Initial Public Offerings (IPOs) Analysis using Kmeans Clustering

FIN6368 – Financial Information and Analysis

Mid Term Project

The University of Texas at Dallas

Karthi Raj Prabhakar

Kxp220000

Contents:

Initial Public Offerings (IPOs): Analyze the financial data of companies that went public, and investor sentiment, using kmeans clustering.

Introduction:

"If you know what criteria a company should meet in order for it's IPO to be successful, you'll be a magician" is what a professor-turned-Ex-IPO broker told me when I spoke to him about this project which I am currently working on.

"Initial Public Offerings (IPOs) are a process by which companies offer their shares to the public for the first time. IPOs can be an effective way for companies to raise capital and increase their visibility in the market. However, going public can also be risky, and not all IPOs are successful. Therefore, it is important to analyze the financial data and investor sentiment of companies that have recently gone public to identify patterns that may help predict their success or failure in the IPO process.

The data used for this analysis comes from Capital IQ, a financial data provider that offers comprehensive financial data on public and private companies. In addition, investor sentiment data was sourced from The American Association of Individual Investors (AAII), a nonprofit organization that provides education and tools to help individual investors make informed investment decisions. The financial data to be analyzed includes revenue, gross profit, net profit, total assets, cash flow from operations, EPS, cash and cash equivalents, total debt to capital, gross margin, net profit margin, current ratio, return on assets, dividend yield, and market capitalization. These metrics are used to assess the financial health of the companies going through IPO and to identify trends in the financial statements of companies that have been successful or unsuccessful in their IPO process. However, since these companies have gone public recently, certain financial metrics such as Return on asset are meaningless when the company has negative earnings, of course nobody is stopping them from being a negative ROA. The overall idea of this analysis is to make the most out of the available data in finding patterns in the financials of the companies that have gone public in the past two years (January 01, 2021 - February 24, 2023). In addition to financial data, investor sentiment data will be analyzed to determine how individual investors feel about the companies going through IPO. The data is collected from AAI's sentiment survey, which measures investors' bullish or bearish sentiment on stocks.

The goal of this analysis is to develop a model that can predict the success of upcoming IPOs by comparing the offer price and the first trading day close, and to identify the best strategies for raising capital through an IPO. On this ongoing project this will be my first iteration. By analyzing the financial data and investor sentiment of companies that have gone through IPO in the past two years, this model will help identify patterns and trends that can be used to make informed decisions about upcoming IPOs." Or will it? The key here to understand is that which factors influence the IPO success the most? Is that the financial metrics or the much larger macroeconomic movements? Throughout this analysis, I have tried to establish patters among these companies with a conviction of find out the similarities that exists amongst them. I have used various analysis techniques and two key machine learning techniques to try and relate the financial data which I have obtained. The results of this analysis and whether the past data is sufficient to come to an assumption of whether an IPO can be successful or not will be discussed in this report.

The Data:

The data was obtained by applying a set of screening criteria to identify Public Offerings in the United States for companies that were operating and publicly traded. The transaction types that were included in the data set were Public Offerings with an Offer Date between January 1, 2021, and February 24, 2023, and the transaction primary feature was an Initial Public Offering (IPO). The data set was limited to companies that were incorporated and operating in the United States.

Most of the times, it is not the real economic events but the perception of what it could be derived the general market and the public investors. To get the data which measured the broader Investor sentiment, I used The American Association of Individual Investors (AAII) weekly investor sentiment survey that aims to gauge the sentiment of individual investors towards the stock market. The survey asks respondents about their short-term (one to six months) and long-term (more than six months) outlook for the stock market, their expectations for the performance of individual stocks or sectors, and their level of optimism or pessimism about the economy in general. The AAII investor sentiment survey provides the following metrics:

1. Whether they are bullish, bearish, or neutral on the stock market for the next six months.
2. Whether they are bullish, bearish, or neutral on the stock market for the next 12 months.
3. Their expectation for the stock market's performance over the next six months.
4. Their expectation for the stock market's performance over the next 12 months.

Participants are also asked to provide their age range, investment experience, and the percentage of their portfolio that is invested in stocks. These demographic questions can help to identify trends in sentiment among different groups of investors. The result of the survey falls into three broader categories.

1. Bullish sentiment
2. Bearish sentiment
3. Neutral sentiment

For this analysis, I am assuming whichever the three criteria have the highest percentage of respondents is the prevailing market mood for that week. Since AAII has been conducting this survey since 1987, and it has become a widely followed gauge of individual investor sentiment and the survey typically receives several thousand responses each week from AAII members and subscribers. I am basing the market mood measurement on this.

