**S&P 500 Stock Data and Sector-Based Performance Analysis**

**1. Introduction**

The S&P 500 is a benchmark index comprising 500 of the largest publicly traded companies in the United States, spanning a wide range of industry sectors. Despite being part of the same index, each stock experiences unique price movements driven by various internal and external factors. This project investigates whether these fluctuations are random or influenced by systematic factors—particularly the sector in which a company operates.

Our central research questions are:

* Does industry sector impact stock performance?
* Are some sectors inherently more volatile or more heavily traded?
* Can we identify patterns or trends that suggest sectoral influence on stock prices?

To address these questions, we analyze S&P 500 stock price data sourced from Yahoo Finance and industry classifications from Wikipedia. We explore relationships between sector and average stock price, assess the trading volume’s correlation with closing prices, and use forecasting and classification techniques to model future price movements for a sample stock.

**2. Data Sources and Preparation**

**2.1 Stock Price Data**

Daily stock data for S&P 500 companies was collected from Yahoo Finance, covering approximately one year (four quarters). This dataset includes open, close, high, and low prices, along with daily trading volume. Over 55,000 individual observations were gathered. The data was cleaned by handling missing values, converting date formats, and standardizing column names to lowercase to ensure consistency.

**2.2 Company and Sector Data**

Company metadata was obtained from the Wikipedia page listing S&P 500 constituents. This included ticker symbols, company names, GICS sector classifications, headquarters locations, and the date of data retrieval. The dataset was parsed using pandas.read\_html(), and then cleaned by removing duplicates, standardizing text formatting (e.g., trimming whitespace and ensuring capitalization consistency), and renaming columns to facilitate merging with the stock price dataset.

**2.3 Merging Datasets**

The stock data and metadata were merged using ticker symbols as a common key. Since the stock price dataset contains multiple daily entries per ticker, and the metadata contains only one entry per ticker, we performed a one-to-many left join. This allowed company metadata to be replicated across all corresponding rows in the stock price dataset.

**2.4 Data Dictionary**

The final dataset included variables such as symbol, company, sector, location, and data\_scraped, along with stock price variables like high, low, open, close, and volume. These fields captured key financial and descriptive information necessary for analysis.

|  |  |  |
| --- | --- | --- |
| **Field Name** | **Data Type** | **Description** |
| symbol | Text | Stock ticker symbol |
| company | Text | Full name of the company |
| sector | Text | GICS sector classification (e.g., Technology, Healthcare) |
| location | Text | Headquarters location (U.S. or international) |
| data\_scraped | Date | Date when the company data was retrieved from Wikipedia |
| high | Numeric | Highest price of the day |
| low | Numeric | Lowest price of the day |
| open | Numeric | Opening price of the day |
| close | Numeric | Closing price of the day |
| volume | Numeric | Number of shares traded on the given day |

**3. Data Analysis**

**3.1 Sector-Based Closing Price Analysis**

To understand how closing prices vary across sectors, we began with univariate analysis to examine the distribution of all closing prices. The results showed a right-skewed distribution with most prices under $1,000. A few outliers extended beyond $10,000.

**Figure 1: Distribution of Stock Closing Prices**  
A graph showing a distribution of stock closing prices

AI-generated content may be incorrect.

We then grouped the data by sector and calculated the average closing price for each. The Consumer Discretionary sector showed the highest average closing prices, followed by categories such as Agricultural & Farm Machinery, Home Improvement Retail, and Research & Consulting Services. On the other hand, sectors such as Drug Retail, Casinos & Gaming, and Advertising had the lowest average prices.

**Figure 2: Average Closing Price by Sector**  
A graph showing a closing price

AI-generated content may be incorrect.

To assess the statistical significance of sectoral differences, we conducted a two-sample t-test comparing the Healthcare and Technology sectors. The null hypothesis assumed no significant difference in average closing prices between the two sectors, while the alternative hypothesis suggested otherwise. The test yielded a t-score of -25.11 and a p-value of 0.0000, leading us to reject the null hypothesis. This indicates that industry sector does indeed have a significant impact on stock prices.

**3.2 Relationship Between Closing Price and Volume**

Next, we investigated the relationship between trading volume and closing price using K-Means clustering. By setting k=3, we identified three distinct clusters. The first cluster included stocks with high closing prices but low trading volumes, typically representing niche, high-value stocks. The second cluster featured stocks with low prices and high volumes, indicative of widely traded but lower-value equities. The third cluster contained stocks with both low volume and low price, which are likely average-performing or overlooked in the market.

