

**Master of Data Analytics**

**Data Analytics Case Study**

Summer 2025

DAMO 611-7

Part 1

Bruno Sassi - NF1008627

Maria Aguilar - NF1009877

Luciana Popa - NF1005554

Tiago Ceolato – NF1006634

**Professor**

Milad Rezamand

August 17, 2025

**Table of Contents**

[Introduction 1](#_Toc1538784521)

[Problem Definition 1](#_Toc1599514120)

[Research Question: 1](#_Toc987547816)

[Hypotheses: 1](#_Toc203005868)

[Justification 1](#_Toc321274973)

[Limitations 1](#_Toc273851982)

[Data Collection 1](#_Toc1671156293)

[Variables Collected 1](#_Toc2116970759)

[SQL Exploration 1](#_Toc181075660)

[Python Processing and Preparation 1](#_Toc1980876126)

[Data Understanding 1](#_Toc945216351)

[Data Visualization 1](#_Toc613237674)

[Summary of Data Visualizations 1](#_Toc1994006407)

[References 1](#_Toc744394891)

[Appendix A 1](#_Toc712737383)

# Introduction

Predicting stock prices with accuracy is crucial for classical financial statements. A remarkable stock is Alphabet Inc., with the ticker symbol GOOGL/GOOG, part of Google’s stock. Also, it is important to know that the Alphabet’s revenue comes through Google Ads and YouTube, with other contributions such as Google Cloud and AI tools. (Wang, 2025) The next-day closing price forecasting helps to make informed buy/sell decisions on the stock, identifying profitability and mitigating risks under operations. Alphabet Inc. (GOOGL) is part of the most influential technology firms globally, and it impacts the broader market sentiment.

## Problem Definition

The highly traded and closely watched stock of Alphabet Inc. presents challenges due to volatility and liquidity, making prediction valuable. Price movement carries signals in historical price behavior, trading volume, and other statistical indicators.

### Research Question:

This project aims to answer the following question: Can historical stock prices, trading volume, and technical indicators predict the next day's closing price of GOOGL?

### Hypotheses:

The following hypothesis will be tested based on the research question:

* H1: Lagging prices are highly predictive of the next day’s closing price, with decreasing returns, meanwhile lag increases.
* H2: Averages, RSI, and MACD significantly improve the accuracy of the model.
* H3: Machine learning models surpass traditional methods in predicting stock prices.

## Justification

The choice of Alphabet Inc. (GOOGL) stock is motivated by its global significance, high trading volume, and influence on market sentiment. Accurate prediction of its daily closing price is relevant for a range of stakeholders, including retail investors, institutional portfolio managers, and algorithmic trading firms. The dataset spanning 2020–2025 provides both post-pandemic recovery dynamics and recent AI-driven growth, ensuring sufficient variation to test the hypotheses.

## Limitations

Despite its strengths, the study has several limitations. First, the focus on a single stock restricts generalizability to other assets or sectors. Second, the dataset reflects only market prices and trading activity, without incorporating external macroeconomic or sentiment data, which may also affect returns. Third, daily returns are inherently noisy, with fat-tailed distributions that limit the explanatory power of linear models. Finally, corporate events such as stock splits or dividend adjustments may introduce structural breaks that need careful consideration. These limitations highlight the challenges of stock prediction and motivate the use of advanced, machine learning–based approaches in later phases.

## Data Collection

The dataset consists of historical stock data for Alphabet Inc. (ticker: GOOGL), obtained from Investing.com (n.d.), a platform that provides publicly accessible financial and market information. The selected period, spanning from August 2020 to August 2025, was chosen to capture both short- and long-term trends, ensuring a comprehensive basis for model training and testing.

### Variables Collected

The dataset comprises standard financial indicators widely used in market analysis, covering the period from August 2020 to August 2025. These variables are fundamental for both descriptive and predictive analytics in stock market research.

* **Date (Trading Date):** Records the calendar day of each observation, serving as the time reference for sequencing price movements and analyzing temporal trends.
* **Price (Closing Price):** The final price at which the stock traded on a given day. It is one of the most widely used measures of market performance, as it reflects the consensus valuation at market close (Nasdaq, n.d.-a).
* **Open (Opening Price):** The first traded price of the day, setting the tone for daily market sentiment and often compared to the previous closing price to detect overnight shifts in investor expectations (Nasdaq, n.d.-b).
* **High (Daily High):** The maximum price reached during the trading session, indicating the upper bound of investor willingness to buy on that day.
* **Low (Daily Low):** The minimum price reached during the trading session, reflecting the lower threshold of selling pressure.
* **Volume:** Represents the total number of shares traded during the day. Trading volume is critical for assessing market activity and liquidity; price movements supported by high volumes are generally considered more reliable signals (Hayes, 2025).
* **Change % (Percentage Change):** The day-to-day variation in the closing price relative to the previous day’s close. This metric is essential for measuring volatility and short-term performance, often serving as a proxy for investor reactions to new information or market events (Kenton, 2025; The Investorpedia Team, 2025).

### SQL Exploration

By analyzing the collected data using SQL, it is possible to extract meaningful insights regarding general characteristics, trends, and key highlights. The dataset comprises 1,256 rows, covering the period from August 2020 to August 2025, which defines the timeframe of analysis. Additionally, by retrieving the top 10 rows from the database, we can observe how the information is structured and prepared for deeper exploration.

