EDA Report

Team Lead:

Recorder:

Spokesperson:

# Background and Introduction

Understanding and anticipating the investment decisions of influential figures has long been a subject of interest in financial markets. Warren Buffett, renowned for his disciplined value investing approach, focuses on identifying fundamentally strong yet undervalued companies (Hagstrom, 2013). Investors and financial professionals closely track his moves through SEC filings and media reports, but these sources provide insights only in retrospect. Team Lambda aspires to bridge that gap by leveraging machine learning to predict which stocks Buffett is likely to buy next, based on historical data and company-level financial indicators. By deploying predictive modeling and analysis, we aim to create a decision-support tool that empowers individual investors with insights traditionally available only after the fact.

The intersection of behavioral finance and predictive analytics presents a compelling opportunity to model Buffett’s investment patterns. While previous research has explored stock price prediction and portfolio performance modeling, few studies have specifically attempted to emulate the buy decision-making process of a known value investor using explainable machine learning models (Fischer & Krauss, 2018). Our approach is novel in that it does not merely assess market trends but seeks to identify the financial characteristics that make a stock appealing to Buffett before he purchases.

We hypothesize that historical financial metrics such as a low price-to-earnings (P/E) ratio, strong revenue growth, and high dividend yield are significant predictors of his stock selections. This is because Buffett prioritizes companies that exhibit financial strength, are undervalued relative to their intrinsic worth, and demonstrate steady, long-term growth potential (Buffett & Cunningham, 2020). If successful, we predict this work could enhance traditional financial analysis by providing predictive insights into high-profile investment strategies, offering a practical tool for investors who seek to align their decisions with Buffett’s time-tested approach.

# Methods

Dataset and cleaning :

Individual stock data, sourced from Bloomberg, was pulled as individual files representing quarterly snapshots from Q1 2007 to Q4 2024. These files were merged on column headers for a dataset containing all stock data. Warren Buffett’s portfolio activity is sourced from Dataroma, filtered on activity representing a buy (initial purchases and additional purchases of existing holdings), then a label was created to indicate 1 for all purchase activity. Economic data, sourced from Bloomberg, was pulled in three month increments to match the quarterly label for a merge. The dates were changed from short form (eg. 03/31/24) into two columns, Quarter (Q1) and Year (2024). These three datasets were then combined. Stock data and purchase activity merged on Ticker Quarter and Year. Stock data and economic data merged on Quarter and Year.

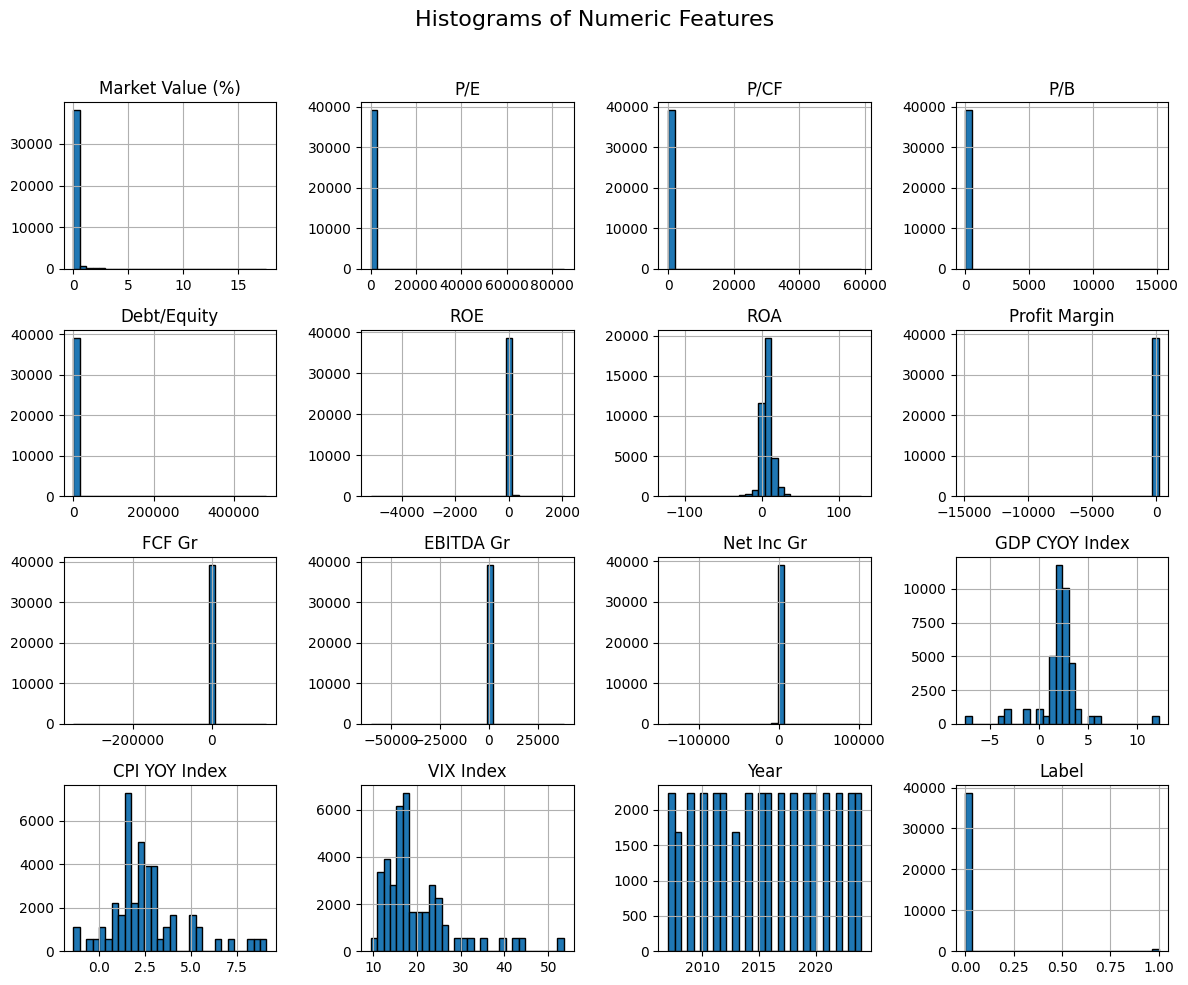
**What methods did you use to perform EDA? Why did you make the choices that you did?**

