IEE 579 Time Series Analysis and Forecasting

Stock analysis on NVIDIA Dataset

TEAM MEMBERS:

Chandana Pulikanti	*1233365902*
Roshita Sunil Chandnani	*1233694646*
Divija Lekkalapudi	*1233702277*

I. INTRODUCTION

In the realm of financial analysis, time series data plays a critical role in understanding market trends and making informed investment decisions. NVIDIA Corporation, a global leader in GPU and AI technologies, has a prominent presence in the semiconductor and technology sectors. The stock data reflects the market's response to NVIDIA's financial performance, industry developments, and macroeconomic factors. As a key player in the tech industry, NVIDIA's stock is influenced by earnings announcements, product launches, and broader market trends. The dataset sourced from Yahoo Finance offers an accurate historical record for detailed analysis.

Historical stock data for NVIDIA and similar companies has been widely used in academic research, financial modeling, and algorithmic trading. Prior studies have leveraged such datasets for:

- Predicting stock prices using machine learning models.
- Studying the impact of global events on technology stocks.
- Designing investment strategies using technical indicators like moving averages and RSI.

While this data was collected for financial analysis, its applications extend to academic explorations and market risk assessments.

We have collected a dataset of NVIDIA Corporation to analyze its historical stock prices and trading behaviors. Through this analysis, we aim to uncover trends, identify patterns, and predict future stock price movements, facilitating better investment strategies and risk management.

II. Dataset Description

The dataset comprises 427 data points covering daily trading information for NVIDIA Corporation from 2023-01-03 to 2024-09-13. The following columns are included:

- **Date**: The trading date.
- **Open**: Stock's opening price.
- **Close**: Stock's closing price.
- **High**: Highest price reached during the trading session.
- **Low**: Lowest price reached during the session.
- **Volume**: The total number of shares traded during the session.
- **Dividends**: Cash dividends paid to shareholders (if any).
- **Stock Splits**: Adjustments in share price due to stock splits.
- **MA 50**: 50-day moving average, used to smooth short-term fluctuations.
- **MA 200**: 200-day moving average, reflecting long-term price trends.

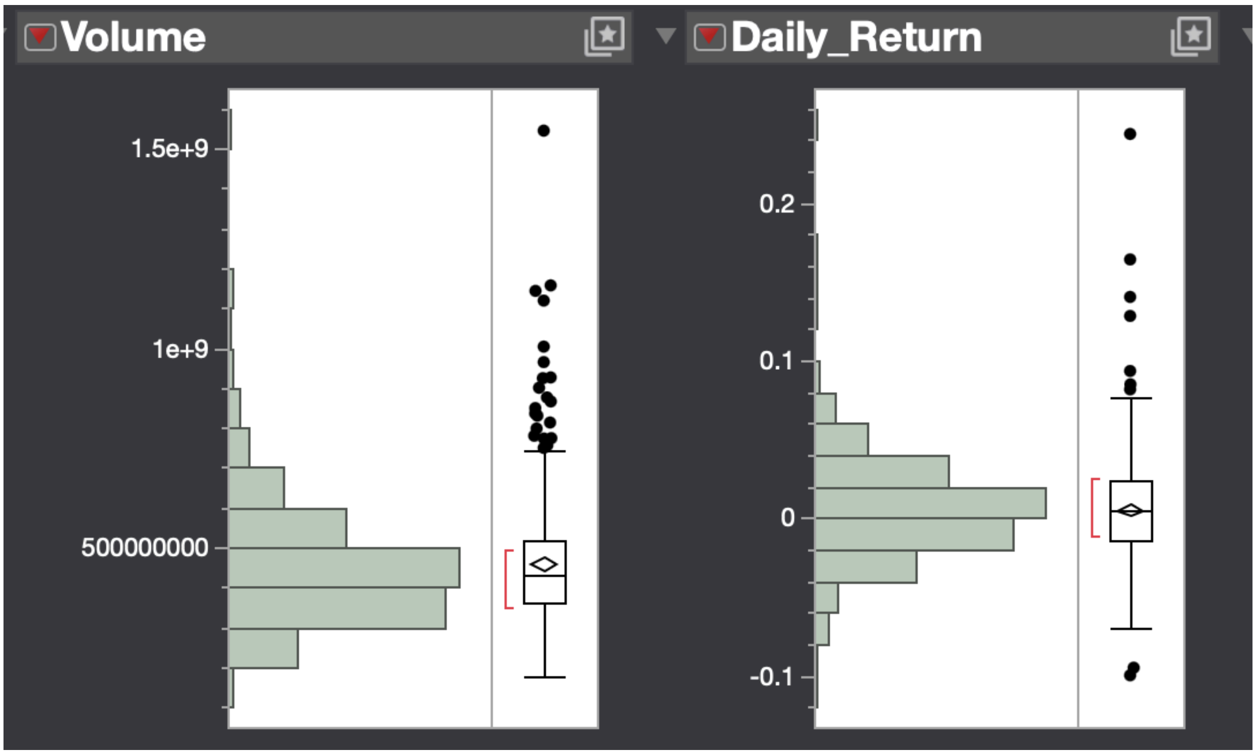
The dataset provides critical information for exploring NVIDIA's stock price dynamics, identifying trends, and applying predictive analytics.

III. DATA PREPROCESSING

1. Missing Data Handling (Interpolation)

Our data was clean and didn't have any missing values. So interpolation techniques weren't needed.

2. Outlier Detection



Volume:

- There are visible points beyond the whiskers in the box plot, indicating **potential outliers** in the volume data. These could represent days with unusually high trading activity, possibly driven by significant news or events.

Daily Return:

- The box plot also shows points beyond the whiskers, suggesting **potential outliers** in the daily return values. These could represent highly volatile days caused by earnings announcements, market corrections, or other economic events.
- The distribution of daily returns is centered around 0, indicating no strong bias toward gains or losses over time. However, the tails suggest rare but extreme market movements, which can significantly influence investment decisions.

We choose to retain the outliers as they might give us valid insights.

3. Formatting the Date Column

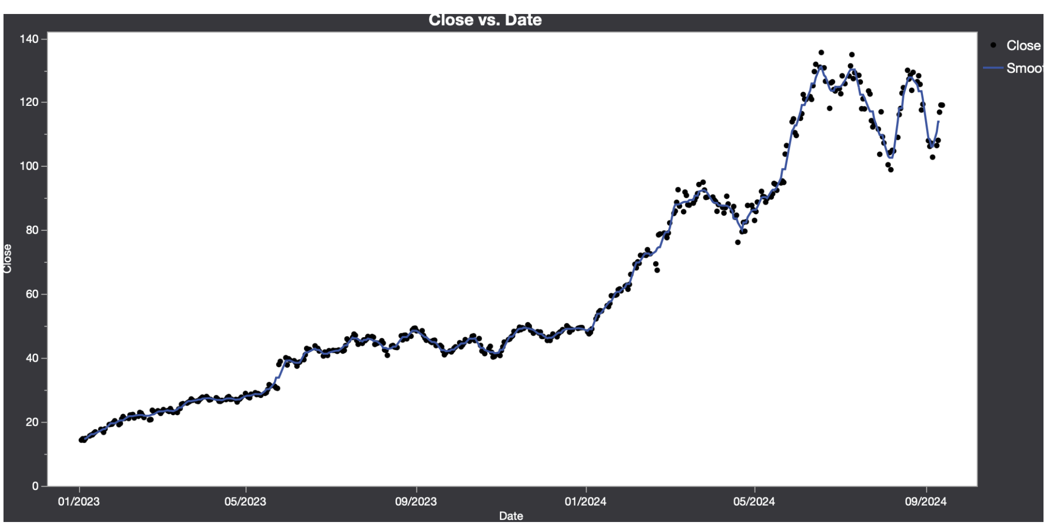
The original Date column contained mixed data types, including timestamps.The data was standardized by converting the Date column into a continuous and consistent format suitable for chronological analysis. This ensured smooth plotting and time-based computations.