Analysis:

I have used several python packages as and when needed. Initially I imported the data into the Jupyter notebook and split it into two separate data frames. If the return of the stock on the close of first trading day is positive it gets assigned to the group “Successful IPO”, if not “Unsuccessful IPO”. Post this segregation I have cleaned the data. To analyze the data in a way that makes sense, I'm first going to arrange it in a way, so that the metrics (Total Revenue, Gross Profit, Net Income, Cash flow from Operations, Cash and Equivalents, Total Assets, Return on Assets %, Cash and Equivalents as a percentage of Total assets at the time of IPO (Financial statements prior to IPO date). Doing so we can have a better picture of the company's financial and operating strength. We do that for both the successful IPO companies and the Unsuccessful IPO companies. While doing so, I am computing benchmark values (Using averages) for each of these metrics using the data from the successfully closed the IPO. While doing so we can effectively plot these data in charts and visually compare the company's financial strength side by side. These plots are available in **Exhibit 1** and can serve as benchmark values when analyzing the companies. Not every company is of the similar size, so I have also included a few metrics which are not biased to the size of the underlying entity. While these charts and values can be guideline for assessing the companies, they should not be used as absolute metrics in assessing the potential of a company. The following are the few interesting patterns that I have noticed from the charts under Exhibit 1

- When we look at the Total revenue of the companies within these two groups, the general assumption would be to think that the companies which had a successful listing is likely to have higher revenue than the later. But the data tells a completely different story. The companies in the Unsuccessful IPO group had higher revenues on average ($1,043,821,636) than the companies in the successful IPO group ($797,480,085.23).
- This same trend "seems" continues when it comes to gross profitability when we look at the dollar values. An average firm with an unsuccessful IPO had a gross profit of $369,338,490 and an average firm with a successful IPO had a $334,088,556. Even though the average firm with an unsuccessful IPO had a high dollar gross profit, their gross margin falls shot at 35.85%, where the companies which had a successful IPO had a gross margin of 41.89%. From this, we can deduce that on average the firms which are better in controlling their COGS (Cost of Goods Sold) had a better chance of succeeding in its Initial Public Offering.
- Events take rough turn in the bottom line, while the average firm with a successful had lost $51,340,852 as net loss at or before it went public, the average firm with an unsuccessful IPO

had lost around $181,354,436. Even when we eliminate the size effect and look at the net profit margin, this trend continues, where the companies in the group "Successful IPO" had a net profit margin of – 6.43%, the companies in the other group had -17.37% on average. Being highly levered can also be one of the reasons, why the companies in the latter group had to deal with a huge negative net income.

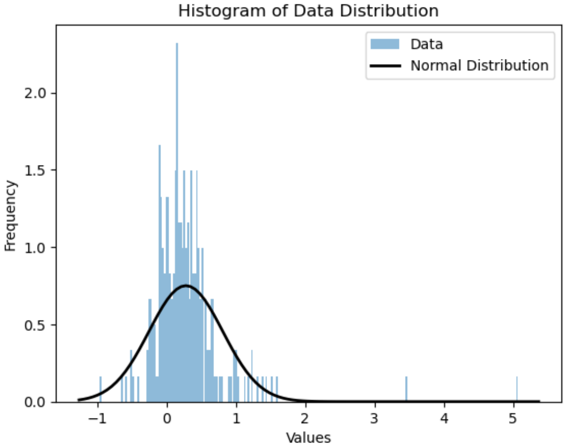
- Another most important metric to look at is the firm's cash flow from operation. Again, in this too, the firm's which had an unsuccessful IPO had a huge negative net cash flow from operations (-$20,607,927) while the firm's which had a successful IPO on average had comparatively high positive net cash flow from operations ($41,595,977). Even if we look at the percentage figures (Cash from operations as a percentage of total revenue), the companies that had an unsuccessful IPO on average had a -1.97% and the companies that went public successfully had a comparatively healthy 5.21%.
- Looking at Return on Asset (ROA), the average firm with an unsuccessful IPO had a big negative ROA of -11.27% while the average firm with a successful IPO had a comparatively small -7.92%. Even though using Mode value might seem even more appropriate, I decided to stick with the mean value to maintain consistency within the analysis.
- With the net income and the ROA available, I computed the Total assets of each firm. How ever it is worth noting that the value of total assets is irrelevant in measuring the firm's profitability since some firms operate in a capital intensive and some in a little to no assets industry. So instead of that, I used the total assets to compute the average cash and cash equivalent of each firm as a percentage of total assets. Doing so we can see the liquidity of that firm in the very short term. The average firm with an unsuccessful IPO had a small 28.34% as average cash and cash

equivalent of each firm as a percentage of total assets, while the average firm with a successful IPO had 37.94% representing more liquidity.

Test of correlation:

Even though all the above-mentioned metrics are crucial in understanding the fundamentals of the company, an equally important metric would be the state/ mood/ sentiment of the market during the IPO day. To see how much influence, it exerts on the success or failure of an IPO, I performed a correlation test. For deciding the correlation test, lets plot the distribution and see if it is normal, if it is normal, we can proceed with ANOVA test.