To perform Exploratory Data Analysis (EDA), we started by analyzing the dataset's structure, including checking for missing values, duplicates, and data types. We handled missing values by imputing numerical columns with the mean and categorical columns with the mode to ensure data completeness. We also removed duplicate rows to maintain data integrity. For feature selection, We focused on financial metrics (e.g., P/E, Profit Margin), macroeconomic indicators (e.g., GDP CYOY Index, VIX Index), and categorical/time-based features (e.g., Sector, Year) that are most relevant to predicting investor decisions.

We used visualizations such as histograms, scatter plots, and correlation heatmaps to understand the distribution of variables, relationships between features, and potential multicollinearity. For example, scatter plots helped identify trends between macroeconomic indicators like CPI YOY Index and PPI YOY Index, while the correlation matrix highlighted redundant features. These methods were chosen to ensure a comprehensive understanding of the dataset and to identify patterns and relationships that could inform feature engineering and predictive modeling.

# Results

1. P/E, P/CF, P/B, Debt/Equity, FCF Gr, , Profit Margin,Market value % ,these features are highly right skewed , with most values concentrated near zero and few extreme outliers on the higher end. We will consider handling outliers or applying transformations to reduce skewness.



**Summary Statistics for Numeric Variables**

The table provides descriptive statistics for all numeric variables, including measures of central tendency (mean, median), spread (standard deviation, min, max), and distribution shape (skewness, kurtosis). Variables like P/E, P/CF, and Debt/Equity are highly right-skewed with extreme outliers, as indicated by their high skewness and kurtosis values. In contrast, variables like GDP CYOY Index and CPI YOY Index exhibit relatively normal distributions with low skewness and kurtosis.

**Summary Statistics for Categorical variables : Frequencies, %, Proportions :**

Summary for Sector: The Sector variable is dominated by Software & IT Services, which accounts for 12.5% of the dataset, followed by Residential & Commercial REIT (5.18%) and Machinery, Equipment & Components (4.82%). The remaining sectors are more evenly distributed, with many contributing less than 4% each. This indicates a concentration of data in a few key sectors.

Summary for Quarter: The Quarter variable is evenly distributed across all four quarters, with Q3 and Q2 each accounting for 25.71% of the data, while Q4 and Q1 contribute slightly less at 24.29% and 24.28%, respectively. This suggests that the dataset captures data consistently across time periods.

The dominance of certain sectors, such as Software & IT Services (12.5%), indicates investor preferences or market trends that can be leveraged as key predictors in the model.

These insights will help the model identify sector-specific and time-based factors influencing stock selection.

# Discussion

# Appendix

References:

Buffett, W., & Cunningham, L. A. (2020). *The Essays of Warren Buffett: Lessons for Corporate America*. The Cunningham Group.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270(2)*, 654-669.

Hagstrom, R. G. (2013). *The Warren Buffett Way*. Wiley.

Data Dictionary:

|  |  |
| --- | --- |
| Column Name | Description |
| Market Value (%) | The stock's weight in the portfolio, indicating its relative importance. |
| P/E | Price-to-Earnings ratio, a valuation metric showing how much investors pay per unit of earnings. |
| P/CF | Price-to-Cash Flow ratio, a valuation metric showing the stock price relative to cash flow. |
| P/B | Price-to-Book ratio, a valuation metric comparing the stock price to book value. |
| Debt/Equity | Financial leverage ratio, showing the proportion of debt relative to equity. |
| ROE | Return on Equity, a profitability metric indicating how efficiently equity generates returns. |
| ROA | Return on Assets, a profitability metric showing how efficiently assets generate returns. |
| Profit Margin | The percentage of revenue that turns into profit after expenses. |
| FCF Gr | Free Cash Flow Growth, indicating the growth rate of free cash flow over time. |
| EBITDA Gr | EBITDA Growth, showing the growth rate of earnings before interest, taxes, depreciation, and amortization. |
| Net Inc Gr | Net Income Growth, indicating the growth rate of net income over time. |
| GDP CYOY Index | Year-over-Year GDP growth, reflecting economic growth trends. |
| CPI YOY Index | Year-over-Year Consumer Price Index, a measure of inflation. |
| PPI YOY Index | Year-over-Year Producer Price Index, another measure of inflation. |
| VIX Index | Volatility Index, reflecting market uncertainty and investor sentiment. |
| Sector | The industry or sector to which the stock belongs. |
| Year | The year of the data record, used for time-based analysis. |
| Quarter | The quarter of the year (Q1, Q2, Q3, Q4), capturing seasonality. |
| Label | The target variable indicating whether the stock was purchased (1) or not (0). |