nvidia_stock_data

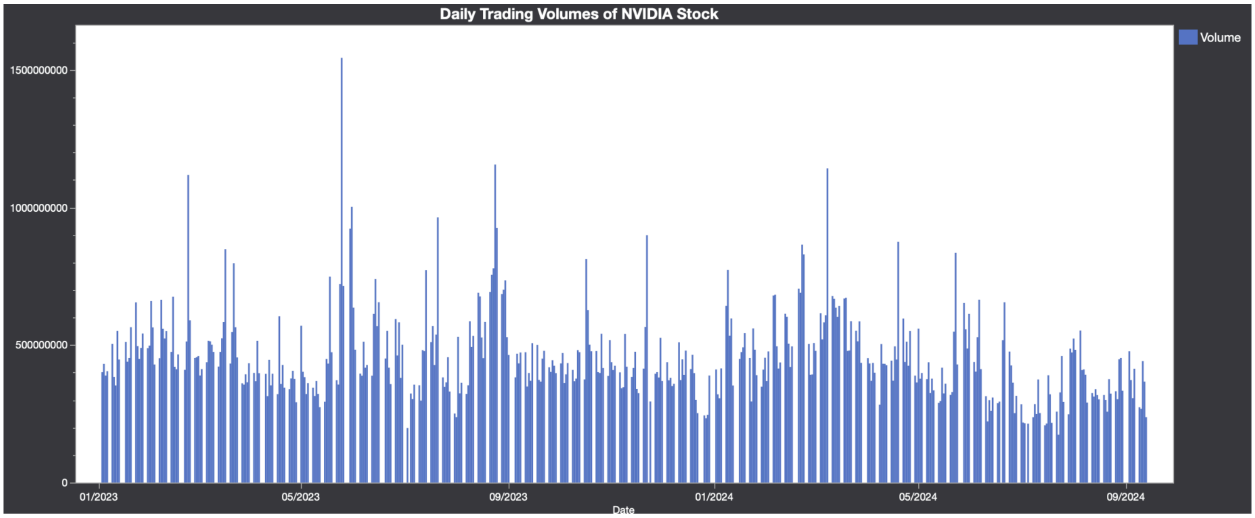
	Date	Open	High	Low	Close	Volume
1	03/01/2023	14.84	14.99	14.09	14.31	401277000
2	04/01/2023	14.56	14.84	14.23	14.74	431324000
3	05/01/2023	14.48	14.55	14.14	14.26	389168000
4	06/01/2023	14.46	15	14.02	14.85	405044000
5	09/01/2023	15.27	16.05	15.13	15.62	504231000
6	10/01/2023	15.5	15.95	15.46	15.9	384101000
7	11/01/2023	15.83	16.02	15.55	15.99	353285000
8	12/01/2023	16.09	16.63	15.48	16.5	551409000
9	13/01/2023	16.27	16.91	16.15	16.89	447287000
10	17/01/2023	16.89	17.72	16.89	17.69	511102000
11	18/01/2023	17.66	17.86	17.27	17.37	439624000
12	19/01/2023	17.02	17.19	16.72	16.75	452932000
13	20/01/2023	17	17.84	16.81	17.83	564967000
14	23/01/2023	18.05	19.23	17.81	19.18	655163000
15	24/01/2023	18.81	19.48	18.81	19.25	496204000

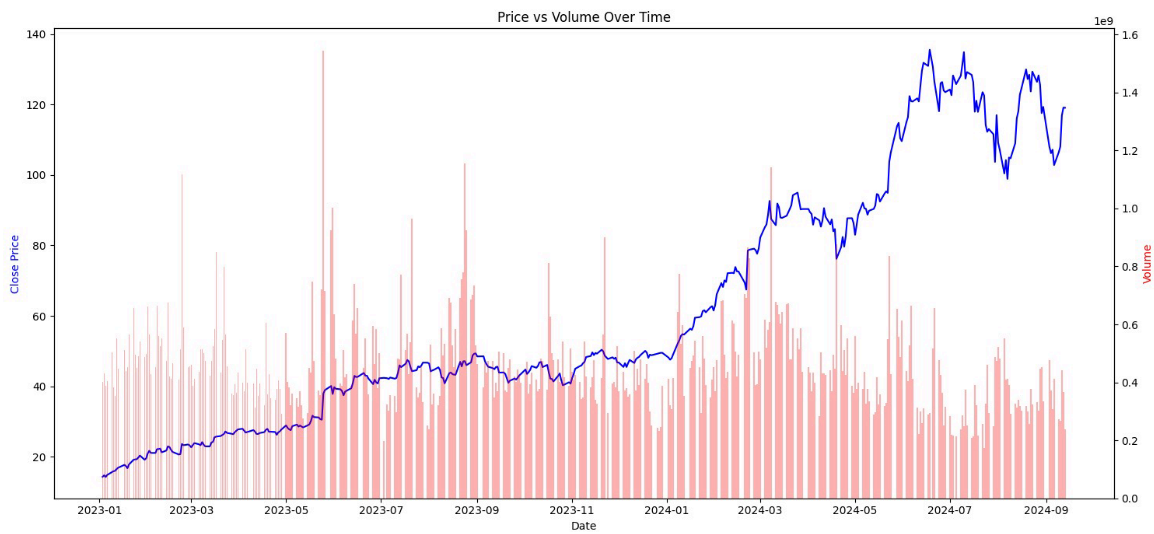
III. Exploratory Data Analysis (EDA)

1. A **line chart** of daily closing prices to observe overall trends.



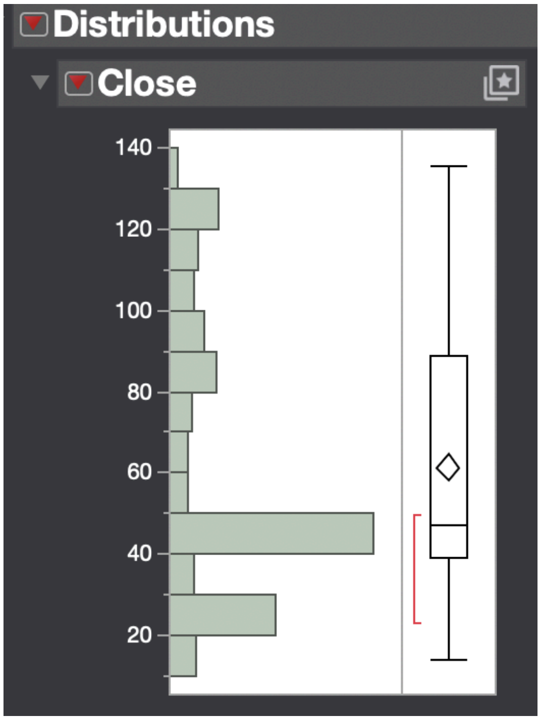
2. **Bar charts** depicting daily trading volumes.





3. Statistical tables summarizing descriptive measures, such as mean, median, standard deviation, and range for all numerical columns.

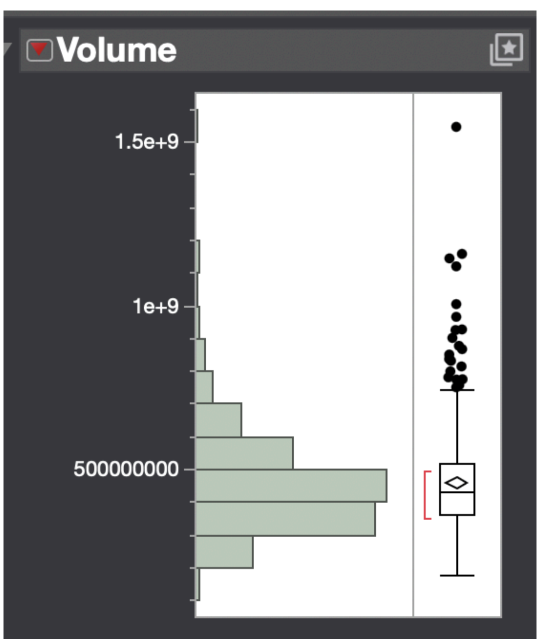
SUMMARY STATS OF CLOSE



Summary Statistics

Mean	61.31178
Std Dev	34.10034
Std Err Mean	1.6502311
Upper 95% Mean	64.555389
Lower 95% Mean	58.068171
N	427
N Missing	0