*“--* ***Initial exploration***

*-- Checking how many rows were imported:*

*SELECT COUNT(\*) AS total\_rows*

*FROM alphabet\_a;*

*-- Checking 10 first rows:*

*SELECT \* FROM alphabet\_a LIMIT 10;*

*-- Checking min and max dates:*

*SELECT MIN(trade\_date) AS first\_date,*

*MAX(trade\_date) AS last\_date*

*FROM alphabet\_a;”*

When examining the dataset’s statistical properties, we can observe the price range within the period analyzed. The maximum stock price was approximately 193% higher than the minimum, highlighting the asset’s volatility. The average stock price across the period was $131.34, a figure that is useful for subsequent analysis, but should not be interpreted isolated without further consideration of potential outliers. Another relevant finding is that the average trading volume was around 32 million.

*“--* ***Basic Statistics***

*-- Min, max an avg price:*

*SELECT MIN(price) AS min\_price,*

*MAX(price) AS max\_price,*

*ROUND(AVG(price),2) AS avg\_price*

*FROM alphabet\_a;*

*-- Avg quantity negociated:*

*SELECT ROUND(AVG(volume),0) AS avg\_volume*

*FROM alphabet\_a;”*

Another interesting observation concerns the largest trades during the period, which can serve as a starting point for investigating potential underlying causes. The top 10 trade volumes reveal that the month of February consistently records higher activity, with most significant trades surpassing 88 million in volume.

When analyzing the top five daily price increases and decreases, we observe that the largest positive change was 10.22%, while the most significant decrease was 9.51%, reinforcing the volatility of Alphabet’s stock.

*“--* ***Market Highlights***

*-- Top 10 bigger trades:*

*SELECT trade\_date, volume*

*FROM alphabet\_a*

*ORDER BY volume DESC*

*LIMIT 10;*

*-- Days with highest increase and decrease:*

*SELECT trade\_date, change\_pct*

*FROM alphabet\_a*

*ORDER BY change\_pct DESC*

*LIMIT 5;*

*SELECT trade\_date, change\_pct*

*FROM alphabet\_a*

*ORDER BY change\_pct ASC*

*LIMIT 5;”*

The consistently rising on yearly average price since 2021 suggests that Alphabet demonstrated resilience during and after the pandemic, reflecting solid business fundamentals. Furthermore, examining the top five monthly averages shows that the distribution of average prices remains relatively balanced across those months.

*“--* ***Trends***

*-- Avg monthly price*

*SELECT DATE\_FORMAT(trade\_date, '%Y-%m') AS month,*

*ROUND(AVG(price),2) AS avg\_price*

*FROM alphabet\_a*

*GROUP BY month*

*ORDER BY avg\_price DESC*

*LIMIT 5;*

*-- Avg yearly price*

*SELECT YEAR(trade\_date) AS year,*

*ROUND(AVG(price),2) AS avg\_price*

*FROM alphabet\_a*

*GROUP BY year*

*ORDER BY year;”*

As part of the SQL analysis, we also examined the daily return of Alphabet’s stock. The top 10 percentage returns highlight the exact dates when the asset delivered its strongest performances. The highest daily return occurred on April 26, 2024 (10.22%), followed by April 9, 2025 (9.68%). It is important to note that the greatest percentage return does not necessarily coincide with the highest stock price, since this calculation reflects relative changes compared to the previous trading day.

The days with extreme returns (both gains and losses) point to periods of intensified market sensitivity, which may be associated with earnings announcements or macroeconomic events.

*“--* ***Daily return***

*SELECT trade\_date,*

*price,*

*ROUND(100 \* (price / LAG(price) OVER (ORDER BY trade\_date) - 1), 2) AS return\_pct*

*FROM alphabet\_a*

*ORDER BY return\_pct DESC*

*LIMIT 10;”*

The SQL exploration provided a comprehensive overview of Alphabet’s stock performance between 2020 and 2025. The results highlight the asset’s volatility, its resilience after the pandemic, and specific moments of abnormal trading activity. This combination of statistical and trend analysis establishes a solid foundation for subsequent visual exploration and hypothesis testing in the next phase of the project.

### Python Processing and Preparation

To ensure consistency and reproducibility, the raw dataset obtained from Investing.com was cleaned and enriched using Python. Several transformations were necessary before the data could be analyzed and visualized:

* **Cleaning:** Converted the Date column to datetime format, standardized the trading volume (Vol.) by parsing millions and billions into numeric values, and stripped the percentage symbol from Change %.
* **Derived Features:** Created additional variables essential for analysis and modeling, including daily returns, moving averages (10, 20, 50, and 200 days), technical indicators such as the Relative Strength Index (RSI) and the Moving Average Convergence Divergence (MACD), as well as a 20-day rolling volatility measure.
* **Lagged Features:** Introduced lagged price variables (1-day and 5-day) to capture short-term dependencies.
* **Export:** The cleaned and enriched dataset was saved as a single CSV file (alphabet\_prepared.csv), which serves as the input for Tableau dashboards and the predictive modeling stage.

The following Python code excerpt illustrates the preparation process:

*“import pandas as pd*

*#Load dataset*

*df = pd.read\_csv("alphabet\_raw.csv")*

*#Cleaning*

*df['Date'] = pd.to\_datetime(df['Date']) df['Vol.'] = df['Vol.'].str.replace('M','e6').str.replace('B','e9').astype(float) df['Change %'] = df['Change %'].str.replace('%','').astype(float) / 100*

*#Derived features*

*df['Return'] = df['Price'].pct\_change() df['MA50'] = df['Price'].rolling(50).mean() df['MA200'] = df['Price'].rolling(200).mean()*

*#Technical indicators (example: RSI)*

*delta = df['Price'].diff() gain, loss = delta.clip(lower=0), -delta.clip(upper=0) rs = gain.rolling(14).mean() / loss.rolling(14).mean() df['RSI'] = 100 - (100 / (1 + rs))*

*#Export clean dataset*

*df.to\_csv("alphabet\_prepared.csv", index=False)”*

This automated pipeline guarantees that the dataset is replicable, consistent, and ready for both visualization and predictive modeling, aligning with the course requirement of minimizing manual adjustments.