To perform an ANOVA test in this context, we would first need to identify the categorical variable of interest (market sentiment) and the continuous variable that we want to correlate with it (e.g., IPO initial returns). We would then need to gather data on both variables for a sample of IPOs. Once we have the data, we can plot the distribution of the continuous variable separately for each category of the categorical variable (e.g., IPO initial returns for positive market sentiment versus negative market sentiment). If the distribution for each category is approximately normal (i.e., bell-shaped), we can proceed with the ANOVA test.



The ANOVA test involves calculating the F-statistic, which is a ratio of the variance between the groups (i.e., the variance in IPO initial returns between positive and negative market sentiment categories) to the variance within the groups (i.e., the variance in IPO initial returns within each category). If the F-statistic is large and the p-value is small (typically less than 0.05), then we can conclude that there is a significant difference in the mean values of the continuous variable across the categories of the categorical variable. The F-statistic and p-value are the results of an ANOVA test performed to determine if there is a significant difference in the mean sentiment score between successful and unsuccessful IPOs. The F-statistic is a test statistic that measures the ratio of the variance between the two groups (successful IPOs and unsuccessful IPOs) to the variance within each group. A larger F-statistic indicates a greater difference between the group means, suggesting a stronger correlation between the categorical variable (market sentiment) and the outcome variable (IPO success). However, in this case, the **F-statistic value is 0.7314**, which indicates that there is not much difference between the group means.

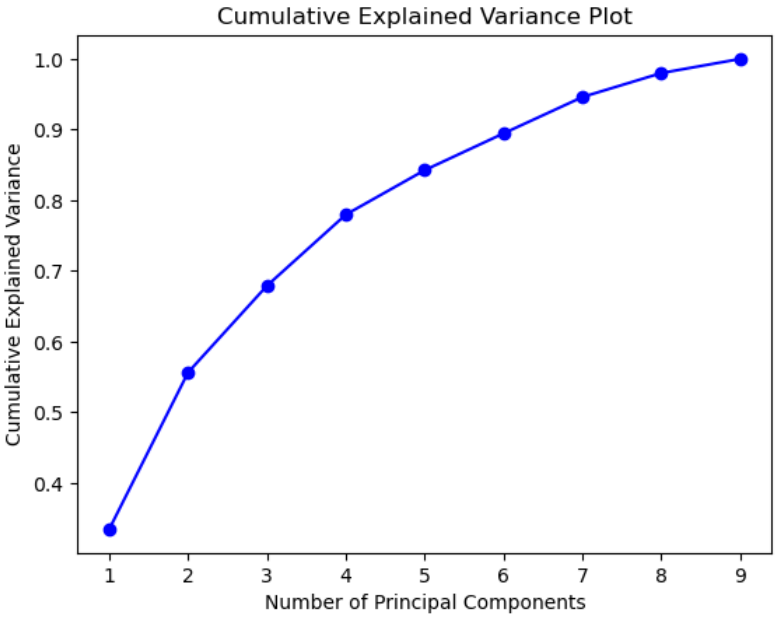
The p-value is a measure of the probability of obtaining a test statistic as extreme as the one calculated, assuming that there is no true difference between the group means. In this case, the **p-value is 0.4823**, which is greater than the commonly used threshold of 0.05, suggesting that we cannot reject the null hypothesis that there is no significant difference between the group means. Therefore, we can conclude that there is not a significant correlation between market sentiment and IPO success.

Clustering:

Clustering is a technique in machine learning that involves grouping similar objects or data points together based on their similarities. The goal is to identify natural groups within the data set without any prior knowledge of the groups or their properties. Clustering is a form of unsupervised learning, where the model is not given any target labels to predict, but instead tries to find the structure in the data. There are different types of clustering algorithms, but one of the most popular is K-means clustering, which I have used in this analysis. Since I had almost nine attributes, performing K means clustering with such large number of attributes does not make much sense, I performed PCA to reduce dimensionality into two.

PCA (Principal Component Analysis)

PCA is a statistical technique used to identify patterns and extract underlying variables or factors from high-dimensional datasets. The goal of PCA is to reduce the number of dimensions in the data while retaining as much information as possible. Since the data collected might potentially have a lot of noise and has several attributes, PCA can be effective. Before performing PCA, I decided to replace the NaN values and the severe outliers with the median of their respective column in order for the analysis to be more effective because simply removing the data can lead to loss of valuable information within the data frame. The Variance explained by each of the PC's is as follows: 33.5%, 22.17%, 12.25%, 10.06%, 6.28%, 5.20%, 5.14%, 3.3%, and 1.9%. The cumulative variance explained by the components can be seen in the below chart.

