Pharma_Cluster_Analysis

Nikhil Prema Chandra Rao

2024-10-06

Clustering Pharmaceutical Firms Based on Financial Data

Introduction

This research will investigate the pharmaceutical industry's financial structure by evaluating data from 21 businesses. The dataset covers essential quantitative financial metrics including market capitalization, return on equity, and revenue growth. The primary purpose is to employ cluster analysis to discover groupings of enterprises with comparable financial characteristics. This allows us to acquire insights on the fundamental structure of the pharmaceutical industry, as well as the similarities and variances across businesses.

The project uses two main clustering techniques: K-means clustering and hierarchical clustering. The study includes visualization tools to help explain the results, as well as validation procedures such as silhouette analysis and dimensionality reduction using PCA.

The project addresses the following objectives:

1. **Clustering Firms**: Group the firms based on quantitative financial variables using K-means clustering.
2. **Justification of Clustering Approach**: Explain the process of selecting variables, applying scaling, and choosing the clustering method, as well as the rationale behind the number of clusters chosen.
3. **Interpretation of Clusters**: Understand the characteristics of each cluster in terms of the financial variables used.
4. **Naming the Clusters**: Assign meaningful names to the clusters based on the defining financial metrics of each group.

Data loading and Preprocessing

The dataset contains financial information on the following variables:

- **Market Cap** (a)
- **Beta** (b)
- **PE Ratio** (c)
- **ROE** (d)
- **ROA** (e)
- **Asset Turnover** (f)
- **Leverage** (g)
- **Revenue Growth** (h)
- **Net Profit Margin** (i)

```r
library(readr)
library(cluster)
library(ggplot2)
library(dplyr)
library(tidyverse)
library(GGally)
library(corrplot)
library(reshape2)
library(plotly)
library(gridExtra)

pharma_data <- read_csv("Pharmaceuticals.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.
```

```r
str(pharma_data)
```

```
## spc_tbl_ [21 x 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Symbol               : chr [1:21] "ABT" "AGN" "AHM" "AZN" ...
##  $ Name                 : chr [1:21] "Abbott Laboratories" "Allergan, Inc." "Amersham plc" "AstraZen
##  $ Market_Cap           : num [1:21] 68.44 7.58 6.3 67.63 47.16 ...
##  $ Beta                 : num [1:21] 0.32 0.41 0.46 0.52 0.32 1.11 0.5 0.85 1.08 0.18 ...
##  $ PE_Ratio             : num [1:21] 24.7 82.5 20.7 21.5 20.1 27.9 13.9 26 3.6 27.9 ...
##  $ ROE                  : num [1:21] 26.4 12.9 14.9 27.4 21.8 3.9 34.8 24.1 15.1 31 ...
##  $ ROA                  : num [1:21] 11.8 5.5 7.8 15.4 7.5 1.4 15.1 4.3 5.1 13.5 ...
##  $ Asset_Turnover       : num [1:21] 0.7 0.9 0.9 0.9 0.6 0.6 0.9 0.6 0.3 0.6 ...
##  $ Leverage             : num [1:21] 0.42 0.6 0.27 0 0.34 0 0.57 3.51 1.07 0.53 ...
##  $ Rev_Growth           : num [1:21] 7.54 9.16 7.05 15 26.81 ...
##  $ Net_Profit_Margin    : num [1:21] 16.1 5.5 11.2 18 12.9 2.6 20.6 7.5 13.3 23.4 ...
##  $ Median_Recommendation: chr [1:21] "Moderate Buy" "Moderate Buy" "Strong Buy" "Moderate Sell" ...
##  $ Location             : chr [1:21] "US" "CANADA" "UK" "UK" ...
##  $ Exchange             : chr [1:21] "NYSE" "NYSE" "NYSE" "NYSE" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Symbol = col_character(),
##   ..   Name = col_character(),
##   ..   Market_Cap = col_double(),
##   ..   Beta = col_double(),
##   ..   PE_Ratio = col_double(),
##   ..   ROE = col_double(),
##   ..   ROA = col_double(),
##   ..   Asset_Turnover = col_double(),
##   ..   Leverage = col_double(),
##   ..   Rev_Growth = col_double(),
##   ..   Net_Profit_Margin = col_double(),
```

```
##   ..   Median_Recommendation = col_character(),
##   ..   Location = col_character(),
##   ..   Exchange = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

The result shows the structure of the 'pharma_data' dataframe, which has 21 rows (firms) and 14 columns (variables). The columns contain a combination of numerical data (e.g., Market Cap, Beta, ROE) and character data (e.g., Symbol, Name, Location, etc.). It gives an overview of each column's kind and the first few values.

```
# Select quantitative variables (a)-(i)
quant_data <- pharma_data %>%
  select(Market_Cap, Beta, PE_Ratio, ROE, ROA, Asset_Turnover, Leverage, Rev_Growth, Net_Profit_Margin)

# Scale the data to normalize it before clustering
scaled_data <- scale(quant_data)
head(scaled_data)
```

```
##      Market_Cap        Beta    PE_Ratio         ROE        ROA Asset_Turnover
## [1,]  0.1840960 -0.80125356 -0.04671323  0.04009035  0.2416121  -5.121077e-16
## [2,] -0.8544181 -0.45070513  3.49706911 -0.85483986 -0.9422871   9.225312e-01
## [3,] -0.8762600 -0.25595600 -0.29195768 -0.72225761 -0.5100700   9.225312e-01
## [4,]  0.1702742 -0.02225704 -0.24290879  0.10638147  0.9181259   9.225312e-01
## [5,] -0.1790256 -0.80125356 -0.32874435 -0.26484883 -0.5664461  -4.612656e-01
## [6,] -0.6953818  2.27578267  0.14948233 -1.45146000 -1.7127612  -4.612656e-01
##        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
```

In this portion of the code, we choose the quantitative variables from the 'pharma_data' dataframe that are useful for clustering. The variables used are 'Market_Cap', 'Beta', 'PE_Ratio', 'ROE', 'ROA', 'Asset_Turnover', 'Leverage', 'Rev_Growth', and 'Net_Profit_Margin'.

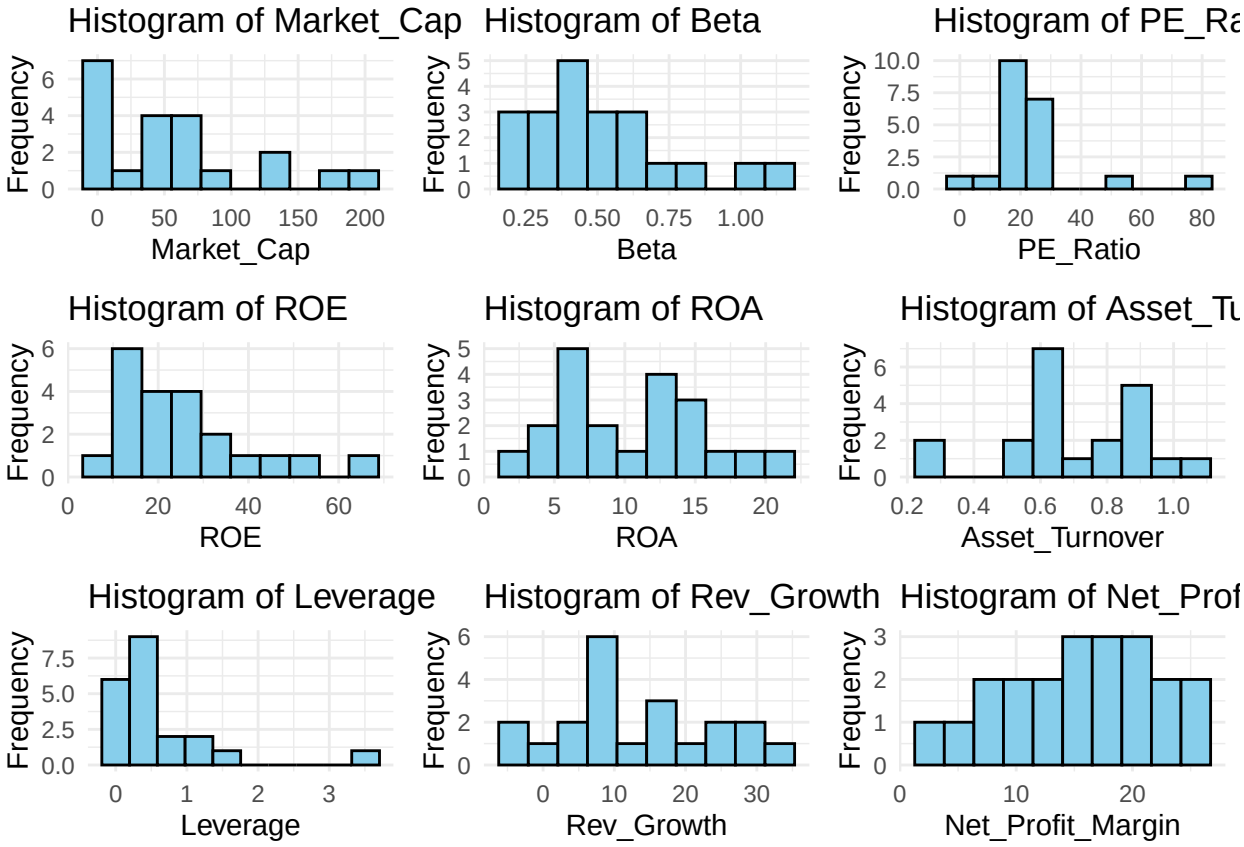
Next, we use the'scale()' method to standardize the data. Scaling is necessary in clustering because it guarantees that each variable contributes equally to distance computations, avoiding variables with longer ranges from dominating the clustering process.

The result displays the first six rows of the scaled data, with each value normalized (mean = 0, standard deviation = 1). For example, the top row values show how each firm's 'Market_Cap', 'Beta', and so on compare to the mean values of these variables across all businesses. Positive values indicate a value greater than the mean, whereas negative values indicate a value less than the mean, allowing for simpler comparisons across variables.

Exploratory Data Analysis

1. Histograms for each variable