# Data Understanding

The purpose of the data understanding stage is to conduct a thorough examination of the collected dataset, to identify the key metrics, and to uncover patterns that are relevant for the research question. This step provides the foundation for both hypotheses testing and the development of predictive models. The dataset, which covers Alphabet Inc. (GOOGL) stock between August 2020 and August 2025, contains 1,256 daily observations with standard financial variables such as opening price, closing price, daily high and low, trading volume, and daily percentage change.

An initial descriptive analysis reveals that the closing price fluctuated between a minimum of approximately USD 83 and a maximum of USD 244, representing a 193 percent variation over the period. The average closing price during these five years was USD 131.34, although this value should be interpreted cautiously, as it is influenced by periods of sharp volatility. Trading activity was also significant, with a daily average of about 32 million shares, and with peaks above 88 million during specific market events. The day-to-day changes in closing price confirm the stock’s volatility, as the most pronounced increase reached 10.22 percent while the most severe decline was 9.51%.

In terms of temporal patterns, the long-term trajectory indicates consistent growth in average annual prices beginning in 2021. This trend reflects Alphabet’s resilience after the pandemic and the strengthening of its business fundamentals. Monthly averages, on the other hand, reveal that February tends to concentrate higher trading activity and more pronounced price movements, which is likely associated with the timing of quarterly earnings announcements. Several extreme daily returns, such as those observed on April 26, 2024 (10.22%), and April 9, 2025 (9.68%), coincide with corporate events or broader macroeconomic shocks. The sharp increase on April 26, 2024, aligns with Alphabet’s official announcement of the “Internet Availability of Proxy Materials for its 2024 Annual Meeting of Stockholders,” which was publicly released on the company’s investor relations website (Alphabet, 2024). This illustrates how corporate disclosures can act as immediate catalysts for significant shifts in market sentiment and stock performance.

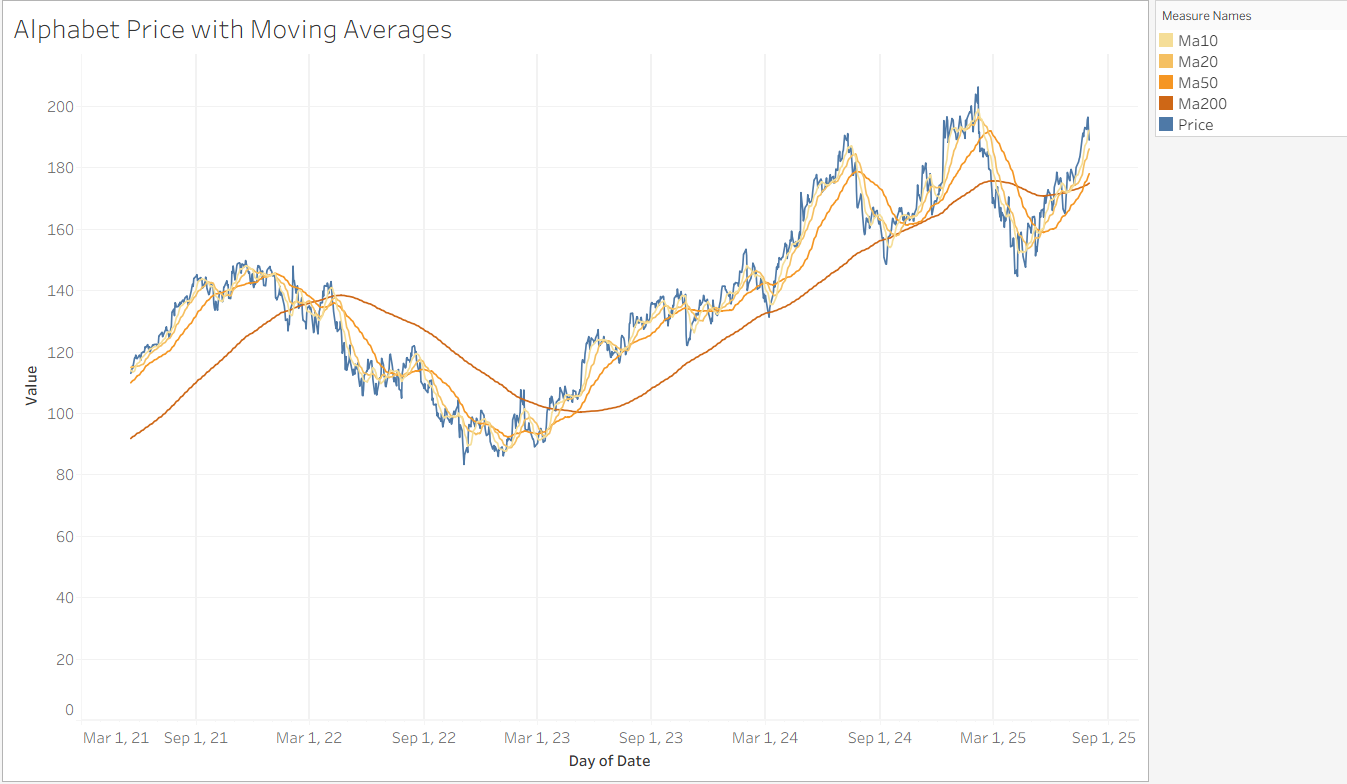
The positive return of April 9, 2025, requires additional contextualization. On that date, Alphabet released its Q1 2025 earnings report, which highlighted a 12 percent increase in revenues, reinforcing the company’s growth trajectory and helping to explain the strong positive market reaction. Interestingly, the date also carries symbolic interpretations outside the financial domain. In the provided context, April 9 can be associated with the letter “D” in the alphabet, reflecting cultural references such as the ABC News Live Prime broadcast schedule. On the same day, the popular puzzle game Wordle featured “WHEAT” as its solution, further anchoring the date in collective memory. While these cultural references are not direct drivers of financial performance, they underscore how symbolic associations and external narratives can intertwine with the financial calendar, providing additional texture to the analysis of market events.