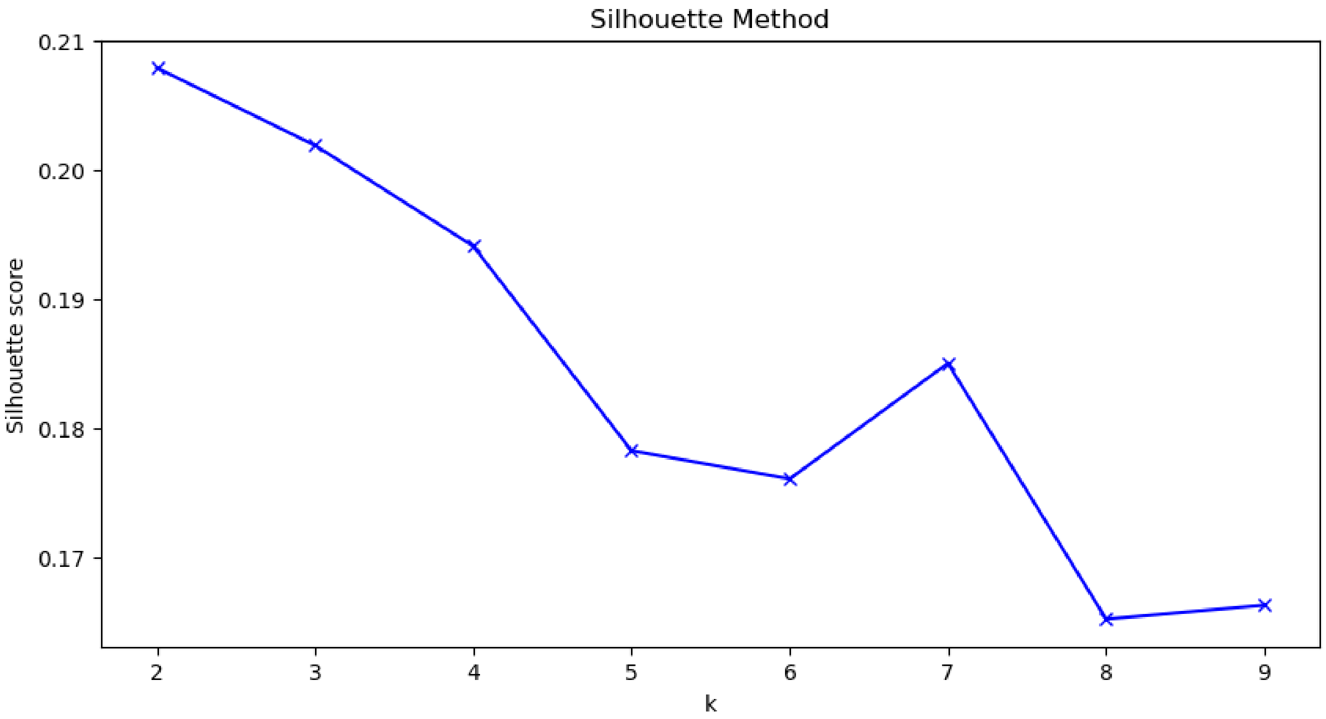


K-Means Clustering:

