

INTELLIHACK 5.0

Task 04_Part 01



Team Cognic AI



EXPLORATORY DATA ANALYSIS

Stock Price Prediction

Introduction

Exploratory Data Analysis (EDA) is a critical step in any data-driven project, particularly in the context of stock price prediction. It involves the systematic examination of data to uncover patterns, trends, relationships, and anomalies that can inform the development of predictive models. In the realm of financial markets, where data is often noisy, non-stationary, and influenced by a multitude of external factors, EDA serves as the foundation for understanding the underlying structure of the data and making informed decisions.

The Role of EDA in Stock Price Prediction

Stock price prediction is a complex task that requires a deep understanding of historical price movements, market behavior, and external factors such as economic indicators, news events, and investor sentiment. EDA plays a pivotal role in this process by:

1. Understanding Data Structure:

- EDA helps us understand the structure of the dataset, including the types of variables (e.g., numerical, categorical), their distributions, and their relationships. For stock price data, this involves analyzing time series components such as trends, seasonality, and volatility.

2. Identifying Patterns and Trends:

- By visualizing historical stock prices and related metrics (e.g., opening price, closing price, high, low, volume), EDA reveals patterns and trends that can inform predictive models. For example, identifying long-term growth trends or cyclical patterns can help in forecasting future price movements.

3. Detecting Anomalies and Outliers:

- Financial data often contains anomalies or outliers caused by market crashes, sudden price spikes, or data errors. EDA helps detect these anomalies, which can be addressed before model training to improve prediction accuracy.

4. Feature Engineering:

- EDA guides the creation of meaningful features from raw data. For stock price prediction, features such as moving averages, rolling volatility, and price differences can be derived to capture important aspects of market behavior.



5. **Assumptions and Limitations:**

- EDA helps validate assumptions about the data, such as stationarity, normality, and correlation between variables. It also highlights limitations, such as missing data or insufficient historical data, which can impact model performance.

Key Components of EDA in Stock Price Prediction

In this project, the EDA process focuses on the following key components:

1. **Time Series Analysis:**

- Analyzing trends, seasonality, and volatility in stock prices over time.
- Decomposing the time series into trend, seasonal, and residual components to understand underlying patterns.

2. **Volatility Analysis:**

- Examining the variability of stock prices using metrics such as rolling standard deviation and quantile analysis.
- Identifying periods of high and low volatility, which are crucial for risk management and trading strategies.

3. **Correlation Analysis:**

- Investigating relationships between different features (e.g., opening price, closing price, high, low) to identify potential predictors for the target variable (e.g., future closing price).

4. **Distribution Analysis:**

- Analyzing the distribution of stock prices, returns, and other metrics to understand their statistical properties (e.g., skewness, kurtosis).

5. **Anomaly Detection:**

- Identifying and addressing outliers or anomalies in the data that could distort model predictions.



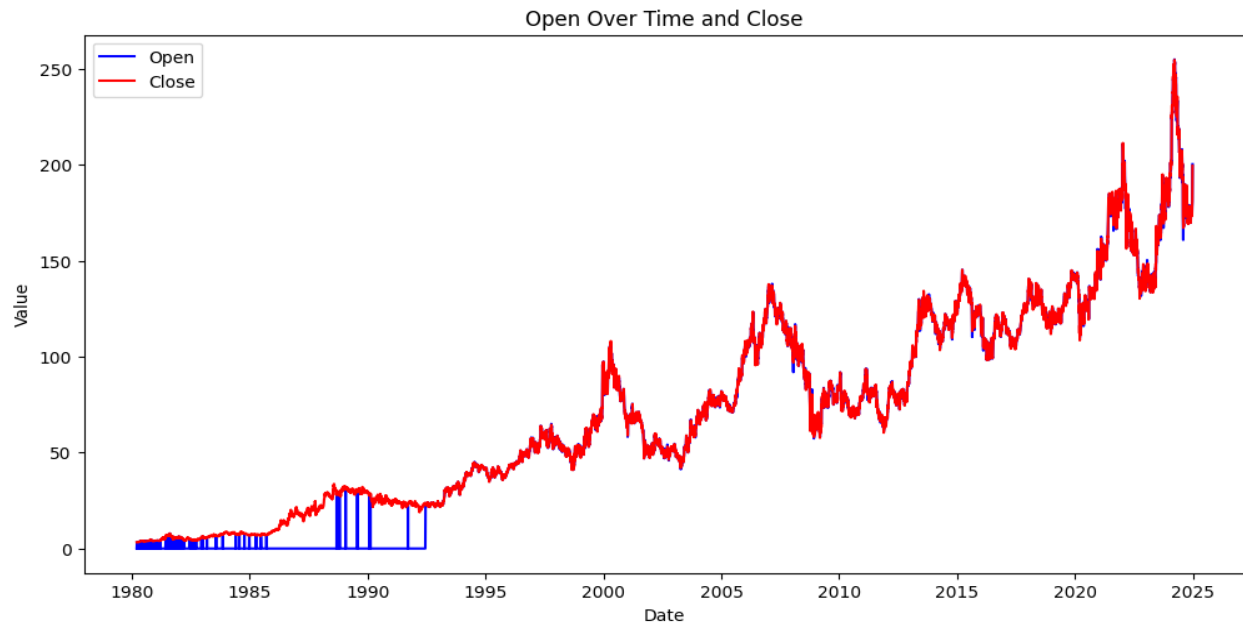
Objective of This EDA Report

The primary objective of this EDA report is to provide a comprehensive understanding of the historical stock price dataset, uncover insights that can guide feature engineering and model selection, and identify potential challenges in the data. By the end of this analysis, we aim to:

- Gain a clear understanding of the dataset's structure and characteristics.
- Identify key trends, patterns, and anomalies in stock price movements.
- Generate actionable insights for building a robust stock price prediction model.

This report will serve as the foundation for the subsequent stages of the project, including feature engineering, model development, and evaluation. Through a combination of visualizations, statistical analysis, and domain knowledge, we will extract meaningful insights from the data and lay the groundwork for a successful stock price prediction system.

Plot Features Over Time



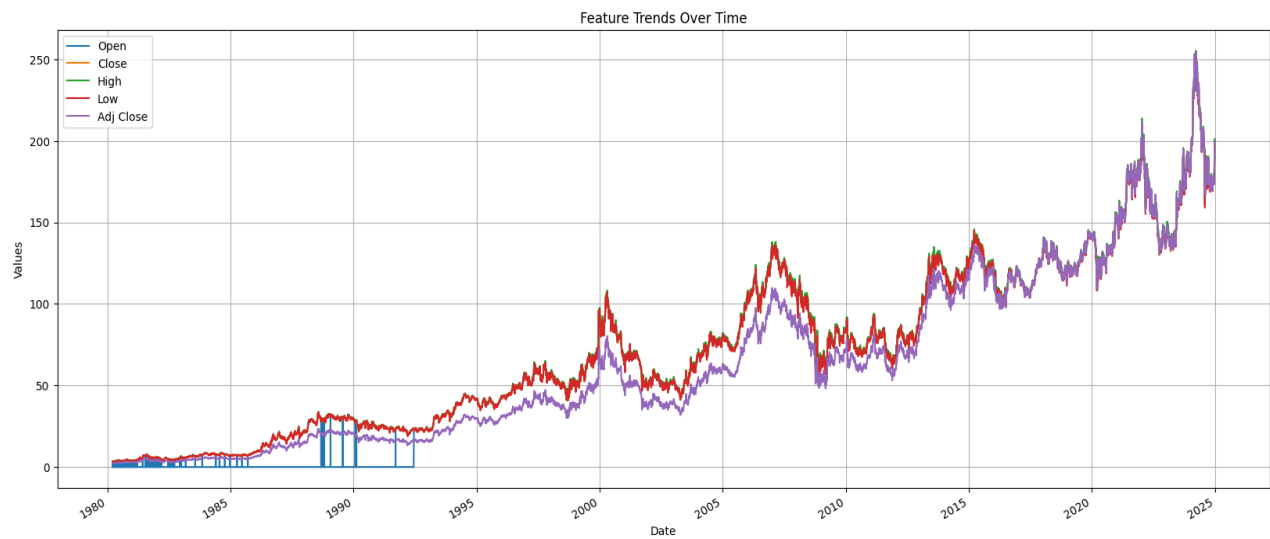
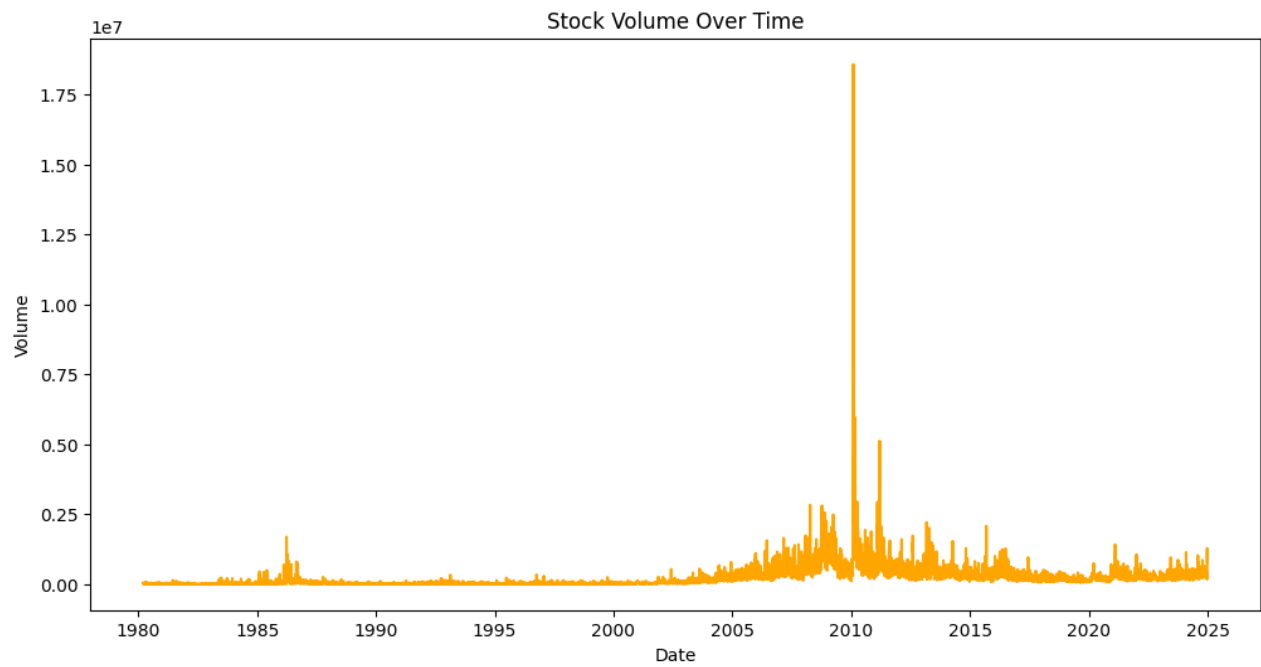
- The graph shows the **Open** and **Close** prices of the stock and **Volume** (number of shares traded) over time..

Key Observations:

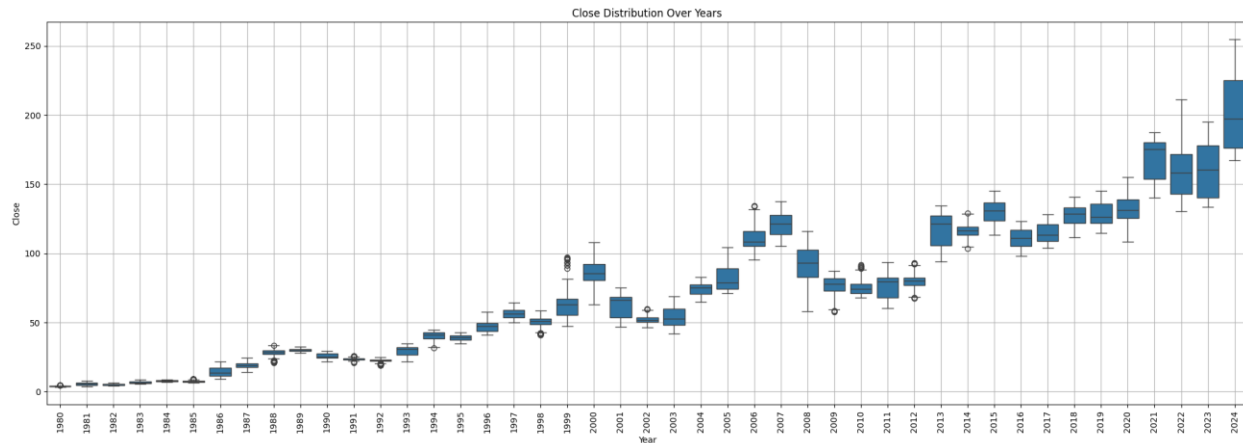
- The **Open** and **Close** lines are very close to each other, meaning the stock price at the start of the day (Open) and the end of the day (Close) are similar.
- The stock price was very low in the 1980s (around 0-50) but started increasing after 1990.
- After 2000, the stock price grew significantly, reaching its highest point around 2020-2025.
- The **Volume** was very low in the 1980s and 1990s (close to 0).
- After 2000, the **Volume** started increasing, and it grew significantly after 2010.
- The highest **Volume** is seen around 2020-2025.

Insights:

- The stock price has grown a lot over time, especially after 2000.
- The small gap between **Open** and **Close** means the stock price doesn't change much during the day.
- More people started trading the stock after 2000, and trading activity increased a lot after 2010.
- The increase in **Volume** matches the increase in stock price, which means more people were interested in buying and selling the stock as its price went up.



Close Price Distribution over Year



- This is a boxplot showing the distribution of **closing prices** over different years.
- **X-axis:** Years (1980 - 2024)
- **Y-axis:** **Closing prices** of the stock.

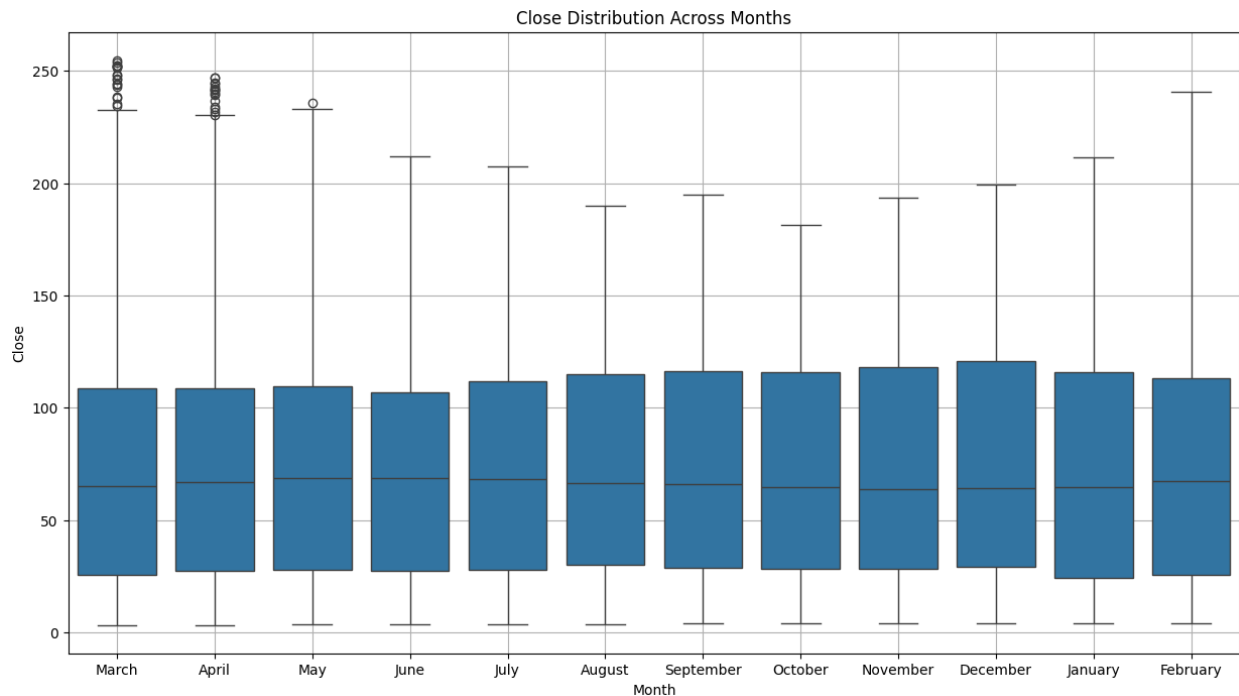
Key Insights

- **Steady Increase:** The closing prices show a clear upward trend over the decades, indicating long-term stock growth.
- **Higher Volatility in Recent Years:** The spread has widened significantly, especially after 2020, suggesting increased price fluctuations.
- **Market Crashes & Recoveries:**
 - The 2008 **financial crisis** caused a sharp decline in stock prices.
 - A noticeable drop in 2020, followed by a strong recovery.
- **Outliers:** Many years have outliers, especially during market booms and crashes.