```r
numerical_vars <- names(quant_data)
plot_list <- list()
for (var in numerical_vars) {
  p <- ggplot(pharma_data, aes_string(x = var)) +
    geom_histogram(bins = 10, fill = "skyblue", color = "black") +
    labs(title = paste("Histogram of", var), x = var, y = "Frequency") +
    theme_minimal()

  plot_list[[var]] <- p
}
grid.arrange(grobs = plot_list, ncol = 3)
```

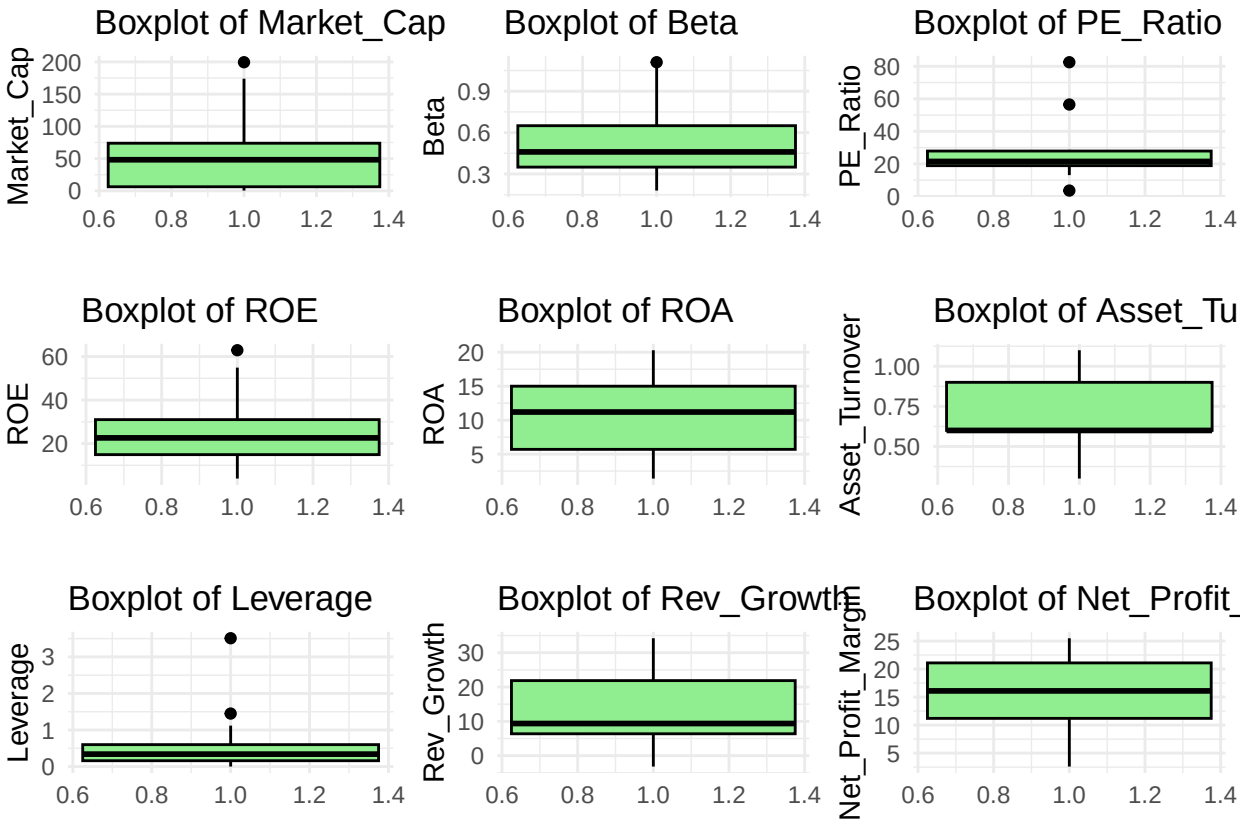


1. **Market Cap**: Smaller businesses predominate in the dataset; the majority of enterprises have a market capitalization below 50.
2. **Beta**: Most businesses have a beta value in the range of 0.25 to 0.75, indicating that their volatility is typically lower than that of the market.
3. **PE Ratio**: Most companies have a low price-to-earnings ratio, indicating that the market might regard them as having weaker growth potential or profitability.
4. **ROE (Return on Equity)**: According to the histogram, the majority of businesses have a ROE of less than 40, which indicates a range of profitability with a few outstanding achievers.
5. **ROA (Return on Assets)**: The majority of businesses have a ROA of five to fifteen, indicating that they typically make profitable use of their assets.
6. **Asset Turnover**: The distribution peaks between 0.6 and 1.0, suggesting that many businesses use their assets to produce income with a modest level of efficiency.

7. **Leverage**: Most businesses appear to use less debt in comparison to equity, as indicated by the leverage values clustering below 1.
8. **Revenue Growth**: While few outliers have seen noticeably higher growth, most businesses have had growth below 10%.
9. **Net Profit Margin**: Most businesses have net profit margins that fall between 10% to 30%, which indicates a modest level of profitability.

2. Boxplots for outliers detection

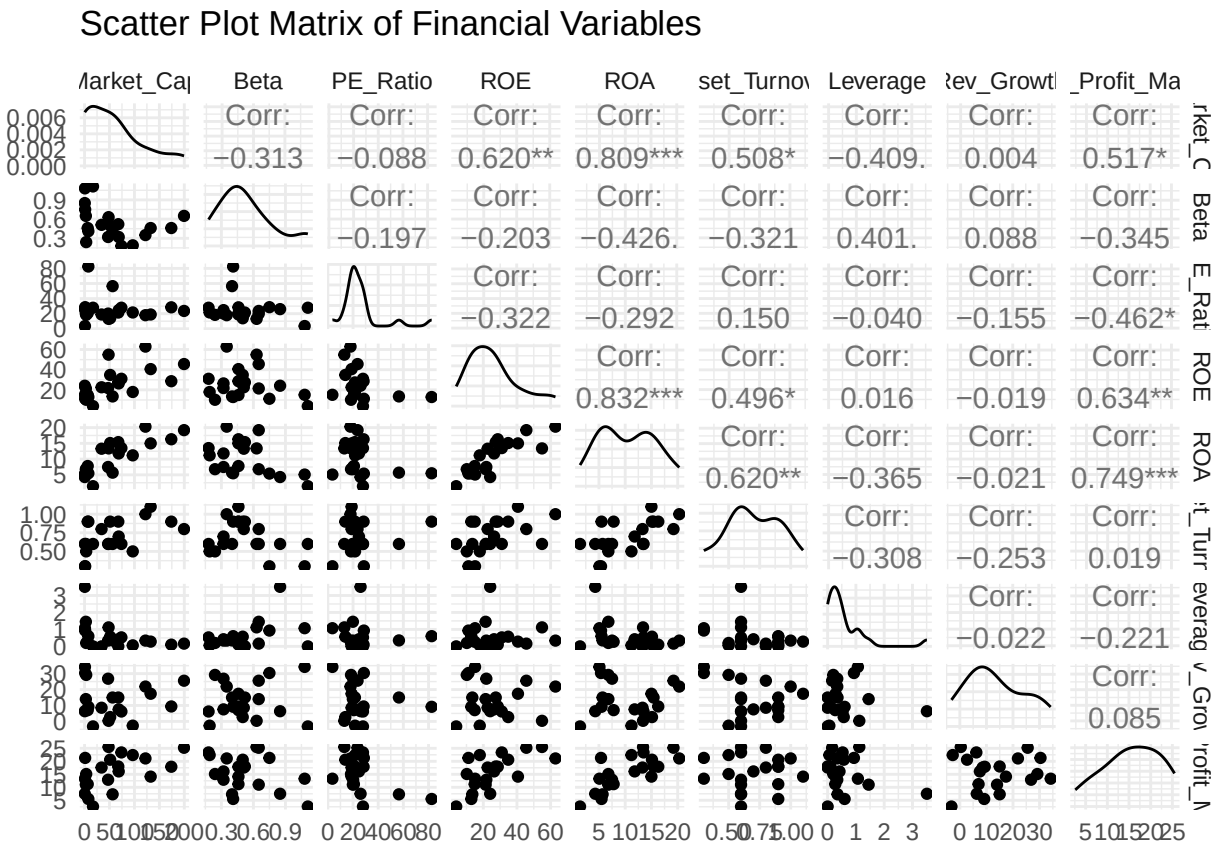
```
boxplot_list <- list()
for (var in numerical_vars) {
  p <- ggplot(pharma_data, aes_string(x = "1", y = var)) +  # x = "1" to create a single box for each v
    geom_boxplot(fill = "lightgreen", color = "black") +
    labs(title = paste("Boxplot of", var), x = "", y = var) +
    theme_minimal()

