(x = "Clusters", y = "Frequency")
```



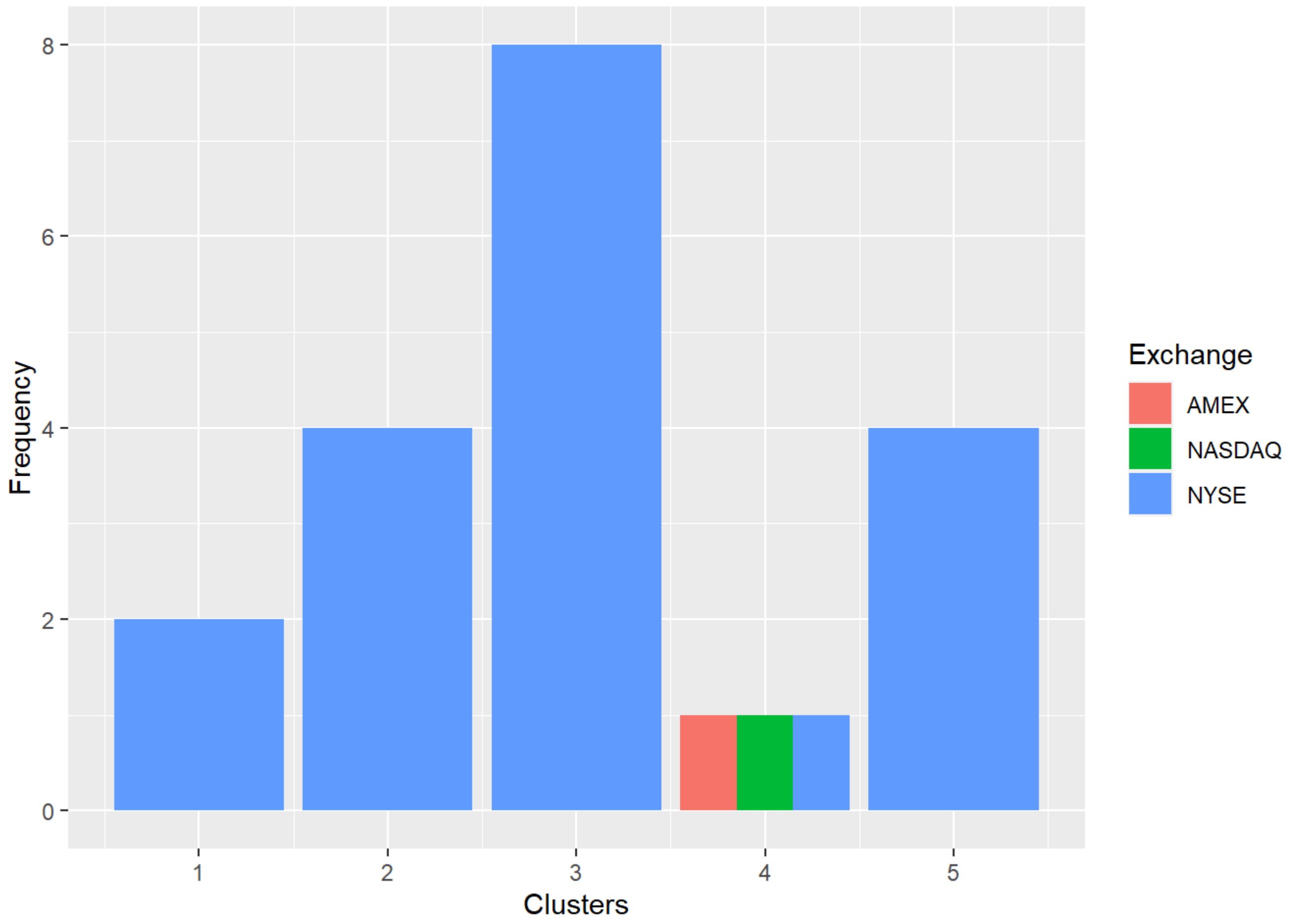
Evaluating countries of clusters□

```
# Creating a bar plot to visualize the distribution of locations across clusters.
ggplot(Pharm.2, aes(cluster, fill = Location)) +
  geom_bar(position = 'dodge') +
  labs(x = 'Clusters', y = 'Frequency')
```



stock-exchanges of clusters

```
ggplot(Pharm.2,mapping = aes(cluster,fill=Exchange))+
  geom_bar(position = 'dodge') +
  labs(x='Clusters',y='Frequency')
```



Interpretation and Observed Pattern

- Cluster 1: Companies listed in multiple exchanges (diversified) and operating in different countries (USA and Germany). Recommendation: Hold more companies with moderate buying on some.
- Cluster 2: Companies listed only on NYSE and operating in specific countries (Canada and USA). Recommendation: Hold half of the companies and moderately buy the other half.
- Cluster 3: Companies listed only on NYSE but operating in different countries (Switzerland, UK, and US). Recommendation: Mixed, with a tendency towards holding most companies.
- Cluster 4: Companies listed only on NYSE and operating in specific countries (UK and US). Recommendation: Moderately buy half and hold the other half of the companies.
- Cluster 5: Companies listed only on NYSE and operating in specific countries (France, Ireland, and US). Recommendation: Moderately buy half and moderately sell half of the companies.

The observed pattern indicates that the recommendations vary based on the geographical presence and exchange listing of the companies. Companies with diversified operations and listings tend to have different recommendations compared to those operating in specific regions or listed on a single exchange.

Solution 4

- Cluster 1: Low cap Highly-Volatile companies, comprises companies with low market capitalization and high volatility, driven by their high Beta values and low profits.
- Cluster 2: Small cap overpriced companies, represents small-cap companies that are deemed overpriced due to their higher price-to-earnings ratios and smaller market capitalization.

- Cluster 3: Mid cap Profitable companies, consists of mid-cap companies that are profitable, with most surpassing the average profit margins and possessing an average market capitalization.
- Cluster 4: Large-cap Under-priced companies, identifies large-cap companies that are undervalued, characterized by their high market capitalization and lower-than-average price-to-earnings ratios, despite exhibiting favorable financial metrics.
- Cluster 5: Small cap Less-Profitable companies, encompasses small-cap companies with lower profitability compared to the average, as evidenced by their smaller market capitalization and below-average profit margins.