What This Means

- **Investors** should be aware of periods of high volatility.
- The **median closing price** has been increasing, indicating strong long-term performance.
- The presence of **outliers** suggests moments of extreme highs/lows, possibly due to market corrections or economic events.

Monthly Variation and Monthly Trends



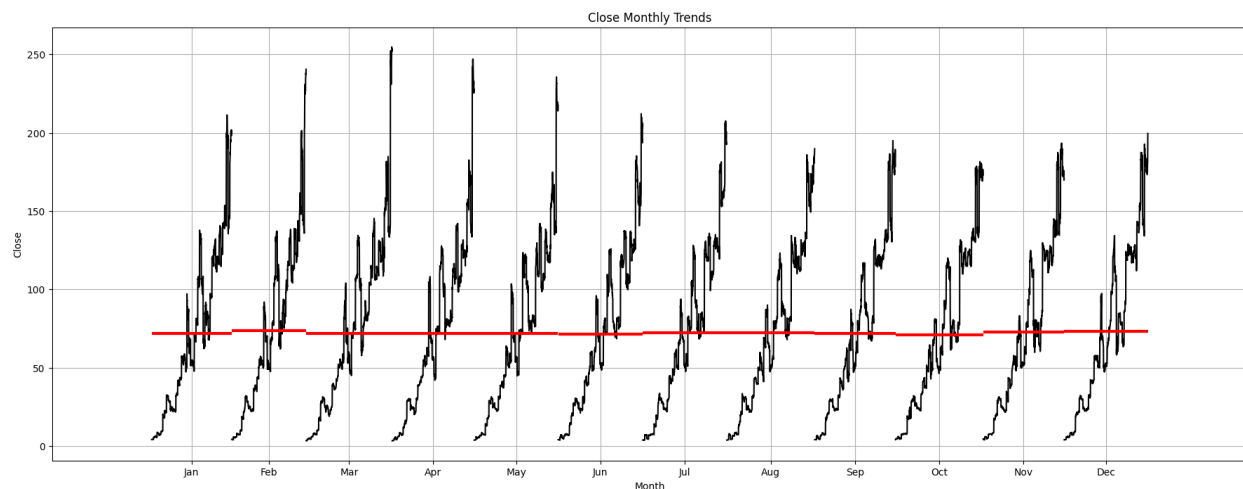
- The graph shows the distribution of closing prices (likely the adjusted closing prices) across each month of the year.

Visualizations of key patterns and relationships:

- The graph allows us to observe how the closing prices are distributed across different months. It can help identify if certain months tend to have higher or lower closing prices compared to others.
- The height of the bars indicates the frequency of closing prices within a specific range for each month.

Analysis of trends, seasonality, and anomalies:

- **Trend:** If certain months consistently show higher or lower closing prices, it could indicate a seasonal trend. For example, if December consistently has higher closing prices, it might suggest a year-end rally.
- **Seasonality:** The graph can help identify seasonal patterns in the stock price. For instance, if the closing prices are consistently higher in certain months (e.g., December) and lower in others (e.g., September), it could indicate seasonality.
- **Anomalies:** Any month that significantly deviates from the general pattern (e.g., an unusually high or low bar) could indicate an anomaly. For example, if March shows an unusually high frequency of high closing prices, it might be worth investigating further.



- The graph shows the monthly trends of the **"Close"** price of the stock over time.
- The **x-axis** represents the **months** of the year (January to December).
- The **y-axis** represents the **stock price** (Close price).

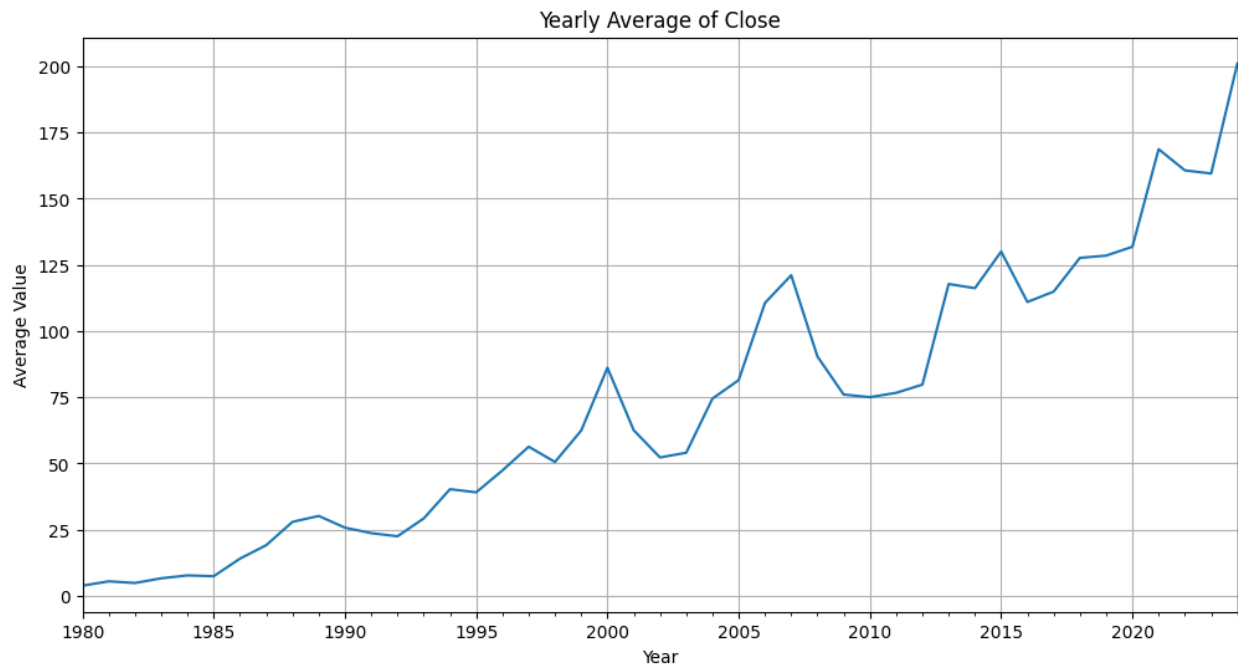
Visualizations of Key Patterns and Relationships:

- The graph shows how the stock's closing price behaves in different months of the year.
- Each line represents a different year, and the lines are grouped by month.

Analysis of Trends, Seasonality, and Anomalies:

- **Trends:** The closing price tends to fluctuate throughout the year. Some months show higher prices, while others show lower prices.
- **Seasonality:** There might be some seasonal patterns:
 - The stock price tends to be higher in certain months (e.g., towards the end of the year, like November or December).
 - The price might be lower in other months (e.g., mid-year, like June or July).
- **Anomalies:** Some years show unusual spikes or drops in certain months, which could be due to external factors like market events or news.

Yearly Average and Quartey Average



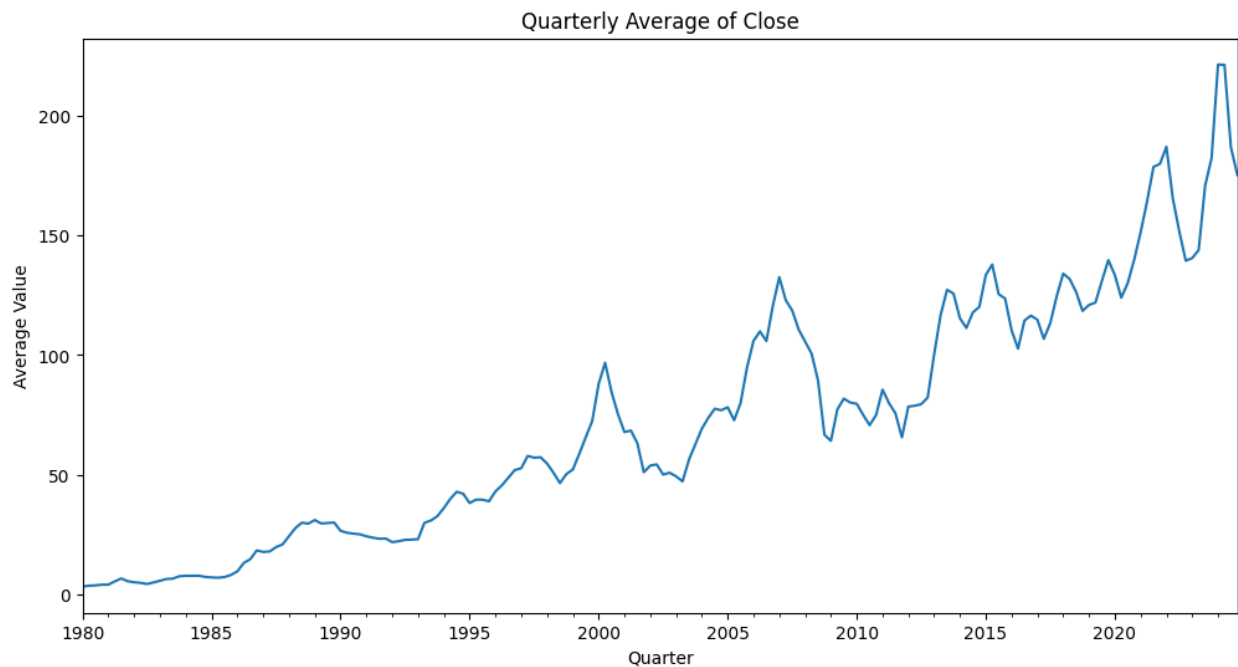
- The graph shows the yearly average of the closing price (likely the adjusted closing price) from 1980 to 2020.

Visualizations of key patterns and relationships:

- The graph allows us to observe the long-term trend of the stock's closing price over several decades. It helps identify whether the stock price has generally increased, decreased, or remained stable over time.
- The line or bars (depending on the graph type) show the average closing price for each year, allowing us to compare the performance of the stock across different years.

Analysis of trends, seasonality, and anomalies:

- **Trend:** The graph can reveal long-term trends in the stock price. For example, if the average closing price consistently increases over the years, it indicates a bullish trend. Conversely, a consistent decrease would indicate a bearish trend.
- **Seasonality:** While this graph focuses on yearly averages, it might not directly show seasonality. However, if there are recurring patterns (e.g., certain years consistently show higher or lower averages), it could hint at cyclical behavior.
- **Anomalies:** Any significant spikes or drops in the average closing price for a particular year could indicate anomalies. For example, if the average closing price in 2008 is significantly lower than surrounding years, it might reflect the impact of the financial crisis.



- The graph shows the **quarterly average** of the "**Close**" price of the stock over time.
- The **x-axis** represents the **quarters** (Q1, Q2, Q3, Q4) across different years.
- The **y-axis** represents the **average closing price** of the stock.

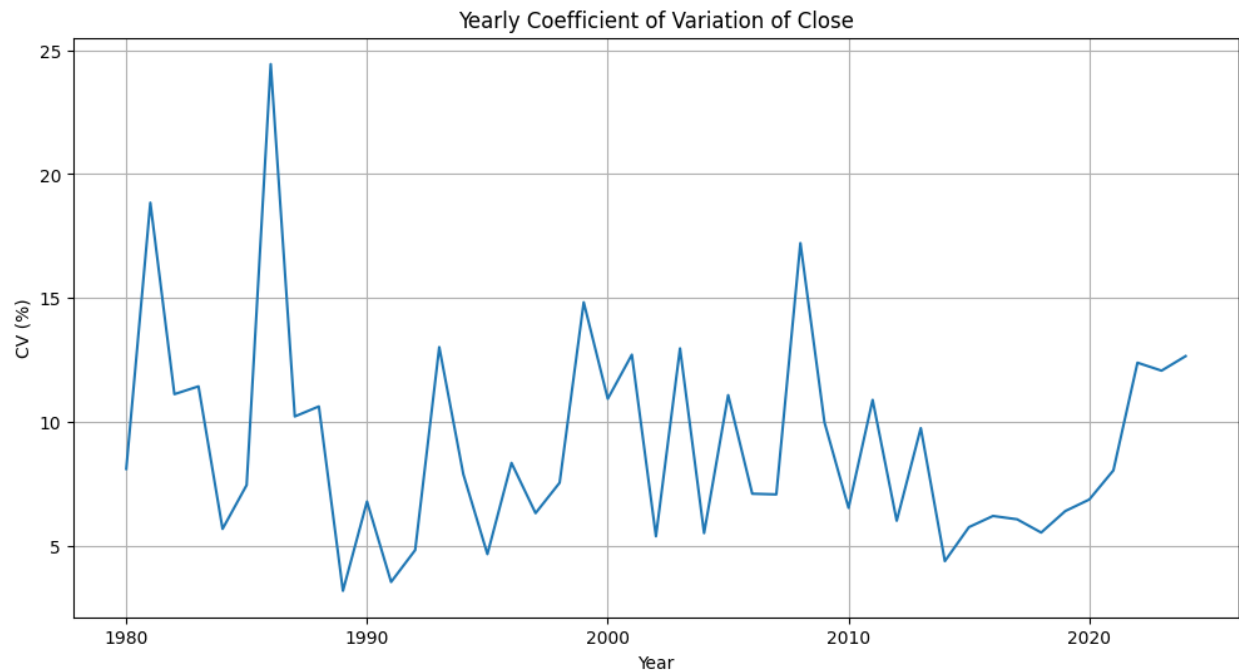
Visualizations of Key Patterns and Relationships:

- The graph shows how the average closing price of the stock changes over each quarter (3-month period) across multiple years.
- Each point or line represents the average closing price for a specific quarter.

Analysis of Trends, Seasonality, and Anomalies:

- **Trends:** The average closing price generally increases over time, especially after the year 2000. This indicates that the stock has grown in value over the long term.
- **Seasonality:** There might be some seasonal patterns within each year:
 - For example, the stock price might be higher in certain quarters (e.g., Q1, Q4) and lower in others (e.g., Q2, Q3). There might be seasonal trends within each year, such as higher prices in begin and end of the year, lower prices in Q2 mid-year.
- **Anomalies:** There are no major anomalies (unusual spikes or drops) in the graph, but the steady increase in price after 2000 is a notable trend.