  boxplot_list[[var]] <- p
}
grid.arrange(grobs = boxplot_list, ncol = 3)
```



1. **Market Cap**: The distribution of the data is right-skewed, with a small number of outliers suggesting substantially higher market capitalization than the bulk of enterprises.
2. **Beta**: With a few outliers showing increased volatility, the majority of corporations have beta values that are close to 0.5.

3. **PE Ratio**: A few companies have noticeably higher P/E ratios than the majority, which is indicated by the distribution's heavy right-skewedness and several outliers.
4. **ROE (Return on Equity)**: With a few severe outliers indicating some very lucrative corporations, the data is generally distributed.
5. **ROA (Return on Assets)**: The boxplot displays a distribution that is balanced, with the majority of values falling between 5 and 15, which denotes overall company efficiency.
6. **Asset Turnover**: With no notable outliers, most businesses have asset turnover values of 0.75 or lower, which indicates moderate efficiency.
7. **Leverage**: The majority of businesses have lower debt levels, but a small number of them (outliers) have extremely high leverage.
8. **Revenue Growth**: A few outliers exhibit significantly higher growth than the average of 10% for most businesses.
9. **Net Profit Margin**: A few outliers, or more lucrative businesses, are included in the data, which is centered around 20% for net profit margin.

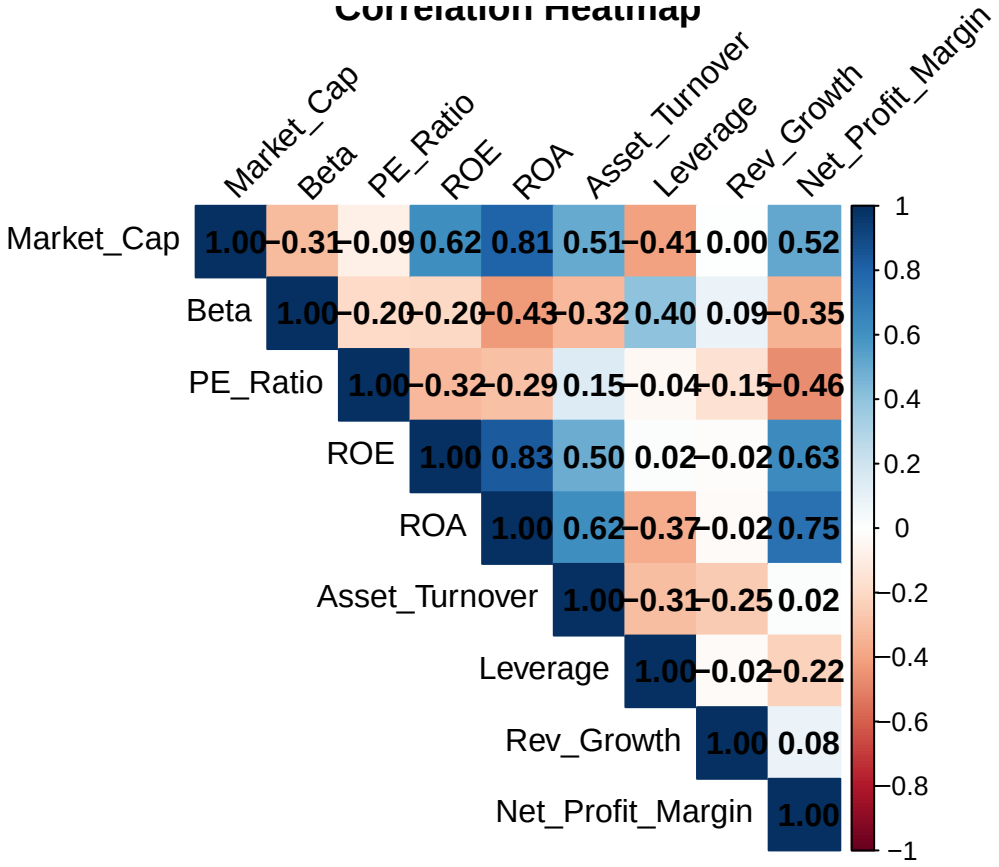
3. Pair Plot and Correlation Matrix



The correlations between various financial variables are represented visually in the scatter plot matrix. The correlation coefficient is shown above each scatterplot, which displays the correlation between two variables. For example, there is a large positive correlation (0.832) between **ROE** and **ROA**, suggesting that businesses with high returns on equity also typically have high returns on assets. Similar to **ROA** (0.426) and **Asset Turnover** (0.508), **Market Cap** exhibits moderately favorable correlations, indicating that larger companies typically make better use of their assets. There appears to be less correlation between firm size and valuation multiples for other factors, such as **PE Ratio** and **Market Cap**. The distribution of each variable is displayed separately by the density curves that run down the diagonal.

4. Correlation Matrix Heatmap

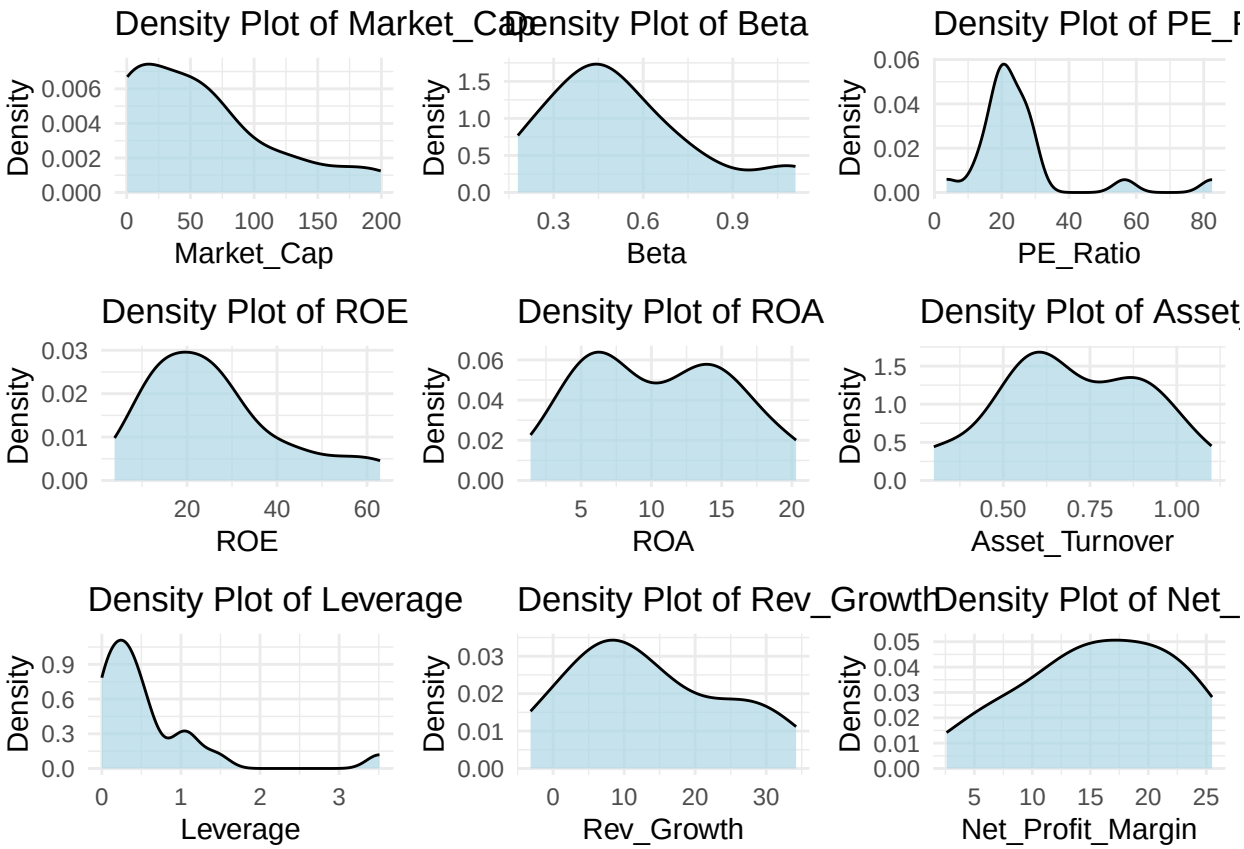
```
correlation_matrix <- cor(quant_data)
corrplot(correlation_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45, addCoef.c
```



The correlation heatmap you gave shows the links between several financial data, with color intensities indicating the strength of the associations. Positive correlations are indicated in blue, while negative correlations are shown in red. For example, "ROE" (Return on Equity) and "ROA" (Return on Assets) have a strong correlation (0.83), meaning that when ROE grows, so does ROA. Conversely, "PE_Ratio" and "Net_Profit_Margin" show a negative connection (-0.46), indicating that firms with a higher P/E ratio may have lower profit margins. This image allows you to quickly determine which factors are interconnected.

5. Density Plots for each variable

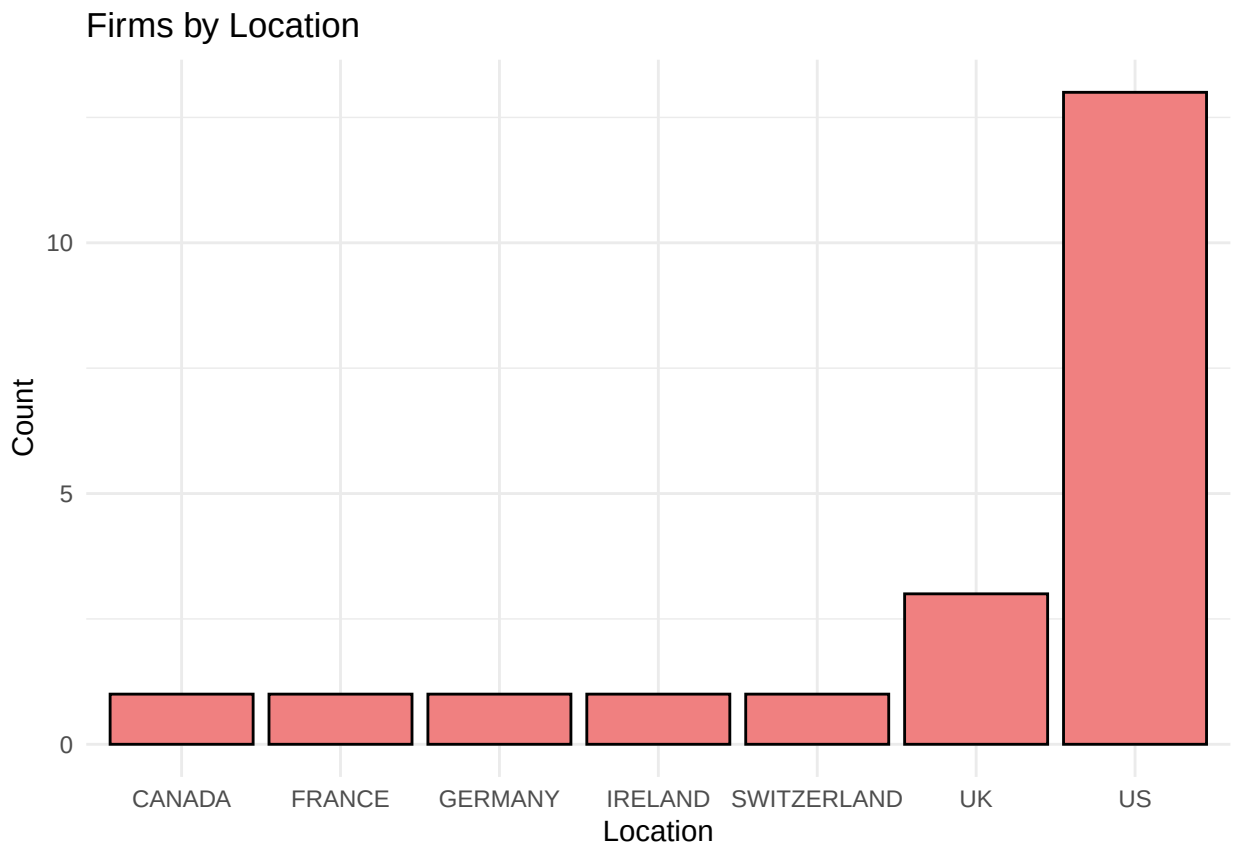
```
plots <- list()
for (var in numerical_vars) {
  Densp <- ggplot(pharma_data, aes_string(x = var)) +
    geom_density(fill = "lightblue", alpha = 0.7) +
    labs(title = paste("Density Plot of", var), x = var, y = "Density") +
    theme_minimal()