The exploratory analysis also points to meaningful relationships among variables. There is clear evidence that abnormally high trading volumes tend to accompany significant price movements, which reinforces the notion that liquidity plays a central role in market volatility. Prices such as open, high, and low show strong co-movement with the closing price, indicating that lagged price variables may be highly predictive. Furthermore, the observed increases in the daily price range during periods of intense trading activity suggest that higher volatility is systematically linked with liquidity spikes.

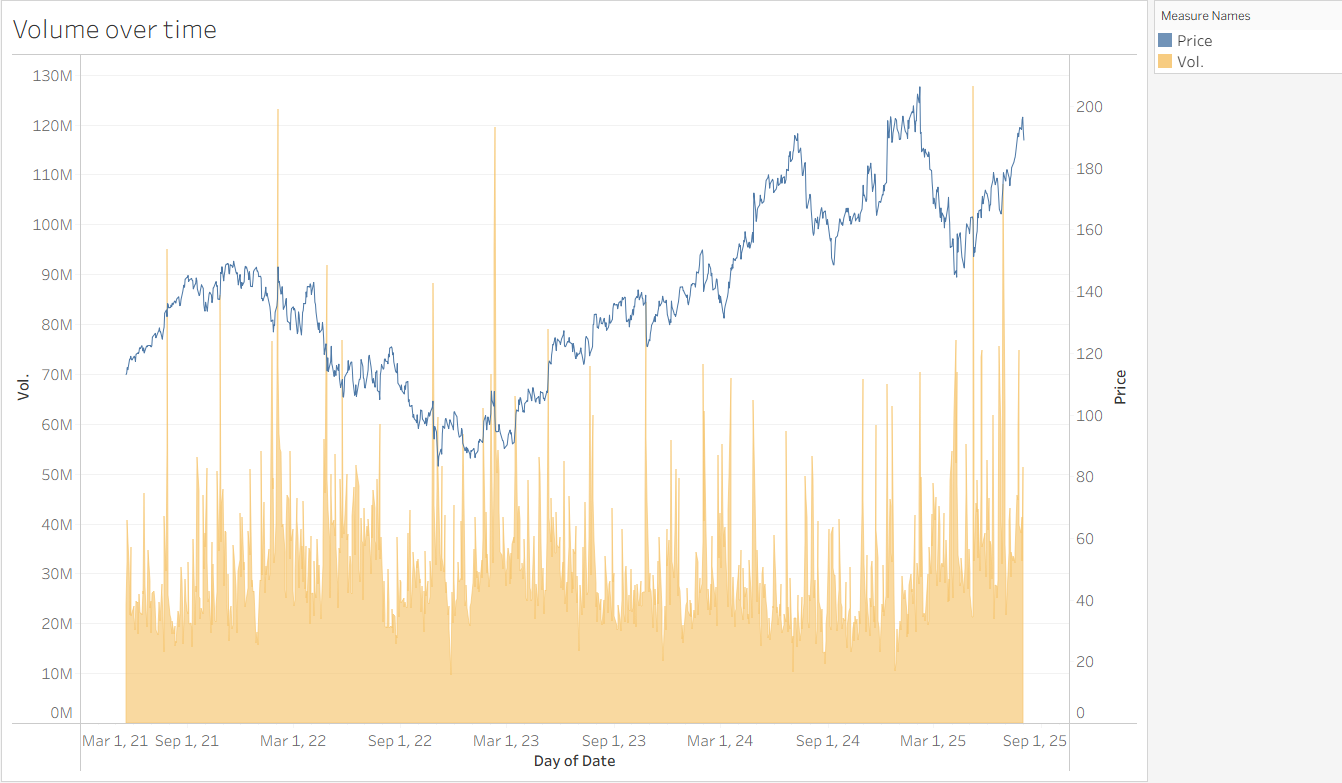
Taken together, these findings support the hypotheses initially formulated. The historical behavior of lagged prices appears to be a strong predictor of the following day’s closing value. Indicators such as moving averages, RSI, and MACD are expected to capture persistent volatility and improve forecasting accuracy. Finally, the presence of nonlinearities and complex interactions between price and volume highlights the potential advantage of machine learning methods over traditional statistical models in predicting Alphabet’s stock price.