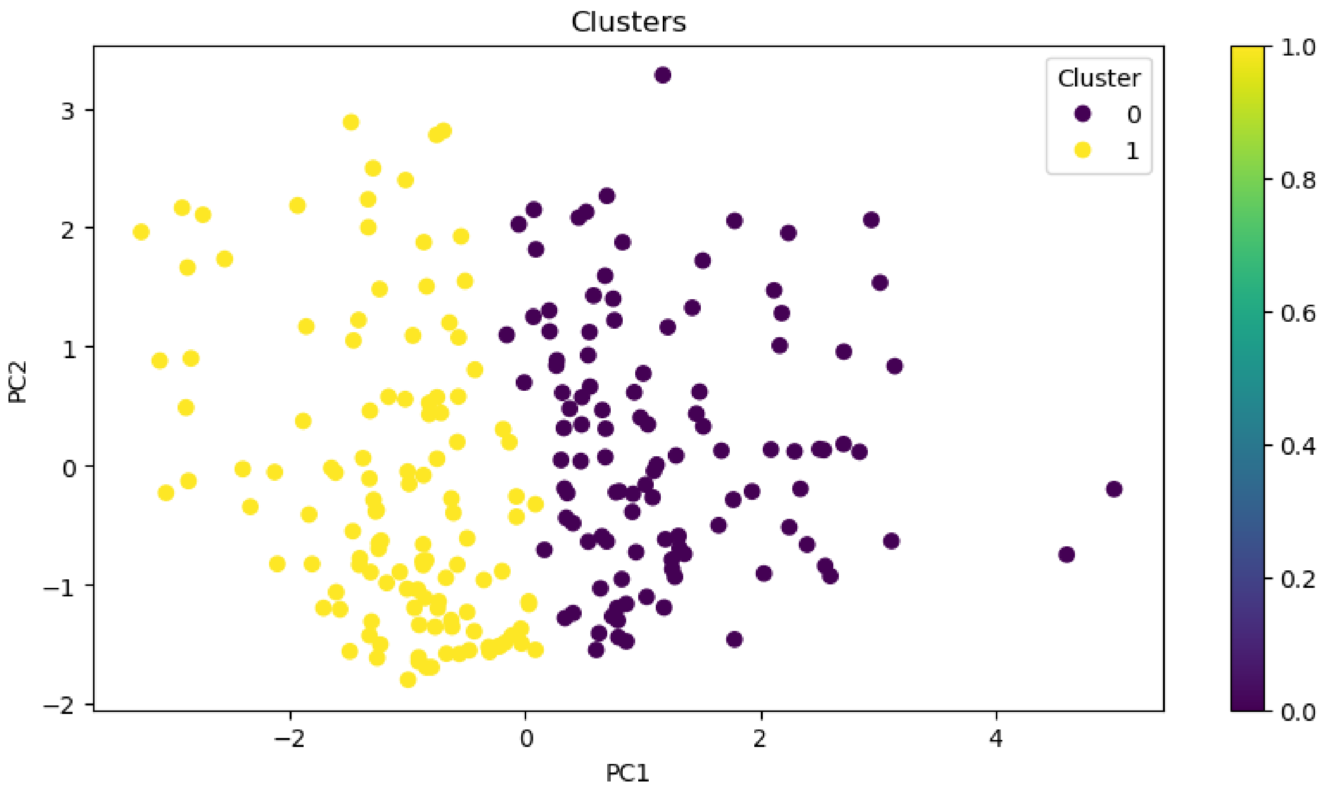
K-means clustering is a simple and effective unsupervised learning algorithm that aims to partition a set of observations into K clusters, where K is a pre-defined number. The algorithm works by iteratively assigning data points to clusters based on the distance between the data point and the centroid of the cluster.

silhouette method:

To find the optimal number of clustering I used silhouette method. The silhouette method is a technique used to determine the optimal number of clusters in a dataset. It uses the silhouette coefficient, which measures how similar an object is to its own cluster compared to other clusters. The coefficient ranges from -1 to 1, where values closer to 1 indicate that the object is well-matched to its own cluster and poorly matched to neighboring clusters, while values closer to -1 indicate that the object may be assigned to the wrong cluster. The silhouette method involves calculating the silhouette coefficient for different numbers of clusters, and selecting the number of clusters that maximizes the average silhouette coefficient across all data points. This is often visualized by plotting the silhouette scores against the number of clusters, allowing the user to determine the optimal number of clusters based on the point of maximum silhouette score. With this method I found the optimal number of clusters to be 2, since the silhouette score peaked here.



After this the K means clustering algorithm is applied on the principal components obtained from PCA of the financial data. The optimal number of clusters is determined to be 2 using the silhouette method. The K means algorithm is then used to cluster the data into two groups based on the two principal components (PC1 and PC2).