  plots[[var]] <- Densp
}
grid.arrange(grobs = plots, ncol = 3)
```



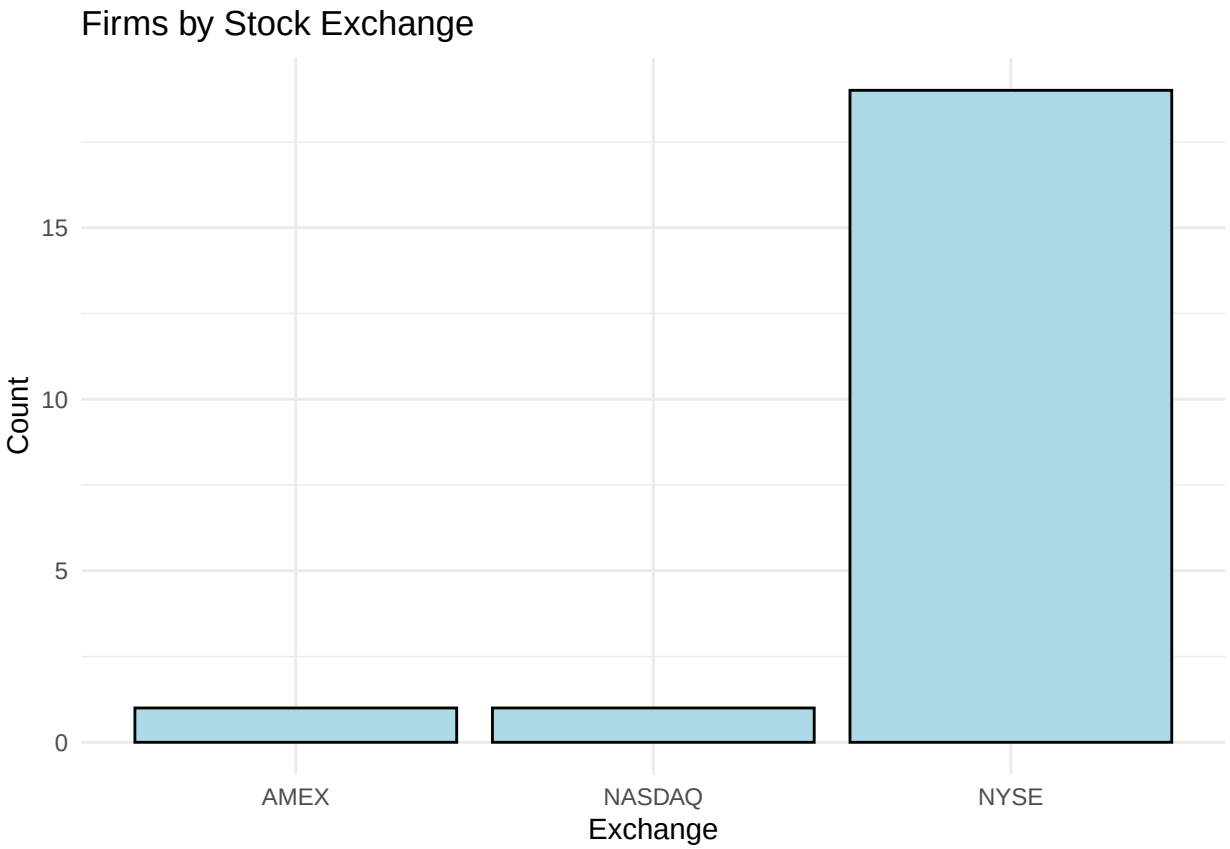
The graphic shows different density charts for various financial indicators, illustrating the distribution of each variable. The smooth curves give information on the frequency of data points at various levels. For example, the "Market_Cap" plot has a right-skewed distribution, showing that the majority of firms have smaller market capitalizations, while a few have considerably higher values. The "Leverage" plot has a significant peak at lower values, indicating that most businesses have little leverage. Each figure contributes to a better understanding of how the data is distributed by showing patterns such as skewness, peaks, and the dispersion of financial metric values.

6. Bar Plot for Categorical Variables (Location and Exchange)

```
ggplot(pharma_data, aes(x = Location)) +
  geom_bar(fill = "lightcoral", color = "black") +
  labs(title = "Firms by Location", x = "Location", y = "Count") +
  theme_minimal()
```




```
ggplot(pharma_data, aes(x = Exchange)) +
  geom_bar(fill = "lightblue", color = "black") +
  labs(title = "Firms by Stock Exchange", x = "Exchange", y = "Count") +
  theme_minimal()
```



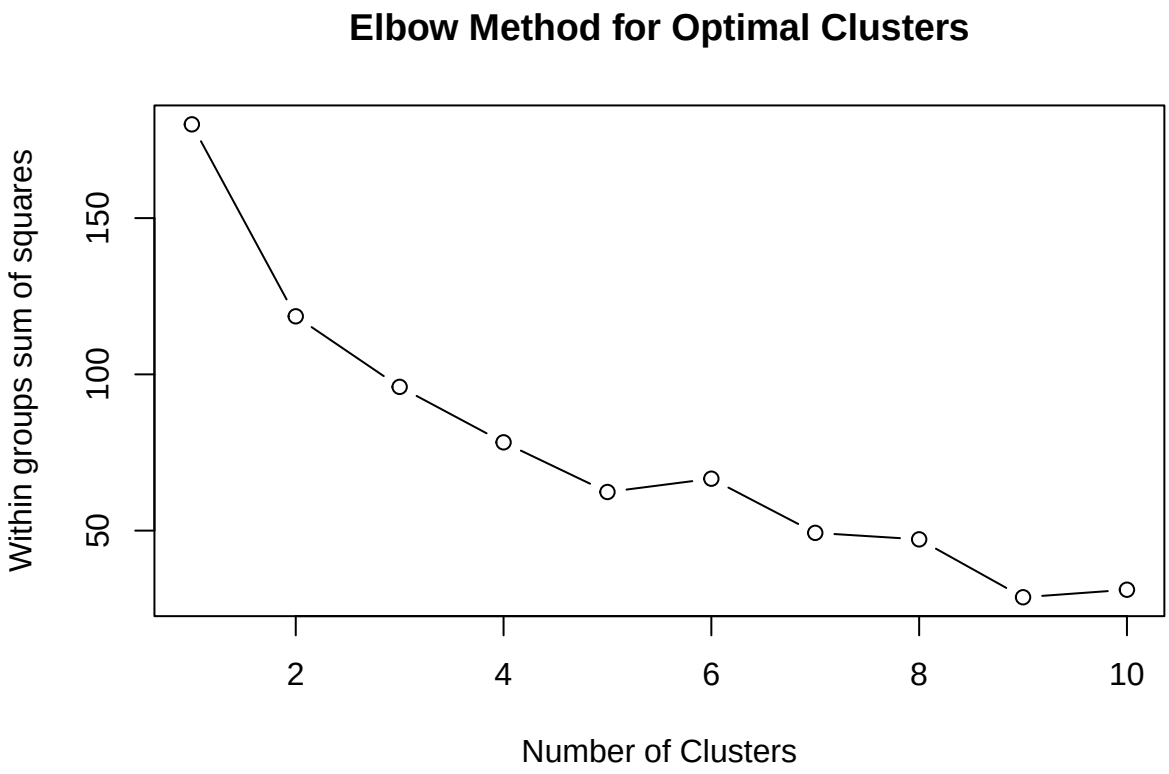
The first code snippet creates a bar plot that shows the distribution of pharmaceutical enterprises by geographical region, with each bar reflecting the number of firms in a certain place and colored in light coral.

The second code snippet generates a similar bar plot depicting the number of businesses listed on several stock exchanges, with bars tinted light blue to highlight the number of enterprises affiliated with each.

Clustering Methodology

K-Means Clustering

```
scaled_data <- scale(quant_data)
scaled_data <- as.matrix(scaled_data)
wss <- numeric(10)
for (i in 1:10) {
  kmeans_result <- kmeans(scaled_data, centers=i)
  wss[i] <- kmeans_result$tot.withinss
}
plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares",
     main="Elbow Method for Optimal Clusters")
```



This figure use the Elbow Method to calculate the ideal number of clusters in a dataset for clustering algorithms such as K-means. The y-axis indicates the within-group sum of squares (a measure of variance), while the x-axis shows the number of clusters. As the number of clusters grows, the variance reduces. The "elbow" point, when the rate of reduction slows dramatically (around 3 or 4 clusters in this example), is frequently regarded as the ideal number of clusters since it balances model complexity and performance.

This line of code will generate a plot with the number of clusters on the x-axis and the within-cluster sum of squares (WSS) on the y-axis. You should be able to see the "elbow" point in the graph, which indicates the optimal number of clusters. In your case, you mentioned that you decided on 3 clusters based on this elbow plot.