# Data Visualization

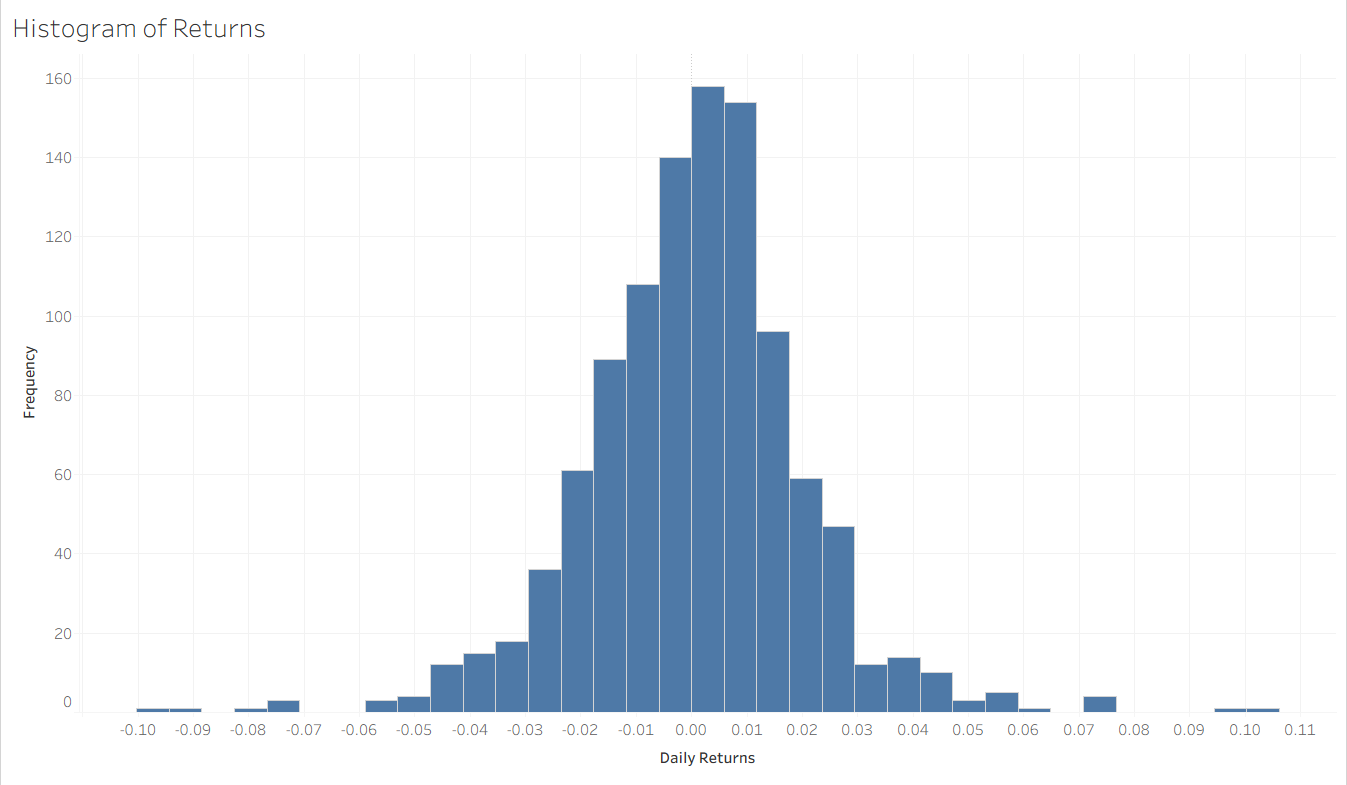
To complement the data exploration, a set of visualizations was developed using Tableau and Python. These visualizations illustrate key trends, patterns, and relationships within the dataset and provide evidence to support the hypotheses defined earlier.

**Figure 1.** Price with Moving Averages (MA10, MA20, MA50, MA200)

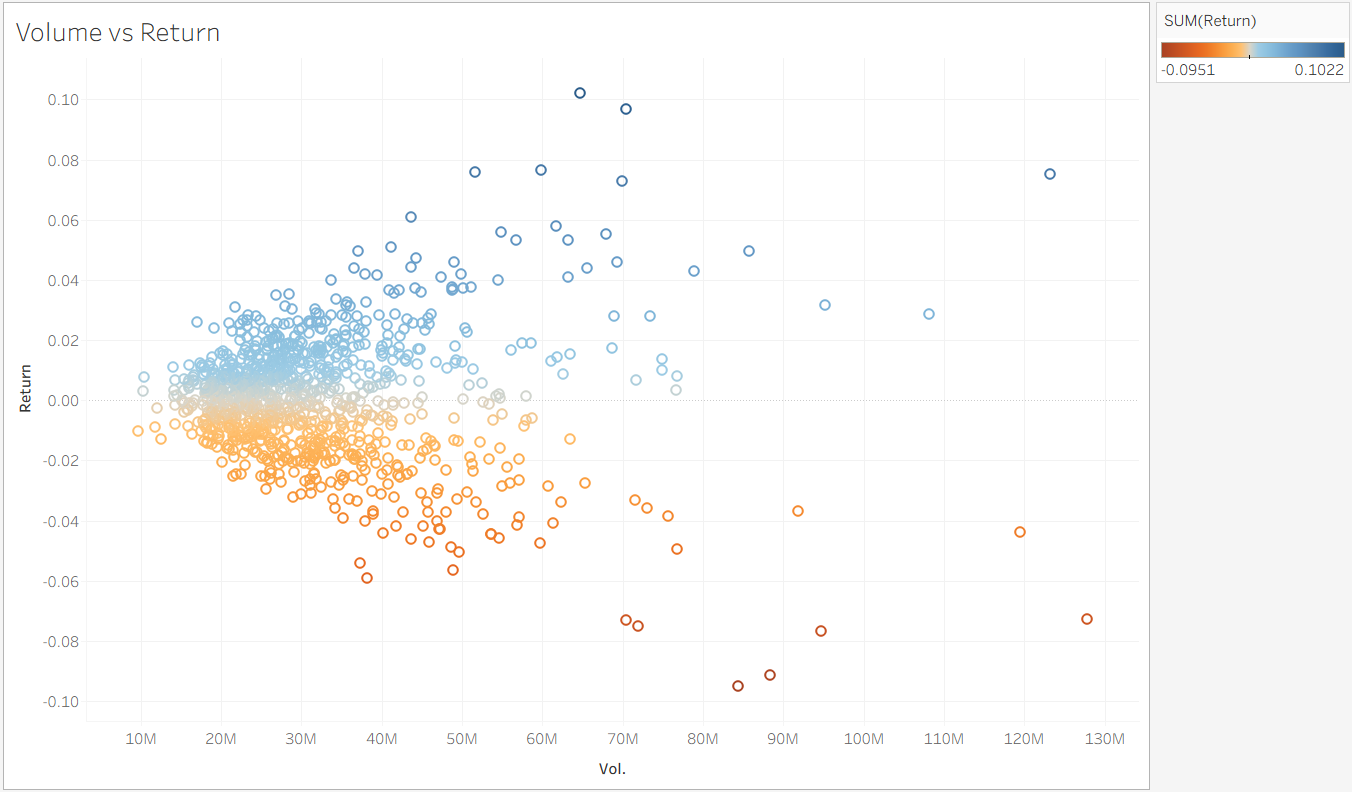
This chart plots Alphabet’s daily closing price alongside four moving averages with different time horizons. The short-term averages (MA10, MA20) closely follow the stock’s daily oscillations, while the MA50 and MA200 reveal medium- and long-term trends. The visualization highlights the cyclical nature of Alphabet’s stock price and shows that lagged prices and smoothed averages are strongly correlated, supporting Hypothesis 1 (lagging prices are predictive of the next day’s closing price).