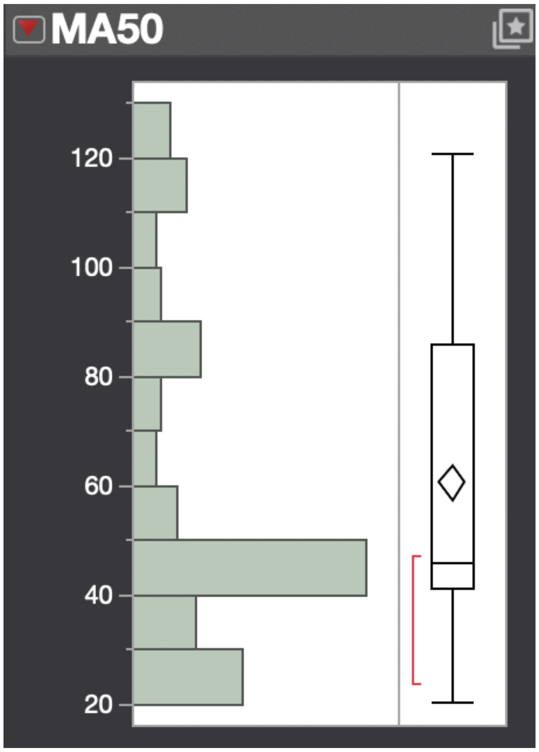
SUMMARY STATS OF VOLUME



Summary Statistics

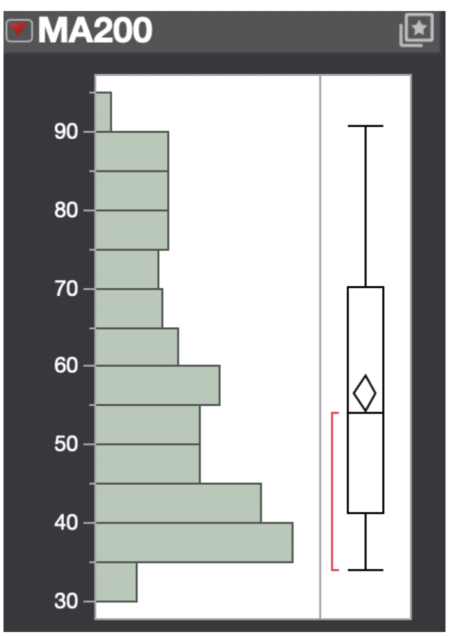
Mean	459383636
Std Dev	157785297
Std Err Mean	7635765.7
Upper 95% Mean	474392102
Lower 95% Mean	444375170
N	427
N Missing	0

SUMMARY STATS OF MA50



Summary Statistics	
Mean	60.527252
Std Dev	30.692092
Std Err Mean	1.5786309
Upper 95% Mean	63.631277
Lower 95% Mean	57.423228
N	378
N Missing	49

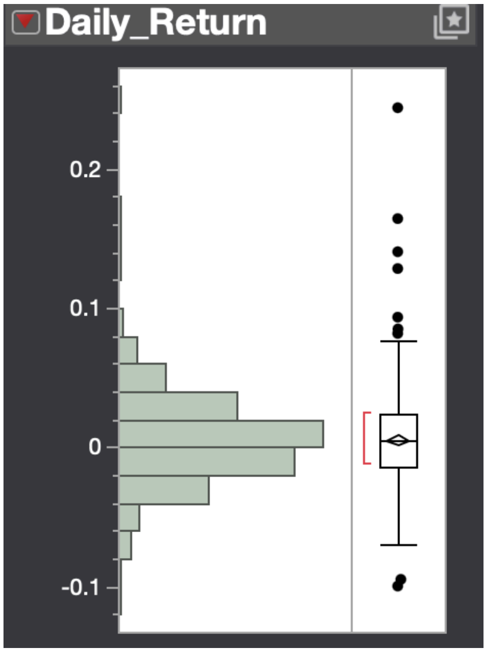
SUMMARY STATS OF MA200

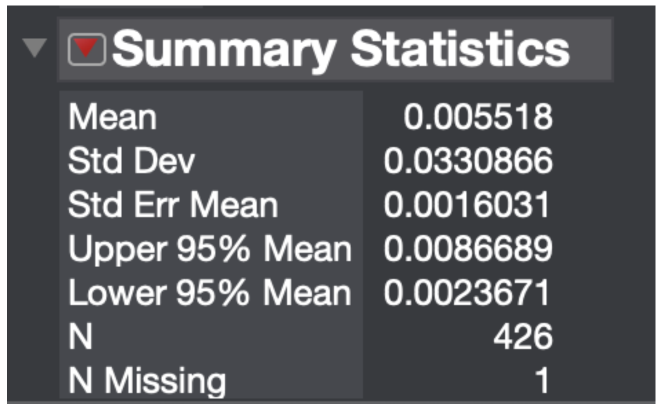


Summary Statistics	
Mean	56.637088
Std Dev	16.815836
Std Err Mean	1.1136559
Upper 95% Mean	58.831513
Lower 95% Mean	54.442663
N	228
N Missing	199

- The distribution plots for **Open**, **Close**, **High**, **Low**, and **Volume** reveal a slightly skewed distribution with a few outliers, especially in the **Volume.**

SUMMARY STATS OF DAILY RETURNS



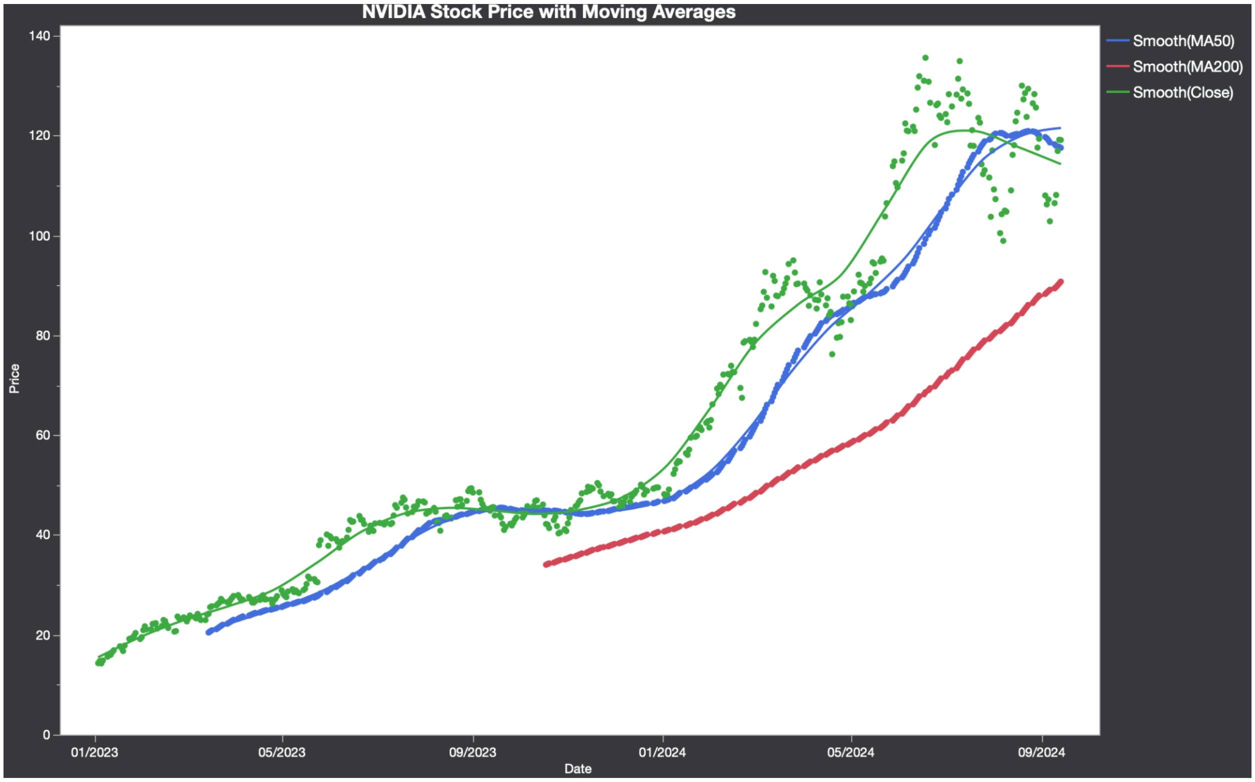


Summary Statistics

Mean	0.005518
Std Dev	0.0330866
Std Err Mean	0.0016031
Upper 95% Mean	0.0086689
Lower 95% Mean	0.0023671
N	426
N Missing	1

  - **Daily return** was calculated as it can examine daily returns to uncover recurring patterns, such as higher volatility during earnings releases, or macroeconomic announcements.
  - We can rely on returns rather than raw prices to achieve stationarity, making the series easier to model and predict.
  - The distribution of daily returns is centered around 0, indicating no strong bias toward gains or losses over time. However, the tails suggest rare but extreme market movements, which can significantly influence investment decisions.

4. An **overlay plot** combining moving averages (MA 50 and MA 200) with closing prices, revealing significant price trends and crossovers.