```
set.seed(123)
kmeans_result <- kmeans(scaled_data, centers = 3)
pharma_data$cluster <- kmeans_result$cluster
table(pharma_data$cluster)
```

```
##
## 1 2 3
## 7 9 5
```

Code Explanation

The provided R code begins by plotting the elbow method graph, which displays the relationship between the number of clusters (ranging from 1 to 10) and the within-cluster sum of squares (WSS), helping to identify the optimal number of clusters by observing the "elbow" point in the graph. After determining that 3 clusters are appropriate based on this analysis, the code performs K-means clustering on the scaled dataset

with the kmeans() function, setting a random seed for reproducibility. It then adds the cluster membership information to the original pharma_data dataset by creating a new column called cluster, indicating the cluster assignments for each observation. Finally, the code uses the table() function to generate a frequency table that summarizes the number of observations assigned to each cluster.

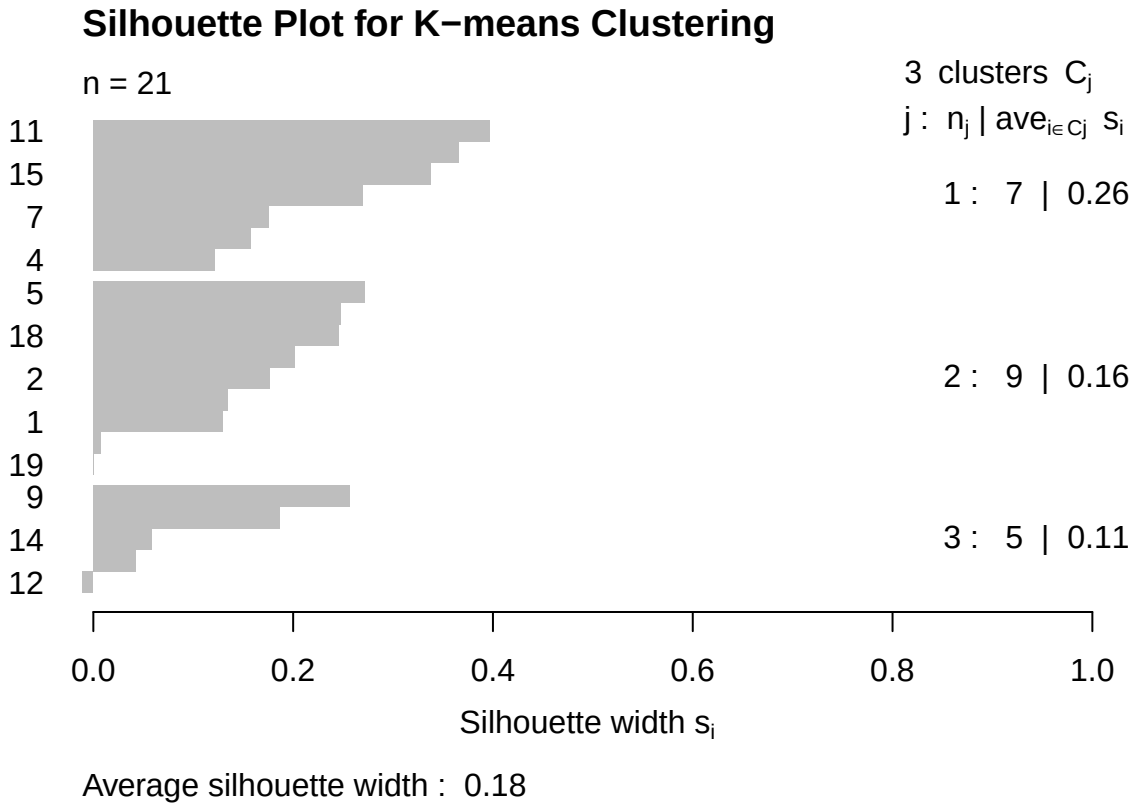
Output Summary

Interpretation of the Output: Cluster 1: There are 7 observations (or firms) that belong to cluster 1. Cluster 2: There are 9 observations that belong to cluster 2. Cluster 3: There are 5 observations that belong to cluster 3.

This output shows the distribution of firms among the three clusters that were determined by the K-means algorithm based on the features in scaled_data. Understanding these distributions can help in analyzing the characteristics of each cluster and interpreting the results of the clustering process.

Silhouette Analysis for Validation

```
silhouette_score <- silhouette(kmeans_result$cluster, dist(scaled_data))
plot(silhouette_score, main = "Silhouette Plot for K-means Clustering")
```



This is a silhouette plot for K-means clustering, showing the quality of clustering for 21 data points divided into 3 clusters. The x-axis represents the silhouette width, which measures how similar each point is to its own cluster versus other clusters. A higher silhouette width indicates better clustering. The plot lists the number of data points (n_j) in each cluster and the average silhouette width (s_i) for each cluster. The

overall average silhouette width is 0.18, suggesting that the clustering is weak, as values closer to 1 indicate better-defined clusters. Cluster 1 has the highest silhouette width (0.26), while cluster 3 has the lowest (0.11).

The provided R code performs silhouette analysis to evaluate the quality of the K-means clustering solution obtained in the previous step. The silhouette() function calculates the silhouette scores for each observation based on the cluster assignments from kmeans_result$cluster and the distance matrix computed from the scaled dataset using the dist() function. The silhouette score measures how similar an observation is to its own cluster compared to other clusters; scores range from -1 (poor clustering) to 1 (good clustering).

Post-Clustering Analysis

1. Cluster Size Information

```
cluster_sizes <- table(kmeans_result$cluster)
print(cluster_sizes)
```

```
##
## 1 2 3
## 7 9 5
```

Interpretation: The output indicates that Cluster 1 contains 7 firms, Cluster 2 contains 9 firms, and Cluster 3 contains 5 firms.

2. Intra-cluster Variance

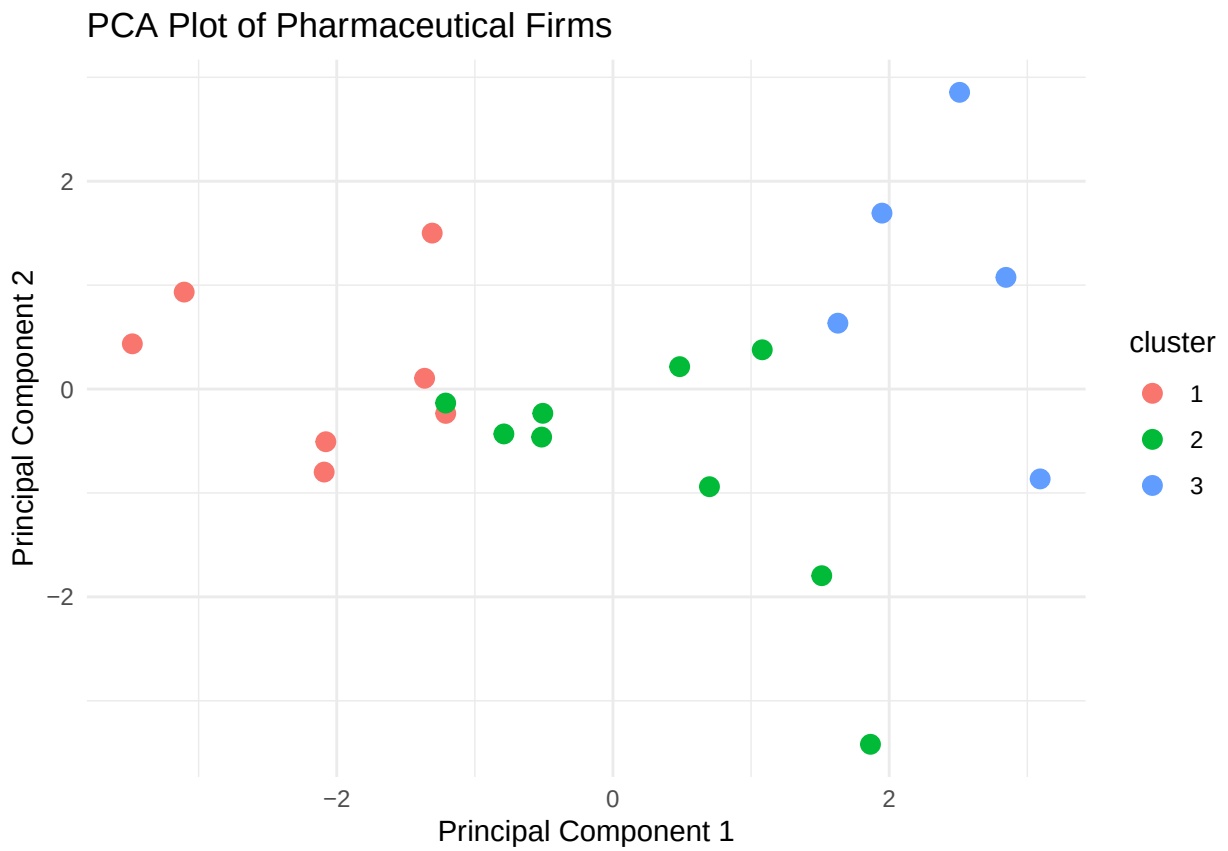
```
cluster_variances <- kmeans_result$withinss
print(cluster_variances)
```