**Figure 2.** Volume over Time (with Price Overlay)

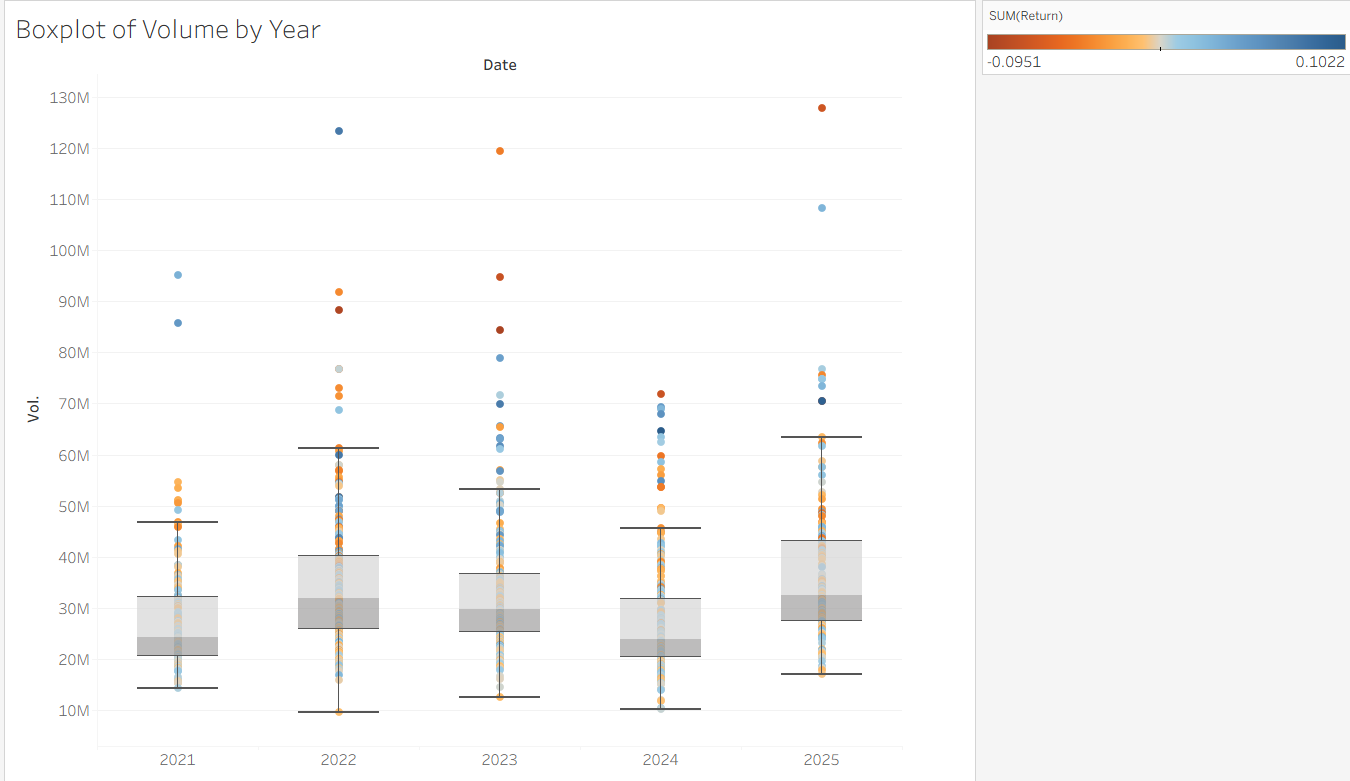
This visualization combines daily trading volume (area) with closing price (line). Peaks in volume often coincide with sharp price movements, reinforcing the importance of liquidity as a driver of volatility. The plot also shows consistently higher trading activity during specific months, such as February, which aligns with earnings announcements and supports earlier SQL findings.

