**Project Final Report: Clustering Analysis of the Stock Market**

[[GitHub - NirNagary/ML-final-project](https://github.com/NirNagary/ML-final-project)](https://github.com/NirNagary/ML-final-project)

**Anna Vardanyan[annakosh49@gmail.com], Nir Nagary[nir.nagary@gmail.com]**

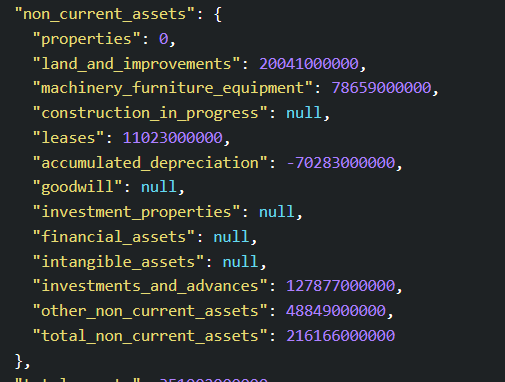
# Introduction

In the modern financial world, the Multitude of investment choices available can overwhelm even the most experienced investors. Among the countless trade activities on the Nasdaq Stock Exchange, New York Stock Exchange (NYSE), and over the counter (OTC) markets, discerning the financial viability and investment potential of over 10,000 publicly traded stocks is a Substantial task. The main question that confuses investors is how to recognize which companies are a good choice for long-term investment within the constantly changing market.  
This project seeks to simplify the complex dynamics of the stock market through application of machine learning clustering algorithms. Our research is motivated by the objective to categorize publicly traded companies into distinctive groups based on quantifiable measures of financial health. By using a comprehensive approach that combines companies' profiles information, historical income statements, balance sheet, and cash flow data into accessible clusters, we aim to provide investors with a powerful lens through which to assess investment potential.  
  
Our approach is implementing a variety of machine learning methods, including K-means, EM Algorithm, Hierarchical Clustering, and DBSCAN on the dataset. We make sure our groupings are accurate by examining measures like  
 Silhouette, Cohesion, and Separation scores. These advanced methods help us find patterns and connections in the complex world of financial data that might not be obvious at first glance.  
  
This report takes you through the steps of preparing the data, choosing important features, and analyzing it with algorithms. We aim to share insights and findings that could be useful for investors who are trying to make sense of the constantly changing stock market. We believe our work offers valuable strategies and knowledge for anyone looking to make informed investment decisions with a solid understanding of the market's dynamics. Our code can be found in this [this GitHub repo.](https://github.com/NirNagary/ML-final-project)

# Dataset and Features

**Data Collection**   
In our comprehensive analysis of the stock market, we used data from two primary sources: the Yahoo Finance API and Twelve data. Yahoo Finance, which is freely accessible, provided us with crucial information on market capitalization.  
Twelve data, a paid service we were able to access through a personal project, offered a rich set of financial data including balance sheets, income statements, company profiles, and cash flow reports. This diverse dataset allowed us to construct a multidimensional view of company performance. The data mining process, particularly from Twelve data, was an intensive task that spanned over 22 hours of downloading.

fig1



**Preprocessing**

Following the data collection, our preprocessing steps included:

**Data Mining -** We carefully analyzed the features within each report and determined which features are combinations of others. For example, in fig 1 you can see part from the balance sheet report all the features are summed in the 'total\_non\_currrent\_assets' so we keep only that feature. We removed duplicates and finally merged all the reports into one dataset based on the year and symbol (company), so we left with 44 features.

**Data Cleaning & Filtering –**   
Our examination revealed a significant presence of zero or null values throughout our dataset. Further analysis indicated that companies may choose not to report all their expenses or incomes in specific fields for various reasons. Recognizing the significance of the 'net income' feature, which is calculated as "Net Income = Total Revenues - Total Expenses" and encapsulates both the gains and losses of a company within, we decided to exclude all records showcasing zero or null values in this feature. This decision was informed by the understanding that accurate net income figures are crucial for a comprehensive analysis, reflecting both positive and negative financial outcomes.

As a result, to make our dataset more consistent and trustworthy, we decided to replace all the missing (null) values with zeros throughout the dataset. This step helps us fix the problem of incomplete data and keeps everything uniform, making our analysis more dependable.

Next, we plan to include only those companies in our analysis that have data spanning at least three years, up to a maximum of four years, for our experiments.

**Feature Engineering –**

In our study, we focused on key financial health indicators including Debt to Equity Ratio, Profit Margin, Return on Equity, and Return on Assets. We identified the most relevant indicators through comprehensive online research and found specific formulas for each. Subsequently, we developed functions corresponding to these formulas and applied them to our dataset to assess the financial health of the companies.  
Next, we conducted a correlation analysis and eliminated features with high correlations to optimize the dataset for our analysis.

