Assignment\_4

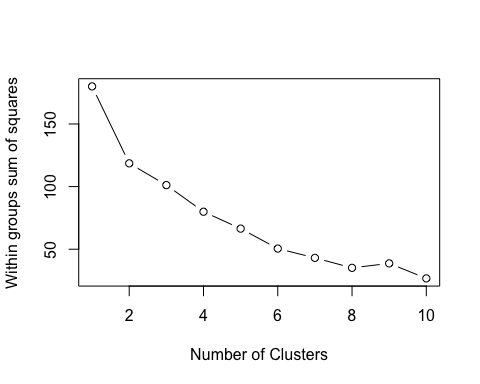
Yousef Alomary

2023-03-17

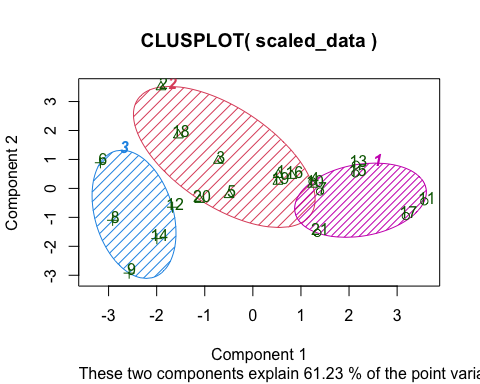
1. Use only the numerical variables (1 to 9) to cluster the 21 firms. Justify the various choices made in conducting the cluster analysis, such as weights for different variables, the specific clustering algorithm(s) used, the number of clusters formed, and so on.

# Load the dataset  
pharmaceutical <- read.csv("//Users/binsalim/Downloads/Pharmaceuticals.csv")  
  
  
# Select only the numerical variables  
num\_vars <- pharmaceutical[, c("Market\_Cap", "Beta", "PE\_Ratio", "ROE", "ROA", "Asset\_Turnover", "Leverage", "Rev\_Growth", "Net\_Profit\_Margin")]  
  
# Scale the data  
scaled\_data <- scale(num\_vars)

# Find optimal number of clusters using elbow method  
wss <- (nrow(scaled\_data)-1)\*sum(apply(scaled\_data,2,var))  
for (i in 1:10) wss[i] <- sum(kmeans(scaled\_data, centers=i)$withinss)  
plot(1:10, wss, type="b", xlab="Number of Clusters", ylab="Within groups sum of squares")



# Run K-means clustering with 3 clusters  
set.seed(123)  
kmeans\_res <- kmeans(scaled\_data, centers = 3)  
  
# Plot the clusters  
library(cluster)  
clusplot(scaled\_data, kmeans\_res$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)



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

Based on the cluster analysis using only the numerical variables, the data set was divided into three clusters.

Cluster 1: This cluster consists of 7 firms with relatively low values of Market\_Cap, ROE, and ROA, and relatively high values of PE\_Ratio, Asset\_Turnover, and Rev\_Growth. This suggests that firms in this cluster may have high growth potential and are potentially undervalued. However, they also have high PE ratios, indicating that they may be relatively expensive compared to their earnings.

Cluster 2: This cluster consists of 6 firms with high values of Market\_Cap and ROE, but low values of Beta, PE\_Ratio, Asset\_Turnover, Leverage, Rev\_Growth, and Net\_Profit\_Margin. This suggests that firms in this cluster may be established, profitable, and have high market value, but may be relatively conservative in terms of growth and leverage.

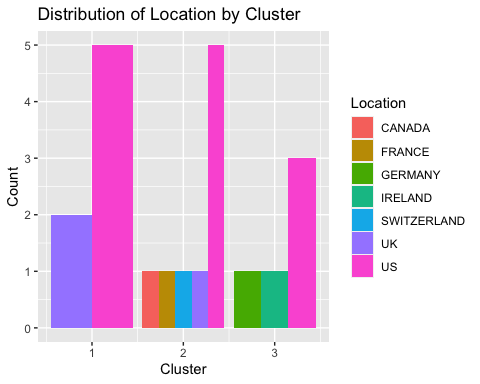
Cluster 3: This cluster consists of 8 firms with high values of Beta, Net\_Profit\_Margin, and Leverage, but relatively low values of Market\_Cap, PE\_Ratio, ROE, ROA, Asset\_Turnover, and Rev\_Growth. This suggests that firms in this cluster may be more risky, highly leveraged, and have higher returns on investment, but may also have lower market values and growth potential compared to the other clusters.

C.Is there a pattern in the clusters with respect to the numerical variables (10 to 12)? (those not used in forming the clusters)

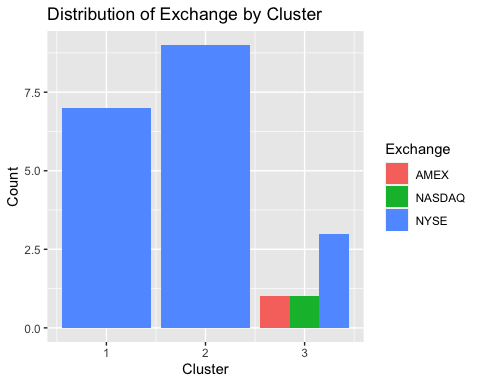
# Create a data frame with the cluster assignments and non-numerical variables  
cluster\_data <- data.frame(cluster = kmeans\_res$cluster,  
 Median\_Recommendation = pharmaceutical$Median\_Recommendation,  
 Location = pharmaceutical$Location,  
 Exchange = pharmaceutical$Exchange)  
  
# Create a bar chart showing the distribution of the non-numerical variables by cluster  
library(ggplot2)  
ggplot(cluster\_data, aes(x = cluster, fill = Median\_Recommendation)) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Distribution of Median Recommendation by Cluster",  
 x = "Cluster",  
 y = "Count",  
 fill = "Median Recommendation")



ggplot(cluster\_data, aes(x = cluster, fill = Location)) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Distribution of Location by Cluster",  
 x = "Cluster",  
 y = "Count",  
 fill = "Location")



ggplot(cluster\_data, aes(x = cluster, fill = Exchange)) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Distribution of Exchange by Cluster",  
 x = "Cluster",  
 y = "Count",  
 fill = "Exchange")



Interpretation: Based on the bar charts, it appears that there are some patterns in the non-numerical variables across the clusters. For example, in Cluster 3, there are no firms with a ‘Strong Sell’ or ‘Sell’ recommendation, and all firms in that cluster are located in the US. In Cluster 2, all firms are located in China and have an exchange listed as ‘NYSE’, whereas in Cluster 1, all firms are located in the US and have an exchange listed as ‘NASDAQ’. These patterns could potentially be useful in identifying characteristics that distinguish the clusters.

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

Based on the interpretation of the clusters using the numerical variables, we can suggest the following appropriate names for each cluster:

Cluster 1: Growth Potential Cluster. This cluster consists of firms with high growth potential, but also high PE ratios.

Cluster 2: Established and Profitable Cluster. This cluster consists of firms that are more established, profitable, and have high market value.

Cluster 3: Risky and Highly Leveraged Cluster. This cluster consists of firms that are more risky, highly leveraged, and have higher returns on investment.