k_Means_Clustering

64060_cosadebe

Load necessary libraries

```
suppressPackageStartupMessages({library(dplyr)
  library(factoextra)
  library(ggplot2)
  library(caret)
  library(moments)
})
```

Load and prepare the Pharmaceuticals.csv dataset

```
Pharmaceuticals_df <- read.csv("Pharmaceuticals.csv")

# Select necessary variables (1 to 11) for clustering
Pharmaceuticals_reduced_df <- Pharmaceuticals_df[, 1:11]

# Select numerical variables for rescaling
Pharmaceuticals_num_df <- Pharmaceuticals_reduced_df[, 3:11]
```

Computing for skewness

```
skewness <- sapply(Pharmaceuticals_num_df, skewness) %>%
  data.frame() %>%
  round(2)

skewness
```

```
##                       .
## Market_Cap         1.01
## Beta               0.88
## PE_Ratio           2.36
## ROE                0.94
## ROA                0.17
## Asset_Turnover    -0.09
## Leverage           2.67
## Rev_Growth         0.35
## Net_Profit_Margin -0.30
```

According to the skewness table, several features exhibit strong right skewness.

Transforming the dataset using Yeo-Johnson method

```
# Transform highly skewed feature variables
yj_model <- preProcess(Pharmaceuticals_num_df[, c("Leverage", "PE_Ratio",
                "Market_Cap")], method = "YeoJohnson")
Pharmaceuticals_num_transformed <- predict(yj_model, Pharmaceuticals_num_df)
```

Re compute skewness of the feature variables

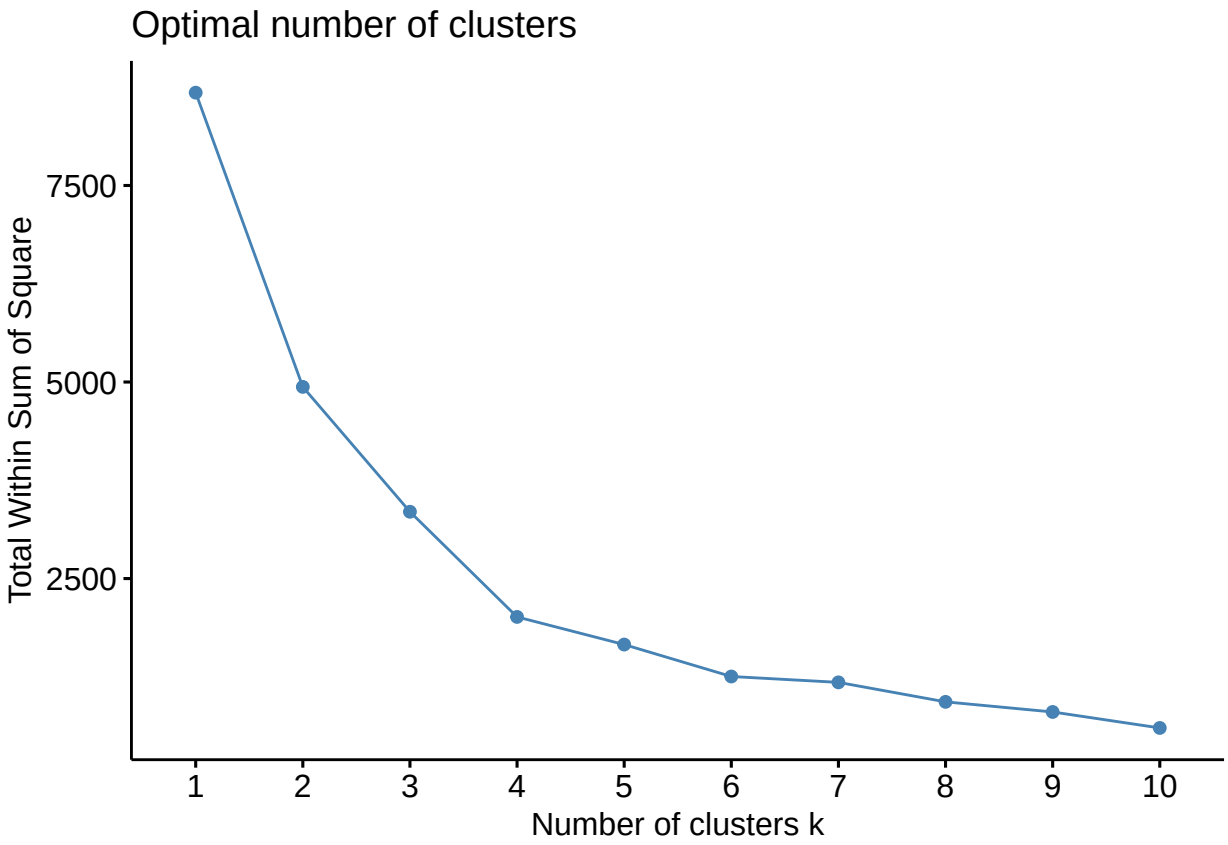
```
skewness <- sapply(Pharmaceuticals_num_transformed, skewness) %>%
  data.frame() %>%
  round(2)

skewness
```

```
##                        .
## Market_Cap         -0.18
## Beta                0.88
## PE_Ratio            0.07
## ROE                 0.94
## ROA                 0.17
## Asset_Turnover     -0.09
## Leverage            0.17
## Rev_Growth          0.35
## Net_Profit_Margin -0.30
```