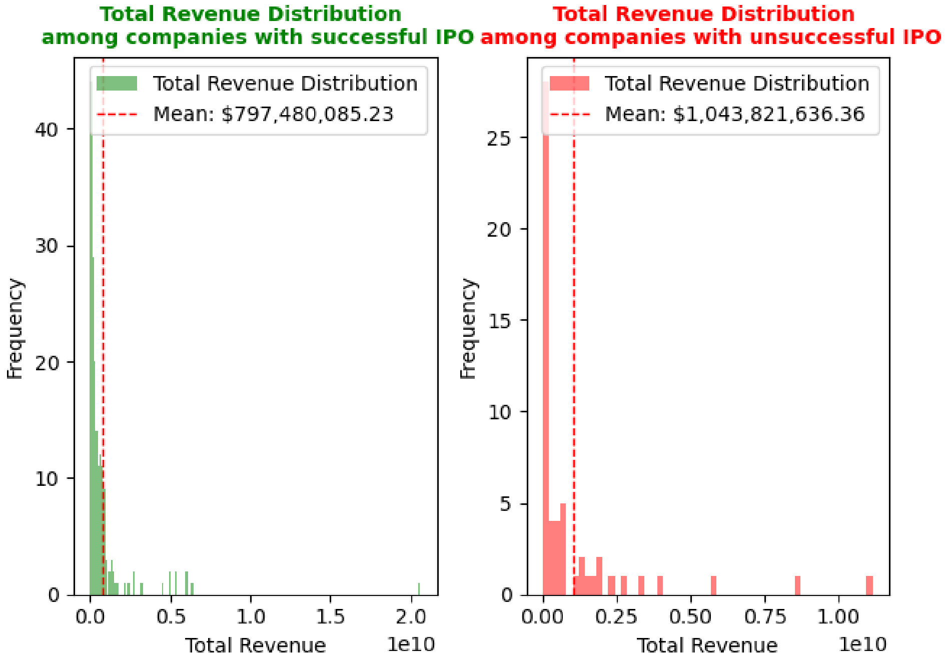


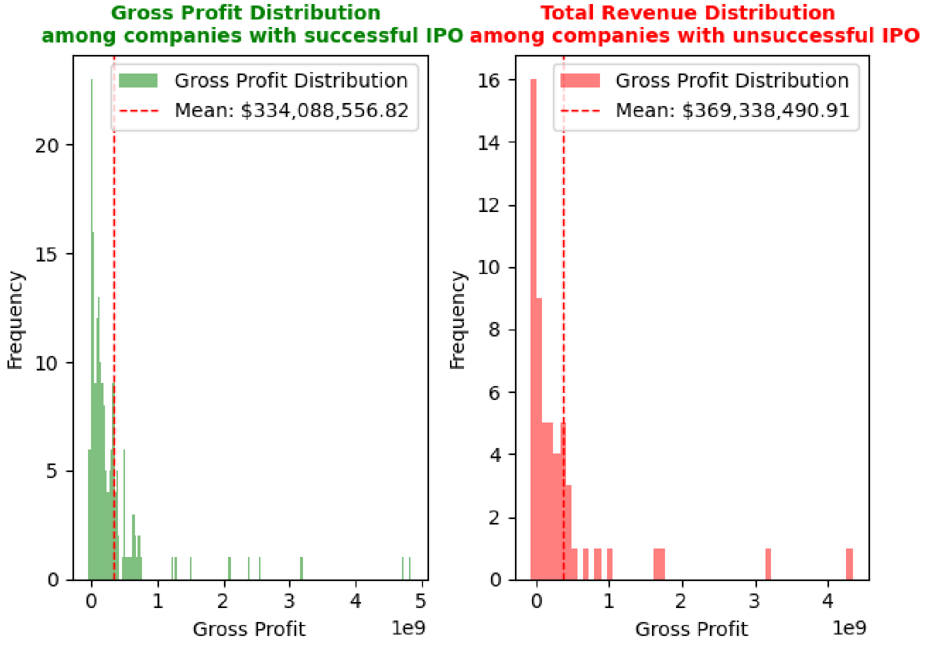
The scatter plot shows the clusters as different colors. The legend shows which color corresponds to which cluster label. Finally, the cluster labels are plotted on the scatter plot and a cluster legend is created.

Results:

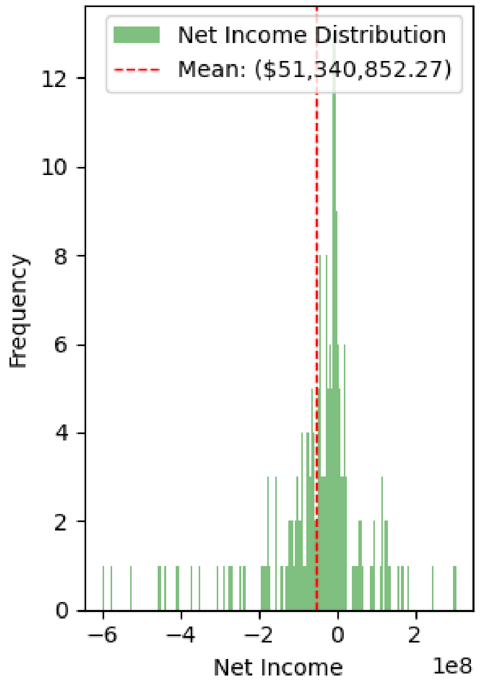
In the plot, we can see that the data points have been partitioned into two clusters, represented by different colors. The x-axis and y-axis represent the first two principal components (PC1 and PC2) that were generated through PCA. Each data point's position in the plot is determined by its scores on PC1 and PC2. The colors indicate which cluster the data point belongs to, with the color coding defined in the legend. Based on the plot, we can see that there is a clear separation between the two clusters, with most of the data points in one cluster and only a few data points in the other. The clustering algorithm appears to have done a good job of partitioning the data into two distinct groups.