Our dataset includes a categorical variable named 'industry,' which comprises 165 distinct values. To streamline the dimensionality in the 'one-hot encoding' process, we opted to group these industries into broader categories based on our domain expertise and judgment. This approach allowed us to effectively reduce the complexity and improve the manageability of our dataset.

**Data Normalization –**

In the initial phase of data normalization, we organized our data into distinct data frames based on their scale and type.

This included grouping all monetary value features into one data frame, financial indicators into another, and categorical features into a third. For the categorical features, we applied one-hot encoding to facilitate their use in our models.

Subsequently, we normalized each data frame using row-wise min-max normalization to standardize the range of our data, ensuring consistency across different scales. Following this, we employed a Quantile Transformer on both the monetary value features and the financial indicators.

We selected the Quantile Transformer due to its effectiveness in handling datasets with negative values and significant outliers. This method is particularly suitable for our needs for the following reasons: Normal Distribution: It ensures data follows a gaussian distribution, minimizing the impact of outliers. As we mentioned we have a significant amount of zero values or a very high positive or negative number of values.

Compatibility with Negative Values: Unlike other normalization techniques, it can process both negative and positive values smoothly.

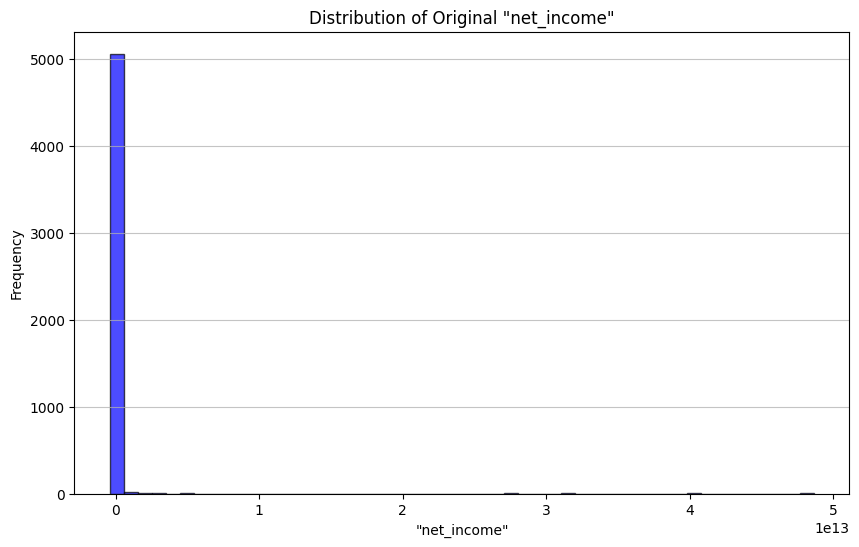
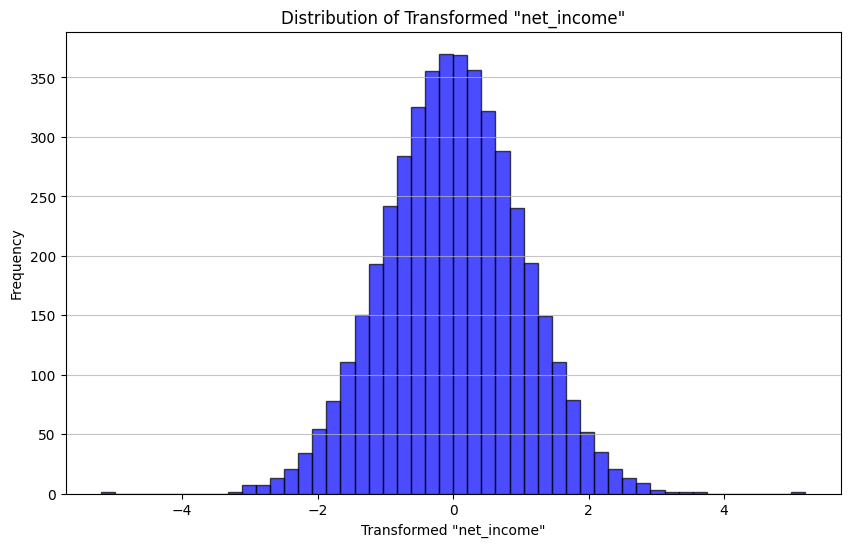


fig2



Outlier Robustness: Its ability to reduce the influence of extreme values makes it ideal for our dataset, enhancing model accuracy. outliers. Distribution demonstrated in histogram in fig2.

A blue and white graph

Description automatically generatedTop of Form

**Feature Reduction –**

Feature Reduction - We conducted Principal Component Analysis (PCA) with both 2 and 3 components across our various datasets. We ultimately refined our dataset to contain a total of 33 features. Details on the experiments will be elaborated in the following section.

