Stock Market Analysis and Forecasting Using S&P 500 Data

(COMP3125 Individual Project)

Nguyen Thach Tran
School of Computing and Data Science

I. Introduction

Stock market analysis is important for understanding trends, guiding investment decisions, and identifying high-performing sectors. This project focuses on analyzing S&P 500 stock sectors to determine which is performing the best, examine price trends over the past year, identify the most stable companies, and predict closing prices using historical data. By combining descriptive analysis and predictive modeling, this study provides insights into sector performance and short-term stock behavior.

II. Datasets

A. Source of dataset

The dataset used in this project was obtained from Kaggle, a reliable platform widely used for data analysis and academic research. It contains historical daily prices and sector information for S&P 500 companies and is updated daily, making it suitable for trend analysis and predictive modeling.

S&P 500 Stocks (daily updated)

B. Character of the datasets

The dataset used in this analysis is composed of three CSV files: sp500_companies.csv, sp500_stocks.csv, and sp500_index.csv, although the study primarily focuses on the first two. All files are in standard comma-separated format, with sp500_companies.csv containing approximately 505 rows of company-level metadata, such as ticker symbols, long names, sectors, and subsectors, while sp500_stocks.csv contains several hundred thousand rows of daily price data, including open, high, low, close, volume, and date for each S&P 500 stock. The dataset reflects the structure of the S&P 500 index, which includes 500 companies but 505 stocks due to multiple share classes (e.g., GOOG and GOOGL). Before analysis, the stock and company datasets were merged using the ticker symbol to attach each price record to its respective sector and company name. Basic cleaning steps included removing rows with missing closing prices. Several engineered features were created to support forecasting, such as lagged closing prices and moving averages, computed using rolling window formulas. These additions enhanced the dataset's analytical value and allowed for sector comparison, stability measurement, and predictive modeling.

III. Methodology

A. Method A: Sector Performance Calculation

To compare the performance of all sectors in the S&P 500, the percentage change in the average closing price was computed for each sector. This method is straightforward and provides a simple measure of long-term growth.

Percentage Change = ((Last Close/First Close) - 1) * 100

Reason for Choosing: This method is easy to interpret and allows sectors to be ranked based on overall price appreciation. Since the task only requires comparing performance, a simple percentage change is sufficient.

- Pandas for grouping data and performing calculations.

B. Method B: Stability Analysis Using Standard Deviation

Standard deviation measures how much prices fluctuate around the average.

- Low standard deviation = more stable stock
- High standard deviation = more volatile stock

Reason for Choosing: Standard deviation is one of the most used volatility metrics in finance. It provides a direct way to identify companies with steady price movement, which is helpful for determining stability within a sector.

- Pandas for groupby operations and computing standard deviation.
- NumPy for numerical support.

C. Method C: Price Forecasting (Regression + Exponential Smoothing)

1. Linear Regression Model

A regression model uses past data to predict future prices based on relationships between variables. In this analysis (using NVDA as an example), the model includes:

- Lagged closing prices
- Moving averages (5-day, 10-day, 20-day)
- Basic features such as Open, High, Low, and Volume

2. Exponential Smoothing (ETS Model)

Exponential Smoothing is a time-series method that gives more weight to recent observations. It is effective for forecasting short-term trends because it adjusts smoothly as new data arrives.

3. To increase reliability, the final prediction is an average of both models:

Combined Forecast=0.5*(Regression)+0.5*(ETS)

Reason for Choosing: Regression incorporates technical indicators, while ETS captures time-series trend behavior. By combining them, the model benefits from both

Identify applicable funding agency here. If none, delete this text box.

approaches, improving stability and reducing error for short-term stock forecasting.

- scikit-learn (Linear Regression)
- statsmodels (Exponential Smoothing)
- Pandas and NumPy for feature engineering

IV. Results

A. Result 1: Best Performing Sector

The analysis of all sectors in the S&P 500 based on the percentage change in average closing prices over the period shows that the Communication Services sector is currently performing the best. The table below summarizes the percentage change of all sectors, sorted from highest to lowest

TABLE I. PERCENTAGE CHANGE OF ALL SECTORS

Rank	Sector	Percentage Change (%)
1	Communication Services	966.255416
2	Industrials	668.820152
3	Consumer Cyclical	638.986837
…	…	….
11	Energy	122.914509

This indicates that companies in the Communication Services sector experienced the largest growth in stock prices compared to other sectors during the observed period.