- The upward trend of both moving averages indicates that NVIDIA stock has generally been increasing over time.

Time Series Basic Diagnostics

Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.9917	422.925	<.0001*	1	0.9917
2	0.9841	840.307	<.0001*	2	0.0319
3	0.9760	1251.88	<.0001*	3	-0.0239
4	0.9692	1658.71	<.0001*	4	0.0702
5	0.9623	2060.65	<.0001*	5	-0.0103
6	0.9561	2458.40	<.0001*	6	0.0427
7	0.9492	2851.38	<.0001*	7	-0.0370
8	0.9424	3239.69	<.0001*	8	-0.0041
9	0.9353	3623.07	<.0001*	9	-0.0149
10	0.9266	4000.26	<.0001*	10	-0.1086
11	0.9189	4372.04	<.0001*	11	0.0516
12	0.9105	4737.98	<.0001*	12	-0.0413
13	0.9022	5098.11	<.0001*	13	-0.0193
14	0.8933	5452.03	<.0001*	14	-0.0305
15	0.8848	5800.10	<.0001*	15	0.0045
16	0.8768	6142.72	<.0001*	16	0.0405
17	0.8680	6479.35	<.0001*	17	-0.0636
18	0.8593	6810.03	<.0001*	18	0.0030
19	0.8506	7134.87	<.0001*	19	0.0126
20	0.8423	7454.19	<.0001*	20	0.0045
21	0.8342	7768.18	<.0001*	21	0.0236
22	0.8263	8077.00	<.0001*	22	-0.0020
23	0.8184	8380.68	<.0001*	23	0.0104
24	0.8116	8680.10	<.0001*	24	0.0613
25	0.8055	8975.75	<.0001*	25	0.0401

INSIGHTS DRAWN

Autocorrelation (ACF)	High autocorrelation at lag 1 (0.9917) and gradually decreasing values with increasing lag suggest that the time series is highly correlated with its past values autocorrelation decreases slowly, suggesting a non-stationary time series.
Partial Autocorrelation (PACF)	Lag 1 has a strong partial autocorrelation (0.9917), but the remaining values fluctuate and quickly diminish. suggests that the series is dominated by a trend or other components, which may require differentiation to achieve stationarity.
Ljung-Box Test Results	p-Values (< 0.0001): These small p-values suggest that there is significant autocorrelation present in the series, confirming that the data is not white noise. This further supports the need for transformation or differencing to remove trends and make the series stationary.
Augmented Dickey-Fuller Test	Zero Mean ADF (1.5778059):no deterministic trend, series is **not stationary** Single Mean ADF (-0.426172): series is still non-stationary. Trend ADF (-2.287012):fails to reject the null hypothesis of non-stationarity, indicating the presence of a trend in the series
time series is **non-stationary**	NEXT STEPS: **Transformation:** Log or square root transformations can stabilize variance **Differencing:** Applying first-order differencing to remove the trend and achieve stationarity. **Decomposition:** Decomposing the series into trend, seasonality, and residual components.

IV. TRANSFORMATIONS:

- **Stabilizing Variance:**

Financial time series often exhibit heteroscedasticity, where variance changes over time. These transformations help stabilize variance, making the series more homoscedastic and easier to model.

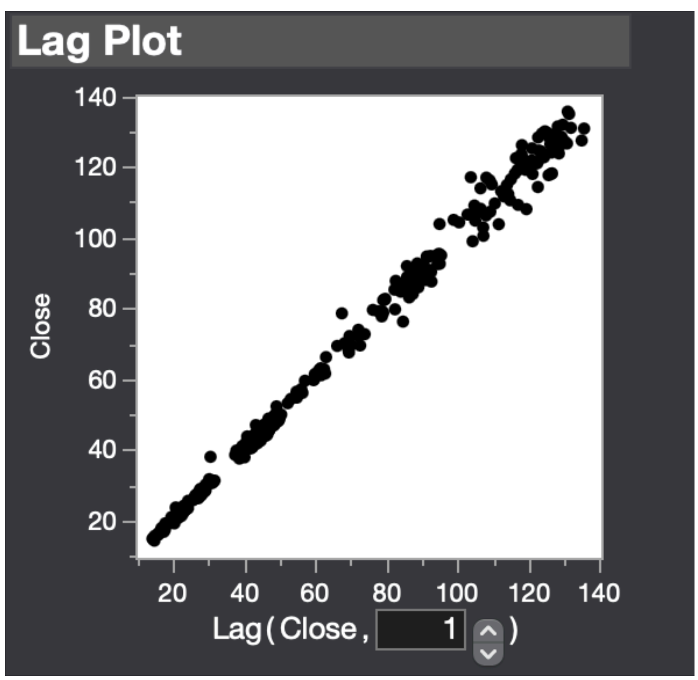
- **Improving Stationarity:**

Non-stationary series, characterized by trends or varying means, require transformation to meet the assumptions of many forecasting models like ARIMA. The log transformation particularly helps reduce the impact of large spikes in data, such as those caused by significant market events.

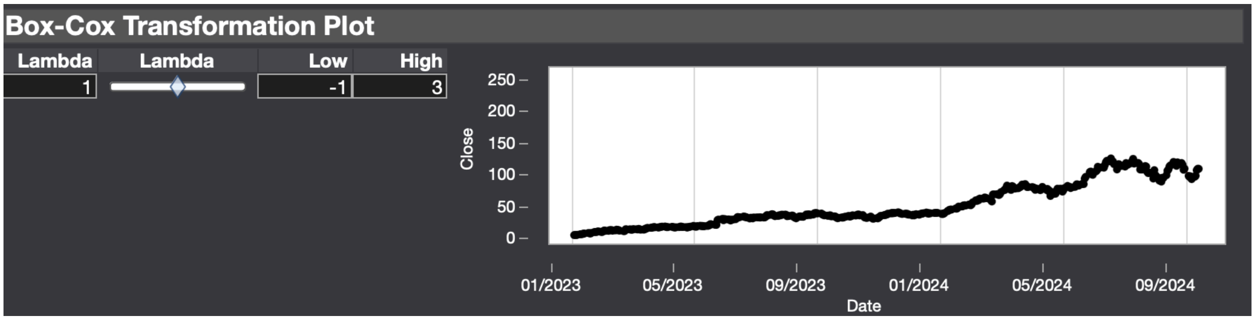
- **Enhanced Predictive Accuracy:**

By reducing skewness and normalizing the data, these transformations improve the performance of statistical and machine learning models, as they perform better on stabilized and normalized datasets.

Lag Plot: Linear alignment of points in the Lag Plot confirms a high degree of autocorrelation in the time series.



Box-Cox Transformation: A Box-Cox transformation with λ=1 (log transformation) was applied to handle skewness and stabilize variance, which is critical for improving model performance.



V. Methodologies used in Stock Market Analysis

- In stock price forecasting, ***ARIMA*** is a foundational time series model that captures trends, seasonality, and past forecast errors.
- When seasonality is present, ***SARIMA*** extends ARIMA by adding seasonal components to model periodic price fluctuations.
- GARCH models are used to capture time-varying volatility, essential for understanding market risk and volatility.
- Exponential Smoothing (ETS) provides short-term trend forecasting by weighting recent observations more heavily.
- Vector Autoregression (VAR) models multiple interdependent time series, useful for analyzing relationships between different financial assets.
- Machine learning models like Random Forest and Decision Trees handle non-linear relationships and uncover the most influential factors affecting stock prices. Support Vector Machines (SVM) can be applied to predict stock movements by finding optimal decision boundaries.
- RNNs and LSTMs are deep learning models that capture long-term dependencies, making them ideal for forecasting stocks with complex temporal patterns.

VI. Exploration of Models

We considered a variety of time series modeling approaches based on the characteristics of the data:

Method 1: Exponential Smoothing (SES, Seasonal Smoothing, Winters Method)

Exponential smoothing methods are suitable for capturing trends and seasonality in time series data.We tested different configurations: simple exponential smoothing, seasonal smoothing, and Winters additive method.

Challenges/Results:

While these methods captured trends reasonably well, they struggled with complex seasonality and produced higher errors (MAPE, MAE) compared to other models.

Method 2: ARIMA Models

ARIMA models were chosen for their strength in handling autocorrelation and differenced series to address non-stationarity.We applied transformations (1−B)(1−B5)^1 to make the series stationary and used ACF/PACF to identify potential ARIMA configurations (e.g., ARIMA(1,1,1) and ARIMA(1,1,2)).

