**BAIS:3250 Data Wrangling Final Project Report**

**S&P 500 Stock Analysis Final Project Analysis Report**

1. **Introduction**

The S&P 500 represents 500 of the largest publicly traded U.S. companies and serves as a key indicator of overall market performance. Analyzing trends within this index offers valuable insights for investors and analysts, especially amid rising market volatility and economic uncertainty.

This project explores recent stock behavior in the S&P 500 using real-world data on prices, trading volume, and sector classifications. Through data wrangling, statistical analysis, and machine learning, the study examines sector trends, volatility, unusual trading activity, and buy/sell signals across a six-month period.

The goal is to show how data-driven methods can uncover actionable insights and enhance understanding of market dynamics.

1. **Data**

This project integrates data from three key sources, TradingView, Yahoo Finance, and Wikipedia, to analyze trends and anomalies across companies in the S&P 500. TradingView and Yahoo Finance served as the two primary sources, providing the list of tickers and detailed historical stock data, while Wikipedia was used to enrich the dataset with standardized sector classifications. Each source was cleaned, standardized, and merged into a final structured dataset used for the project’s analysis.

* 1. **TradingView (Scraped)[[1]](#footnote-1)**

I sourced the list of S&P 500 companies from TradingView’s U.S. stock screener, which offers a regularly updated view of active tickers and company names. Using Selenium, I automated the process to navigate the page and extract the complete set of records. After scraping, I cleaned the data by removing null values, trimming whitespace, and standardizing ticker formats. Tickers not part of the S&P 500 were filtered out, resulting in a reliable and current list that served as the foundation for all further analysis.

* 1. **Yahoo Finance (yfinance package)**

To obtain quantitative time-series data for each S&P 500 company, I used the yfinance Python library to access Yahoo Finance’s API. For each valid ticker symbol collected from TradingView, I retrieved six months of historical data at a daily frequency, including:

* Open price
* High price
* Low price
* Close price
* Trading volume

These metrics were essential for calculating stock returns, volatility, and spotting volume anomalies. The downloaded data was restructured into a long format dataframe, with each row representing a single stock on a single trading day. The data was also checked for missing values and cleaned to ensure consistency across tickers.

If the API failed to return data for a given ticker, due to delisting or data availability constraints, we excluded it from the final dataset and logged it for reference.

* 1. **Wikipedia (Scraped) [[2]](#footnote-2)**

To perform sector-level analysis, I added GICS sector classifications by scraping the “List of S&P 500 companies” page from Wikipedia. I focused specifically on extracting each company’s GICS sector, which was not available through TradingView or Yahoo Finance.

Using Selenium, I navigated the page and scraped only the relevant columns: Ticker and GICS Sector. I cleaned the ticker symbols to ensure they matched the formatting used in both TradingView and Yahoo Finance. I then merged the sector classifications into the main dataset using the ticker as the join key.

This step allowed me to group companies by sector and explore patterns in returns, volatility, and trading activity across different industries.

* 1. **Data Cleaning and Integration**

After collecting data from TradingView, Yahoo Finance, and Wikipedia, I performed a multi-step cleaning and integration process to create an analysis-ready dataset named sp500\_stocks. This dataset combines daily price data, trading volume, and sector classifications for S&P 500 companies over a six-month period.

I began by cleaning the ticker list scraped from TradingView, removing null entries, trimming whitespace, and standardizing formats to ensure compatibility with Yahoo Finance. Invalid or incomplete tickers were excluded. Using the cleaned list, I retrieved daily OHLC (Open, High, Low, Close) and volume data through the yfinance library. The data was reformatted into long-form, with each row representing one stock on one trading day. Tickers without available data were skipped using error handling.

To enrich the dataset, I scraped Wikipedia’s list of S&P 500 companies to extract each firm’s GICS sector. After cleaning and aligning the tickers, I merged this sector information into the price data using a left join on the Ticker column.

Next, I engineered several variables to support analysis. These included daily return ((Close - Open) / Open), daily volatility (High - Low), absolute price movement, and both total and 30-day average trading volume. I also calculated two key ratios: Volume Ratio (daily volume / total average volume) and Volume Spike Ratio (daily volume / 30-day average volume), which helped flag unusual activity.

Finally, I reordered columns, removed rows with missing or invalid values, and confirmed the dataset was complete. The resulting sp500\_stocks dataset contains over 60,000 observations and serves as the foundation for all analysis in this project. A full description of each variable is provided in Table 1.