Coefficient of Variation



- The graph shows the yearly coefficient of variation (CV) of the closing price (likely the adjusted closing price) from 1980 to 2020. The x-axis represents the years, and the y-axis represents the CV in percentage terms.
- The CV measures the relative volatility of the stock's closing price for each year. A higher CV indicates greater volatility relative to the mean closing price, while a lower CV indicates more stability.

Visualizations of key patterns and relationships:

- The graph allows us to observe how the volatility of the stock's closing price has changed over time. It helps identify periods of high or low relative volatility.
- The line or bars (depending on the graph type) show the CV for each year, allowing us to compare the volatility of the stock across different years.

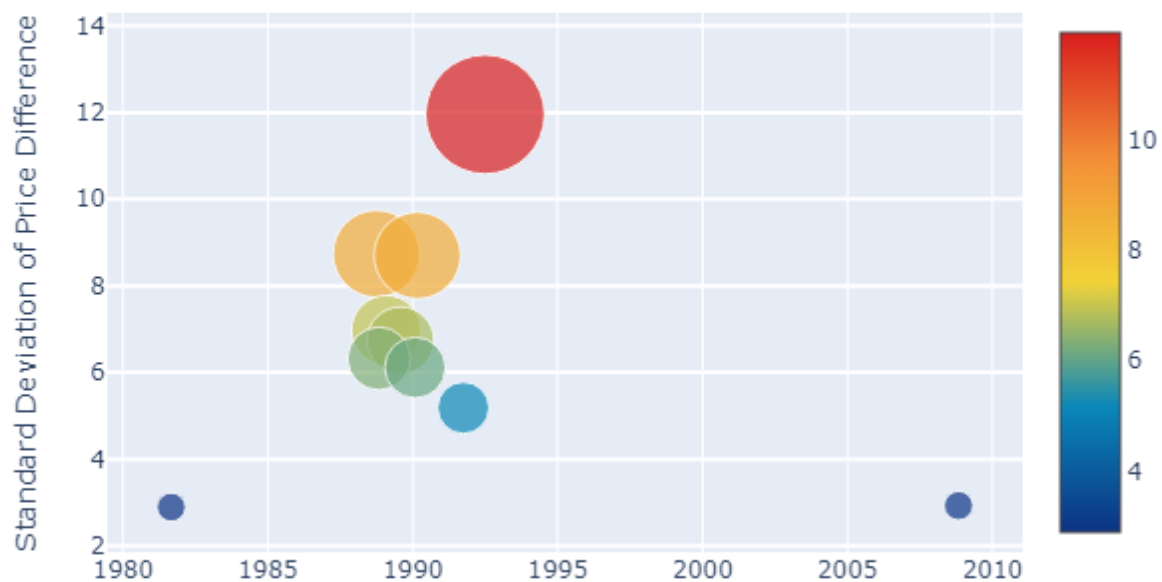
Analysis of trends, seasonality, and anomalies:

- **Trend:** The graph can reveal trends in the stock's volatility over time. For example, if the CV consistently increases over the years, it indicates that the stock has become more volatile relative to its mean price. Conversely, a decreasing trend would indicate that the stock has become more stable.

- **Seasonality:** While this graph focuses on yearly CV, it might not directly show seasonality. However, if there are recurring patterns (e.g., certain years consistently show higher or lower CV), it could hint at cyclical behavior in volatility.
- **Anomalies:** Any significant spikes or drops in the CV for a particular year could indicate anomalies. For example, if the CV in 2008 is significantly higher than surrounding years, it might reflect the impact of the financial crisis, which caused extreme market volatility.

Top Months by Volatility

Top 10 Months by Standard Deviation of Price Change Within a Day



The graph titled "**Top 10 Months by Standard Deviation of Price Change Within a Day**" visualizes the **top 10 months** with the highest **volatility** in daily price changes. Volatility is measured by the **standard deviation** of the daily price differences ($\text{price_diff} = \text{Close} - \text{Open}$). Each marker represents a month, and its size and color correspond to the standard deviation of the daily price changes during that month.

- **X-axis:** Represents time (from 1980 to 2010).
- **Y-axis:** Represents the standard deviation of the daily price differences (price_diff).
- **Marker Size and Color:** Represent the magnitude of the standard deviation (larger and darker markers indicate higher volatility).

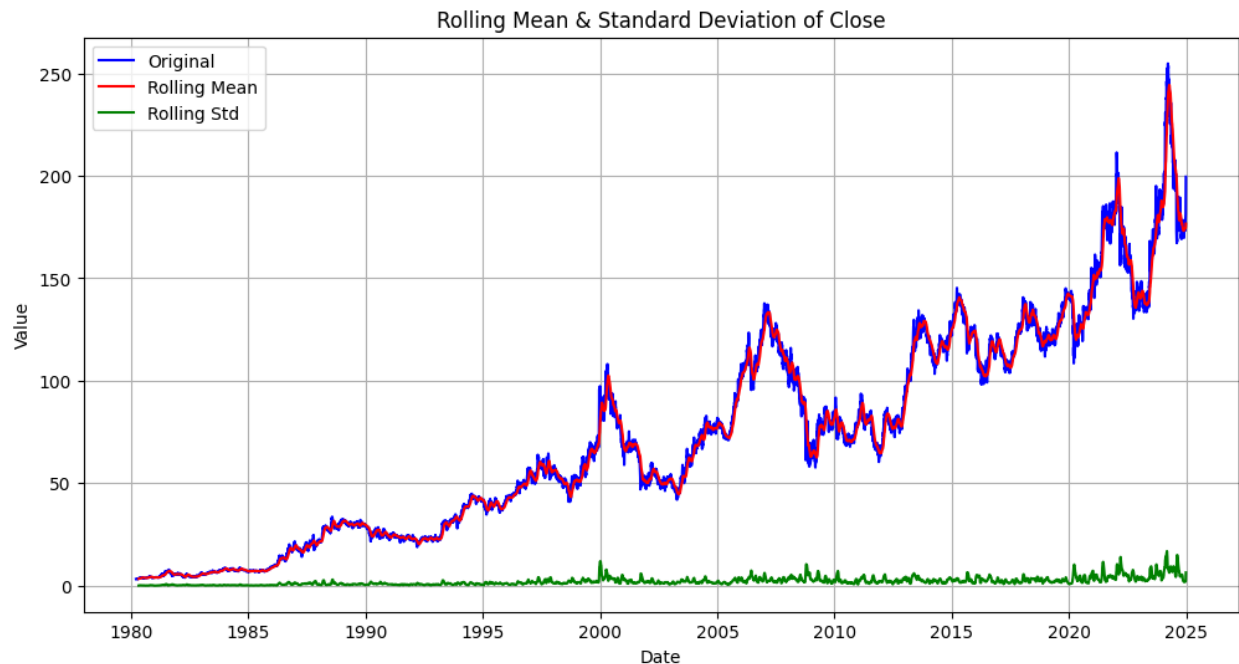
Key Insights from Visualization

1. **High Volatility Months:**
 - The graph highlights the **top 10 months** with the highest volatility in daily price changes. These months are likely associated with significant market events or economic crises.
2. **Temporal Distribution:**
 - The high-volatility months are not evenly distributed over time. They tend to cluster around specific periods, such as the early 1980s, late 1990s, and late 2000s.
3. **Magnitude of Volatility:**
 - The standard deviation values range from **2 to 14**, with the highest values occurring in the late 1990s and early 2000s. This indicates extreme price movements during those months.
4. **Anomalies:**
 - The months with the highest volatility likely correspond to major market events, such as the **Dot-com Bubble (late 1990s)** and the **2008 Financial Crisis**.

Analysis of Trends, Seasonality, and Anomalies

1. **Trends:**
 - The graph does not show a clear long-term trend in volatility. Instead, it highlights **sporadic spikes** in volatility during specific months.
 - The highest volatility months are concentrated in the late 1990s and early 2000s, suggesting that these periods were particularly turbulent for the stock.
2. **Seasonality:**
 - There is no evidence of **seasonality** in the graph. The high-volatility months do not follow a regular pattern (e.g., occurring in specific months or quarters).
3. **Anomalies:**
 - **Late 1990s and Early 2000s:** The months with the highest volatility likely correspond to the **Dot-com Bubble** and its subsequent burst, which caused extreme price movements in technology and growth stocks.
 - **2008:** The high volatility in 2008 is likely associated with the **2008 Financial Crisis**, which led to widespread market instability.

Rolling Statistics and Stationary



- The graph shows three lines:
 - **Original:** The actual closing price over time.
 - **Rolling Mean:** The rolling average (mean) of the closing price over a specified window (30 days). This smooths out short-term fluctuations and highlights long-term trends.
 - **Rolling Standard Deviation:** The rolling standard deviation of the closing price over the same window. This measures the volatility of the stock price over time.
- The x-axis represents the date (from 1980 to 2025), and the y-axis represents the value of the closing price, rolling mean, and rolling standard deviation.

Visualizations of key patterns and relationships:

- The **rolling mean** helps identify the long-term trend of the stock price by smoothing out short-term noise. It shows whether the stock price is generally increasing, decreasing, or remaining stable over time.
- The **rolling standard deviation** helps identify periods of high or low volatility. When the standard deviation is high, it indicates that the stock price is fluctuating significantly, and when it is low, the stock price is more stable.
- The **original closing price** provides context for the rolling mean and standard deviation, showing the actual price movements.

Analysis of trends, seasonality, and anomalies:

- **Trend:** The rolling mean line helps identify the long-term trend of the stock price. For example, if the rolling mean consistently increases over time, it indicates a bullish trend. Conversely, a decreasing rolling mean indicates a bearish trend.
- **Volatility:** The rolling standard deviation line helps identify periods of high or low volatility. For example, during market crashes or periods of economic instability, the rolling standard deviation may spike, indicating increased volatility.
- **Anomalies:** Any significant deviations between the original closing price and the rolling mean could indicate anomalies. For example, if the original price spikes far above the rolling mean, it might indicate a short-term price surge or an outlier event.

ADF Test

ADF Statistic: -0.46737151245213665

p-value: 0.8982279985700212

Critical Values: {'1%': np.float64(-3.4309306294476727), '5%': np.float64(-2.8617966068504166), '10%': np.float64(-2.5669065867160596)}

Fail to reject the null hypothesis: Data is non-stationary.

What the test represents:

- The ADF test checks whether a time series is stationary or non-stationary. The null hypothesis (H_0) is that the data is non-stationary (has a unit root), while the alternative hypothesis (H_1) is that the data is stationary.

Key metrics from the test:

- **ADF Statistic:** -0.46737151245213665
- **p-value:** 0.8982279985700212
- **Critical Values:**
 - 1%: -3.430
 - 5%: -2.861
 - 10%: -2.566

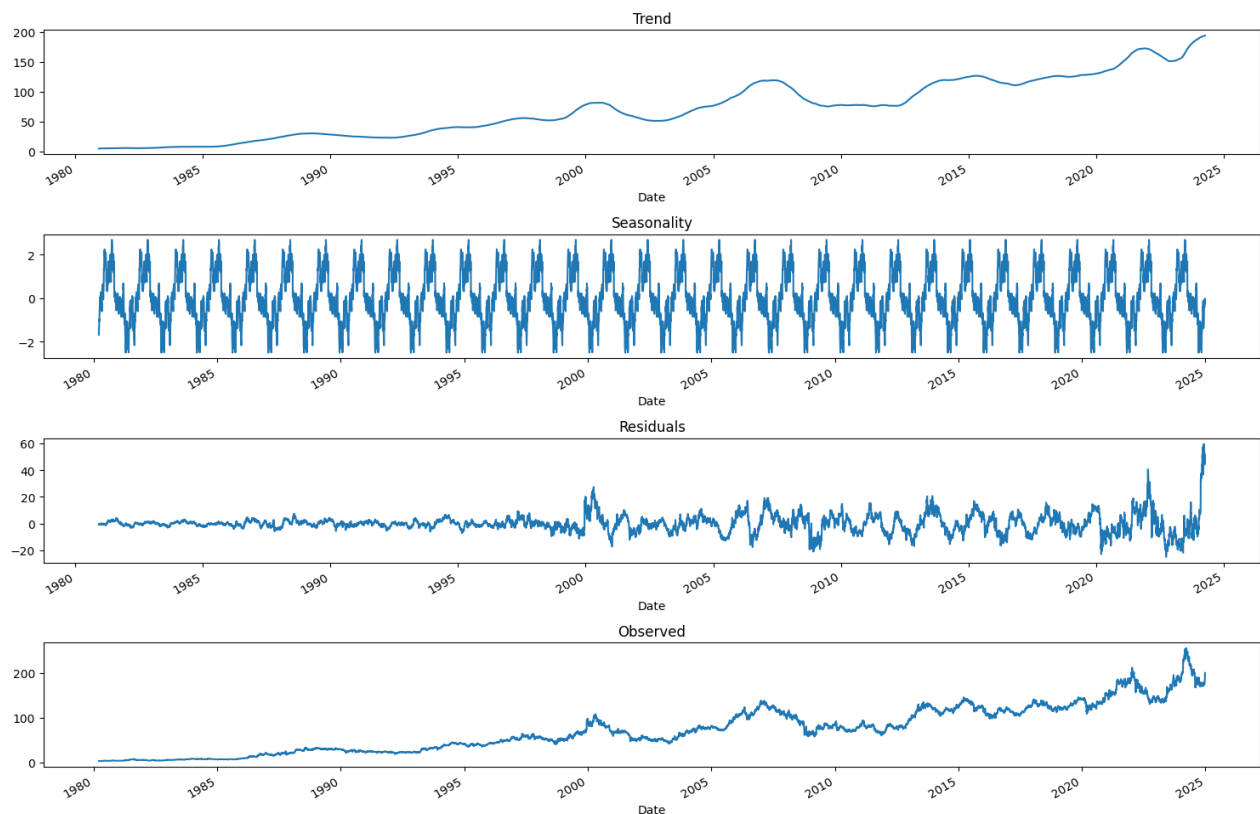
Interpretation of the results:

- The **p-value (0.898)** is much greater than 0.05, so we **fail to reject the null hypothesis**. This means the data is **non-stationary**.
- The **ADF statistic (-0.467)** is also less negative than all critical values, further supporting the conclusion that the data is non-stationary.

Implications for modeling:

- Non-stationary data can be challenging to model because its statistical properties (mean, variance, etc.) change over time. This can lead to unreliable predictions.
- To make the data stationary, you may need to apply transformations such as:
 - **Differencing:** Subtract the previous value from the current value to remove trends.
 - **Log transformation:** Apply a logarithmic transformation to stabilize variance.
 - **Detrending:** Remove trends or seasonality from the data.

Decomposition Analysis



1. Trend Component

Representation

- This graph represents the **long-term movement** of the closing price.
- It captures **gradual increases or decreases** over time by filtering out short-term fluctuations.

Key Patterns & Relationships

- A **general upward trend** is observed, indicating **long-term growth**.
- Periods of **stagnation or decline** may represent **market corrections, financial crises, or company-specific downturns**.

Analysis

- A **consistent rise** in stock price suggests the company is **growing and gaining value**.
- The **dips and plateaus** could be linked to **economic recessions, policy changes, or investor sentiment shifts**.
- The **most recent period** shows a **sharp increase**, which might be due to **market optimism, high earnings, or external economic factors**.