```
## [1] 25.26414 42.25037 31.94053
```

These values indicate the variance within each of the three clusters. Lower values suggest that the firms within that cluster are more similar to each other. Cluster 1 has the lowest variance, indicating firms in this cluster are more closely grouped together.

3. PCA for Dimensionality Reduction and Visualization

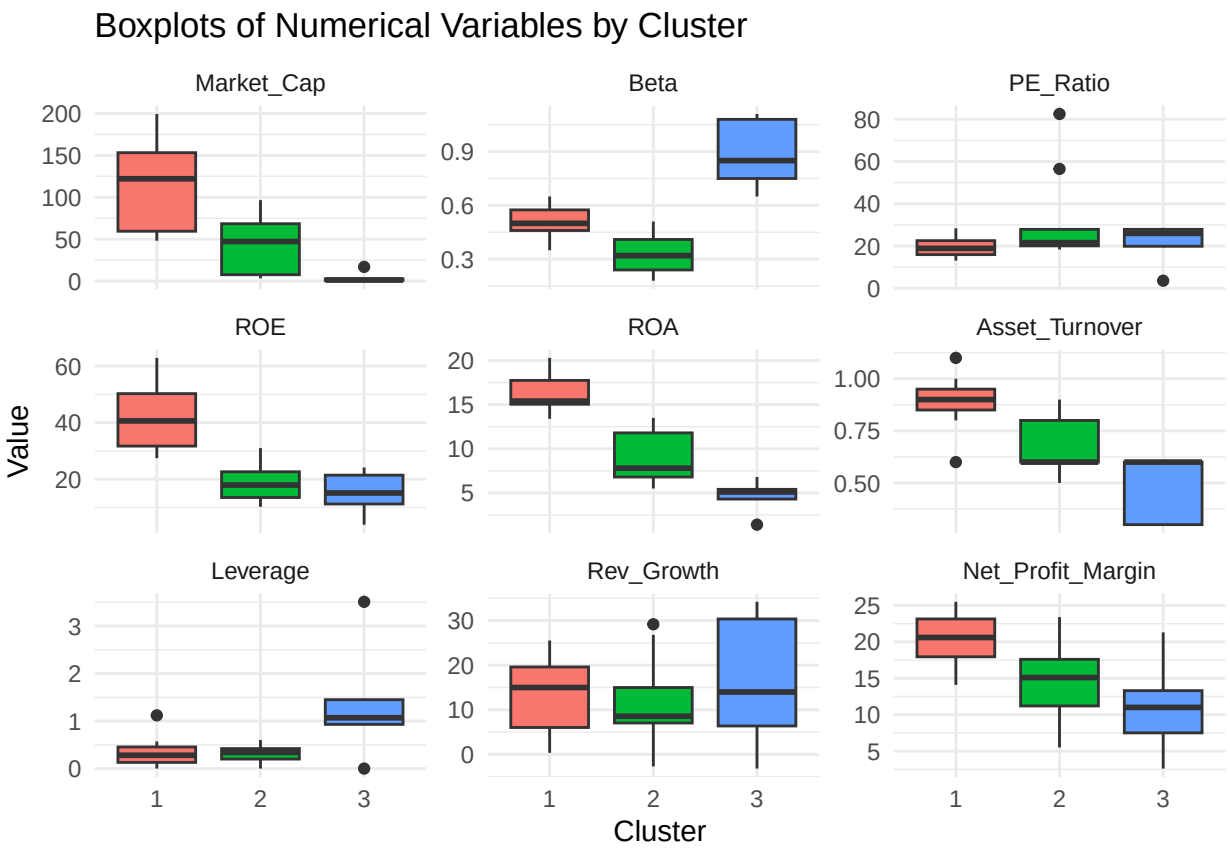
```
pca_result <- prcomp(scaled_data)
pca_data <- data.frame(pca_result$x[, 1:2], cluster = factor(kmeans_result$cluster))
ggplot(pca_data, aes(x = PC1, y = PC2, color = cluster)) +
  geom_point(size = 3) +
  labs(title = "PCA Plot of Pharmaceutical Firms", x = "Principal Component 1", y = "Principal Componen
  theme_minimal()
```



The PCA plot of pharmaceutical firms displays the distribution of the firms based on the first two principal components, which capture the most variance in the data. Each point represents a firm, color-coded by cluster: red for Cluster 1, green for Cluster 2, and blue for Cluster 3. The separation of clusters in the plot indicates distinct groupings based on their financial metrics, with firms in Cluster 1 (red) appearing more spread out along the negative side of Principal Component 1, while Clusters 2 (green) and 3 (blue) cluster towards the positive side, suggesting differences in their financial characteristics and performance.

4. Cluster Profiles with Boxplots

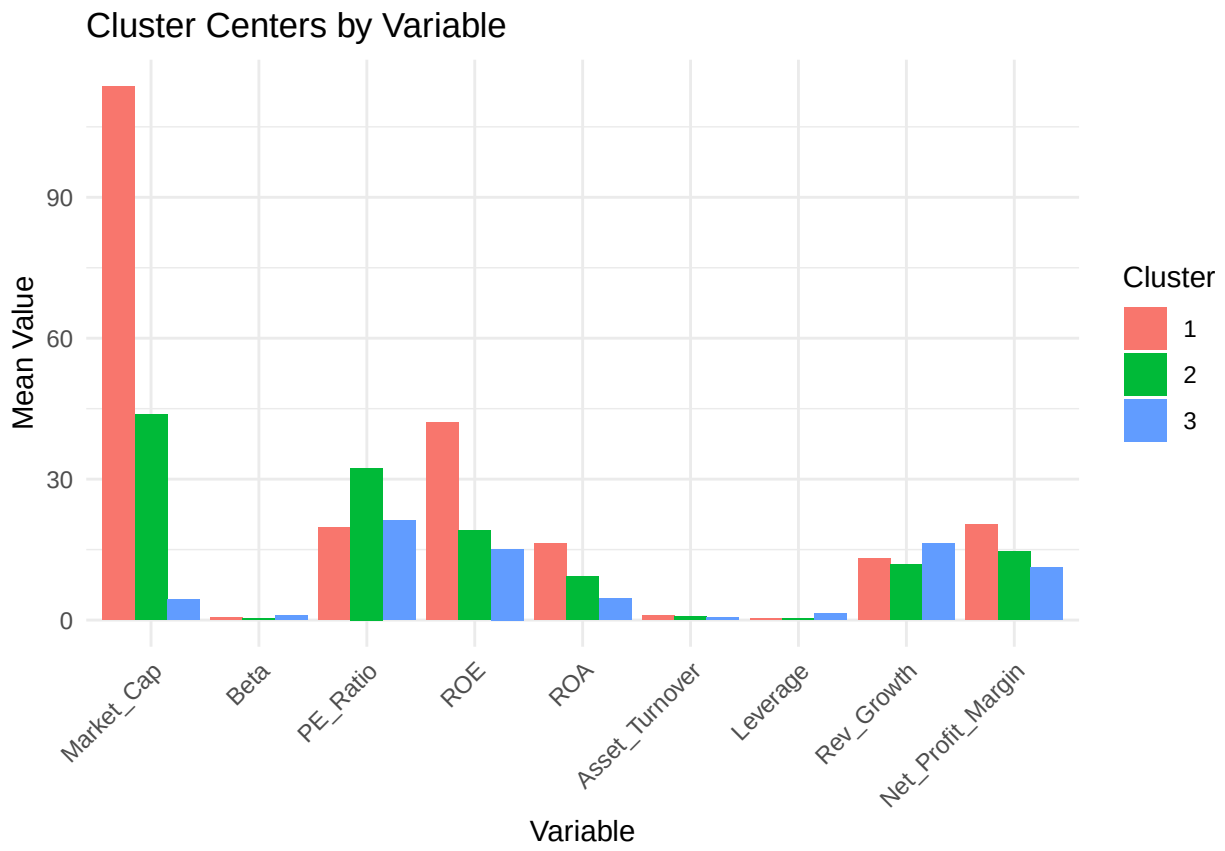
```
pharma_long <- melt(pharma_data[, c("cluster", numerical_vars)], id.vars = "cluster")
ggplot(pharma_long, aes(x = factor(cluster), y = value, fill = factor(cluster))) +
  geom_boxplot() +
  facet_wrap(~ variable, scales = "free_y") +
  labs(title = "Boxplots of Numerical Variables by Cluster", x = "Cluster", y = "Value") +
  theme_minimal() +
  theme(legend.position = "none")
```



The boxplots show how various numerical variables are distributed throughout the three groups of pharmaceutical enterprises. Each variable, such as Market Cap, Beta, ROE, and others, is represented by a distinct boxplot for each cluster (1, 2, and 3), providing for a clear comparison of financial features. For example, Cluster 1 (red) has greater Market Cap and ROE values, suggesting better financial success than Clusters 2 (green) and 3 (blue), which have more fluctuation in their measures. The boxplots successfully emphasize the disparities in financial measures between the clusters, implying separate performance characteristics for each group of enterprises.

5. Cluster Centers Visualization

```
cluster_means <- aggregate(. ~ cluster, data = pharma_data[, c("cluster", "Market_Cap", "Beta", "PE_Rati
cluster_means_melt <- melt(cluster_means, id.vars = "cluster")
ggplot(cluster_means_melt, aes(x = variable, y = value, fill = factor(cluster))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Cluster Centers by Variable", x = "Variable", y = "Mean Value", fill = "Cluster") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The bar chart shows the mean values of several financial indicators for each of the three groups of pharmaceutical companies, allowing for a clear comparison of their features. Each cluster is depicted with a distinct color: Cluster 1 (red) has much higher mean values for Market Cap, ROE, and PE Ratio, implying that these companies are larger and more profitable. Clusters 2 (green) and 3 (blue) have lower mean values across most variables, indicating that they are smaller and maybe less financially resilient. This image successfully illustrates the financial differences across the clusters, supporting the profiles developed in prior research.

6. Feature Importance (Variance explained by each variable in clustering)

```
var_explained <- apply(kmeans_result$centers, 2, function(x) var(x))
var_explained_sorted <- sort(var_explained, decreasing = TRUE)
print(var_explained_sorted)
```

```
##              ROA             Beta              ROE        Market_Cap
##        1.2449839        1.2058250        0.9345132        0.8910515
##    Asset_Turnover         Leverage Net_Profit_Margin         PE_Ratio
##        0.8757366        0.6086793        0.5015973        0.1813260
##       Rev_Growth
##        0.0436618
```

The output shows the variance explained by each variable in clustering. ROA and Beta are the most significant variables, indicating they contribute more to differentiating the clusters than others.