# Methodology

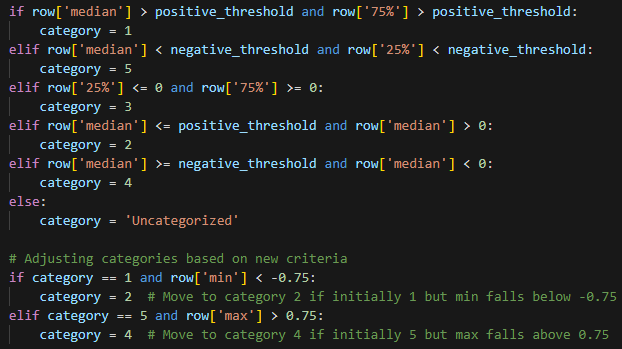
We explored several clustering algorithms, including K-Means, hierarchical clustering, and Expectation-Maximization (EM), each tested across a range of hyperparameters. Our objective was clear: to identify an algorithm capable of finding meaningful patterns that align with our expectations of market behavior. DBSCAN emerged as the most fitting choice, demonstrating superior performance in clustering validity as evidenced by silhouette scores of 0.645.

A heatmap of epsilon and minimum sample sizes was instrumental in fine-tuning DBSCAN’s hyperparameters, ultimately guiding us to an epsilon of 3.2 and a minimum sample size of 10.

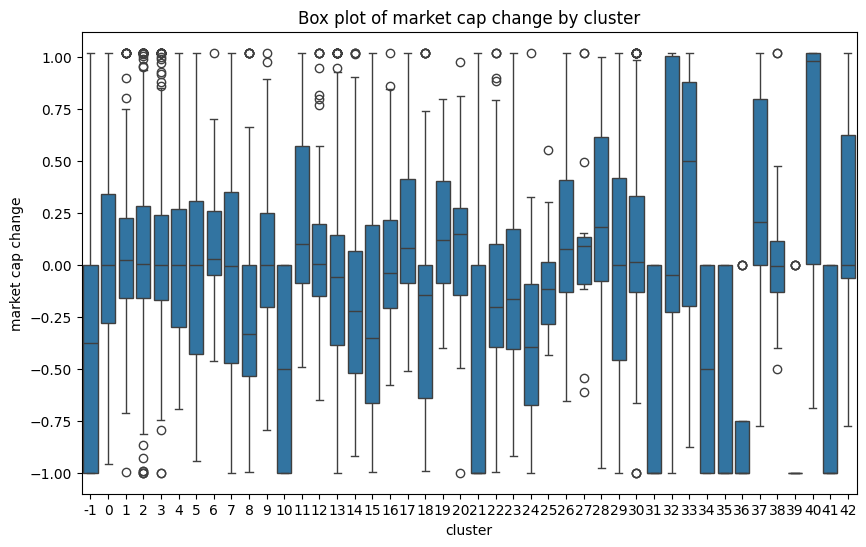
A noteworthy aspect of our approach was the exclusion of the 'market cap change' indicator from the clustering process. This deliberate omission was designed to test the algorithm’s ability to naturally segregate companies in a manner indicative of market cap fluctuations, thereby offering insights into potential investment strategies. The result was 42 clusters.

Although the resulting clusters unveiled patterns that aligned closely with our analysis of market capitalization changes, we wanted a more pragmatic segmentation.

We consolidated these into five principal clusters based on their market cap change propensity, ranging from the most favorable (Group 1) to the least (Group 5), with Group 3 embodying the median market behavior.

The following logic was applied to create the final group clusters:

Leading to this cluster:



Group 3

Group 1

Group 5

.

# Experiments/Results/Discussion

**Experiments:**

Our journey to arrive at the optimal solution was marked by an extensive series of experiments. Initially, our strategy encompassed testing a variety of datasets, upon which we applied all our selected models. This phase involved creating distinct datasets to assess their clustering effectiveness, gauged primarily by the silhouette score—a measure of how well an object is matched to its own cluster and separated from other clusters.

**Scaling\Normalization-** We explored various scaling techniques, including Standard Scaler, Robust Scaler, Min-Max Scaler, and Quantile Transformation (QT), but these did not yield satisfactory results. Through further investigation, we discovered the potential of row normalization combined with QT and realized its suitability for our data as the most sensible approach for our analysis.

**Data Set/** **Different Features -** The datasets varied mainly in their inclusion or exclusion of categorical variables. However, through iterative testing, we gained the insight to also evaluate datasets comprised solely of our financial indicators. This decision expanded our exploration to at least 6 different datasets, each subjected to a thorough analysis with all the chosen clustering models. Each of the 6 datasets underwent Principal Component Analysis (PCA), utilizing both 2 and 3 components for visualization and evaluation purposes.