Challenges/Results:

ARIMA performed well and provided low AIC and error metrics, but it required significant preprocessing (differencing and transformations).

Method 3: Regression Analysis

Regression was explored to model the relationship between time (as a proxy for trend) and closing prices.It served as a baseline model for comparison.

Challenges/Results:

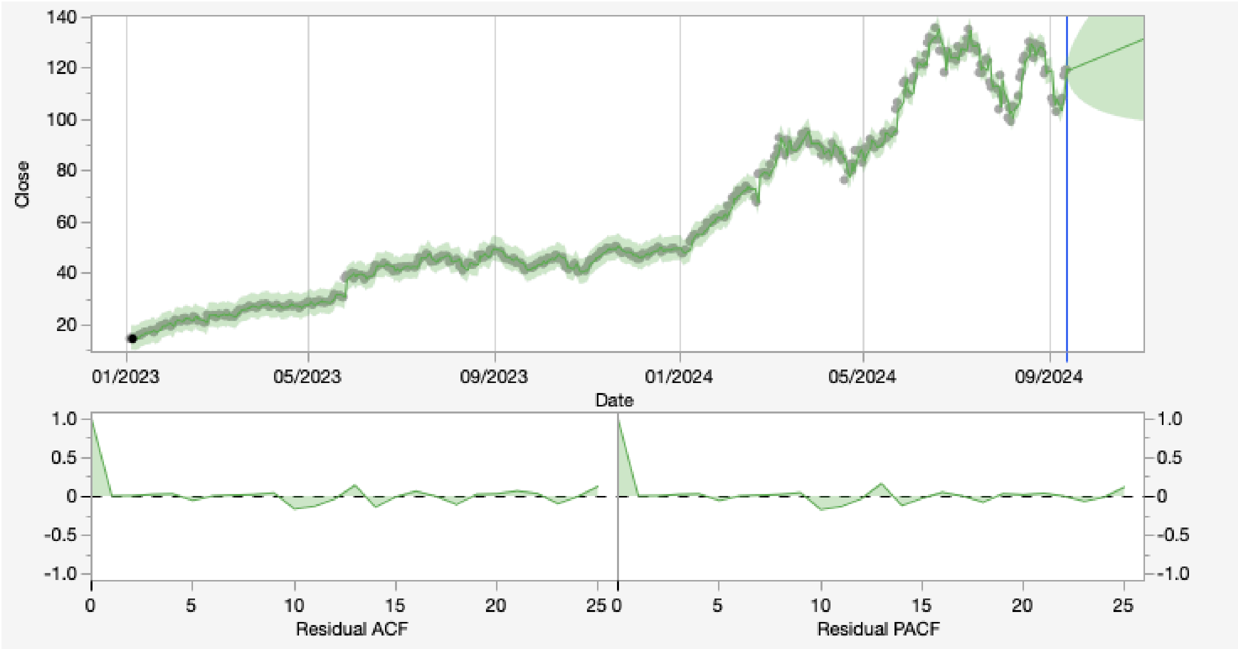
The regression model showed a good R^2 but residual diagnostics revealed non-stationarity and autocorrelation, indicating it was not well-suited for capturing time series dynamics.

VII. Time Series Model Selection and Evolution for NVIDIA Stock Price Analysis

This analysis follows a structured approach to model and forecast NVIDIA's stock prices using progressively advanced time series models. The goal is to demonstrate how each model builds on the previous one, leading to more precise predictions that capture both short-term price fluctuations and seasonal market trends.

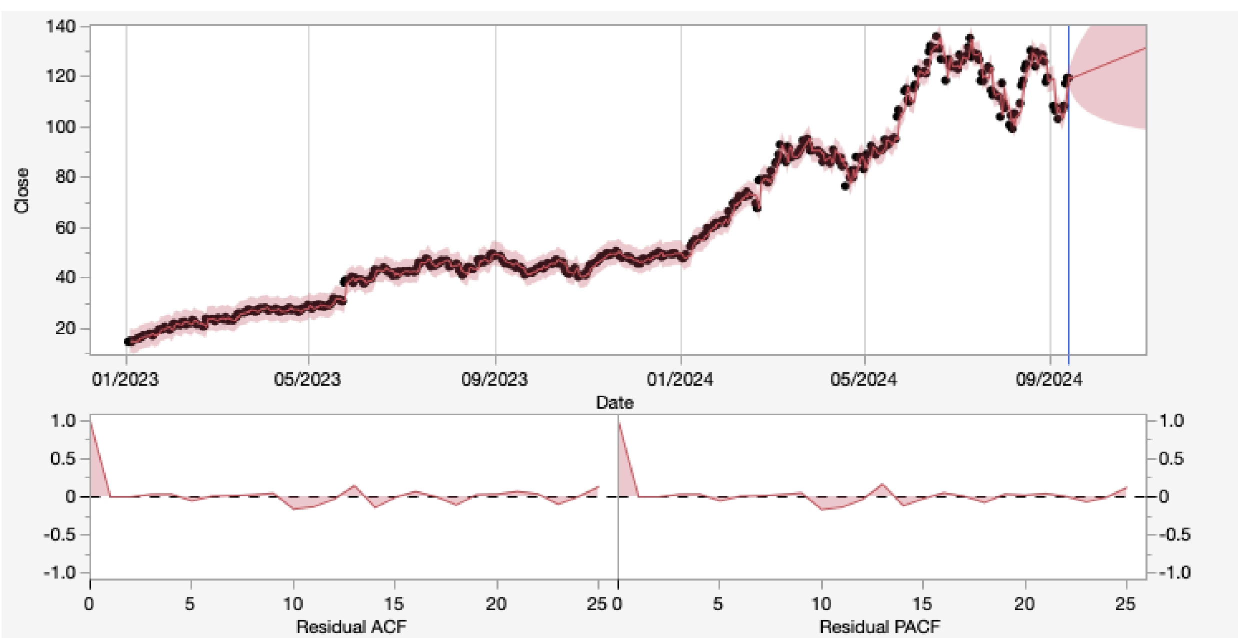
ARIMA(1,1,1) Model:

- The initial model selected was ARIMA(1,1,1), chosen due to the non-stationarity of the price series, requiring first-order differencing (d=1).
- Autocorrelation and partial autocorrelation functions revealed significant first-order relationships, justifying the inclusion of one autoregressive term (p=1) and one moving average term (q=1).
- **Performance:** The model demonstrated strong results, with an R-squared of 0.9945 and an AIC of 1270.79, effectively capturing the main price dynamics.



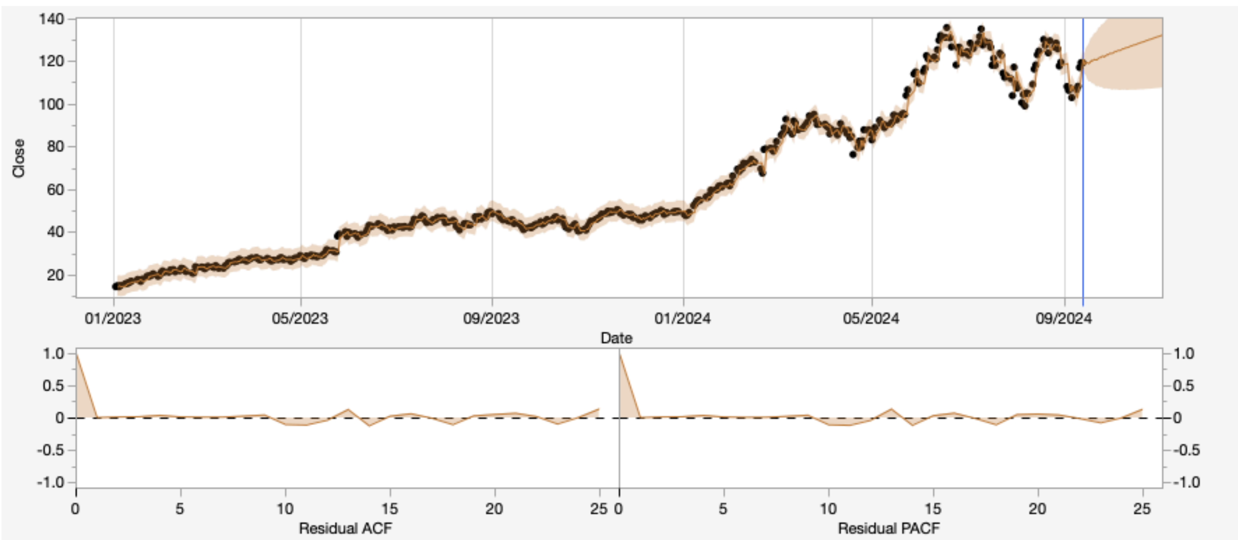
ARIMA(2,1,1) Model:

- Building on the ARIMA(1,1,1) model, a more complex ARIMA(2,1,1) model was explored, driven by patterns observed in the residuals of the simpler model.
- The second autoregressive term (p=2) was statistically significant, suggesting more intricate temporal dependencies in the stock price movement.
- **Performance:** The R-squared remained high at 0.9946, but the AIC increased slightly to 1267.76, indicating a trade-off between model complexity and predictive power.