B. Results 2: Sector Trend

The average closing price of companies in the Communication Services sector shows an upward trend over the past year. The line plot below visualizes this trend:

- Observation: The sector's average price has steadily increased, reflecting positive market performance.
- Visualization: The plot displays dates on the x-axis and average close price on the y-axis.

C. Results 3: Most Stable Companies in Top Sector

Within the top-performing sector, the most stable companies were identified using the standard deviation of daily closing prices. Lower standard deviation indicates more consistent performance. Table II lists the top 10 most stable companies in the Communication Services sector, with AT&T Inc. (T) showing the lowest volatility.

TABLE II. MOST STABLE COMPANIES IN COMMUNICATION SERVICES

Rank	Symbol	Company Name	Standard Deviation
1	T	AT&T Inc.	4.13
2	FOX	Fox Corporation	4.37
3	FOXA	Fox Corporation	4.82
…	…	…	…
10	TTWO	Take-Two Interactive Software, Inc	60.61

D. Results 4: Forecasting Next 5 Days (NVIDIA)

Using historical data of NVDA, a combined linear regression and Exponential Smoothing (ETS) model was applied to forecast the next 5 days' closing prices. The predicted prices are as follows:

TABLE III. PREDICTED CLOSING PRICES FOR NVDA

Date	Predicted Closing Price (USD)
2024 – 12 – 21	$131.73
2024 – 12 – 22	$135.23
2024 – 12 – 23	$135.43
2024 – 12 – 24	$137.08
2024 – 12 – 25	$135.76

V. Discussion

Although the analysis successfully identified the top-performing sector, the most stable companies, and provided short-term forecasts, several limitations exist. First, the feature-based regression and ETS models for stock price prediction assume that past price patterns and trends can largely explain future movements. However, stock prices are influenced by unpredictable market events, news, and macroeconomic factors, which may reduce the accuracy of the forecasts. Second, the analysis focused on the most recent 30 trading days, which may limit the robustness of predictions in longer-term horizons. Additionally, while standard deviation provides a simple measure of stability, it does not capture other important aspects such as sudden spikes or volatility clustering.

Future improvements could include incorporating additional features, such as trading volume anomalies, sentiment analysis from financial news, and macroeconomic indicators, to improve prediction accuracy. More sophisticated models, such as ARIMA, LSTM, or ensemble methods, could also be explored for multi-day forecasting. Moreover, using a rolling window with longer historical data could enhance model reliability and reduce sensitivity to short-term fluctuations.

VI. Conclusion

This project analyzed the performance and stability of companies within the S&P 500 index and predicted short-term stock prices for a company in the top-performing sector. The Communication Services sector was found to be the best performing, showing significant gains over the analysis period. Within this sector, AT&T, Fox Corporation, and Comcast were identified as the most stable companies based on their historical price volatility. Using a combination of feature-based regression and Exponential Smoothing (ETS), NVDA's next five trading days' closing prices were forecasted, showing a gradual upward trend consistent with sector performance.

The findings demonstrate that combining historical price patterns with statistical modeling can provide useful

short-term insights for investors and analysts. However, these methods should be applied with caution due to market unpredictability. This analysis can assist portfolio management decisions, sector selection, and risk assessment for financial planning.

ACKNOWLEDGMENT

The author would like to thank all data providers and maintainers of the S&P 500 dataset, which made this analysis possible. Python libraries such as pandas, NumPy, scikit-learn, statsmodels, and seaborn were utilized for data processing, statistical modeling, and visualization.

REFERENCES

[1] [7] L. Larxel, "S&P 500 Stocks Dataset," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks. Accessed: Dec. 6, 2025.

[2] GeeksforGeeks, "Exponential Smoothing for Time Series Forecasting," 2025. [Online]. Available: https://www.geeksforgeeks.org/exponential-smoothing-for-time-series-forecasting/. Accessed: Dec. 6, 2025.

[3] A. Ajiono and T. Hariguna, "Comparison of Three Time Series Forecasting Methods on Linear Regression, Exponential Smoothing and Weighted Moving Average," Int. J. Informatics and Information Systems, vol. 6, no. 2, pp. 89–102, 2023. doi: 10.47738/ijiis.v6i2.165.

[4] M. H. Abdelati and H. A. Abdelwali, "Optimizing simple exponential smoothing for time series forecasting in supply chain management," Indonesian J. of Innovation and Applied Sciences (IJIAS), vol. 4, no. 3, pp. 247–256, 2024.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[6] [5] M. Peixeiro, *Time series forecasting in Python*, Simon and Schuster, 2022.

[7] [6] H. Bhasin, *Machine Learning for Beginners*, BPB Publications, 2020.