**Figure 3.** Histogram of Daily Returns

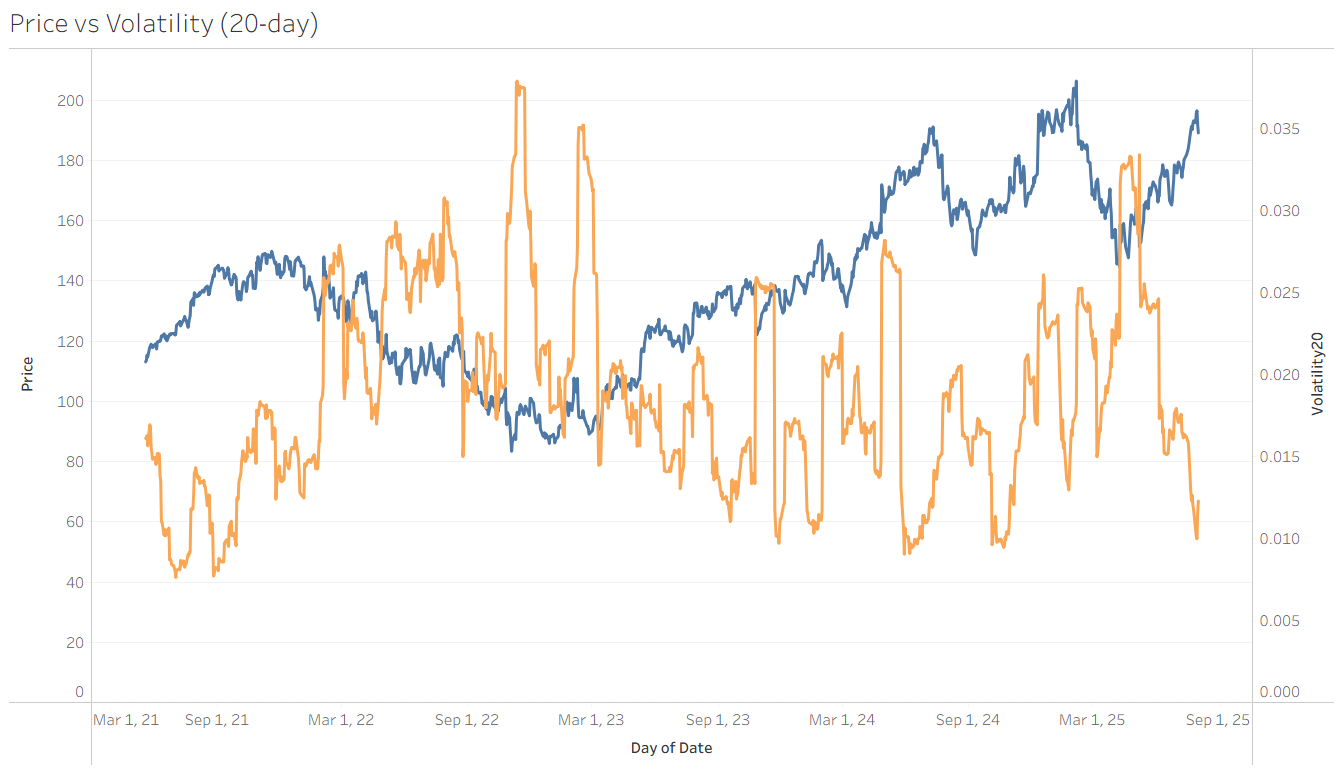
The histogram shows the distribution of Alphabet’s daily returns. Most observations cluster around zero, reflecting stable day-to-day performance, but the fat tails indicate frequent extreme events in both directions. The asymmetric distribution of returns underscores the risks associated with volatility and justifies the inclusion of technical indicators such as RSI and MACD (Hypothesis 2).

**Figure 4.** Scatterplot of Volume vs. Daily Return

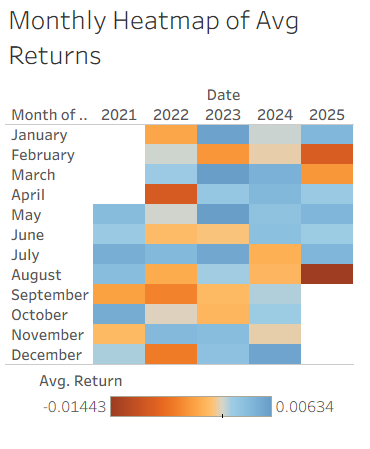
This scatterplot investigates the relationship between trading volume and daily returns. The color gradient highlights positive (blue) and negative (orange) returns. While most daily returns occur at moderate volumes, extreme returns are typically associated with unusually high trading activity. This visualization emphasizes the connection between liquidity and volatility, confirming that volume plays a critical role in market dynamics.

**Figure 5.** Boxplot of Volume by Year

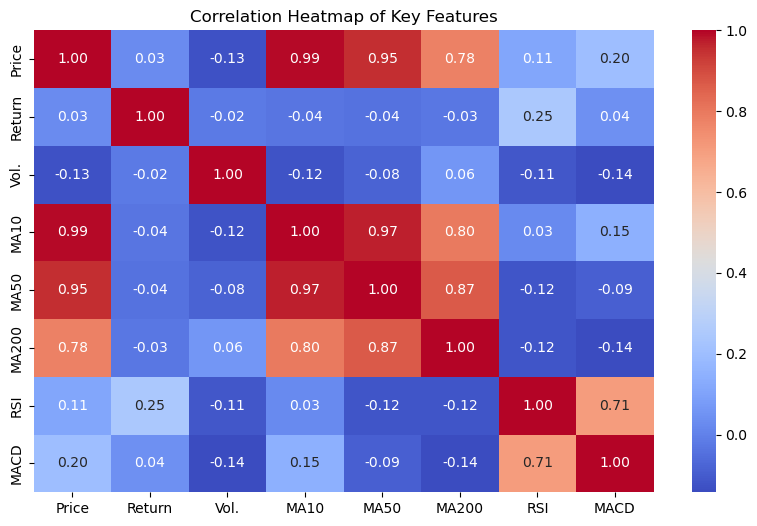
The boxplot summarizes the distribution of daily trading volume across years. While median volumes remain relatively stable, outliers indicate sporadic periods of very high activity, particularly in 2022 and 2025. This visualization contextualizes the scatterplot by showing how annual liquidity patterns evolve and helps identify years with unusual market sensitivity.