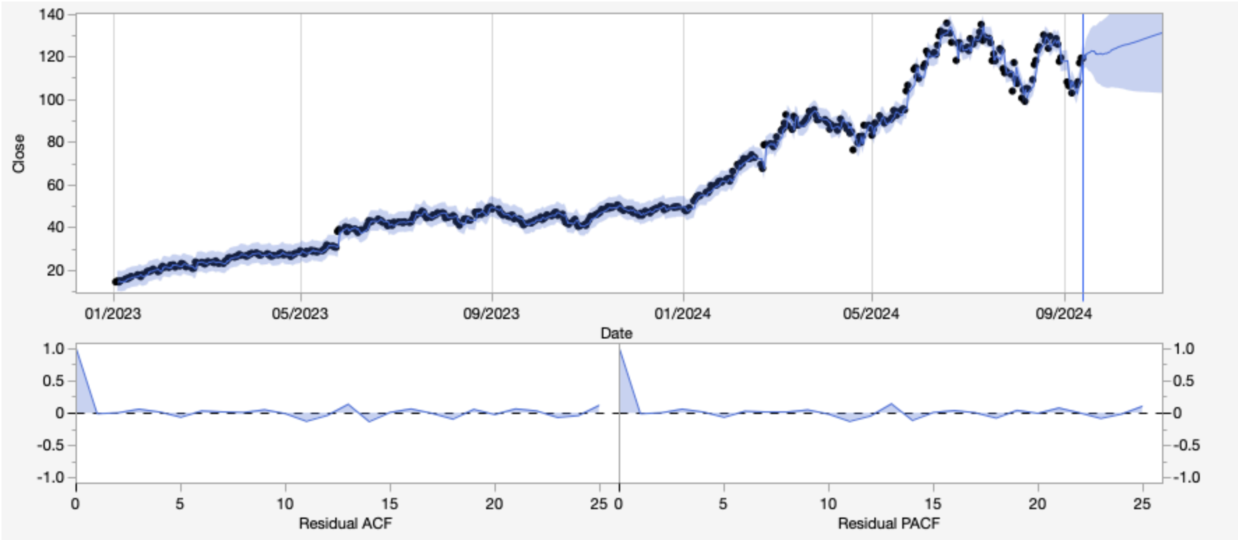
SARIMA(1,1,1)(1,0,1)10 Model:

- To capture seasonal patterns, a SARIMA(1,1,1)(1,0,1)10 model was introduced, incorporating a ten-day seasonal component. This model aimed to address bi-weekly cycles in the stock market, potentially linked to recurring events like options expiration dates.
- The model was designed to accommodate both short-term daily price fluctuations and longer-term seasonal patterns.
- **Performance:** The model's AIC was 1273.09, and residual diagnostics showed improvement, although the seasonal moving average term was not statistically significant, suggesting that further refinement could enhance the model.



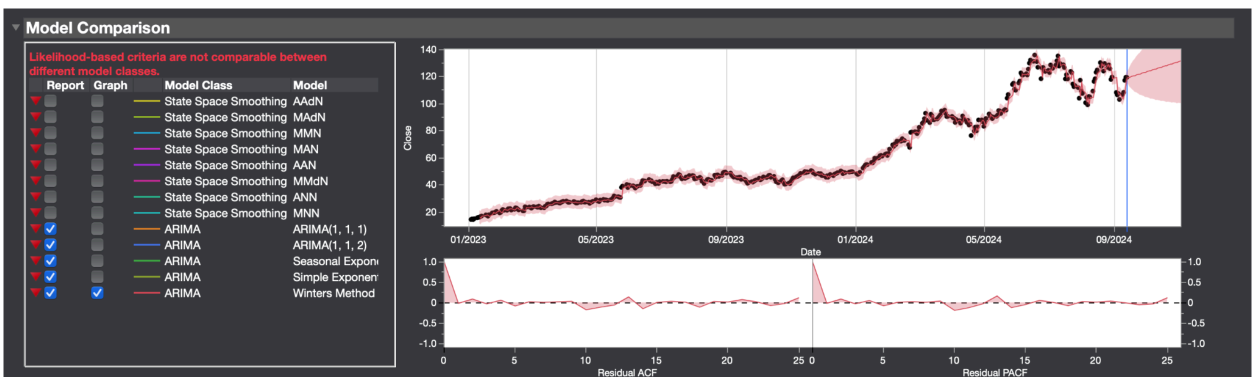
SARIMA(1,1,1)(1,0,1)5 Model:

- In the final model, the seasonal component was adjusted to a five-day cycle, corresponding to the regular weekly trading pattern in financial markets. This model aimed to capture weekly effects on stock prices.
- **Performance:** The model's performance metrics were similar to those of the ten-day seasonal model, with high R-squared values, offering different insights and potentially more relevant predictions for weekly trading strategies.



WINTER METHOD(Additive):

- Similar to the seasonal exponential smoothing model, this model performs well but is slightly outperformed by ARIMA models and shows higher variance.



Conclusion:

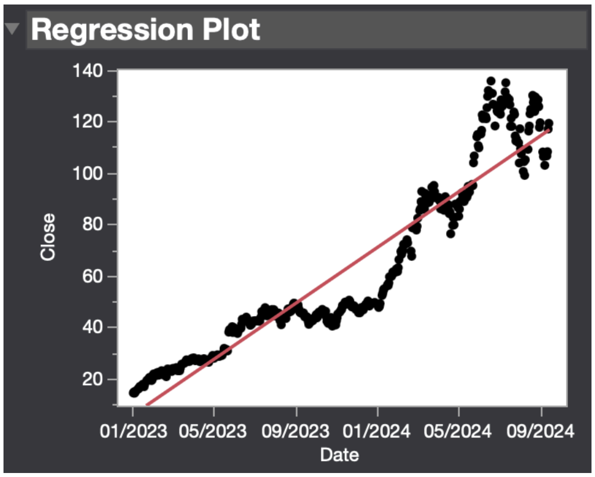
- The ARIMA models focused primarily on capturing price dynamics, while the SARIMA models introduced seasonal elements that helped capture market cycles and volatility, with the five-day seasonal model aligning well with weekly trading patterns.
- The analysis suggests that incorporating seasonal effects, particularly those related to the trading week, significantly enhances the model's forecasting capabilities.

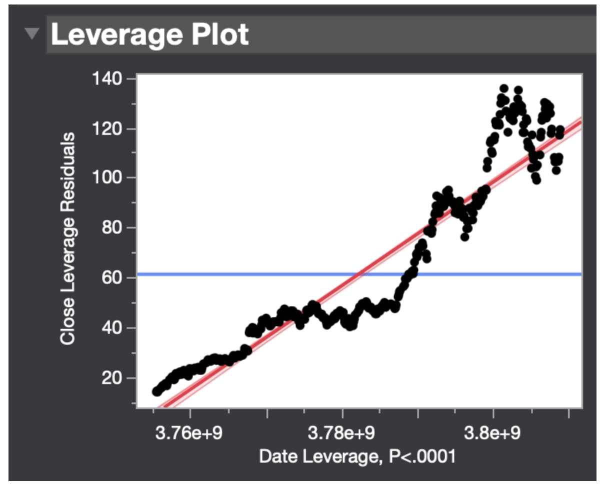
Regression Model:

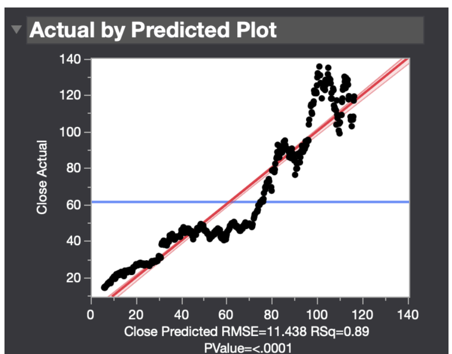
- Positive slope for Date shows a significant upward trend.
- Good R^2, but poor residual diagnostics highlighted its limitations for time series forecasting.