This segmentation reflects expected market dynamics, where lower-priced stocks attract higher trading activity, while high-priced stocks tend to be traded less frequently.

**Figure 3: K-Means Clustering of Closing Price vs Volume**  
A graph with numbers and lines

AI-generated content may be incorrect.

**4. Forecasting and Predictive Modeling**

**4.1 ARIMA Forecasting**

To explore price forecasting, we applied the ARIMA (Autoregressive Integrated Moving Average) model to 3M Company (MMM) stock. We used the past 10 months of closing prices to fit the model, specifying parameters of p=5, d=1, and q=0 to address lagged dependencies and ensure stationarity.

The model produced a forecast over the next 42 trading days. Residual diagnostics confirmed a good fit, with no major anomalies detected. Overall, the ARIMA model effectively captured the underlying price trend and demonstrated potential for short-term forecasting of stock movements.

**Figure 4: ARIMA Forecast of 3M Closing Prices**

A graph showing a line

AI-generated content may be incorrect.

**4.2 Logistic Regression for Stock Movement Classification**

To complement the time-series analysis, we implemented a logistic regression model to classify whether MMM's stock would experience a gain over the next quarter (63 trading days). The target variable was binary: a value of 1 indicated a gain, and 0 indicated no gain or a loss. Predictor features included the previous quarter’s closing price and the percentage change during the current quarter.

Evaluation metrics showed strong performance in predicting gains, with a recall score of 1.00. However, the model struggled with detecting non-gain outcomes, suggesting a class imbalance favoring gain predictions. To improve the model, we recommend using SMOTE (Synthetic Minority Over-sampling Technique) or resampling methods to address imbalance, and exploring more advanced models such as Random Forest or XGBoost for enhanced accuracy and robustness.

**5. Conclusion**

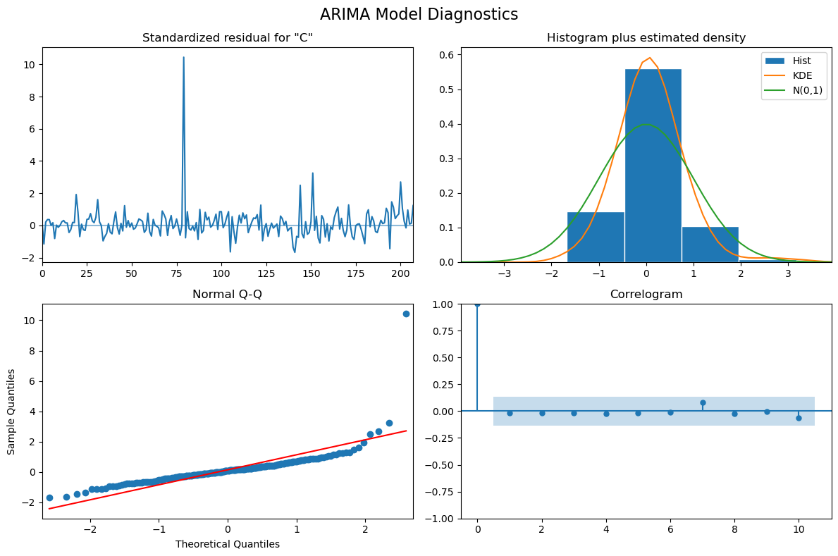
This analysis confirms that industry sector significantly influences stock performance, particularly in terms of average closing prices. Additionally, the relationship between price and trading volume reveals predictable patterns, allowing for useful segmentation using unsupervised learning techniques like clustering.

The ARIMA model demonstrated effectiveness in forecasting future price trends for individual stocks, while logistic regression showed potential for classification tasks, albeit with limitations due to class imbalance. Overall, our findings support the hypothesis that sector affiliation and trading behaviors are important factors in analyzing and predicting stock performance.

**References**

* Yahoo Finance. (n.d.). *Historical Stock Data.* Retrieved from <https://finance.yahoo.com>
* Wikipedia. (2025). *List of S&P 500 Companies.* Retrieved from <https://en.wikipedia.org/wiki/List_of_S%26P_500_companies>
* Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed). OTexts.

A screenshot of a computer

AI-generated content may be incorrect. **Appendix**