2. Seasonality Component

Representation

- This graph represents **repeating patterns** that occur at regular intervals (e.g., daily, monthly, yearly).
- It highlights **predictable cycles** in stock price movements.

Key Patterns & Relationships

- The pattern suggests **recurring fluctuations**, meaning stock prices tend to **rise and fall in a consistent cycle**.
- The **frequency** of the peaks and troughs suggests a **strong seasonal component**.

Analysis

- Seasonality could be due to **quarterly earnings reports, market cycles, or investor behavior**.
- Some industries (like retail) experience **higher prices in specific months** (e.g., holiday season).
- Understanding this pattern helps in **making better investment decisions**, like identifying the best times to buy/sell.

3. Residuals Component

Representation

- Residuals represent the **unexplained variations** in the data after removing trend and seasonality.
- It shows **random noise and anomalies** that are not part of the regular patterns.

Key Patterns & Relationships

- **Mostly stable** throughout, meaning the model captures most of the patterns.
- Some **periods of higher residual values** suggest **unexpected volatility**.
- Towards the **end of the graph**, residuals increase significantly, indicating **sudden market events or outliers**.

Analysis

- Residuals should ideally be **randomly distributed**; any visible patterns might indicate **missing factors in the model**.
- Large residuals could be caused by **unexpected market events** (e.g., economic crashes, sudden policy changes).
- **Increased volatility at the end** may suggest **recent market instability** or **external shocks**.

4. Observed Component

Representation

- This is the **original stock price data** before decomposition.
- It represents the actual closing price with all trends, seasonality, and noise included.

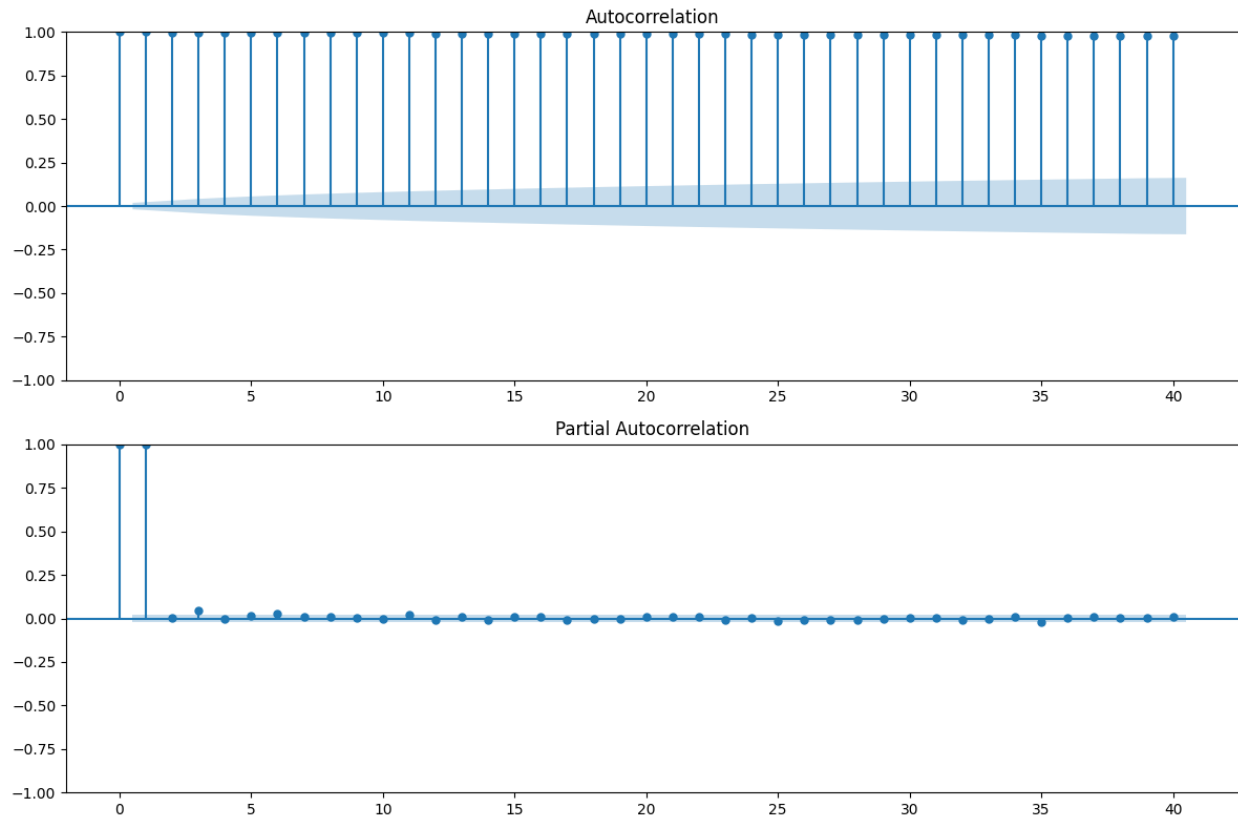
Key Patterns & Relationships

- The price shows **significant growth over time**.
- Some **sharp peaks and dips** suggest **high volatility in certain periods**.
- The **most recent years** display an **increasing trend with stronger fluctuations**.

Analysis

- The overall trend suggests that the stock has been **growing in value**.
- Spikes and drops could be linked to **external market events** (e.g., financial crises, corporate news, economic policies).
- **Understanding these components helps in forecasting** future stock prices by identifying key influencing factors.

Partial and Auto Correlation



- **Autocorrelation (ACF):** This graph shows how correlated the stock's closing price is with its past values at different time lags (e.g., 1 day ago, 2 days ago, etc.).
- **Partial Autocorrelation (PACF):** This graph shows the correlation between the stock's closing price and its past values, but after removing the influence of other time lags.

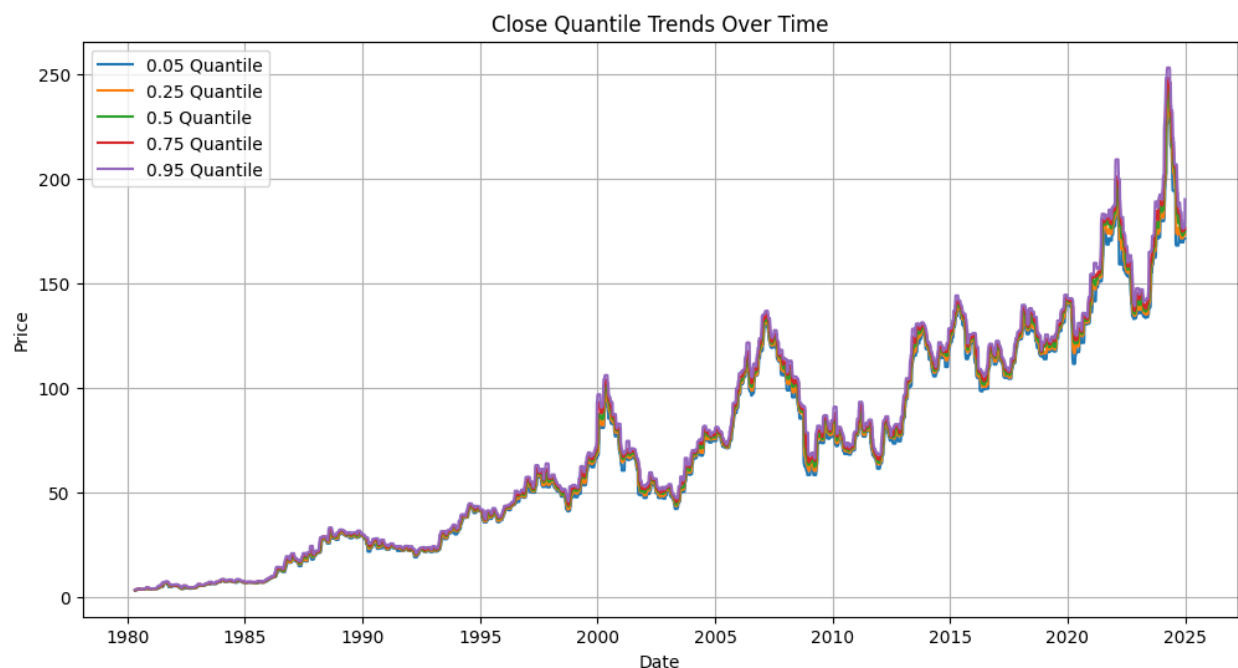
Visualizations of Key Patterns and Relationships:

- **Autocorrelation (ACF):**
 - The **x-axis** represents the **time lags** (e.g., 0 to 40 days).
 - The **y-axis** represents the **correlation coefficient** (ranging from -1 to 1).
 - The blue shaded region represents the **confidence interval**. If the bars go outside this region, the correlation is statistically significant.
- **Partial Autocorrelation (PACF):**
 - The **x-axis** represents the **time lags** (e.g., 0 to 40 days).
 - The **y-axis** represents the **partial correlation coefficient** (ranging from -1 to 1).
 - The blue shaded region represents the **confidence interval**. If the bars go outside this region, the correlation is statistically significant.

Analysis of Trends, Seasonality, and Anomalies:

- **Autocorrelation (ACF):**
 - The stock's closing price is highly correlated with its recent past values (e.g., 1-2 days ago).
 - The correlation decreases as the lag increases, which is typical for time series data.
 - There might be some **seasonal patterns** (e.g., spikes at regular intervals like 7, 14, 21 days), indicating weekly seasonality.
- **Partial Autocorrelation (PACF):**
 - The stock's closing price is strongly correlated with its immediate past values (e.g., 1-2 days ago).
 - After a few lags, the correlation drops sharply, indicating that the influence of older data points diminishes quickly.
 - This suggests that the stock price is influenced more by recent events than by older ones.

Quantile Trends



The graph titled "**Close Quantile Trends Over Time**" visualizes the trends of different quantiles (0.05, 0.25, 0.5, 0.75, 0.95) of the **closing price** of a stock over time. Each quantile represents a specific percentile of the closing price distribution:

- **0.05 Quantile:** The 5th percentile (lower range of prices).
- **0.25 Quantile:** The 25th percentile (lower-middle range of prices).
- **0.5 Quantile:** The 50th percentile (median price).
- **0.75 Quantile:** The 75th percentile (upper-middle range of prices).
- **0.95 Quantile:** The 95th percentile (upper range of prices).

The x-axis represents **time** (from 1980 to 2025), and the y-axis represents the **closing price** of the stock.

Key Insights from Visualization

1. Long-Term Growth Trend:

- The **0.5 Quantile (Median)** and **0.95 Quantile (Upper Range)** show a clear upward trend over time, indicating that the stock's closing price has generally increased over the years.
- The **0.05 Quantile (Lower Range)** also shows growth but at a slower pace, suggesting that even the lowest prices have increased over time.

2. Volatility Over Time:

- The gap between the **0.05 Quantile** and **0.95 Quantile** represents the volatility in the stock's closing price. This gap widens over time, indicating increasing volatility in the stock's price movements.

3. Seasonality:

- The graph does not show strong seasonal patterns, as the quantile trends are relatively smooth without recurring spikes or dips at regular intervals.

4. Anomalies:

- There are occasional sharp drops or spikes in the quantile trends, which could represent market anomalies or significant events (e.g., financial crises, market crashes, or sudden price surges).

5. Convergence and Divergence:

- In some periods, the quantiles converge (e.g., around 1990 and 2010), indicating lower volatility. In other periods, they diverge (e.g., around 2000 and 2020), indicating higher volatility.



Analysis of Trends, Seasonality, and Anomalies

1. Trends:

- The overall trend is **upward**, with the median and upper quantiles showing consistent growth. This suggests that the stock has been a good long-term investment.
- The lower quantile (0.05) also shows growth but at a slower rate, indicating that even during downturns, the stock's price has not fallen drastically.

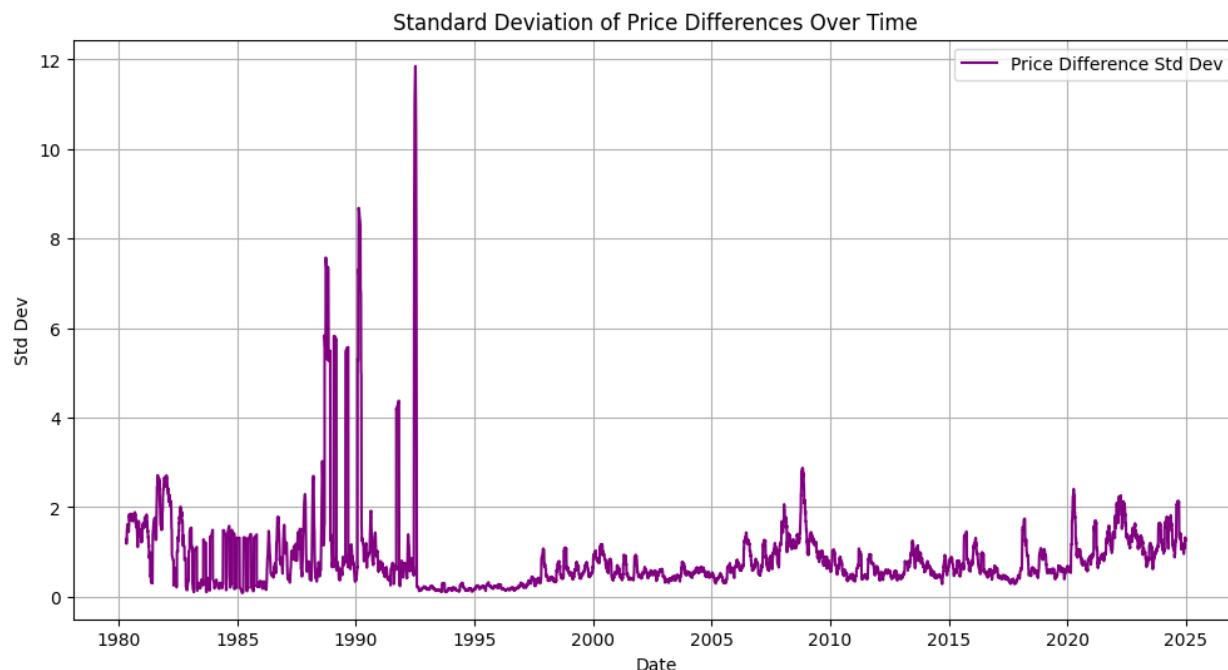
2. Seasonality:

- There is no clear evidence of **seasonality** in the graph. The trends are relatively smooth, and there are no recurring patterns that repeat at regular intervals (e.g., yearly or quarterly).

3. Anomalies:

- **Sharp Drops:** There are noticeable drops in the quantile trends around **2000** and **2008**, which could correspond to market crashes (e.g., the Dot-com Bubble and the 2008 Financial Crisis).
- **Sharp Spikes:** There are spikes in the upper quantiles (e.g., around 2020), which could represent periods of rapid price appreciation or market bubbles.

Price Difference Standard Deviation



What Does the Graph Represent? The graph titled "**Standard Deviation of Price Differences Over Time**" visualizes the **volatility** of the daily price differences between the **closing price** and the **opening price** of a stock over time. The standard deviation (Std Dev) is calculated over a rolling window of 30 days, which smooths out short-term fluctuations and highlights longer-term trends in volatility.

- **X-axis:** Represents time (from 1980 to 2025).
- **Y-axis:** Represents the standard deviation of the daily price differences (price_diff).

Key Insights from Visualization

1. Volatility Trends:

- The graph shows periods of **high volatility** (peaks) and **low volatility** (troughs) over time.
- Volatility tends to increase during periods of market uncertainty or economic crises (e.g., around 2000 and 2008).