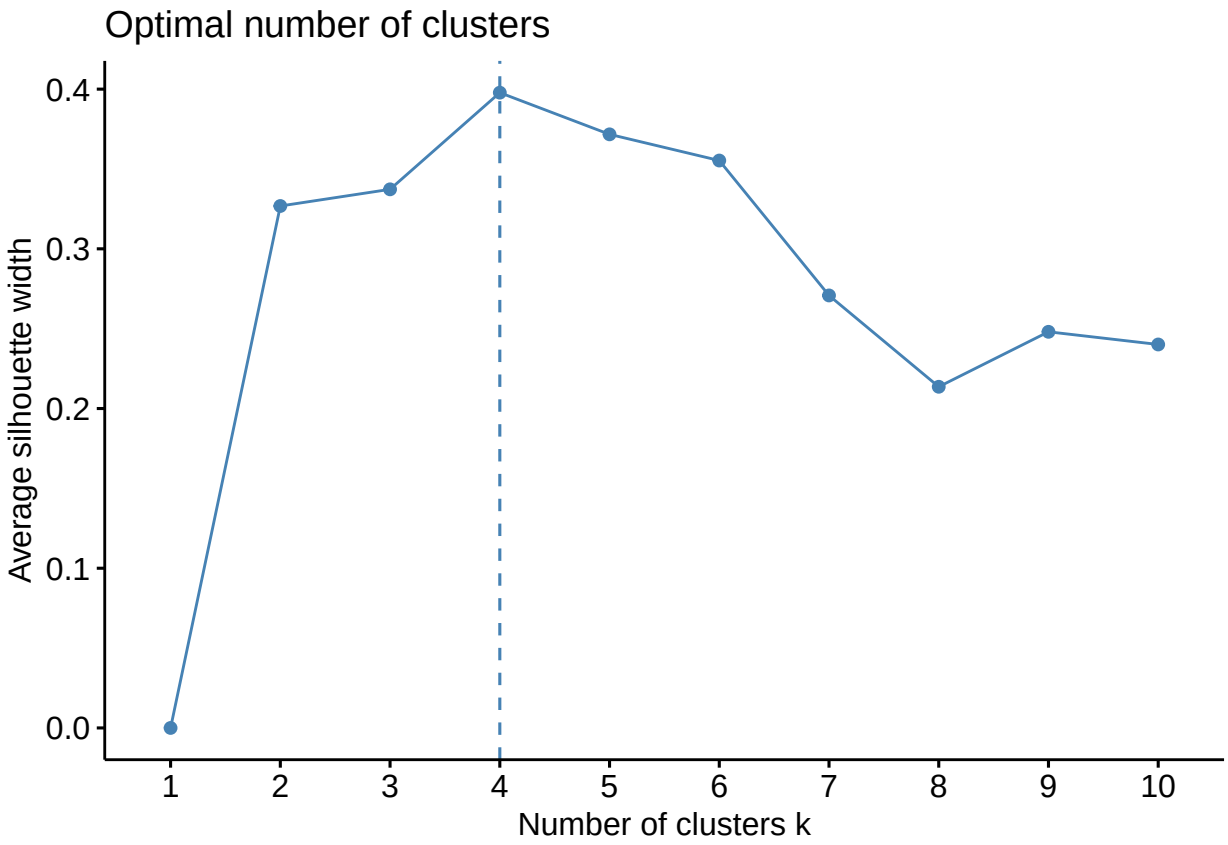
Optimize K value using Elbow method

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



Corroborate k optimization using the silhouette method

```
# Silhouette method
fviz_nbclust(Pharmaceuticals_num_transformed, kmeans, method = "silhouette")
```



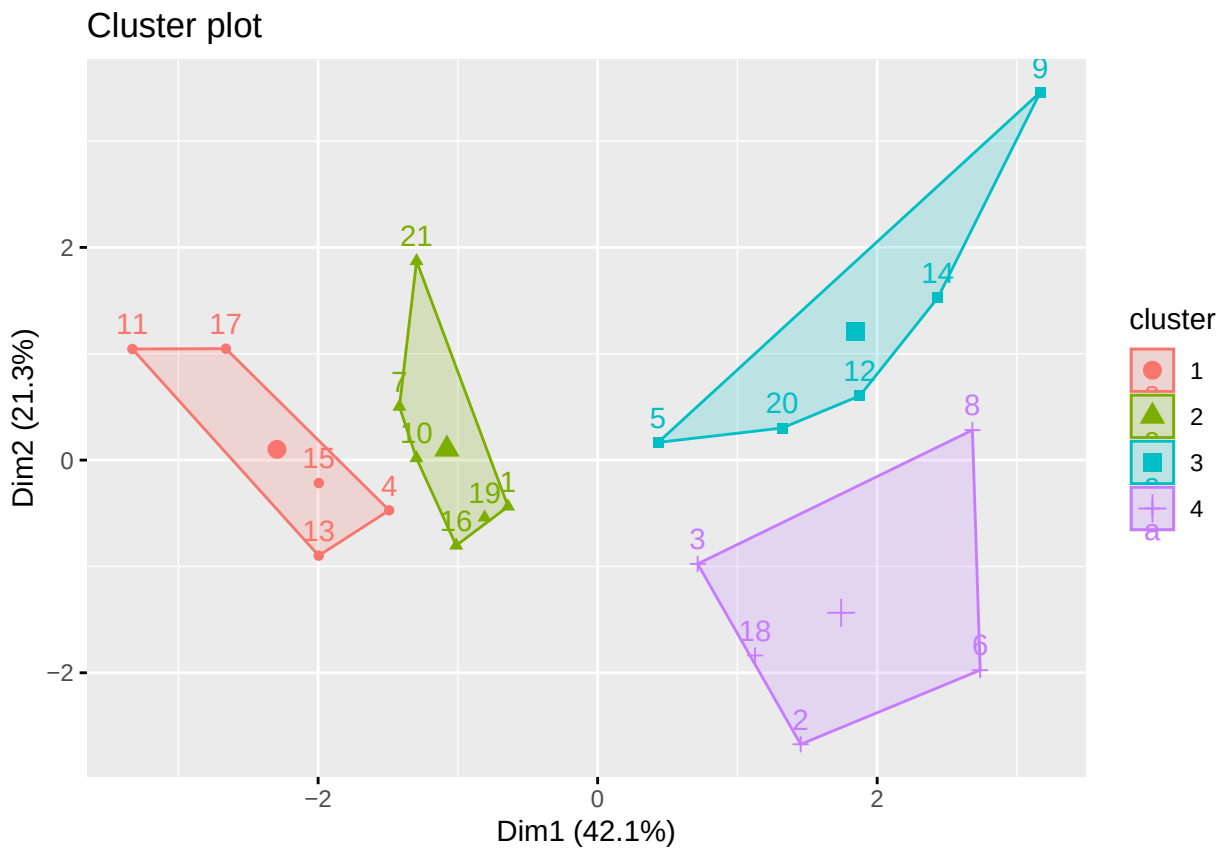
Apply Kmeans Clustering with K = 4

```
# Standardize the data
Pharmaceuticals_norm_df <- scale(Pharmaceuticals_num_transformed)

set.seed(123)

kmeans_result <- kmeans(Pharmaceuticals_norm_df, centers = 4, nstart = 25)

# Visualize the Clustering
fviz_cluster(kmeans_result, data = Pharmaceuticals_norm_df)
```