Analysis of Variance Explained by Variables in Clustering

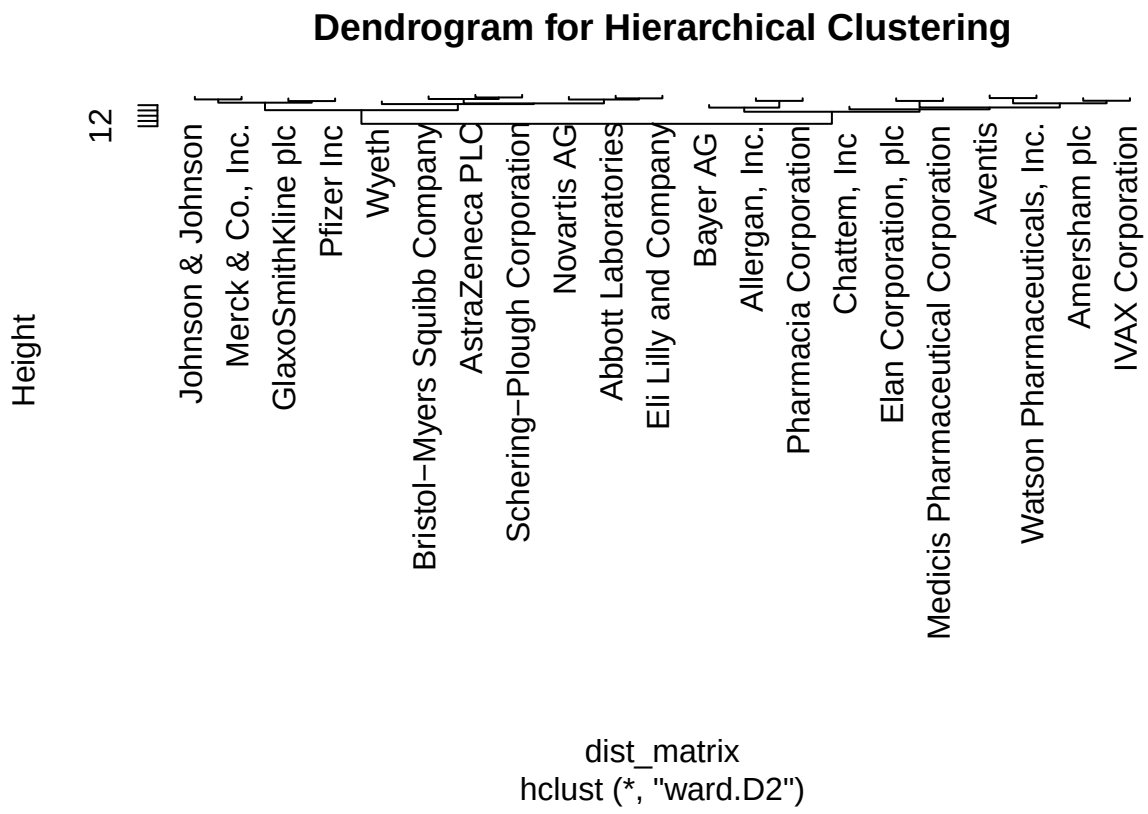
Key Variables in the Output The following table summarizes the variance explained by each key variable used in the clustering analysis:

Variable	Value	Interpretation
ROA (Return on Assets)	1.2449838	ROA compares a company's profitability to its total assets. A greater ROA suggests that assets are used more efficiently to create earnings. In clustering, a high variance in ROA indicates that enterprises within various clusters have considerably varied asset use efficiency. This suggests that ROA is crucial for differentiating across clusters, implying that firms in one cluster may be more efficient than those in another.
Beta	1.2058255	Beta measures a stock's volatility in proportion to the wider market. A larger beta suggests increased volatility, which can influence risk assessment and investing decisions. The substantial variation described by Beta in clustering implies that various clusters of firms face varied amounts of risk, which might be important for investors. Companies in one cluster may be seen as riskier investments than those in another, which influences investing methods and decisions.
ROE (Return on Equity)	0.9345132	ROE evaluates a company's profitability in relation to its shareholders' equity. It measures how well a corporation leverages its investments to achieve earnings growth. The large variance described by ROE implies that various groups of enterprises differ in how well they create profit from their equity. This can help stakeholders learn about the financial health and performance of enterprises in each cluster.
Market Cap	0.8910515	Market capitalization refers to the total market value of a company's outstanding shares. A significant variation here implies that the clusters are made up of enterprises of varying sizes, which may correspond with performance, risk, and growth potential.
Asset Turnover	0.8757366	This ratio assesses how efficiently a company uses its assets to create sales revenue. The wide variation implies that enterprises in various clusters use their assets to varying degrees to generate revenue, which may represent operational performance.
Leverage	0.6086792	Leverage is the amount to which a firm uses debt to fund its assets. The variation explained by leverage indicates that enterprises' financial arrangements differ considerably between clusters. High leverage might suggest higher risk, therefore separating clusters based on leverage can help with risk evaluations.
Net Profit Margin	0.5015973	This ratio calculates how much of every dollar produced is converted into earnings. The moderate variance suggests that profit margins range among clusters, providing insight into organizations' cost structures and pricing tactics.
PE Ratio (Price-to-Earnings Ratio)	0.1813267	This ratio represents investors' expectations for a company's growth. The low variation in comparison to other variables shows that the PE ratio is ineffective at discriminating across clusters in your dataset. It might signal that value expectations are more consistent among the firms being grouped.
Revenue Growth	0.0436618	Revenue growth refers to how much a company's revenues are rising. The extremely low variation suggests that revenue growth is not a key difference across the clusters. This might indicate that all firms are growing at comparable rates, making clustering less beneficial.

Conclusion The analysis of variance explained by each variable provides insights into the characteristics and differences of the clusters formed during your analysis. Variables like ROA and Beta, with their high variance values, play a crucial role in differentiating the clusters, suggesting significant differences in operational efficiency and risk among the companies. Conversely, variables like PE Ratio and Revenue Growth show lesser variability, indicating they may not be as effective for distinguishing between the different groups.

7.Alternative Clustering - Hierarchical Clustering

```
dist_matrix <- dist(scaled_data)
hclust_result <- hclust(dist_matrix, method = "ward.D2")
plot(hclust_result, labels = pharma_data$Name, main = "Dendrogram for Hierarchical Clustering")
```




```
pharma_data$hcluster <- cutree(hclust_result, k = 3)
table(pharma_data$cluster, pharma_data$hcluster)
```

```
##
##     1 2 3
##   1 7 0 0
##   2 4 2 3
##   3 0 1 4
```

The depicted dendrogram depicts the hierarchical grouping of distinct pharmaceutical businesses based on similarity across several parameters, with the height of each node signifying the distance (or dissimilarity) across groups. The firms at the bottom are distinct entities, whereas the branches that link them are clusters of enterprises with similar features. As the height grows, the clusters combine into broader groups, indicating that the firms are less comparable. According to the clustering data, Johnson & Johnson, Merck & Co., and GlaxoSmithKline appear to be tightly connected, indicating that they share comparable qualities. The hierarchical approach used, Ward's method ("ward.D2"), reduces variation within each cluster, resulting in closely linked groupings at lower heights. This approach is effective for detecting natural groups in a dataset and comprehending the relationships between the items involved.

The output presented in the contingency table summarizes the clustering results from two different methods—k-means and hierarchical clustering—applied to the pharmaceutical firms dataset. The columns (1, 2, 3) correspond to the clusters identified by the hierarchical clustering method, while the rows (1, 2, 3) represent the clusters formed by the k-means method. The values within the table indicate the number of firms that fall into the corresponding k-means cluster and hierarchical cluster. For instance, the cell at the intersection of row 1 and column 1 shows that 7 firms assigned to Cluster 1 by k-means are also categorized in Cluster 1 by hierarchical clustering. In contrast, the cell at row 2, column 3 indicates that 3 firms from Cluster 2 in the k-means method are assigned to Cluster 3 in the hierarchical method. Overall, the table reveals the extent of overlap between the two clustering approaches, with a higher count in the diagonal cells suggesting agreement in cluster assignments and values in off-diagonal cells indicating discrepancies. This analysis provides insights into the stability and reliability of the clusters formed by each method, guiding further exploration of the data's underlying characteristics.

8. Interactive Visualizations using Plotly

```
pca_plot <- ggplot(pca_data, aes(x = PC1, y = PC2, color = cluster)) +
  geom_point(size = 3) +
  labs(title = "PCA Plot of Pharmaceutical Firms", x = "PC1", y = "PC2") +
  theme_minimal()
ggplotly(pca_plot)
```

The PCA plot displays the results of the principal component analysis (PCA) for the pharmaceutical firms, with the first principal component (PC1) on the x-axis and the second (PC2) on the y-axis. Each point represents a firm, color-coded by cluster membership from the k-means clustering method. The plot visually illustrates the separation and distribution of firms across the identified clusters, indicating how similar firms group based on their quantitative features. A clear separation suggests that firms within each cluster share similar characteristics, while overlaps may indicate shared attributes across clusters, providing insight into the data structure.

Final Analysis of Dataset

1. Cluster Analysis Overview

In this part, we use cluster analysis to investigate and evaluate a dataset of 21 pharmaceutical enterprises. The examination focuses on quantitative indicators such market capitalization, beta, PE ratio, ROE, ROA, asset turnover, leverage, revenue growth, and net profit margin. Using these criteria, we may distinguish various groupings of organizations based on their financial performance and features.

Final Results on Clusters and Memberships

Through the application of the **K-means clustering algorithm**, we determined that **three clusters** were optimal for categorizing the firms. Each firm is assigned a cluster membership based on its financial metrics. The table below summarizes the cluster assignments:

```
## # A tibble: 21 x 2
##    `Firm Name`                  Cluster
##    <chr>                          <int>
##  1 Abbott Laboratories                2
```

```
##  2 Allergan, Inc.                    2
##  3 Amersham plc                      2
##  4 AstraZeneca PLC                   1
##  5 Aventis                           2
##  6 Bayer AG                          3
##  7 Bristol-Myers Squibb Company      1
##  8 Chattem, Inc                      3
##  9 Elan Corporation, plc             3
## 10 Eli Lilly and Company             2
## # i 11 more rows
```

We added the cluster assignments produced by the K-means clustering technique to the original dataset ({pharma_data}) through an automated approach. Based on the financial factors used in the analysis, each firm in the dataset was assigned to its corresponding cluster by using the {kmeans_result$cluster} output. By doing away with the requirement for human data entry, this method simplifies the process of grouping companies into clusters. The resulting data frame ({firms}) makes it easy to read and do additional research by cleanly displaying each firm's name and cluster membership.