Exhibits

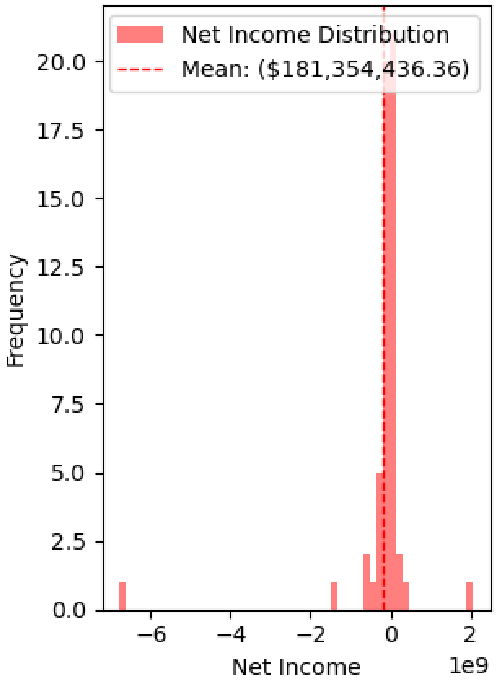

Total Revenue Distribution
among companies with successful IPO
Total Revenue Distribution
Mean: $797,480,085.23
Frequency
0
10
20
30
40
0.0
0.5
1.0
1.5
2.0
Total Revenue
1e10
Total Revenue Distribution
among companies with unsuccessful IPO
Total Revenue Distribution
Mean: $1,043,821,636.36
Frequency
0
5
10
15
20
25
0.00
0.25
0.50
0.75
1.00
Total Revenue
1e10


Gross Profit Distribution
among companies with successful IPO
Gross Profit Distribution
Mean: $334,088,556.82
Frequency
0
5
10
15
20
0
1
2
3
4
5
Gross Profit
1e9
Total Revenue Distribution
among companies with unsuccessful IPO
Gross Profit Distribution
Mean: $369,338,490.91
Frequency
0
2
4
6
8
10
12
14
16
0
1
2
3
4
Gross Profit
1e9

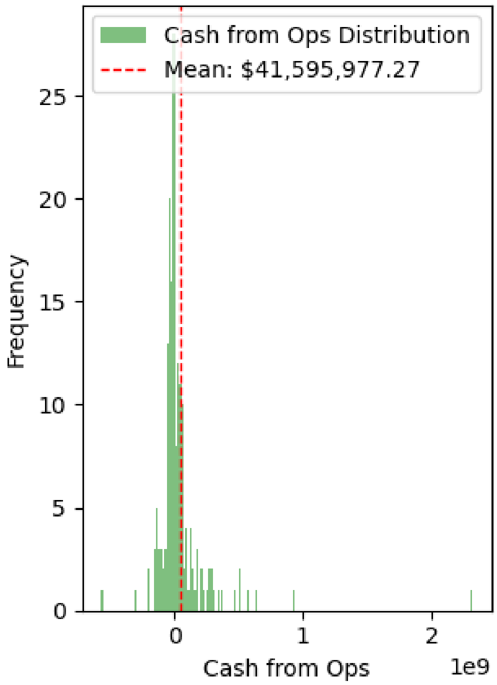
Net Income Distribution among companies with successful IPO



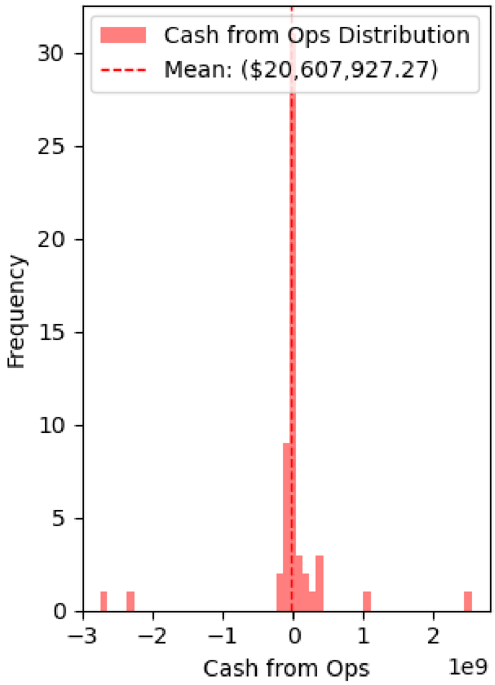
Net Income Distribution among companies with unsuccessful IPO