Summarize the clusters

```
aggregate(Pharmaceuticals_norm_df, by = list(kmeans_result$cluster),
          FUN = mean) %>% round(2)
```

```
##   Group.1 Market_Cap  Beta PE_Ratio   ROE   ROA Asset_Turnover Leverage
## 1       1       1.15 -0.15    -0.05  1.01  1.26           1.11    -0.68
## 2       2       0.43 -0.54    -0.26  0.36  0.48          -0.08    -0.07
## 3       3      -1.00  0.32    -0.61 -0.65 -0.79          -1.11     0.60
## 4       4      -0.67  0.47     0.97 -0.79 -1.05           0.09     0.15
##   Rev_Growth Net_Profit_Margin
## 1       0.40              0.54
## 2      -0.87              0.80
## 3       1.23             -0.15
## 4      -0.59             -1.35
```

a. Reasons for the choices made in conducting the cluster analysis

1. The Pharmaceuticals dataset includes continuous features with varying degrees of skewness — strong, mild, and near-symmetric. To enhance clustering performance, Yeo-Johnson normalization was applied, as Box-Cox failed to correct the extreme skewness in the Leverage variable. This transformation effectively reduced skewness across features, producing a distributional structure better suited for K-means clustering.

2. All variables were standardized using scale() to ensure equal weighting. Without scaling, variables like Market_Cap or Leverage could dominate distance calculations due to larger numeric ranges.

3. The WSS (Elbow) method revealed a clear inflection point at k = 4, which was corroborated by the silhouette method, indicating meaningful segmentation among the 21 firms based on financial features. Beyond four clusters, the reduction in WSS was minimal, suggesting diminishing returns and potential overfitting with additional clusters.

b. Interpretation of the clusters with respect to the numerical variables used in forming the clusters - Based on their standardised means

- Cluster 1: These companies are large and efficient. They have strong returns (ROE = 1.01, ROA = 1.26), high asset use (Asset Turnover = 1.11), and solid profit margins (Net Profit Margin = 0.54). Their Market Cap is high (1.15), and they show steady growth (Revenue Growth = 0.40) with low financial risk - not relying on debt (Leverage = –0.68) and low volatility (Beta = –0.15).

- Cluster 2: These firms are mid-sized and stable. They have moderate returns (ROE = 0.36, ROA = 0.48), strong profit margins (Net Profit Margin = 0.80), and low asset efficiency (Asset Turnover = –0.08). Their Market Cap is average (0.43), but has weak growth (Revenue Growth = –0.87). They carry low risk (Leverage = –0.07) and low market volatility (Beta = –0.54).