2. Long-Term Volatility:

- The overall trend in volatility appears to be **increasing** over time, especially after 2000. This suggests that the stock's price movements have become more unpredictable in recent years.

3. **Seasonality:**

- There is no clear evidence of **seasonality** in the graph. The volatility trends do not show recurring patterns at regular intervals (e.g., yearly or quarterly).

4. **Anomalies:**

- **Sharp Peaks:** There are noticeable spikes in volatility around **2000** and **2008**, which likely correspond to major market events (e.g., the Dot-com Bubble and the 2008 Financial Crisis).
- **Recent Volatility:** The graph shows increased volatility in recent years (e.g., around 2020), possibly due to events like the COVID-19 pandemic.

Analysis of Trends, Seasonality, and Anomalies

1. **Trends:**

- The graph indicates an **upward trend in volatility** over time, especially after 2000. This suggests that the stock's price movements have become more erratic in recent decades.
- The rolling standard deviation helps smooth out short-term fluctuations, making it easier to identify long-term trends.

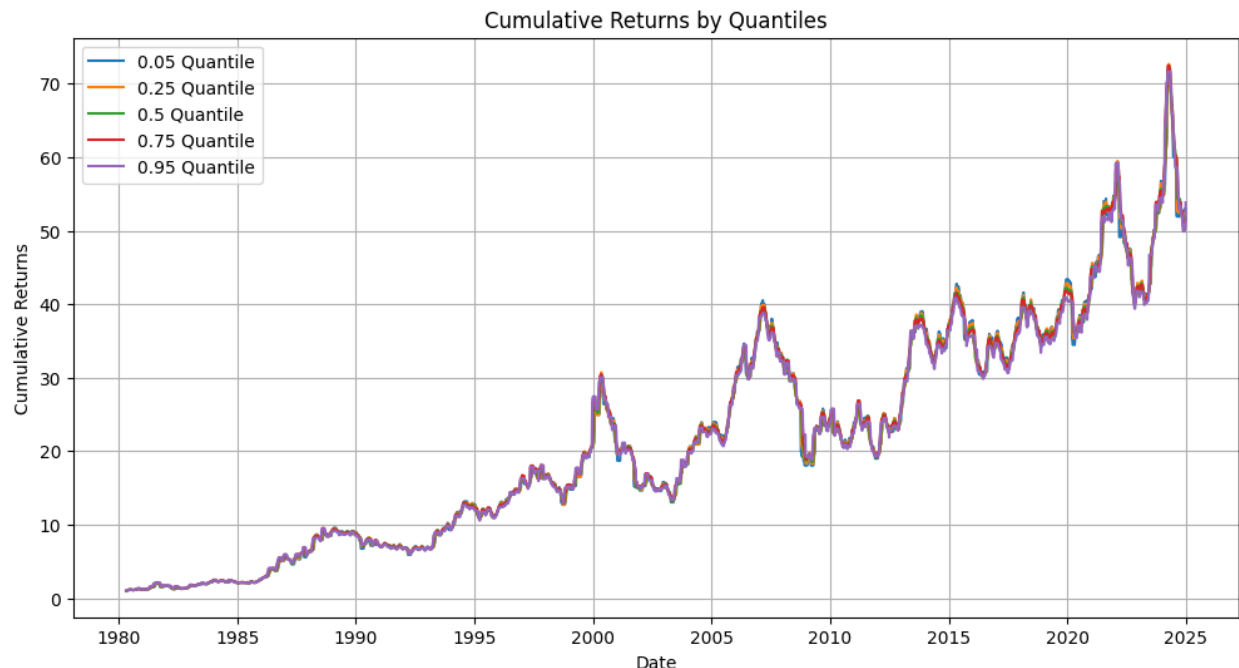
2. **Seasonality:**

- There is no clear **seasonal pattern** in the graph. Volatility does not appear to follow a regular cycle (e.g., higher volatility in certain months or quarters).

3. **Anomalies:**

- **2000 and 2008 Peaks:** The sharp spikes in volatility around these years correspond to major financial crises, indicating that the stock was highly sensitive to market conditions during these periods.
- **2020 Volatility:** The increase in volatility around 2020 likely reflects the impact of the COVID-19 pandemic on the stock market.

Cumulative Returns by Quantiles



The graph titled "**Cumulative Returns by Quantiles**" visualizes the **cumulative returns** of the **closing price** for different quantiles (0.05, 0.25, 0.5, 0.75, 0.95) over time. Cumulative returns show the total growth of an investment over a period, starting from a base value.

- **X-axis:** Represents time (from 1980 to 2025).
- **Y-axis:** Represents the **cumulative returns** (total growth over time).
- **Quantiles:** Each line represents a specific quantile of the closing price distribution.

Key Insights from Visualization

1. Growth Trends:

- The **0.5 Quantile (Median)** and **0.95 Quantile (Upper Range)** show strong **upward trends**, indicating significant long-term growth.
- The **0.05 Quantile (Lower Range)** also grows but at a slower pace, suggesting that even the lowest returns have increased over time.

2. Volatility:

- The gap between the **0.05 Quantile** and **0.95 Quantile** represents the **volatility** in returns. A wider gap indicates higher volatility.

3. Anomalies:

- Sharp drops or spikes in cumulative returns (e.g., around 2000 and 2008) correspond to **market anomalies** like the Dot-com Bubble and the 2008 Financial Crisis.

Analysis of Trends, Seasonality, and Anomalies

1. Trends:

- The overall trend is **upward**, with the median and upper quantiles showing consistent growth.
- The lower quantile also grows but at a slower rate, indicating resilience even during downturns.

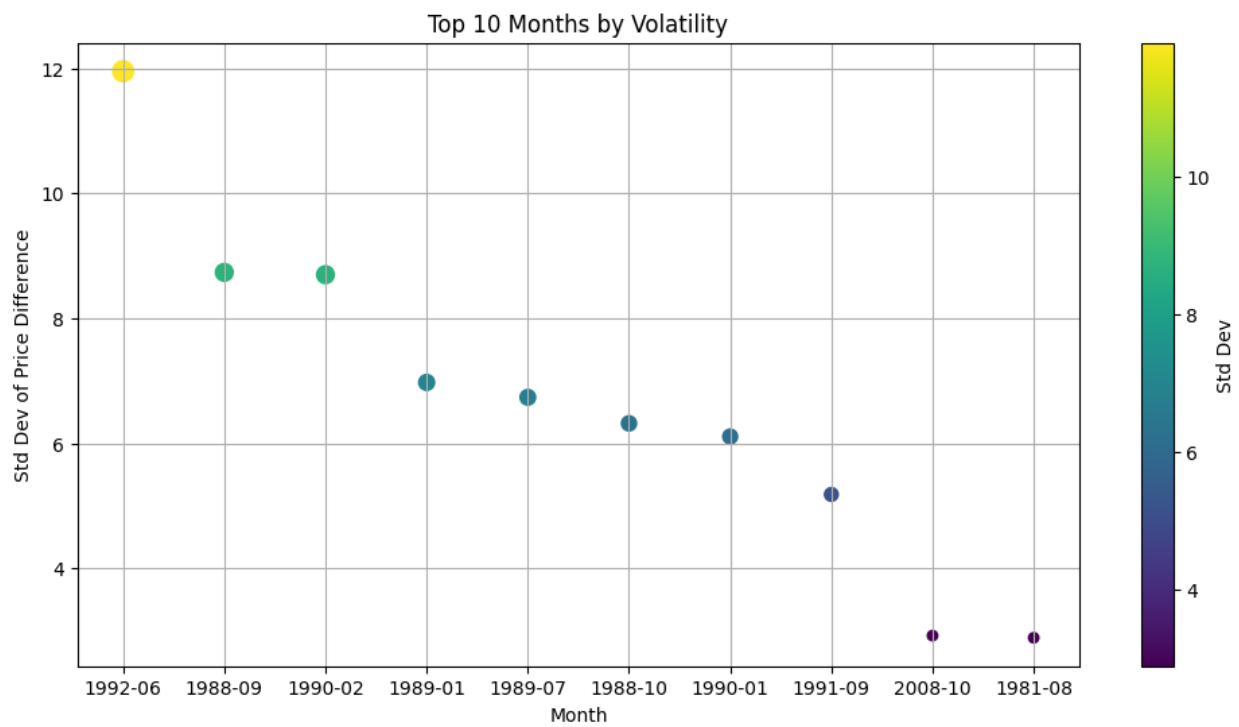
2. Seasonality:

- No clear **seasonality** is observed; the trends are smooth without recurring patterns.

3. Anomalies:

- **2000 and 2008**: Sharp drops correspond to market crashes.
- **2020**: Increased volatility likely due to the COVID-19 pandemic.

Top N Months by Volatility



The graph titled "**Top 10 Months by Volatility**" visualizes the **top 10 months** with the highest **volatility** in daily price changes. Volatility is measured by the **standard deviation** of the daily price differences ($\text{price_diff} = \text{Close} - \text{Open}$). Each marker represents a month, and its size and color correspond to the standard deviation of the daily price changes during that month.

- **X-axis:** Represents the **month** (e.g., 1992-06, 1988-09, etc.).
- **Y-axis:** Represents the **standard deviation** of the daily price differences (price_diff).
- **Marker Size and Color:** Represent the magnitude of the standard deviation (larger and darker markers indicate higher volatility).

Key Insights from Visualization

1. **High Volatility Months:**
 - The graph highlights the **top 10 months** with the highest volatility in daily price changes. These months are likely associated with significant market events or economic crises.
2. **Temporal Distribution:**
 - The high-volatility months are spread across different years, with clusters in the late 1980s, early 1990s, and 2008. This suggests that volatility spikes are not confined to a single period but occur sporadically.
3. **Magnitude of Volatility:**
 - The standard deviation values vary significantly, with the highest values occurring in **1992-06** and **2008-10**. These months likely experienced extreme price movements.
4. **Anomalies:**
 - The months with the highest volatility likely correspond to major market events, such as the **1990-1991 recession** and the **2008 Financial Crisis**.

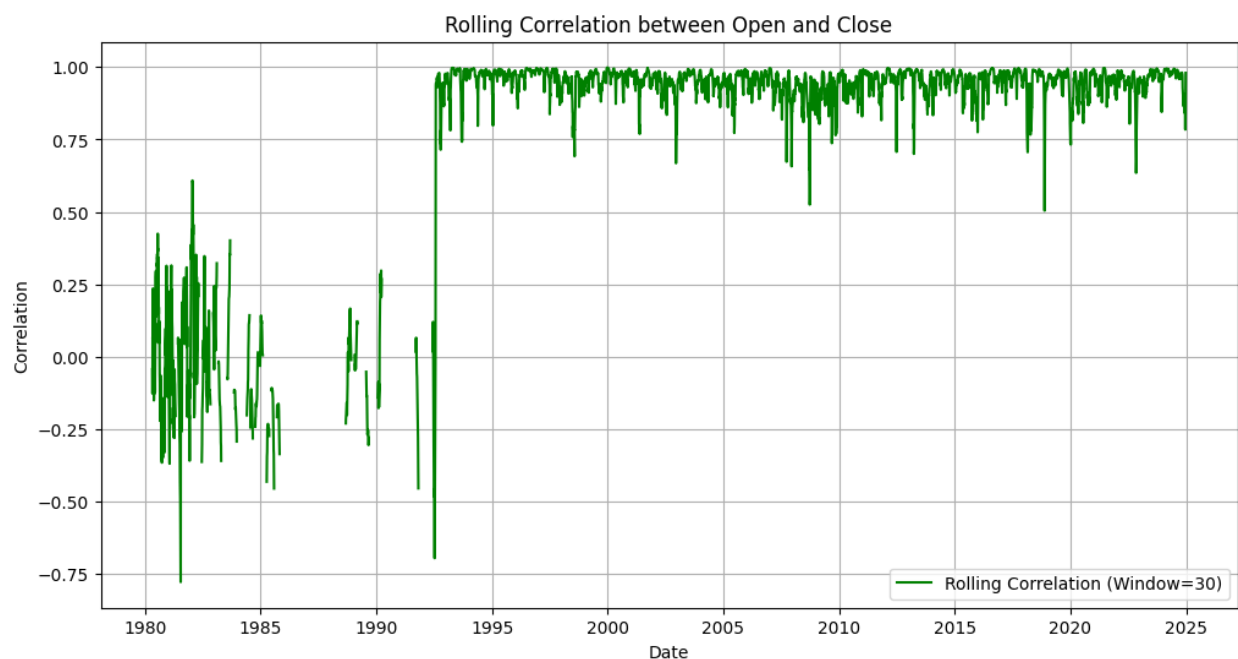
Analysis of Trends, Seasonality, and Anomalies

1. **Trends:**
 - The graph does not show a clear long-term trend in volatility. Instead, it highlights **sporadic spikes** in volatility during specific months.
 - The highest volatility months are concentrated in the late 1980s, early 1990s, and 2008, suggesting that these periods were particularly turbulent for the stock.
2. **Seasonality:**
 - There is no evidence of **seasonality** in the graph. The high-volatility months do not follow a regular pattern (e.g., occurring in specific months or quarters).

3. Anomalies:

- **1992-06:** This month stands out with the highest volatility, possibly due to market uncertainty or a specific event during that period.
- **2008-10:** The high volatility in this month is likely associated with the **2008 Financial Crisis**, which led to widespread market instability.
- **Late 1980s and Early 1990s:** The high volatility in these months may correspond to economic recessions or market corrections during that period.

Rolling Correlation



The graph titled "**Rolling Correlation between Open and Close**" visualizes the **rolling correlation** between the **opening price** and the **closing price** of a stock over time. The correlation is calculated over a rolling window of 30 days, which smooths out short-term fluctuations and highlights longer-term trends in the relationship between the two prices.

- **X-axis:** Represents time (from 1980 to 2025).
- **Y-axis:** Represents the **correlation coefficient** between the opening and closing prices.
 - A correlation of **1** indicates a perfect positive relationship.
 - A correlation of **-1** indicates a perfect negative relationship.
 - A correlation of **0** indicates no relationship.

Key Insights from Visualization

1. Correlation Trends:

- The rolling correlation between the opening and closing prices is generally **positive**, indicating that the closing price tends to move in the same direction as the opening price.
- The correlation fluctuates over time, with periods of **strong positive correlation** (close to 1) and periods of **weaker correlation** (closer to 0).

2. Volatility in Correlation:

- The correlation is not constant and shows significant fluctuations, especially in the early 1980s and around 2008. This suggests that the relationship between the opening and closing prices can vary depending on market conditions.

3. Anomalies:

- There are occasional dips in the correlation, where the relationship between the opening and closing prices weakens or becomes slightly negative. These dips may correspond to periods of market uncertainty or anomalies.

4. Recent Trends:

- In recent years (e.g., after 2010), the correlation appears to stabilize around **0.75**, indicating a consistently strong positive relationship between the opening and closing prices.

Analysis of Trends, Seasonality, and Anomalies

1. Trends:

- The overall trend in the rolling correlation is **positive**, with the correlation coefficient generally staying above **0.5**. This suggests that the opening and closing prices tend to move together over time.
- The correlation has become more stable in recent years, indicating a stronger and more consistent relationship between the two prices.