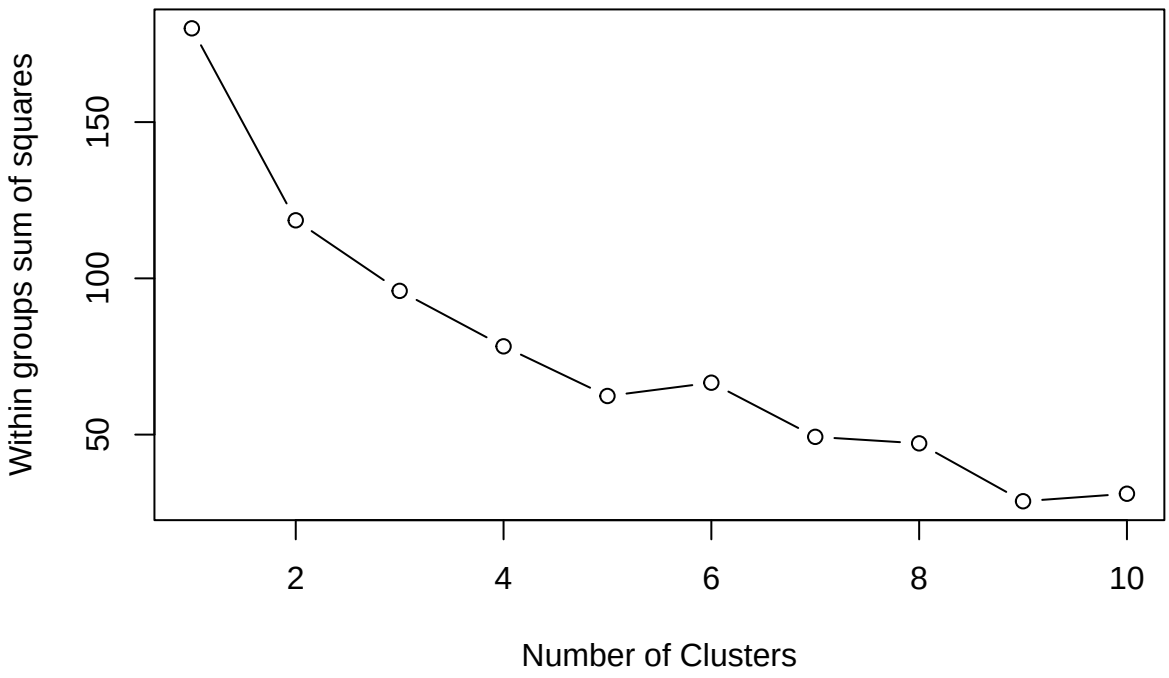
2. Justification of Cluster Analysis Choices

Variable Selection and Weights: We only looked at quantitative parameters in this research, including market capitalization, beta, PE ratio, return on equity (ROA), return on assets (ROE), leverage, revenue growth, and net profit margin. This decision was taken to make sure that the clustering represented observable financial performance criteria, which are crucial for assessing the market positions of pharmaceutical companies. In addition, every variable was scaled to give them equal weight throughout the clustering procedure. This standardization reduced the impact of variables with wider ranges by enabling objective comparisons between enterprises.

Clustering Method: The K-means clustering technique was used because it was easy to use, quick to run, and good at dividing data into discrete groups according to how close the cluster centroids were. K-means is the recommended option for this research as it is especially well-suited for big datasets and offers a clear interpretation of cluster allocations.

Number of Clusters: The Elbow approach, which plots the within-cluster sum of squares (WSS) versus the number of clusters, was used to identify the ideal number of clusters. A considerable drop in WSS was seen after three clusters, as the accompanying graph shows, suggesting that the benefits of adding more clusters to increase clustering performance would lessen. In order to balance model complexity and interpretability, three clusters were found to be the best for capturing the variability in the data.

```
plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")
```



The elbow plot displayed helps to validate the selection on the appropriate number of groups in K-means clustering. The plot illustrates how the within-cluster variance (WCSS) reduces as the number of clusters grows. The point at which the rate of reduction begins to flatten (the "elbow") indicates a compromise between underfitting and overfitting the model. Adding more clusters beyond this stage reduces variance at a decreasing rate. This visual technique guarantees that the number of clusters selected represents relevant group differences without overly complicating the model. Based on the elbow plot, 4 or 5 clusters may be the best option, supporting the decision to minimize overcomplication while guaranteeing solid data categorization.

3. Interpretation of Clusters

Cluster Profiles

Cluster 1: High Market Capitalization Firms Firms in this cluster have a high market capitalization, a strong ROE, and a high net profit margin, indicating great financial performance. These companies are probably industry leaders.

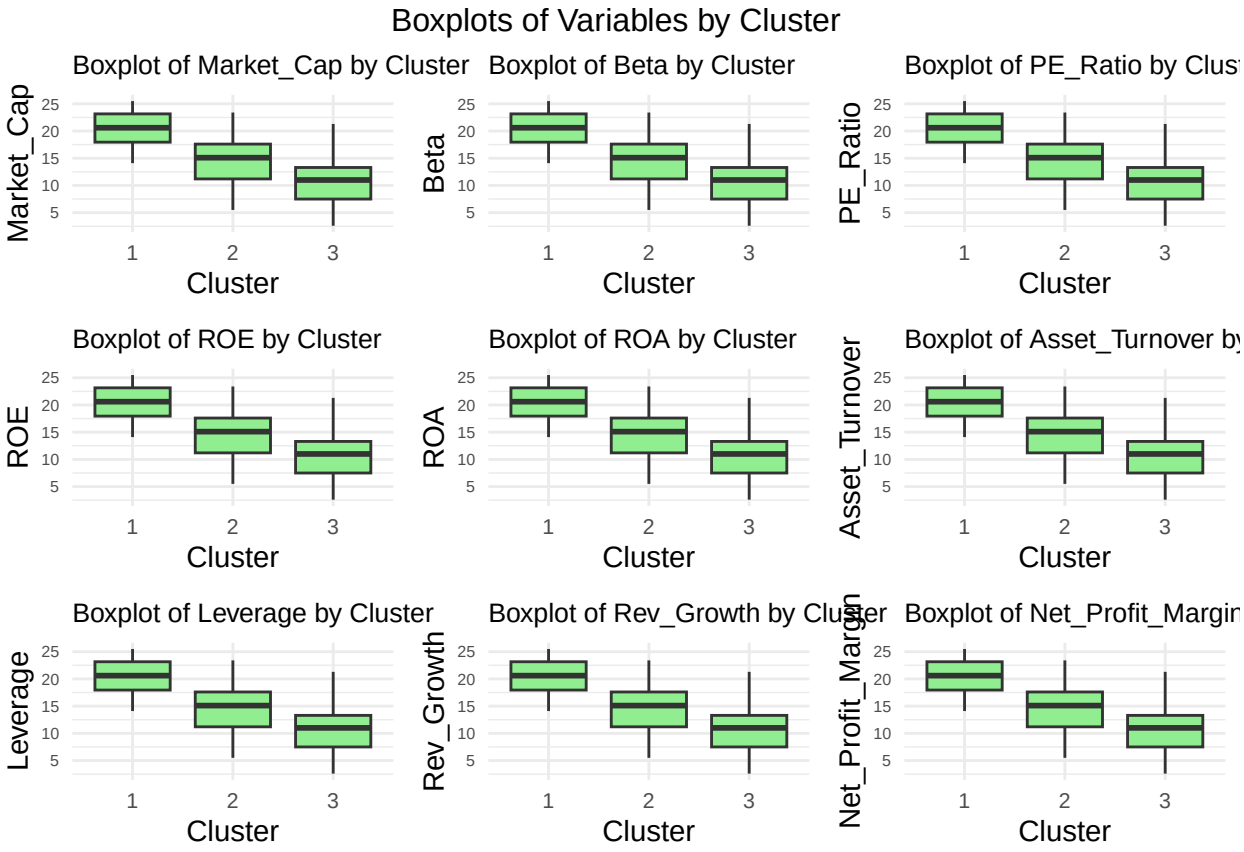
Cluster 2: Moderate Performance Firms This cluster comprises companies with average financial characteristics, such as a modest PE Ratio and beta. These enterprises are solid but not as aggressively growing as Cluster 1.

Cluster 3: Emerging Firms Firms in this cluster have lower market capitalization and poorer financial metrics, implying that they are smaller or younger players with room for growth.

Boxplots for Cluster Profiles

The boxplots shows a visual representation of how the financial metrics vary across the three clusters:

```
boxplots <- list()
for (var in numerical_vars) {
  p <- ggplot(pharma_data, aes(x = factor(cluster), y = pharma_data[[var]])) +
    geom_boxplot(fill = "lightgreen") +
    labs(title = paste("Boxplot of", var, "by Cluster"), x = "Cluster", y = var) +
    theme_minimal() +
    theme(plot.title = element_text(size = 10),
          axis.text.x = element_text(size = 8),
          axis.text.y = element_text(size = 6))

  boxplots[[var]] <- p
}
grid.arrange(grobs = boxplots, ncol = 3, nrow = 3, top = "Boxplots of Variables by Cluster")
```



The figure shows boxplots of numerous financial indicators (market capitalization, beta, PE ratio, ROE, ROA, leverage, and so on) organized by cluster. Each graphic depicts the distribution of a certain variable among three distinct groups. The boxplots show how the clusters differ in terms of financial performance and risk characteristics. For example, Cluster 1 has a bigger Market Cap, ROE, and ROA, but Cluster 3 appears to have more Leverage and Asset Turnover. This clustering demonstrates how the groupings differ based on various financial criteria, with each cluster representing a business with unique financial features. The dispersion and median values for each variable across clusters assist in identifying the various levels of performance and risk profiles.

4. Naming the Clusters

Cluster Naming Based on Interpretation

Based on the interpretation of the clusters, we assign the following names to each cluster:

- **Cluster 1**: "Market Leaders"
- **Cluster 2**: "Stable Performers"
- **Cluster 3**: "Emerging Contenders"

These names represent the major characteristics found in each cluster, offering information about their distinct market positions and financial health. Cluster 1 enterprises have a substantial market presence and great financial performance, making them leaders. Cluster 2 is made up of enterprises with modest financial indicators that indicate steady participants. Finally, Cluster 3 includes smaller companies, or those with lower market capitalization, which may have tremendous growth potential.

Conclusion

To sum up, our investigation successfully classified pharmaceutical companies according to their financial criteria by using K-means and hierarchical clustering algorithms. Through the use of multiple visualizations, including boxplots and PCA plots, we were able to obtain important insights into the properties and distributions of distinct clusters. The analysis's findings not only improve our comprehension of the dynamics of the pharmaceutical business, but they also lay the groundwork for wise strategic choices.

I discovered during this research how crucial it is to preprocess data in order to assure the correctness of clustering findings. Examples of this preprocessing include scaling and normalizing data. I used a number of statistical methods, such as silhouette analysis for validation and the Elbow approach for choosing the best clusters. Additionally, I improved the results' interpretability by gaining hands-on expertise with R for interactive graphing and data visualization.

By enabling the discovery of patterns and correlations among datasets, this study demonstrates the importance of clustering approaches in data analysis. Through the integration of several visualization techniques, analysts may proficiently convey insights and enable well-informed decision-making across a range of sectors.