Exploratory Data Analysis and Hypothesis Testing on S&P 500 Companies

1. Data Overview

The dataset includes 495 S&P 500 companies with variables such as stock price, earnings, market capitalization, and valuation ratios (Price/Earnings, Dividend Yield, Price/Book). Three categorical (Symbol, Name, Sector) and several numerical features describe firm-level financial metrics.

After removing missing entries (0.4% in Price/Earnings, 1.6% in Price/Book), the dataset is clean enough for further analysis.

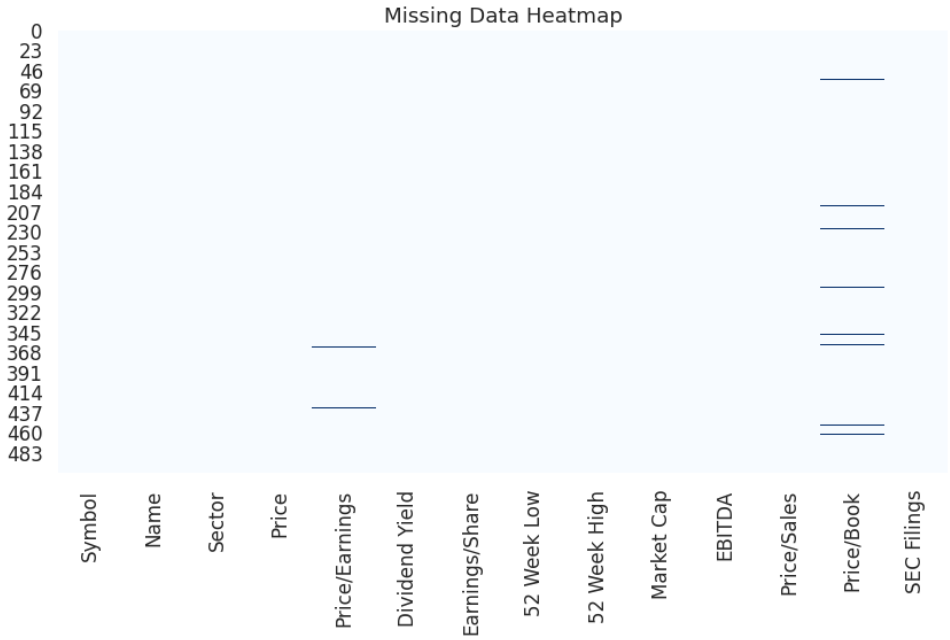

Figure 1: Heatmap of missing data across financial variables.

2. Exploratory Data Analysis

Most financial variables (Price, Market Cap, EBITDA, P/E) possess right-skewed distributions, which indicates domination of a few large-cap firms, like Apple and Microsoft.

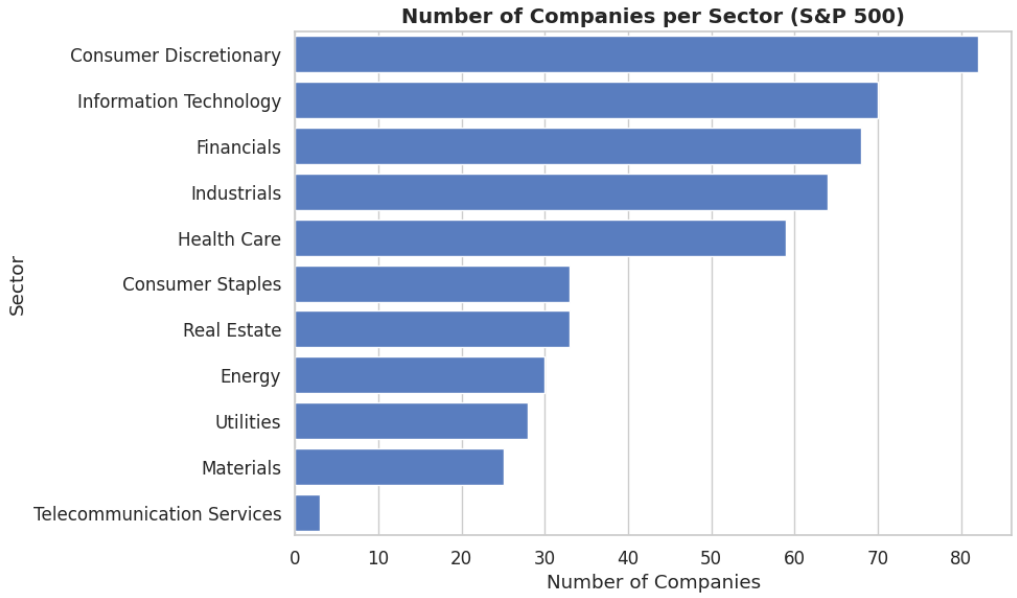

Figure 2: Number of companies per sector (S&P 500).