***Table 1 Data Dictionary***

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Type** | **Source** | **Description** |
| Ticker | Text | All three | Stock symbol used for merging and analysis |
| Company Name | Text | TradingView | Full name of the company |
| Company Sector | Text | Wikipedia | Sector classification from Wikipedia |
| Date | Date | TradingView | Trading day date |
| Stock Price | Numeric | TradingView | Latest available price from TradingView |
| Open | Numeric | Yahoo Finance | Opening price on that day |
| High | Numeric | Yahoo Finance | Highest price on that day |
| Low | Numeric | Yahoo Finance | Lowest price on that day |
| Close | Numeric | Yahoo Finance | Closing price on that day |
| Price % Change (Numeric) | Numeric | TradingView | Cleaned stock price % change as float |
| Price Movement | Numeric | Calculated from Yahoo Finance | Absolute value of daily return, showing magnitude of price change regardless of direction. |
| Daily Volatility | Numeric | Calculated from Yahoo Finance | Intraday volatility measured as High - Low |
| Volume | Numeric | Yahoo Finance | Latest stock volume (converted to number) |
| 30-Day Avg Volume | Numeric | Calculated from Yahoo Finance | Rolling 30-day average of daily trading volume for each stock |
| Volume Spike Ratio | Numeric | Calculated from Yahoo Finance | Ratio of current volume to 30-days average volume |
| Signal | Text | All three | Classification of stock as Buy, Sell, or Hold |
| Predicted Signal | Text | All three | Machine learning model prediction on stock movement, classified as Buy, Sell, or Hold |

1. **Analysis**
   1. **Which stocks are good buy/sell/hold based on short-term price behavior?**

To identify which S&P 500 stocks are potential buy, sell, or hold opportunities based on short-term price movement, I used a machine learning classification approach. I first created a Signal column by categorizing daily percentage price changes (Price % Change (Numeric)) into three classes: Buy (price change > 2%), Sell (price change < -2%), and Hold (between -2% and 2%). This converted a continuous return variable into a categorical target suitable for supervised learning.

Using features derived from Yahoo Finance, including daily volatility, absolute price movement, volume, and volume spike ratio, I trained and evaluated several machine learning classification models with the scikit-learn library. These included Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), and Support Vector Machine (SVM). The data was preprocessed with scaling and encoding before being split into training and test sets.

Model performance was assessed using accuracy and F1-score. The Decision Tree model achieved perfect performance on the test set with an accuracy and F1-score of 1.000, while Logistic Regression and SVM models both scored 0.990 accuracy and 0.989 F1-score. The KNN model also performed well with 0.970 accuracy and 0.969 F1-score, indicating all models were highly effective at classifying stock behavior based on the selected features as shown in Figure 1.

**A math equations with numbers

AI-generated content may be incorrect.*Figure 1. Machine Learning Model’s Accuracy and F1-Score Performance***

The final Predicted Signal forecasts whether stocks will rise, fall, or remain stable the next trading day. Based on this, I **compiled Buy, Sell, and Hold** recommendations (Figures 2–4), offering a model-driven tool for short-term trading decisions in the S&P 500.

**A screenshot of a computer

AI-generated content may be incorrect.*Figure 2. Stocks to Buy***

**A screenshot of a computer

AI-generated content may be incorrect.*Figure 3. Stocks to Sell***

**A screenshot of a computer screen

AI-generated content may be incorrect.*Figure 4. Stocks to Hold***

* 1. **Which industry sectors perform the best/worst in terms of average return and price movement?**

To identify top- and bottom-performing sectors in the S&P 500, I analyzed average daily returns and price movement by GICS sector. As shown in Figures 5 and 6, Energy led with a strong return of +0.89%, followed by Real Estate and Health Care. In contrast, Consumer Discretionary, Utilities, and Communication Services underperformed, with Consumer Discretionary showing the largest average loss at -0.69%.

**A screenshot of a computer screen

AI-generated content may be incorrect.*Figure 5. Average Return by Sector***

**A graph showing a bar graph

AI-generated content may be incorrect.*Figure 6. Bar Chart of Sector Average Return***

I also calculated price movement, defined as the absolute daily percentage change, to measure how much stocks in each sector fluctuate, regardless of direction. This reflects volatility even when overall returns are flat. As shown in Figure 7, Consumer Discretionary had the highest average movement, followed by Information Technology and Health Care, while Utilities and Consumer Staples showed the lowest, indicating greater stability.

**A white text with black text

AI-generated content may be incorrect.*Figure 7. Average Price Movement by Sector***

To visualize this variability, I created a box plot of price movement by sector (Figure 8). It reveals broader distributions and more outliers in sectors like Consumer Discretionary and Tech, while Utilities and Staples are more tightly clustered, reinforcing their defensive profile.

***Figure 8. Box Plot of Sector Average Price Movement* A graph with blue and black bars

AI-generated content may be incorrect.**

Overall, Energy led in returns, but Consumer Discretionary was the most volatile. Utilities and Consumer Staples, though lower performing, were more stable, an appealing trait for risk-averse investors.

* 1. **Which stocks exhibit the highest/lowest volatility based on daily price range?**

To determine which S&P 500 stocks are the most and least volatile, I calculated daily volatility as the difference between each stock’s high and low prices. This simple range-based measure captures intraday price fluctuations and highlights how widely a stock’s value moves within a single trading session.

The top 10 most volatile stocks, shown in Figure 9 and Figure 10, include NVR, Booking Holdings (BKNG), and Fair Isaac (FICO), all companies with high share prices and large daily trading ranges, often exceeding $20 to $100. These large price swings may reflect market sensitivity, news exposure, or speculative activity. On the other hand, the 10 least volatile stocks, presented in Figure 11 and Figure 12, include Walgreens (WBA), Amcor (AMCR), and Ford (F). These stocks had narrow price ranges under $0.25, indicating more stable daily behavior.