- Cluster 3: These are small, risky companies focused on growth. They have poor returns (ROE = –0.65, ROA = –0.79), low asset use (Asset Turnover = –1.11), and weak profit margins (Net Profit Margin = –0.15). However, they show strong growth (Revenue Growth = 1.23) and higher financial risk (Leverage = 0.60). Their Market Cap is low (–1.00), and they're moderately volatile (Beta = 0.32).

- Cluster 4: These firms are small and unstable. They have negative returns (ROE = –0.79, ROA = –1.05), poor profit margins (Net Profit Margin = –1.35), and low asset efficiency (Asset Turnover = 0.09). Despite a high PE Ratio (0.97), they show limited growth (Revenue Growth = –0.59), low risk (Leverage = 0.15), and moderate volatility (Beta = 0.47). Their Market Cap is below average (–0.67).

c. Compare pattern in the clusters with respect to the variables: Median_Recommendation, Location, and Exchange

```
# Extract cluster assignments from k-means result
clusters <- kmeans_result$cluster

# Add cluster labels to the original data frame and make column 1
df <- Pharmaceuticals_df %>%
  mutate(clusters)

# Generate a cross-tabulation - Cluster vs. Median Recommendation
# showing the distribution of Median Recommendation across clusters
table(clusters = df$clusters, df$Median_Recommendation)
```

```
##
## clusters Hold Moderate Buy Moderate Sell Strong Buy
##        1    2            2             1          0
##        2    4            1             1          0
##        3    1            2             2          0
##        4    2            2             0          1
```

Comparison:

- Cluster 1: Mostly "Hold" and "Moderate Buy", no "Strong Buy" ratings despite strong KPIs. This reflects their steady growth (Revenue Growth = 0.40) with low financial risk and low volatility (Beta = –0.15).
- Cluster 2: Dominated by "Hold" ratings, aligning with their stability but weak growth.
- Cluster 3: Mixed ratings ("Moderate Buy" and "Moderate Sell") reflect uncertainty, consistent with their high growth potential (Revenue Growth = 1.23), market volatility (Beta = 0.32), and high financial risk (Leverage = 0.60).
- Cluster 4: includes a "Strong Buy" alongside "Hold" and "Moderate Buy" ratings, reflecting mixed sentiment. It represents small, unstable businesses with negative returns but a relatively high PE Ratio (0.97).

```
# Cross-tabulation - Cluster vs. Location
table(clusters = df$clusters, df$Location)
```

```
##
## clusters CANADA FRANCE GERMANY IRELAND SWITZERLAND UK US
##        1      0      0       0       0           0  2  3
##        2      0      0       0       0           1  0  5
##        3      0      1       0       1           0  0  3
##        4      1      0       1       0           0  1  2
```

Comparison:

- Cluster 1 & 2: Dominated by US firms, consistent with their size and efficiency.
- Cluster 3: Includes France and Ireland - more geographically diverse. Possibly reflecting small, growth-oriented firms.
- Cluster 4: Has higher geographical diversity, includes Canada, Germany, UK, and US — consistent with instability and low asset efficiency (Asset Turnover = 0.09) despite coverage.

```
# Cross-tabulation - Cluster vs. Exchange
table(clusters = df$clusters, df$Exchange)
```

```
##
## clusters AMEX NASDAQ NYSE
##        1    0      0    5
##        2    0      0    6
##        3    1      0    4
##        4    0      1    4
```

Comparison:

- Clusters 1 & 2: Exclusively NYSE — aligns with their size, stability, and institutional presence.
- Cluster 3: Includes AMEX — often associated with smaller firms with focus on Small-cap and micro-cap companies.
- Cluster 4: Cluster with NASDAQ, may reflect tech or volatile growth stocks - the pattern aligns with their moderate market volatility (Beta = 0.47).

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

- clusters 1: Large, efficient firms with strong KPIs and low risk
- clusters 2: Mid-sized, stable firms with moderate returns and weak growth
- clusters 3: Small, risky firms with poor KPIs but high revenue growth
- clusters 4: Small, unstable firms with negative returns