- Less suited to handle the time-dependent structures (autocorrelation) and trends that ARIMA and other time series models are designed to capture.

=> No extreme influential points, confirming that no observations unduly affect the model fit.







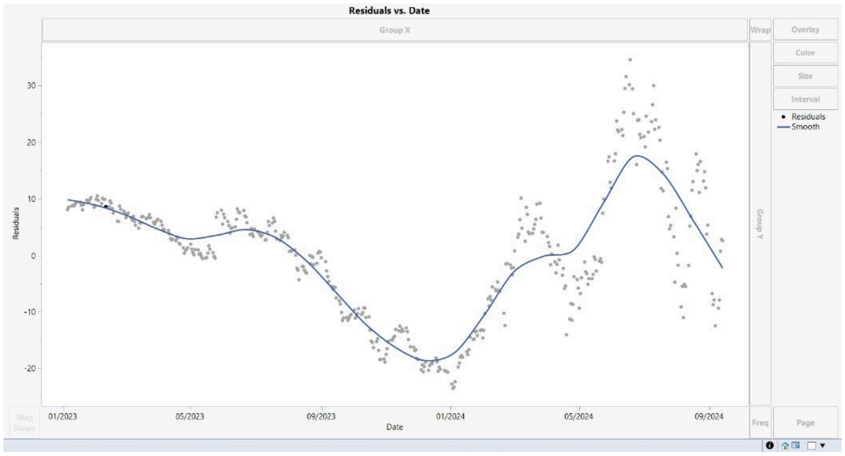
Summary of Fit

RSquare	0.887752
RSquare Adj	0.887488
Root Mean Square Error	11.4382
Mean of Response	61.31178
Observations (or Sum Wgts)	427

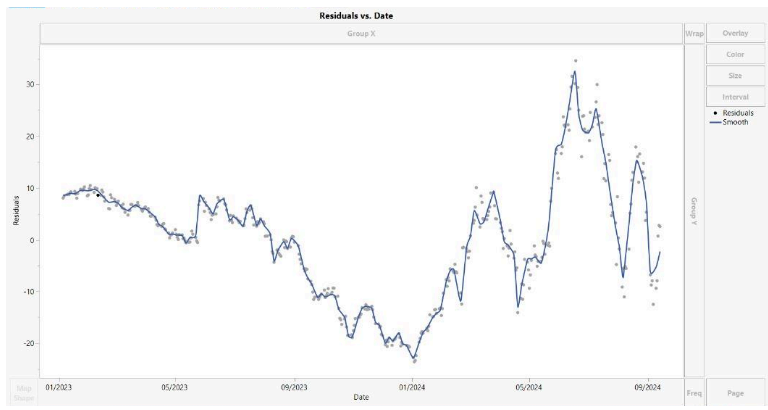
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	439763.11	439763	3361.268
Error	425	55603.82	131	Prob > F
C. Total	426	495366.93		<.0001*

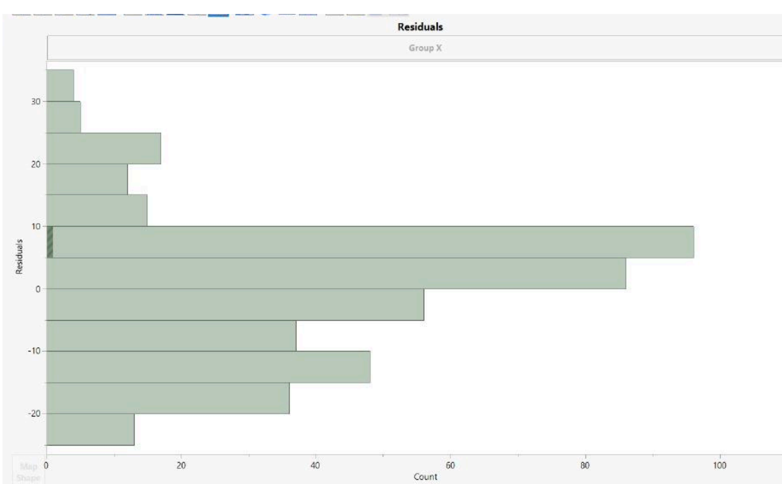
Residual Plot for closing:



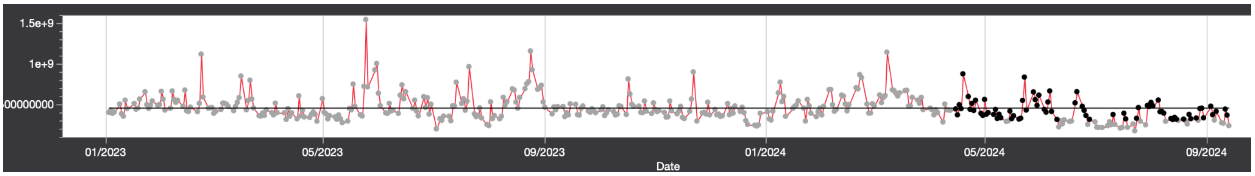
With smoothers added:



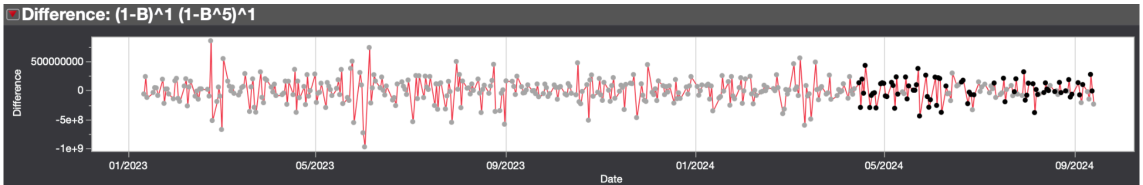
Histogram for distribution of residuals



Time series : Volume vs Date



Difference: (1-B)^1 (1-B^5)^1



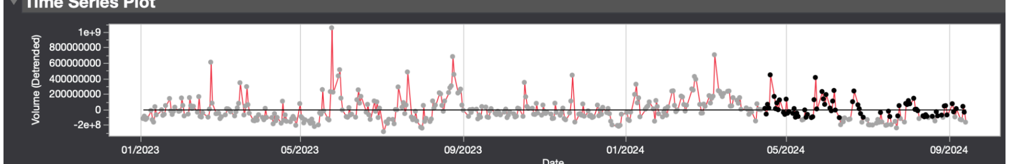
Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	-0.2972	37.4456	<.0001*	1	-0.2972
2	-0.1211	43.6781	<.0001*	2	-0.2297
3	0.0642	45.4366	<.0001*	3	-0.0559
4	0.1583	56.1325	<.0001*	4	0.1576
5	-0.4608	147.043	<.0001*	5	-0.4068
6	0.0748	149.443	<.0001*	6	-0.2107
7	0.0980	153.576	<.0001*	7	-0.1046
8	-0.0586	155.055	<.0001*	8	-0.1039
9	0.0183	155.199	<.0001*	9	0.0879
10	-0.0080	155.227	<.0001*	10	-0.2540
11	0.0590	156.738	<.0001*	11	-0.1010
12	-0.0513	157.884	<.0001*	12	-0.0935
13	0.0579	159.346	<.0001*	13	-0.0333
14	-0.0249	159.617	<.0001*	14	0.0491
15	-0.0675	161.616	<.0001*	15	-0.2759
16	0.0377	162.239	<.0001*	16	-0.0853
17	-0.0189	162.397	<.0001*	17	-0.1518
18	0.0677	164.422	<.0001*	18	0.0579
19	-0.0503	165.542	<.0001*	19	0.0108
20	0.0676	167.570	<.0001*	20	-0.1721
21	-0.0399	168.280	<.0001*	21	-0.0702
22	0.0282	168.634	<.0001*	22	-0.1105
23	-0.0659	170.579	<.0001*	23	0.0144
24	0.0790	173.378	<.0001*	24	0.0769
25	-0.0479	174.411	<.0001*	25	-0.1422

Time Series Volume (Detrended)