**A screenshot of a data

AI-generated content may be incorrect.*Figure 9. Top 10 Most Volatile Stocks***

**A graph of a number of red and brown bars

AI-generated content may be incorrect.*Figure 10. Bar Chart of Top 10 Most Volatile Stocks***

**A screenshot of a computer screen

AI-generated content may be incorrect.*Figure 11. Top 10 Least Volatile Stocks***

**A graph showing a number of green bars

AI-generated content may be incorrect.*Figure 12. Bar Chart of Top 10 Least Volatile Stocks***

This analysis provides a snapshot of short-term risk profiles based on price range. Stocks with high volatility may appeal to active traders seeking large movements, while low-volatility stocks may suit conservative investors focused on price stability.

* 1. **Which stocks have unusual trading volume activity today relative to their normal average volume?**

To detect unusual trading activity, I calculated a Volume Spike Ratio by dividing each stock’s current trading volume by its 30-day average. Stocks with a ratio above 1.5 were flagged as having an unusually high volume. The distribution of these ratios, shown in Figure 13, is heavily right-skewed, with most stocks falling below the 1.0 threshold and a small tail of extreme volume outliers.

***Figure 12.* A graph of a number of columns

AI-generated content may be incorrect.*Histogram of Volume Spike Ratio***

**A screenshot of a computer screen

AI-generated content may be incorrect.**The top 10 stocks with the highest volume spike ratios, such as Insulet Corporation (PODD) and Expedia Group (EXPE), are listed in Figure 14. These stocks traded at two to three times their normal volume, suggesting elevated interest or possible news-driven activity.

***Figure 14. Top 10 Stocks with Unusually High Trading Volume***

To evaluate whether these high-volume stocks experienced different returns, I conducted a two-sample t-test to compare average price changes between stocks with volume spike ratios above and below the 1.5 threshold. As shown in Figure 15, the result (t = 0.475, p = 0.717) showed no significant difference, so I failed to reject the null hypothesis. Therefore, while volume spikes may signal unusual activity, they did not correspond with notable price changes in this case.

**A close up of a text

AI-generated content may be incorrect.*Figure 15. Hypothesis Test: Volume Spike vs. Price Change***

* 1. **Time Series Analysis**

To explore short-term price dynamics, I applied the ARIMA time series model to forecast the three highest-volume stocks on the most recent trading day: TSLA, NVDA, and PLTR.

The ARIMA model was trained on each stock’s historical daily closing prices and used to generate 30-day forward forecasts. As seen in Figures 16–18, the model predicts relatively flat trajectories for all three stocks, suggesting short-term price stabilization after recent periods of volatility. For example, TSLA (Figure 16) showed a volatile decline in recent months but is projected to hover near its current level. Similarly, NVDA (Figure 17) and PLTR (Figure 18) are forecasted to exhibit minimal near-term movement despite prior fluctuations.

A graph of a stock market

AI-generated content may be incorrect.***Figure 16. TSLA Time Series Analysis***

A graph showing a line graph

AI-generated content may be incorrect.***Figure 17. NVDA Time Series Analysis***

A graph of a line graph

AI-generated content may be incorrect.***Figure 18. PLTR Time Series Analysis***

These forecasts provide a useful baseline for traders and analysts evaluating near-term expectations, though further refinements could improve accuracy by accounting for external market factors or higher-frequency data.

1. **Conclusion**

This project provided a multi-dimensional analysis of the S&P 500 using six months of real-world stock data. By integrating price, volume, and sector classification data from TradingView, Yahoo Finance, and Wikipedia, I built a comprehensive dataset that enabled a range of insights, spanning descriptive statistics, volatility tracking, anomaly detection, and predictive modeling.

Key findings include the identification of Energy as the top-performing sector by average return and Consumer Discretionary as the most volatile based on price movement. Although some stocks experienced sharp volume spikes, hypothesis testing revealed no significant correlation between abnormal trading volume and daily price changes. A machine learning classification model accurately predicted short-term Buy/Hold/Sell signals, while ARIMA-based time series forecasting showed that heavily traded stocks such as TSLA, NVDA, and PLTR may stabilize in the near term.

Despite these insights, the analysis has several limitations. It focused solely on daily-level data and excluded external drivers such as earnings reports, macroeconomic events, and sentiment indicators. The forecasting models, while useful, assumed that past trends would persist, an assumption that will not hold in highly dynamic markets.

Future work could extend this analysis by incorporating news-based sentiment, intraday data, or macroeconomic indicators to improve predictive power. More advanced machine learning models or ensemble methods could also enhance classification performance. Finally, exploring portfolio-level strategies across sectors or volatility profiles could translate these insights into actionable investment frameworks.

To summarize, this project demonstrates how structured data analysis can uncover patterns in market behavior, support informed decisions, and highlight the value of combining statistical methods with machine learning in financial research.

1. <https://www.tradingview.com/symbols/SPX/components/> [↑](#footnote-ref-1)
2. <https://en.wikipedia.org/wiki/List_of_S%26P_500_companies> [↑](#footnote-ref-2)