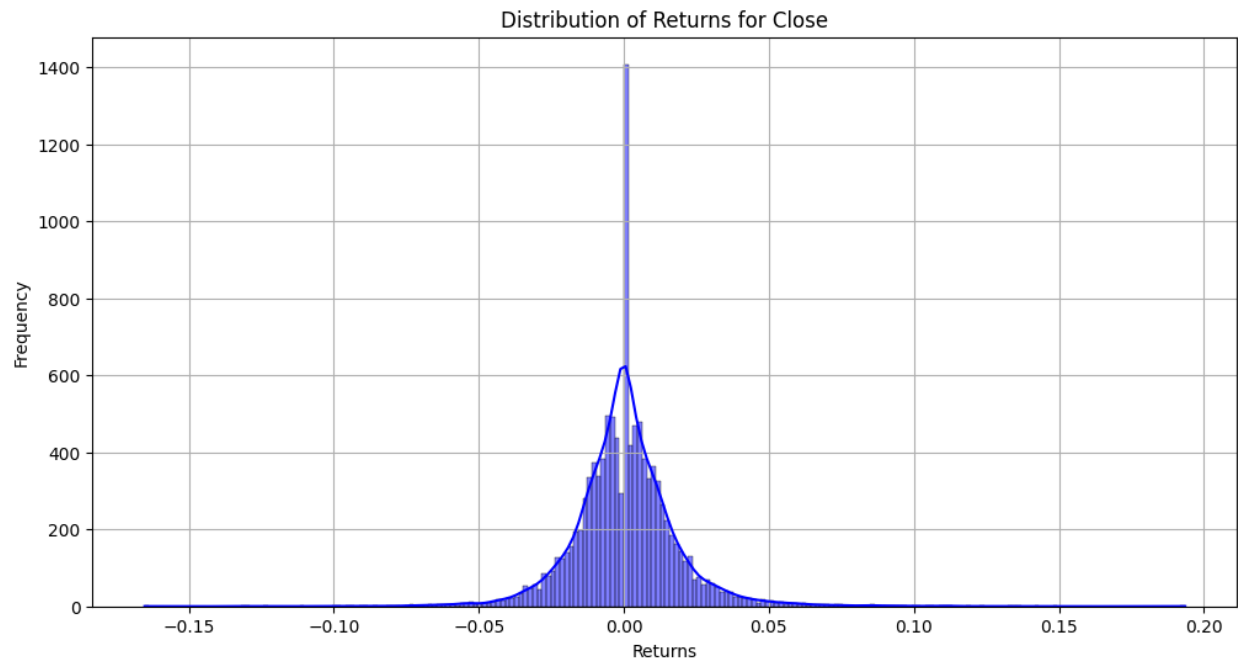
2. Seasonality:

- There is no clear evidence of **seasonality** in the graph. The correlation does not show recurring patterns at regular intervals (e.g., yearly or quarterly).

3. Anomalies:

- **Early 1980s:** The correlation dips significantly during this period, possibly due to market instability or economic events.
- **2008:** The correlation shows a sharp decline around 2008, likely corresponding to the **2008 Financial Crisis**, which caused significant market volatility and disrupted the typical relationship between opening and closing prices.
- **Other Dips:** There are occasional dips in the correlation, which may correspond to smaller market anomalies or periods of uncertainty.

Distribution of Returns



The graph titled "**Distribution of Returns for Close**" visualizes the **distribution of daily returns** for the **closing price** of a stock. Daily returns are calculated as the percentage change in the closing price from one day to the next. The graph uses a **histogram** with a **kernel density estimate (KDE)** overlay to show the frequency of different return values.

- **X-axis:** Represents the **daily returns** (percentage change in closing price).
- **Y-axis:** Represents the **frequency** of returns (how often each return value occurs).

Key Insights from Visualization

1. Shape of the Distribution:

- The distribution of returns is **approximately symmetric** but slightly **right-skewed**, meaning there are more frequent small positive returns and occasional large positive returns.
- The peak of the distribution is around **0.00**, indicating that the most common daily return is close to zero (i.e., no significant change in price).

2. Volatility:

- The spread of the distribution indicates the **volatility** of the stock. A wider spread suggests higher volatility, while a narrower spread suggests lower volatility.
- The tails of the distribution extend to both positive and negative returns, indicating that the stock experiences both significant gains and losses.

3. **Anomalies:**

- The presence of **outliers** in the tails of the distribution (e.g., returns beyond ± 0.15) suggests that the stock occasionally experiences extreme price movements, which could be due to market events or anomalies.

4. **Central Tendency:**

- The mean and median of the distribution are likely close to **0.00**, indicating that, on average, the stock's daily returns are centered around zero.

Analysis of Trends, Seasonality, and Anomalies

1. **Trends:**

- The distribution does not show a clear **trend** because it represents the frequency of returns rather than their progression over time. However, the shape of the distribution provides insights into the stock's typical behavior.

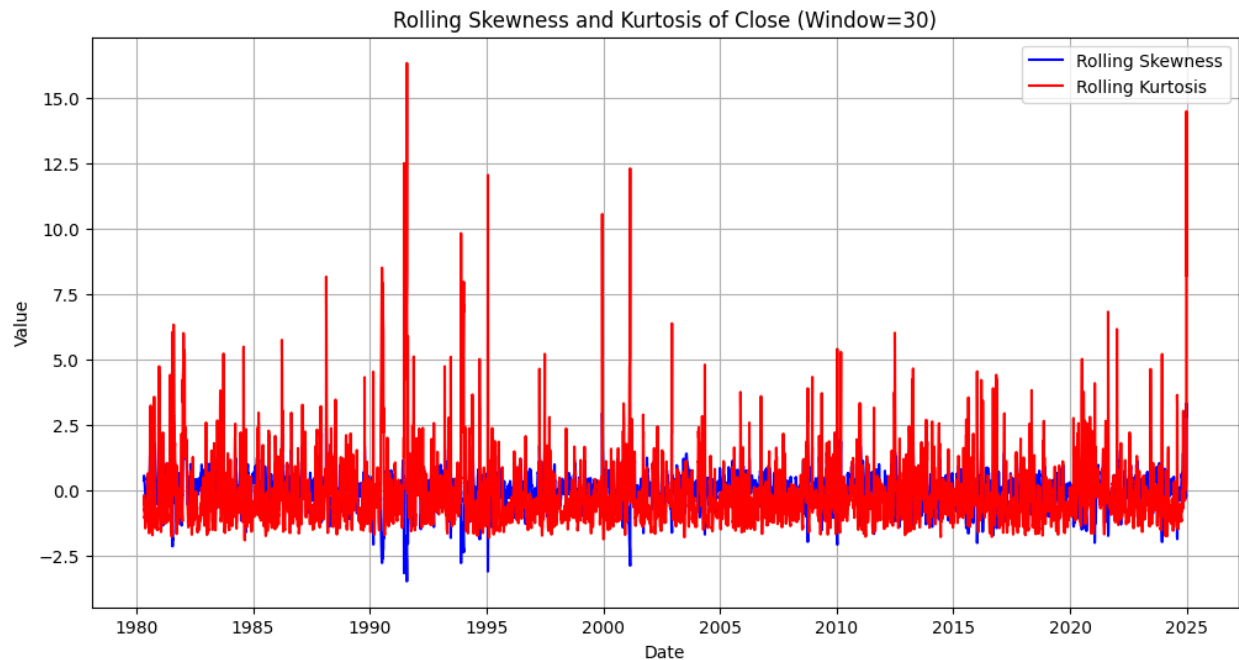
2. **Seasonality:**

- There is no evidence of **seasonality** in the graph. The distribution of returns does not show recurring patterns at regular intervals (e.g., yearly or quarterly).

3. **Anomalies:**

- **Outliers:** The presence of extreme returns in the tails of the distribution suggests that the stock occasionally experiences significant price movements, which could be due to market anomalies or events.
- **Skewness:** The slight right skewness indicates that the stock is more likely to experience large positive returns than large negative returns.

Rolling Skewness and Kurtosis



The graph titled "**Rolling Skewness and Kurtosis of Close (Window=30)**" visualizes the **rolling skewness** and **rolling kurtosis** of the **closing price** of a stock over time. These metrics are calculated over a rolling window of 30 days, which smooths out short-term fluctuations and highlights longer-term trends in the distribution of the closing price.

- **X-axis:** Represents time (from 1980 to 2025).
- **Y-axis:** Represents the **value** of skewness and kurtosis.
 - **Skewness:** Measures the asymmetry of the distribution of closing prices.
 - A skewness of **0** indicates a symmetric distribution.
 - A **positive skewness** indicates a longer tail on the right (more frequent large positive returns).
 - A **negative skewness** indicates a longer tail on the left (more frequent large negative returns).
 - **Kurtosis:** Measures the "tailedness" of the distribution of closing prices.
 - A kurtosis of **3** indicates a normal distribution.
 - A **kurtosis greater than 3** indicates heavier tails (more extreme values).
 - A **kurtosis less than 3** indicates lighter tails (fewer extreme values).

Key Insights from Visualization

1. Skewness Trends:

- The rolling skewness fluctuates around **0**, indicating that the distribution of closing prices is generally symmetric but occasionally exhibits slight asymmetry.
- There are periods of **positive skewness** (e.g., around 2000 and 2020), suggesting that stock experienced more frequent large positive returns during these periods.
- There are also periods of **negative skewness** (e.g., around 2008), suggesting that the stock experienced more frequent large negative returns during these periods.

2. Kurtosis Trends:

- The rolling kurtosis is generally **greater than 3**, indicating that the distribution of closing prices has heavier tails than a normal distribution. This suggests that the stock frequently experiences extreme price movements.
- The kurtosis shows significant fluctuations, with peaks around **2000** and **2008**, indicating periods of extreme volatility.

3. Volatility and Anomalies:

- The peaks in kurtosis correspond to periods of **high volatility** and **market anomalies**, such as the **Dot-com Bubble (2000)** and the **2008 Financial Crisis**.
- The skewness and kurtosis metrics provide insights into the **risk** and **return characteristics** of the stock over time.

Analysis of Trends, Seasonality, and Anomalies

1. Trends:

- The rolling skewness and kurtosis show **no clear long-term trend** but exhibit significant fluctuations over time. This suggests that the distribution of closing prices changes dynamically based on market conditions.
- The kurtosis is consistently **greater than 3**, indicating that the stock frequently experiences extreme price movements.

2. Seasonality:

- There is no evidence of **seasonality** in the graph. The skewness and kurtosis do not show recurring patterns at regular intervals (e.g., yearly or quarterly).

3. Anomalies:

- **2000**: The peaks in skewness and kurtosis around 2000 likely correspond to the **Dot-com Bubble**, which caused extreme price movements in technology and growth stocks.
- **2008**: The peaks in skewness and kurtosis around 2008 likely correspond to the **2008 Financial Crisis**, which led to widespread market instability and extreme price movements.
- **2020**: The increase in skewness and kurtosis around 2020 may reflect the impact of the **COVID-19 pandemic** on the stock market

Cumulative Return Analysis



- This graph represents the **cumulative returns** of a stock over time.
- The **x-axis** shows the date, ranging from **1980** to **2025**.
- The **y-axis** represents cumulative returns, showing the growth of an investment over time.

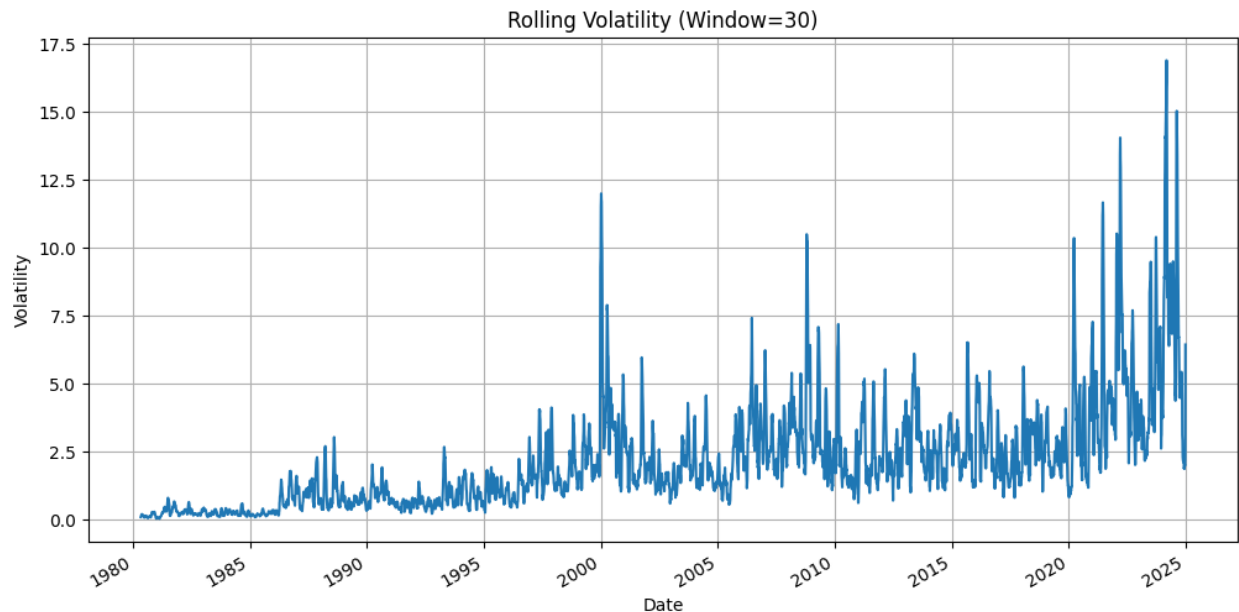
Key Insights from Visualization

- The stock price shows an overall **upward trend**, indicating **long-term growth**.
- There are periods of rapid increase (e.g., around 2000, 2010, and after 2020), which may be due to market booms or company performance improvements.
- There are major dips, such as around 2008 and 2020, possibly related to financial crises or market corrections.
- Recent volatility suggests higher fluctuations in stock price.

Trend & Anomalies

- **Long-Term Trend:** A steady increase in stock value over decades.
- **Anomalies:** Sharp rises and drops indicate potential financial crises or major economic events.

Rolling Volatility Analysis



- The graph shows the **rolling volatility** of the stock's closing price over time. **Volatility** measures how much the stock price fluctuates. Higher volatility means the price changes a lot, while lower volatility means the price is more stable.
- The **x-axis** represents the **Date**, ranging from 2003 to 2025.
- The **y-axis** represents the **volatility** (measured as the standard deviation of the stock price over a 30-day window).