The sector most represented is Consumer Discretionary (82 companies), followed by Information Technology and Financials. These demonstrate the current structure of the U.S. equity market, where consumer and tech businesses represent a substantial share of total capitalization. A correlation analysis reveals strong positive relationships among price-related features (Price, Market Cap, EBITDA) and a negative correlation between Dividend Yield and Price, which suggests that high-growth firms pay smaller dividends.

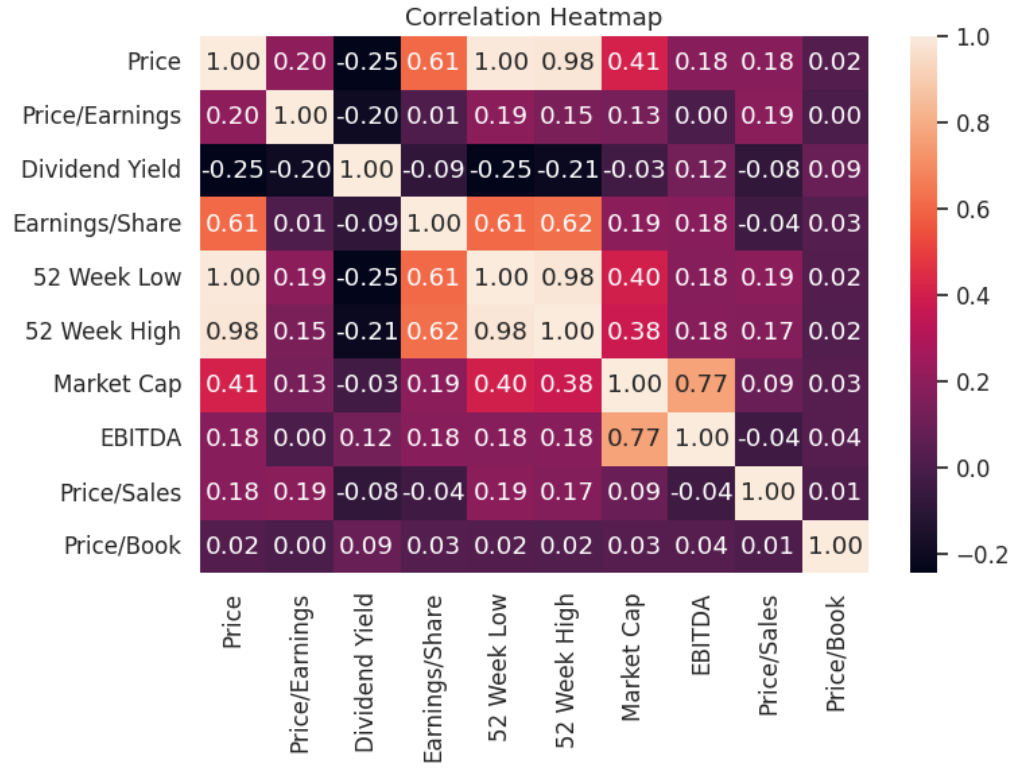

Figure 3: Correlation matrix for numerical variables.

Boxplots show extreme outliers in Price, Market Cap, and P/E Ratio, corresponding to high-growth technology firms and very large companies. Negative outliers in Earnings/Share and EBITDA show firms currently operating at a loss.

3. Hypothesis Testing

We test whether Information Technology firms have higher Price/Earnings ratios than other sectors.

- **Null Hypothesis H_0:** Mean P/E (Technology) = Mean P/E (Non-Technology)
- **Alternative Hypothesis H_1:** Mean P/E (Technology) > Mean P/E (Non-Technology)

Using Welch's two-sample t-test, as we do not have an assumption of equal variances:

$$t = 1.758, \quad p = 0.0414$$

Since $p < 0.05$, we reject H_0. Technology companies indeed have higher average P/E ratios. This result indicates stronger market growth expectations.

4. Predictive Modeling

To predict stock price (Price), we use two explanatory features:

1. Earnings/Share (EPS) — a fundamental performance indicator.
2. Market Cap — reflecting the overall scale of the company and investor sentiment.

Three linear models are compared using 5-fold cross-validation:

Model	**Mean R^2**	**Std R^2**
Linear Regression	0.312	0.05
Ridge Regression	0.329	0.04
Lasso Regression	0.321	0.06

An average R^2 of 0.329 indicates that about one-third of the price variance is explained by earnings and market capitalization. This prediction quality is reasonable in financial datasets, where the valuation of the company is influenced by many qualitative and macroeconomic factors not captured numerically.

5. Conclusions

- Most financial features are skewed.
- Information Technology companies display significantly higher P/E ratios.
- Linear and regularized models provide similar predictive performance.

Using more than two features may be useful in increasing performance.