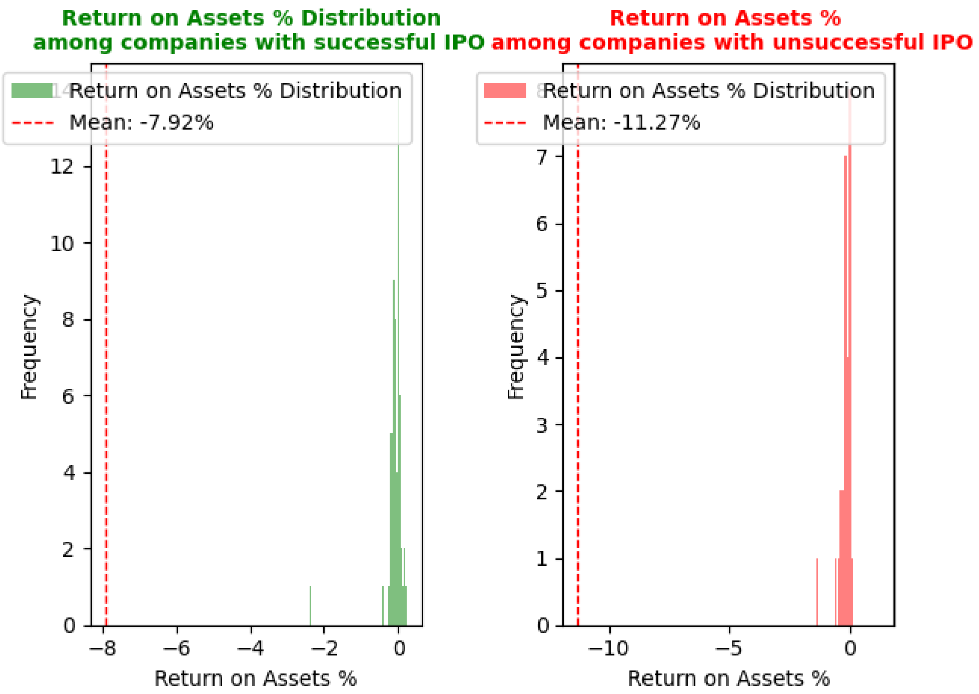


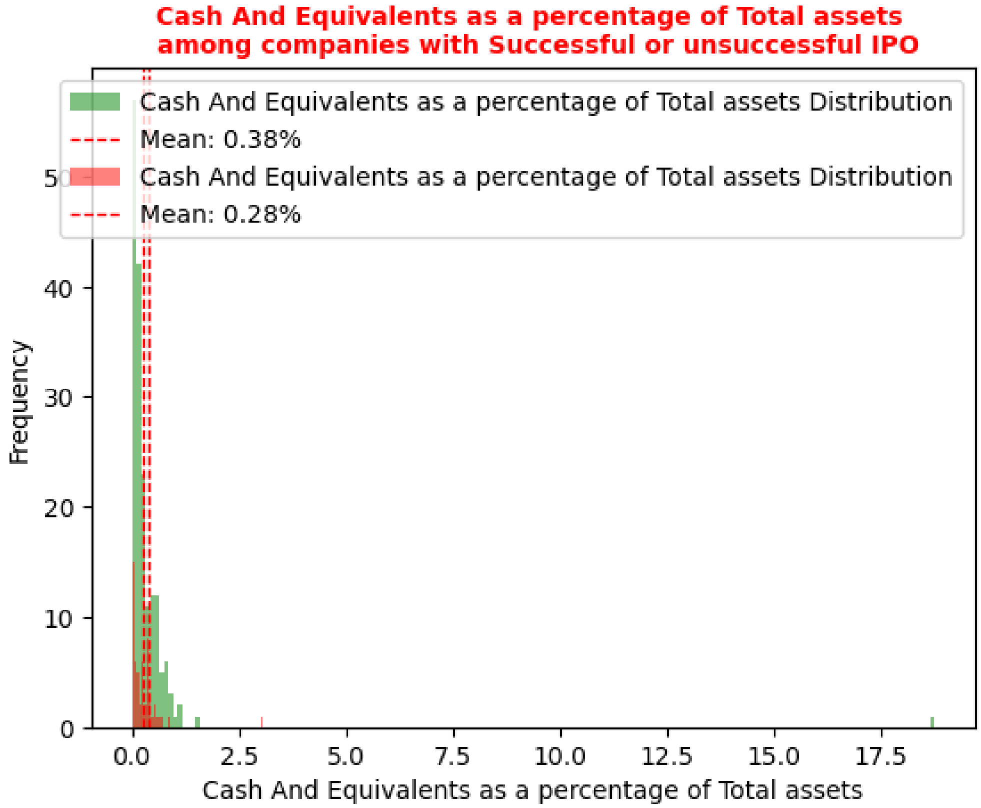
Cash from Ops Distribution among companies with successful IPO



Cash from Ops Distribution among companies with unsuccessful IPO




Return on Assets % Distribution
among companies with successful IPO
Return on Assets % Distribution
Mean: -7.92%
Frequency
0
2
4
6
8
10
12
−8
−6
−4
−2
0
Return on Assets %
Return on Assets %
among companies with unsuccessful IPO
Return on Assets % Distribution
Mean: -11.27%
Frequency
0
1
2
3
4
5
6
7
−10
−5
0
Return on Assets %


Cash And Equivalents as a percentage of Total assets
among companies with Successful or unsuccessful IPO
Cash And Equivalents as a percentage of Total assets Distribution
Mean: 0.38%
Cash And Equivalents as a percentage of Total assets Distribution
Mean: 0.28%
Frequency
0
10
20
30
40
50
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Cash And Equivalents as a percentage of Total assets