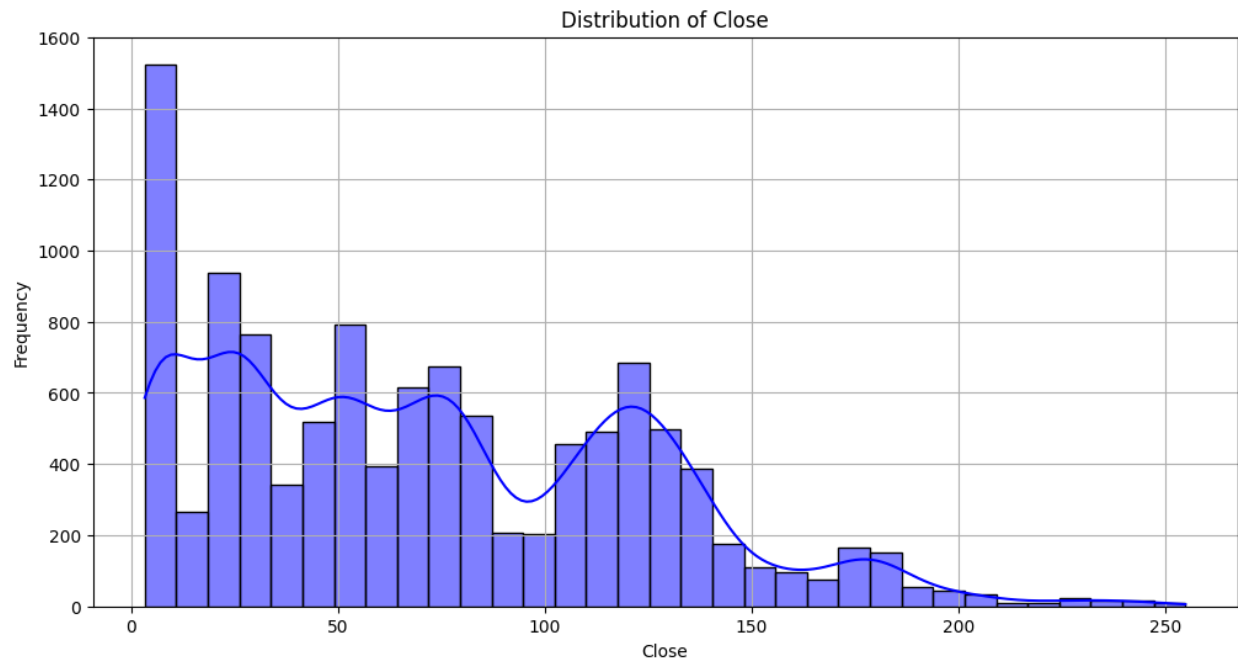
Visualizations of Key Patterns and Relationships:

- The graph shows how the stock's volatility changes over time. A **30-day rolling window** is used, meaning the volatility is calculated based on the past 30 days of data.

Analysis of Trends, Seasonality, and Anomalies:

- **Trends:** The volatility fluctuates over time, with periods of high volatility (e.g., around 2008-2009 and 2020) and periods of low volatility (e.g., around 2012-2015).
- **Seasonality:** There is no clear seasonal pattern in the volatility. However, the spikes in volatility often correspond to major market events (e.g., the 2008 financial crisis and the 2020 COVID-19 pandemic).
- **Anomalies:** The spikes in volatility around 2008-2009 and 2020 are notable anomalies, likely caused by significant market disruptions.

Target Variable Distribution



- The graph shows the **distribution** of the stock's **closing prices**.
- The **x-axis** represents the **closing price** of the stock.
- The **y-axis** represents the **frequency** (how often each closing price occurs).

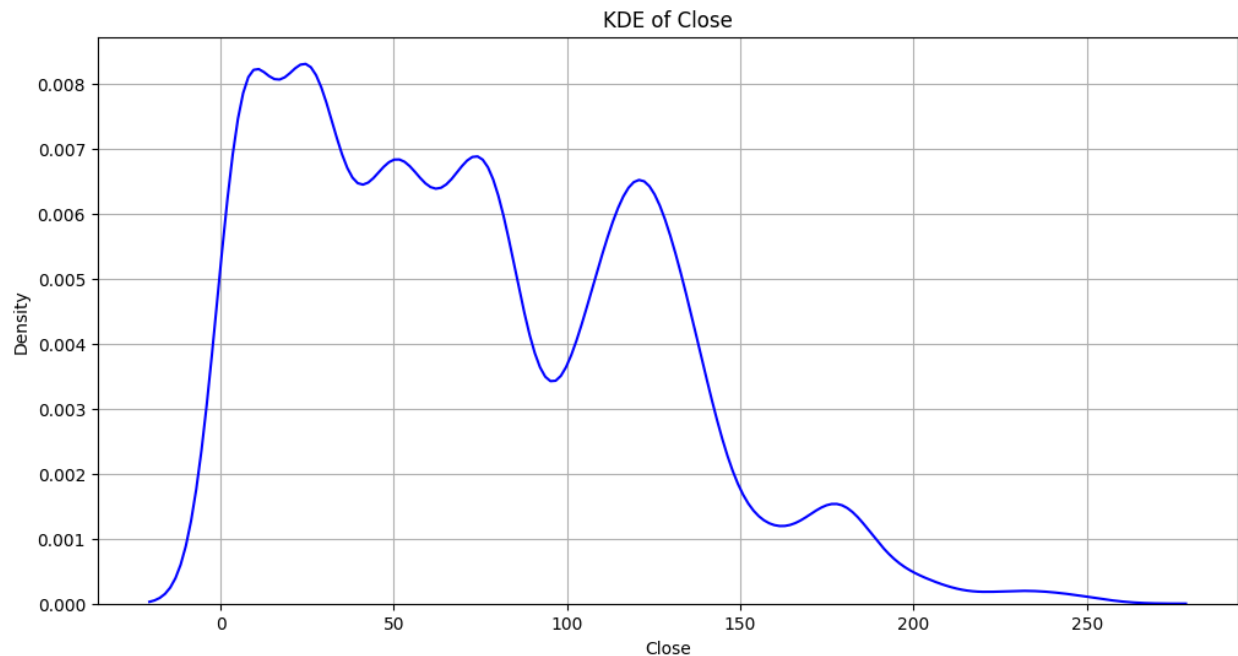
Visualizations of Key Patterns and Relationships:

- The graph is a **histogram**, which shows how often different closing prices occur.
- The shape of the histogram tells us about the distribution of the closing prices.

Analysis of Trends, Seasonality, and Anomalies:

- **Trends:** The closing prices are mostly concentrated in a specific range (e.g., between 50 and 150), indicating that the stock price tends to stay within this range most of the time.
- **Seasonality:** There is no seasonality in this graph because it shows the distribution of prices, not their changes over time.
- **Anomalies:** There are some outliers (e.g., closing prices above 200), which are less frequent and could represent unusual market conditions.

KDE Analysis of the Target



- The graph shows the **Kernel Density Estimate (KDE)** of the stock's **closing prices**.
- The **x-axis** represents the **closing price** of the stock.
- The **y-axis** represents the **density** (probability distribution) of the closing prices.

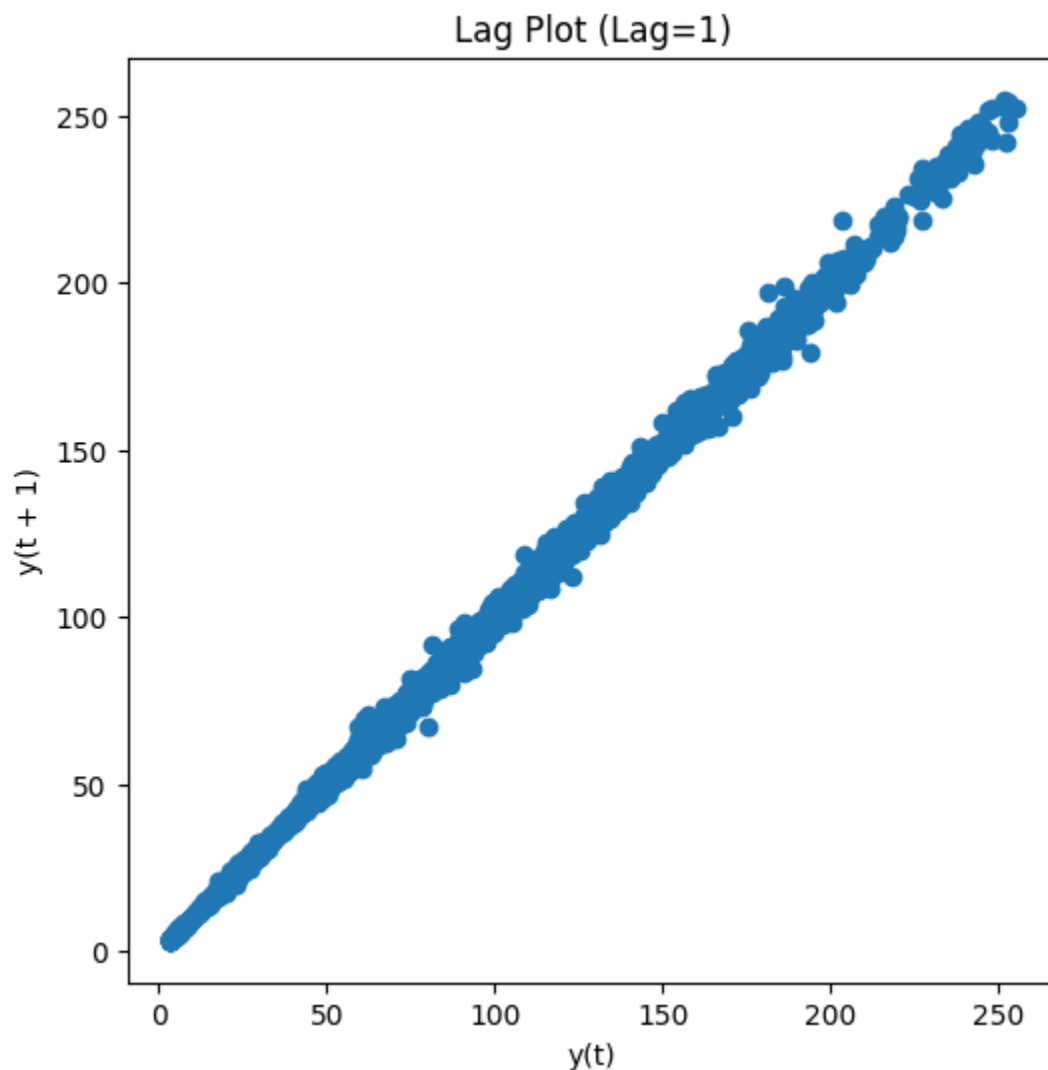
Visualizations of Key Patterns and Relationships:

- The KDE plot is a smoothed version of the histogram, showing the probability distribution of the closing prices.
- The peaks in the graph represent the most likely closing prices.

Analysis of Trends, Seasonality, and Anomalies:

- **Trends:** The KDE plot shows that the closing prices are most likely to be around 100, with a gradual decrease in probability as the prices move away from this range.
- **Seasonality:** There is no seasonality in this graph because it shows the probability distribution of prices, not their changes over time.
- **Anomalies:** The long tail on the right side of the graph (towards higher closing prices) indicates that there are some rare instances where the closing price is much higher than the average.

Lag Plot Analysis (Lag = 1)



- The graph is a **Lag Plot** with a lag of 1, meaning it compares the stock's closing price at time (t) with its closing price at time ($t+1$) (the next day).
- The **x-axis** represents the **closing price at time (t)**.
- The **y-axis** represents the **closing price at time ($t+1$)**.

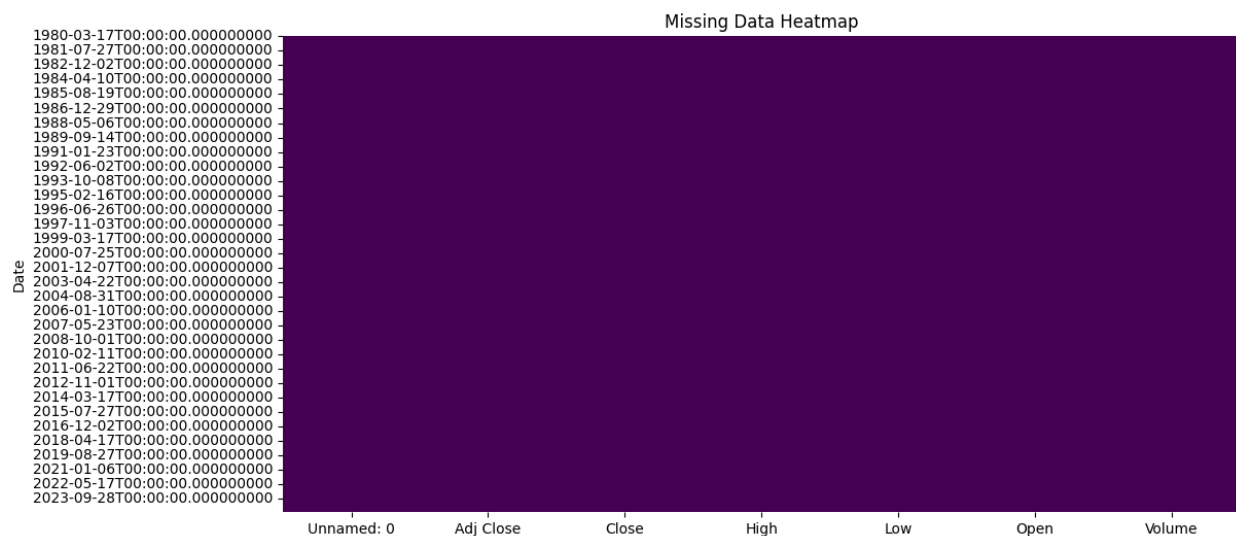
Visualizations of Key Patterns and Relationships:

- The graph shows the relationship between the stock's closing price on one day and its closing price on the next day.
- If the points form a clear pattern (e.g., a straight line), it indicates a strong relationship between the two days' prices.

Analysis of Trends, Seasonality, and Anomalies:

- **Trends:** The points in the graph show a positive trend, meaning that if the stock price is high on one day, it is likely to be high on the next day as well.
- **Seasonality:** There is no seasonality in this graph because it compares prices on consecutive days, not over longer periods.
- **Anomalies:** There are some points that deviate from the general trend, which could represent unusual price movements.

Missing Value Analysis



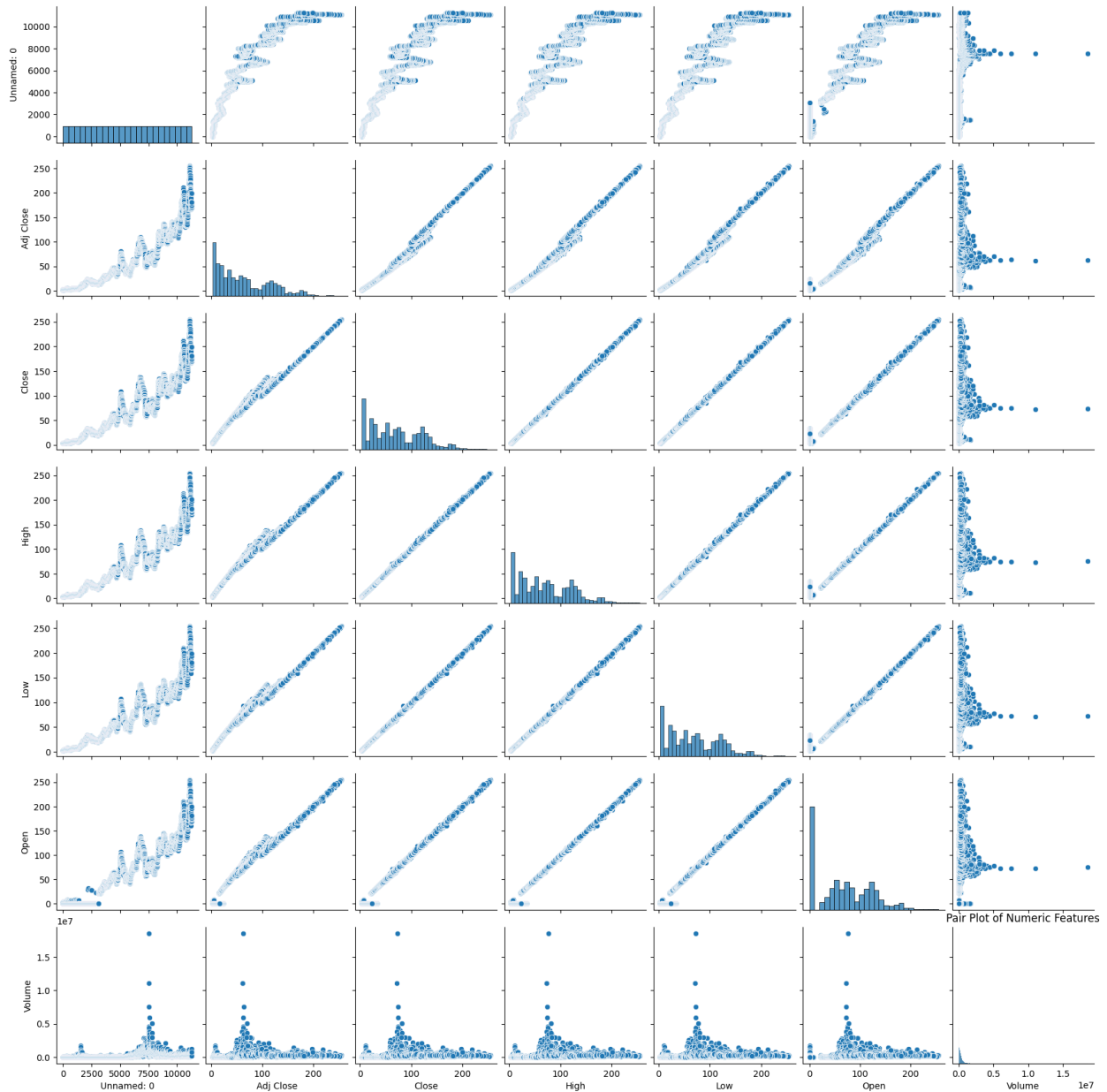
- The graph is a **heatmap** that shows whether there are any missing values in the dataset.
- The **x-axis** represents the **columns** of the dataset (e.g., "Adj Close," "Close," "High," "Low," "Open," "Volume").
- The **y-axis** represents the **dates** (rows) in the dataset.
- The color of each cell indicates whether the data is missing (e.g., yellow for missing, blue for present).