Linear Trend

Parameter	Value
Beta0	8.685e+9
Beta1	-2.17473

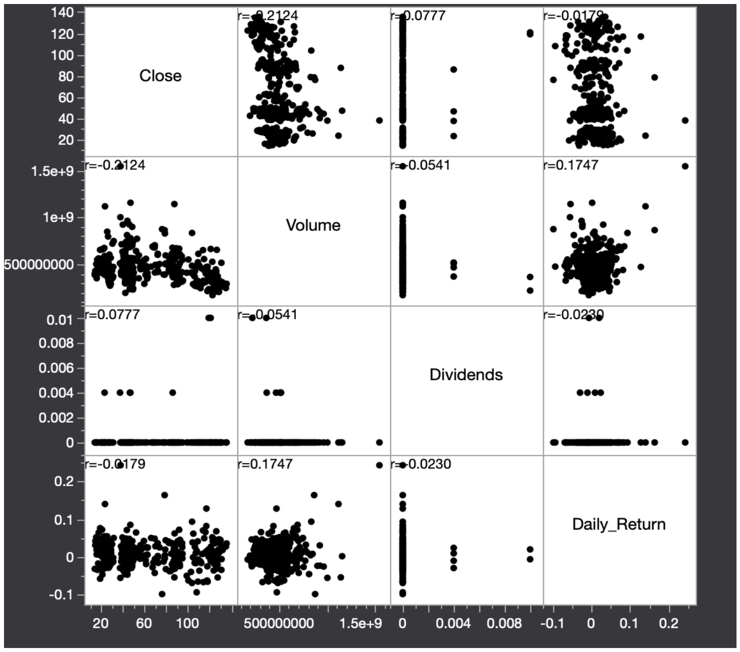
Time Series Plot



Time Series Basic Diagnostics

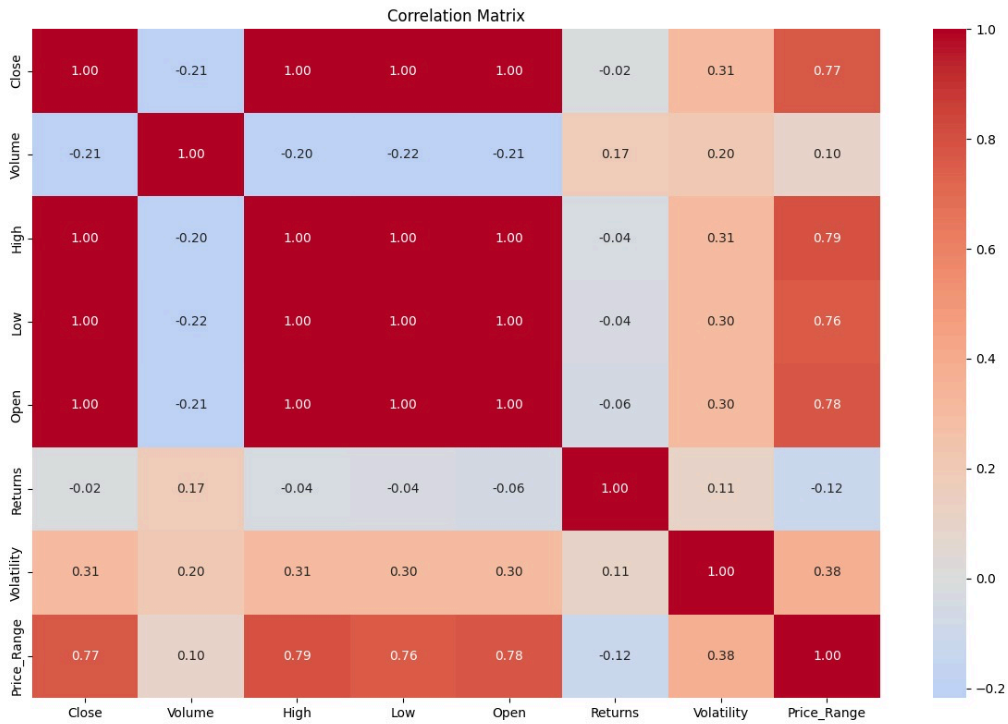
Lag	AutoCorr	Ljung-Box Q	p-Value	Lag	Partial
0	1.0000	.	.	0	1.0000
1	0.5159	114.449	<.0001*	1	0.5159
2	0.3498	167.181	<.0001*	2	0.1139
3	0.2954	204.892	<.0001*	3	0.1060
4	0.2348	228.765	<.0001*	4	0.0359
5	0.2003	246.179	<.0001*	5	0.0392
6	0.1072	251.182	<.0001*	6	-0.0664
7	0.0791	253.913	<.0001*	7	0.0018
8	0.0537	255.175	<.0001*	8	-0.0132
9	0.0392	255.850	<.0001*	9	0.0049
10	0.0402	256.560	<.0001*	10	0.0172
11	0.0477	257.561	<.0001*	11	0.0314
12	0.0155	257.666	<.0001*	12	-0.0332
13	0.0474	258.661	<.0001*	13	0.0531
14	-0.0083	258.692	<.0001*	14	-0.0747
15	-0.0136	258.775	<.0001*	15	-0.0029
16	0.0356	259.338	<.0001*	16	0.0585
17	0.0344	259.866	<.0001*	17	0.0151
18	0.0794	262.689	<.0001*	18	0.0695
19	0.0430	263.517	<.0001*	19	-0.0276
20	0.0406	264.260	<.0001*	20	0.0017
21	0.0209	264.457	<.0001*	21	-0.0377
22	0.0187	264.616	<.0001*	22	0.0036
23	0.0049	264.626	<.0001*	23	-0.0240
24	0.0341	265.156	<.0001*	24	0.0561
25	-0.0085	265.189	<.0001*	25	-0.0474

VII. CORRELATION: MULTIVARIATE CORRELATION:



The correlation matrix uncovers the following dependencies:

- **Close Price vs. Volume** shows a weak negative correlation (-0.2124), indicating that high trading volumes do not necessarily drive significant price changes.
- **Daily Return vs. Volume** has a slight positive correlation (0.1747), hinting that large trading volumes align with volatile price movements.
- **Dividends** are only weakly correlated with all other variables, reflecting that dividend payouts have limited impact on daily market movements.
- These weak correlations suggest that NVIDIA's stock behavior is influenced by external factors, such as market sentiment or macroeconomic trends, more than by internal metrics



Price Range (Correlation: 0.77)

The strong positive correlation with closing prices highlights price range as a key predictor of market activity and potential price movements. Expanding ranges often indicate increased interest and volatility.

Potential Features that can be used:

- **Rolling Averages:** Smooth short-term fluctuations and reveal trends.
- **Momentum Indicators:** Detect shifts in market activity.
- **Range-to-Price Ratios:** Normalize range to capture relative volatility.

Volatility (Correlation: 0.31)

The moderate correlation indicates that volatility captures unique insights into market uncertainty and risk. Higher volatility often signals impending significant price movements.

Potential Features that can be used:

- **Historical Volatility Measures:** Quantify past price fluctuations over fixed periods.
- **Volatility Trend Indicators:** Identify rising or falling volatility trends.

- **Volatility Regime Detection:** Differentiate between stable and turbulent market conditions.

Volume (Correlation: -0.21)

The weak negative correlation suggests that higher trading volumes often occur during price declines, signaling potential reversals and market sentiment shifts.

Potential Features that can be used:

- **Volume Trend Analysis:** Track changes in trading activity over time.
- **Volume-Weighted Price Indicators:** Combine price and volume to identify significant market moves.
- **Relative Volume Metrics:** Compare current volume to historical averages for anomaly detection.

CONCLUSION:

Investment Decisions and Risk Management Strategies

1. **Investment Decisions:**
   - The strong upward trend indicated by the moving averages and ARIMA models suggests a long-term bullish outlook for NVIDIA's stock. This could inform buy-and-hold strategies for investors seeking long-term growth.
   - Insights from daily return patterns (e.g., high volatility during earnings announcements) could help in timing investments or adopting momentum trading strategies.
2. **Risk Management:**
   - Identifying days with unusually high volume and extreme returns (outliers) highlights periods of heightened market activity, which could be linked to external events. These could signal periods of risk or opportunity, prompting investors to adjust portfolios or hedge positions.
   - The weak correlation between dividends and stock prices suggests that relying solely on dividend payouts for valuation may underestimate market risks. Diversifying evaluation metrics could mitigate risks.
3. **Portfolio Optimization:**
   - The high autocorrelation in lagged values underscores the predictability of price movements in the short term. This could aid in short-term trading strategies or fine-tuning portfolio allocations for optimal risk-adjusted returns.

4. **Stress Testing and Scenario Analysis:**
    - The identified extreme movements in daily returns can be used for stress testing portfolios under simulated adverse conditions, preparing investors for potential downturns or unexpected volatility.