**Hyperparameter Optimization -** We conducted experiments with various hyperparameters for our chosen models to optimize their performance. For DBSCAN, key parameters include 'min\_samples' and 'epsilon.' We employed a function to determine the optimal 'epsilon' value using the elbow method. Our observations led us to conclude that an increase in 'min\_samples' tends to bias the clustering towards forming a single cluster. The best epsilons are range from {3.2-4.4} when increasing the features decrease.

In the case of the K-means algorithm, we utilized a function to identify the ideal number of clusters ('k') by evaluating it against the silhouette score. Similarly, for the Expectation-Maximization (EM) algorithm, we assessed the optimal number of components ('n') using both the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC).

For Hierarchical Clustering, we visualized the results with a dendrogram for each type (single, average, complete, ward) to aid in determining the hyperparameters. This graphical representation was instrumental in understanding the cluster formation at different levels of aggregation.

**Result and Discussion:**

* The highest performance was observed in datasets that excluded categorical features, suggesting that the inclusion of company industry and sector does not contribute significantly to discerning patterns of financial health.
* When Principal Component Analysis (PCA) was applied, the results were generally consistent with those of our selected model. However, the 'market cap change' did not align logically with these findings, failing to provide meaningful insights into the data's patterns or structures.
* Our ultimate DBSCAN model was configured with hyperparameters set to a minimum of 10 samples and an epsilon value of 3.2 for the best distributed clustering.

Another approach was discussed, if interested in only the first class (most profitable) epsilon of 4.4 gives the best result.

A diagram of a chart

Description automatically generatedWe decided on epsilon 3.2 to have better broad clustering results, which led to the formation of our final groups. Consequently, when a new company is added to our dataset, it undergoes the entire preprocessing sequence and is then classified into one of five groups based on its financial health, with Group 1 denoting the highest level of financial health.  
  
Box plot of the final groups:

# Conclusion and Future Work

In this project, we embarked on a comprehensive analysis of financial health across various companies, employing a combination of clustering algorithms and preprocessing techniques to discern meaningful patterns within the data.

We consider the application of row normalization combined with QT to be an exceptionally insightful strategy for addressing the presence of zero values in our data.

Our analysis revealed the limited impact of categorical features, such as industry and sector, on financial health clustering.

These findings challenge common assumptions in the field and opens new avenues for exploring alternative indicators of financial health.

**Potential Future Directions:**

* **Exploration of Alternative Features**: Future work could explore other features or engineered indicators that might offer deeper insights into financial health.
* **Application to Predictive Models**: Building on our clustering work, there is an opportunity to develop predictive models that can forecast changes in financial health or market cap, potentially integrating machine learning techniques with traditional financial analysis. For example, next year’s market cap change.
* **Cross-Industry Comparison**: Our methodology could be applied across different industries or sectors to compare financial health indicators, providing a broader perspective on the economic landscape.

As we conclude our project, new questions have emerged regarding the most effective ways to measure and predict financial health.

One crucial lesson from our approach was the pivotal role of data preprocessing in achieving our objectives.

A lingering curiosity persists: What would the impact on our findings be if all companies were to provide complete financial transparency? This question remains at the forefront as we consider the potential in our analysis under a scenario of full financial disclosure.

As new investors, we were incredibly enthusiastic about this project, and we are planning to apply our discoveries in our trading strategies.

# Contributions

|  |  |
| --- | --- |
| **Anna** | **Nir** |
| Data Mining - Analyzing the features in each report. | Data Collection - Getting the data with private API and public API. |
| Data Cleaning - Research the web on each feature to find duplicates. | Data Cleaning – Searched the data for missing values and filter the data that 'net income' doesn’t have 0 values. |
| Feature Engineering – Research the web about financial indicators and created their functions. | Feature Engineering – Conducted correlation between features to reduce dimension. |
| Feature Engineering – Split the industry feature into distinct groups based on domain knowledge. | Data Normalization/ Scaling – Approach of row scaling. |
| Data Normalization/ Scaling – Approach of column QT scaling. | Experiments – Run all models on 3 datasets (mixed ones). |
| Experiments - Run all models on 3 datasets (not mixed ones). | Experiments – Had the idea of the final groups. |
| Write Up – Project proposal, final write up. | Write Up – Part of methodology. |

# Appendices

Appendix: Resources and Acknowledgments

Throughout the course of this project, our research and analysis were supported by a variety of resources, which were invaluable to the completion of our work.

**ChatGPT:** This advanced AI tool played a significant role in the code and error handling aspects of our project.

**Professors PowerPoint Presentations:** The educational materials provided by our professors served as a foundational resource, offering theoretical frameworks and insights that guided our understanding of preprocessing, clustering algorithms and evaluation techniques.

**Web Search:** Extensive online research was conducted to supplement our knowledge base, validate our approaches, and explore the latest trends in financial health analysis.