Visualizations of Key Patterns and Relationships:

- We already handle the missing values in the dataset therefore, heatmap shows that there are **no missing values** in the dataset, as all cells are uniformly colored (likely blue).
- This indicates that the dataset is complete and ready for analysis.

Anomalies: There are no anomalies because the dataset is complete.

Pair Plot



Explanation of the Pair Plot

Each row and column in the grid represents a different **numerical feature**. The diagonal contains **histograms** showing the distribution of individual features, while the **scatter plots** show pairwise relationships between them.

Features in the Plot From the labels, the dataset includes:

- **Unnamed: 0** (index column)
- **Adj Close** (Adjusted closing price of the stock)
- **Close** (Closing price)
- **High** (Highest price of the stock for the day)
- **Low** (Lowest price of the stock for the day)
- **Open** (Opening price of the stock for the day)
- **Volume** (Number of shares traded)

Key Observations

Histograms (Diagonal)

- Each histogram represents the distribution of a single feature.
- The **Adj Close**, **Close**, **High**, **Low**, and **Open** prices all follow an increasing trend, indicating long-term stock growth.
- **Volume** has a right-skewed distribution, meaning a few days had extremely high trading activity.

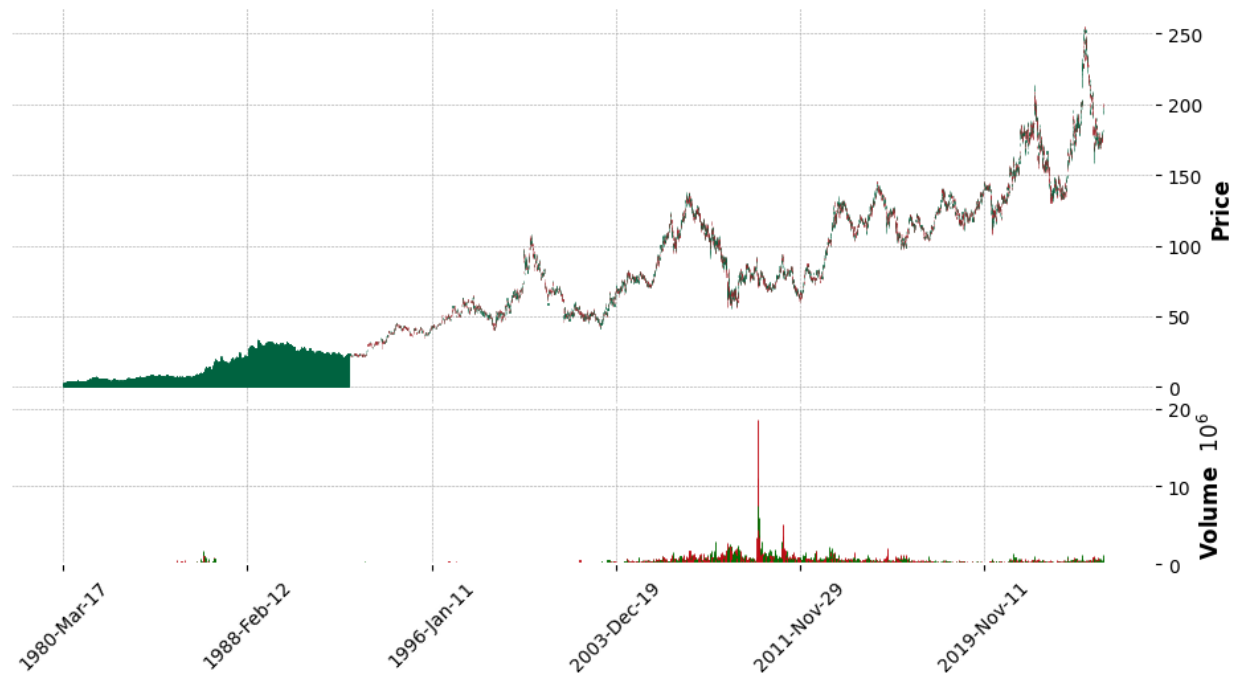
Strong Correlations (Scatter Plots)

- The **Close**, **High**, **Low**, **Open**, and **Adj Close** prices show a perfect linear relationship with each other.
- This is expected, as these values are closely related to stock performance and differ slightly due to market fluctuations.

Volume vs Price

- The scatter plots involving **Volume** appear more scattered, meaning trading volume does not have a strong direct correlation with **stock price**.
- Some extreme **outliers** in **Volume** suggest a few high-trading days, possibly due to major financial events or news.

Candlestick Chart



- The candlestick chart visualizes the price movements of the stock over time. Each candlestick represents a specific time period (e.g., one day), and the color indicates whether the price increased (green/white) or decreased (red/black).
- The **volume bars** at the bottom show the trading volume for each period, which can help confirm the strength of a price movement.

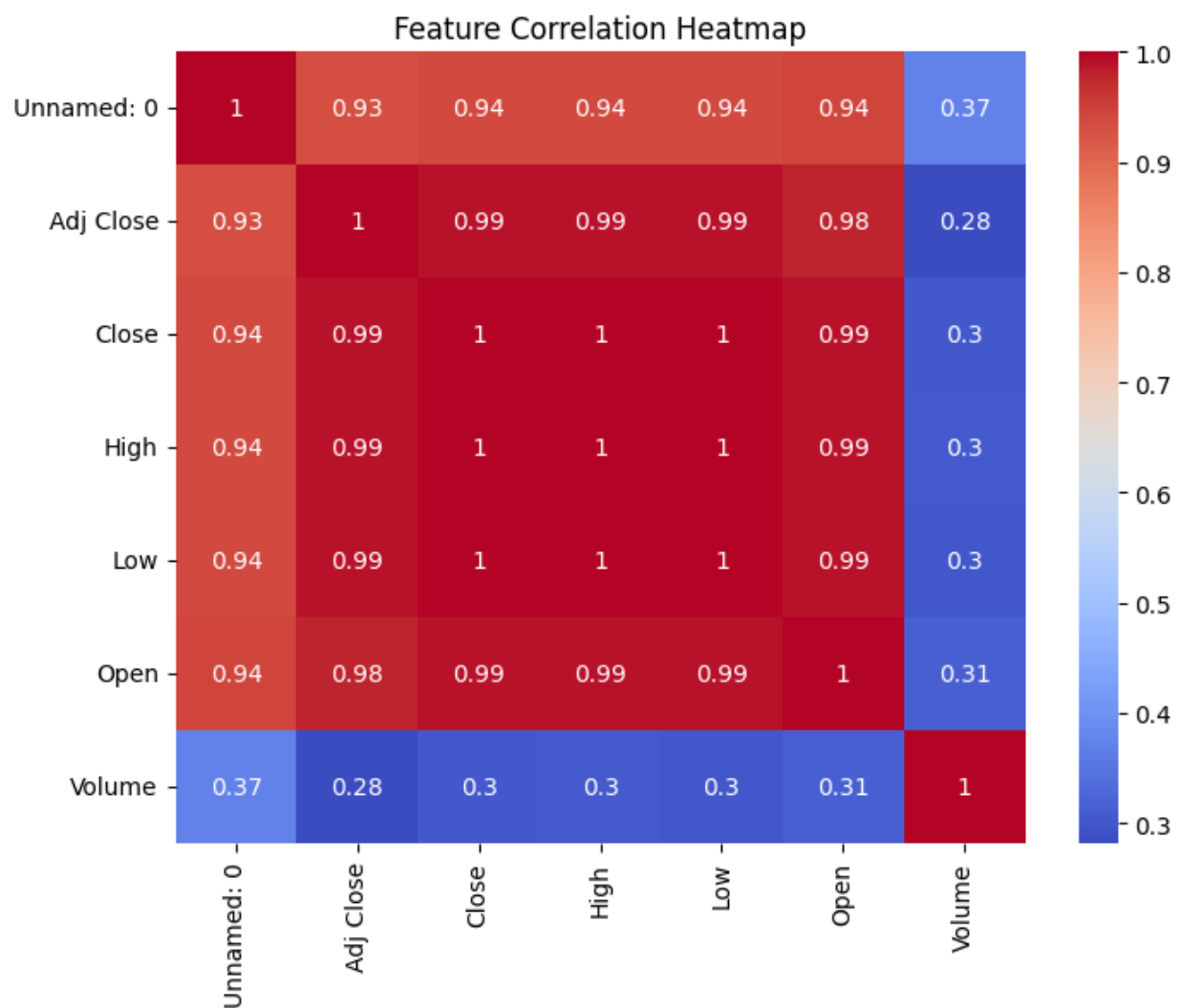
Visualizations of key patterns and relationships:

- **Trends:** The chart helps identify upward (bullish) or downward (bearish) trends in the stock price over time.
- **Support and Resistance Levels:** These are price levels where the stock tends to find support (stops falling) or resistance (stops rising).
- **Candlestick Patterns:** Certain patterns (e.g., Doji, Hammer, Engulfing) can provide insights into potential reversals or continuations in the price trend.
- **Volume Confirmation:** High trading volume during a price movement can confirm the strength of the movement, while low volume may indicate weak momentum.

Analysis of trends, seasonality, and anomalies:

- **Trends:** The chart can reveal long-term trends (e.g., a series of green candlesticks indicating a bullish trend) or short-term fluctuations.
- **Seasonality:** While candlestick charts are not typically used to identify seasonality, recurring patterns could hint at seasonal behavior.
- **Anomalies:** Sudden spikes or drops in price, accompanied by high volume, could indicate significant market events (e.g., earnings reports, news announcements).

Correlation HeatMap



What the heatmap represents:

- The heatmap shows the pairwise correlation coefficients between different features in the dataset. Each cell represents the correlation between two features.
- Features included: **Unnamed: 0**, **Adj Close**, **Close**, **High**, **Low**, **Open**, and **Volume**.

Visualizations of key patterns and relationships:

- **High Correlation (Close to 1):**
 - **Adj Close**, **Close**, **High**, **Low**, and **Open** are highly correlated with each other (correlation coefficients between 0.93 and 1.0).
 - Example: **Close** and **High** have a perfect positive correlation (1.0).
- **Low Correlation (Close to 0):**
 - **Volume** has low correlation with all other features (correlation coefficients between 0.28 and 0.37).


Analysis of trends and relationships:

- **Price Features:** The high correlation between price-related features indicates they are closely related, which is typical for stock price data.
- **Volume:** The low correlation between **Volume** and price-related features suggests that trading volume does not directly influence price movements in a linear way.
- **Multicollinearity:** The high correlation between price-related features may lead to multicollinearity in models. To address this:
 - Remove redundant features (e.g., keep only **Close**).
 - Use dimensionality reduction techniques like PCA.

Conclusion

The Exploratory Data Analysis (EDA) conducted on the historical stock price dataset has provided valuable insights into the structure, patterns, and relationships within the data. These insights will serve as the foundation for building a robust stock price prediction model. Below, we summarize the key findings and their implications for the next steps in the project.

The dataset exhibits **long-term patterns** in stock prices, such as upward trends and cyclical fluctuations. These patterns suggest that the stock price is influenced by both long-term growth and periodic market cycles. To effectively capture these patterns, **Long Short-Term Memory (LSTM) models** are highly recommended. LSTMs are well-suited for time series data, as they can learn and remember long-term dependencies, making them ideal for predicting stock prices based on historical trends.



The correlation analysis revealed that features such as Open, High, Low, and Close are **highly correlated**. This strong correlation indicates that these features provide redundant information, and using all of them in the model may lead to overfitting or unnecessary complexity. Therefore, it is sufficient to use only the **Close price** as the primary feature for prediction. This simplification not only reduces the dimensionality of the dataset but also improves the model's efficiency and interpretability.

The analysis of rolling volatility and quantile trends highlighted periods of **high volatility**, particularly during market crises such as the Dot-com Bubble (2000) and the 2008 Financial Crisis. These findings underscore the importance of incorporating **volatility metrics** (e.g., rolling standard deviation) into the model to account for risk and uncertainty in stock price movements. Additionally, the presence of extreme price movements (outliers) suggests the need for robust preprocessing techniques to handle anomalies and ensure model stability.

The Augmented Dickey-Fuller (ADF) test revealed that the stock price data is **non-stationary**, meaning its statistical properties (e.g., mean, variance) change over time. To address this, **differencing** or **transformation techniques** (e.g., logarithmic transformation) should be applied to make the data stationary before training the model. This step is crucial for improving the accuracy and reliability of time series predictions.

While the Close price is sufficient for basic predictions, additional features such as **moving averages**, **rolling returns**, and **price differences** can be engineered to capture more nuanced aspects of market behavior. These features can enhance the model's ability to identify short-term trends and patterns, leading to more accurate predictions.

Despite the insights gained from this EDA, there are some limitations to consider. The dataset lacks external factors such as **news sentiment**, **economic indicators**, and **market events**, which can significantly impact stock prices. Incorporating such features in future work could improve the model's predictive power. Additionally, the dataset's historical span may not fully capture all market conditions, suggesting the need for continuous updates and model retraining as new data becomes available.

In conclusion, this EDA has provided a comprehensive understanding of the stock price dataset, revealing long-term patterns, feature correlations, and volatility trends. By leveraging these insights, we can build a simplified yet effective prediction model using the Close price as the primary feature and LSTM as the modeling approach. Moving forward, feature engineering, data preprocessing, and the incorporation of external factors will be critical for enhancing the model's performance and ensuring its adaptability to changing market conditions.