**Figure 6.** Price vs. Rolling Volatility (20-day)

This dual-axis plot contrasts the daily closing price with a 20-day rolling volatility measure. Periods of sharp price increases or declines correspond to volatility spikes, reflecting heightened market uncertainty. The inverse relationship between stable growth phases and lower volatility is evident, providing empirical support for using volatility as a predictive feature in stock modeling.

**Figure 7.** Monthly Heatmap of Average Returns

The heatmap shows the average monthly returns of Alphabet’s stock across the five-year period. A clear seasonal pattern emerges: February tends to be one of the weakest months, with consistently negative average returns across multiple years. In contrast, March and April often exhibit stronger performance, likely influenced by the timing of Q1 earnings releases and positive market sentiment early in the year. Other months show mixed results, indicating that while seasonal effects exist, they are not uniform across the calendar. This suggests that month-of-year may provide incremental predictive value but should not be relied upon in isolation.

**Figure 8.** Correlation Heatmap of Key Features

The correlation matrix quantifies the relationships between price, returns, volume, and technical indicators. The results show:

* Price vs. Moving Averages (MA10, MA50, MA200) → near-perfect positive correlations, confirming that lagged price levels are highly predictive of future price behavior (supporting Hypothesis 1).
* Price vs. RSI/MACD → weak-to-moderate correlations, suggesting that momentum indicators contribute complementary but nonlinear predictive signals (supporting Hypothesis 2).
* Returns vs. All Other Variables → correlations are very close to zero, underscoring the noise in daily returns and justifying the need for advanced predictive models (supporting Hypothesis 3).
* RSI vs. MACD → moderate correlation (~0.7), indicating some overlap but not redundancy between the two momentum measures.

## Summary of Data Visualizations

Taken together, the visualizations provide a comprehensive understanding of Alphabet’s stock behavior. Long-term and short-term moving averages confirm the predictive power of lagged prices, while volume and volatility plots highlight the role of liquidity in amplifying price swings. The histogram, scatterplots, and boxplots illustrate the asymmetric and fat-tailed distribution of returns, underscoring the challenges of prediction. Seasonal effects, though not uniform, show that specific months (notably February with weaker performance and March to May with stronger returns) carry distinctive patterns worth considering. Finally, the correlation heatmap emphasizes the dominance of lagged prices in linear relationships and the necessity of more advanced modeling to capture nonlinear interactions among technical indicators and returns. These insights collectively reinforce the three hypotheses and set a solid foundation for the predictive modeling phase.

# References

Alphabet. (2024, April 26). *Alphabet announces Internet availability of proxy materials for its 2024 annual meeting of stockholders*. Retrieved from <https://abc.xyz/2024-0426/>

Hayes, A. (2025, August 13). *Understanding stock volume: A key indicator for investors*. Investopedia. <https://www.investopedia.com/terms/v/volume.asp>

Investing.com. (n.d.). *Alphabet Inc. (GOOGL) historical prices*. Investing.com. <https://www.investing.com/equities/google-inc-historical-data>

Kenton, W. (2025, May 24). *How to calculate a percentage change*. Investopedia. <https://www.investopedia.com/terms/p/percentage-change.asp>

Nasdaq. (n.d.-a). *Closing price definition*. Nasdaq. <https://www.nasdaq.com/glossary/c/closing-price>

Nasdaq. (n.d.-b). *Opening price definition*. Nasdaq. <https://www.nasdaq.com/glossary/o/opening-price>

The Investorpedia Team. (2025, May 1). *How to calculate the percentage gain or loss on an investment*. Investopedia. <https://www.investopedia.com/ask/answers/how-do-you-calculate-percentage-gain-or-loss-investment/>

Wang, Z. (2025). *Proceedings of the 2025 International Conference on Financial Risk and Investment Management (ICFRIM 2025).* Australia: University of New South Wales. doi:10.2991/978-94-6463-748-9\_64

# Appendix A

SQL Exploration Outputs

This appendix presents the outputs generated from the SQL queries conducted during the initial data exploration phase. These figures provide descriptive insights into Alphabet’s stock dataset, including overall size, time coverage, summary statistics, market highlights, and return calculations. They serve as supporting evidence for the findings described in the Data Collection and Data Understanding sections.

**Figure A1.** *Dataset Total Rows*

  
Note. Initial Exploration showing total rows.

**Figure A2.** *Dataset Min. and Max. Dates*  


**Figure A3.** *Dataset first 10 rows to check data*  
A screenshot of a data table

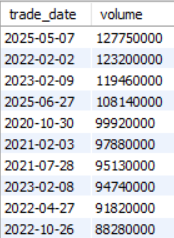
AI-generated content may be incorrect.

**Figure A4.** *Price range and average*  
  
*Note.* Basic Statistics.

**Figure A5.** *Average volume negotiated*  
A close up of a number

AI-generated content may be incorrect.

**Figure A6.** *Top 10 bigger stock trades*

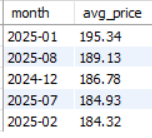
  
*Note*. Market volume highlights.

**Figure A7.** *Days with highest price increase*  
A screenshot of a computer screen

AI-generated content may be incorrect.

**Figure A8.** *Days with highest price decrease*  
A screenshot of a computer

AI-generated content may be incorrect.

**Figure A9.** *Average monthly price*   
  
*Note*. Average price to identify trends.

**Figure A10.***Average yearly price*  


**Figure A11.** *Top 10 daily percentual return*  
  
*Note*. Highlight of peak daily returns.