**Forecasting NVIDIA stock prices**

**PART 1: Forecasting NVIDIA stock prices using a seasonal and trend model.**

**PART 2: Fitting an ARIMA and SARIMAX models on NVIDIA stock prices data.**

**By Charles Luswata**

**(UC Berkeley, Professional Certificate in Machine Learning and AI)**

**Data Description**

The dataset used for this analysis is the historical daily stock prices of NVIDIA (NVDA), a leading technology company specializing in graphics processing units (GPUs). The timeframe of the data, adjusted for splits and dividends, spans several years, typically from the date NVIDIA became publicly traded in 1999 up to the most recent available data in 2024.

**Data Source**: [NVIDIA - 25 Year Stock Price History | NVDA | MacroTrends](https://www.macrotrends.net/stocks/charts/NVDA/nvidia/stock-price-history)

**Forecast**

For the forecast, we utilized a seasonal and trend decomposition model, specifically the seasonal decomposition using LOESS (STL) method provided by python’s statsmodels library. This model separates the time series into trend, seasonal, and residual components:

**Trend**: The underlying long-term movement in NVIDIA stock prices, capturing gradual increases or decreases over time.

**Seasonal**: Regular patterns or cycles that repeat over a specific period, such as yearly, quarterly, or monthly variations.

**Residual**: The remaining variation in the data after removing the trend and seasonal components, representing noise or irregular fluctuations.

**Presentation-Ready Plots**

The plots presented in this study illustrate the components of the decomposition plus:

**Observed vs. Trend:** This plot shows the observed daily closing prices of NVIDIA stock alongside the trend component (by STL and by the 50-day moving average commonly used in financial markets), highlighting the overall movement over time.

**Seasonal Component:** Displays the seasonal variations in NVIDIA stock prices, capturing any recurring patterns or cycles.

**Residuals:** Examines the residuals of the model, highlighting any remaining variability after accounting for trend and seasonal effects.

**Data splits**: The dataset was split into training/history (tdata) set and testing/future (fdata) set, ratio 75:25.

**Uncertainty**

**Model Evaluation:** To evaluate the model on testing data, measures, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), were calculated to quantify the model's performance against the actual stock prices.

**Residual Analysis**: Residuals were examined to help understand the model's predictive errors. Persistent patterns or high variability in residuals may suggest that the model has not captured all the underlying structure in the data.

**Numeric Tests beyond the eye**:

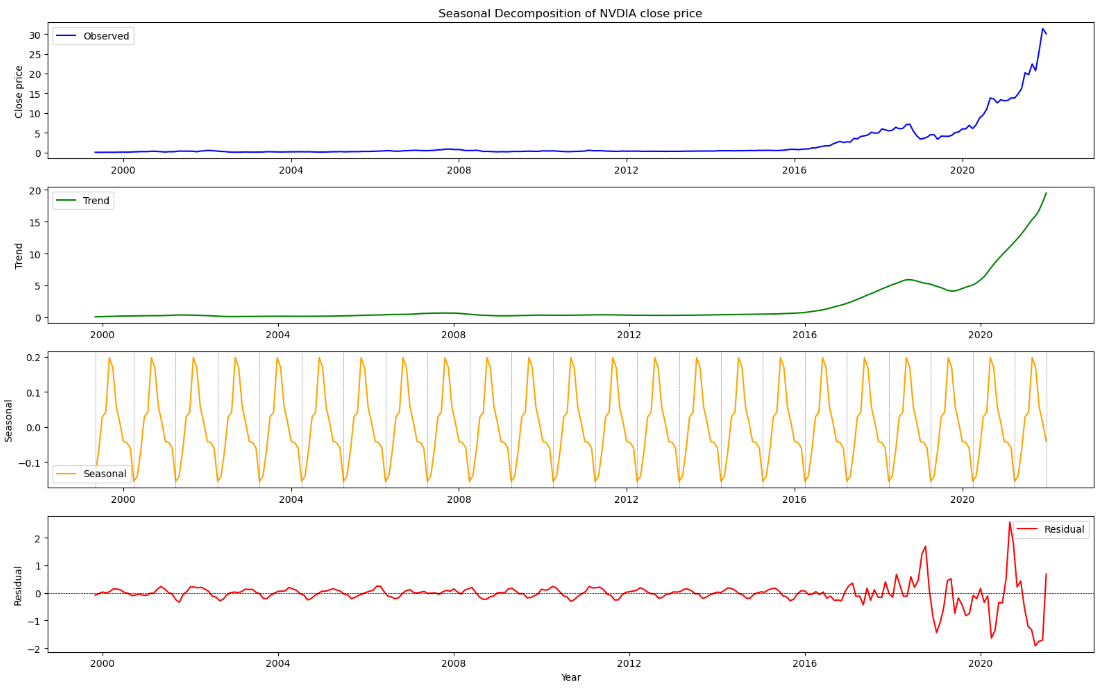
**Autocorrelation**: Correlation of a time series with a lagged version of itself, measures how each observation in a time series is related to its past observations.

**Durbin - Watson Statistic**: Tests for autocorrelation in the residuals of a regression or time series model. It ranges from 0 to 4, where a value near 2 indicates no autocorrelation and Values significantly different from 2 indicate the presence of autocorrelation.

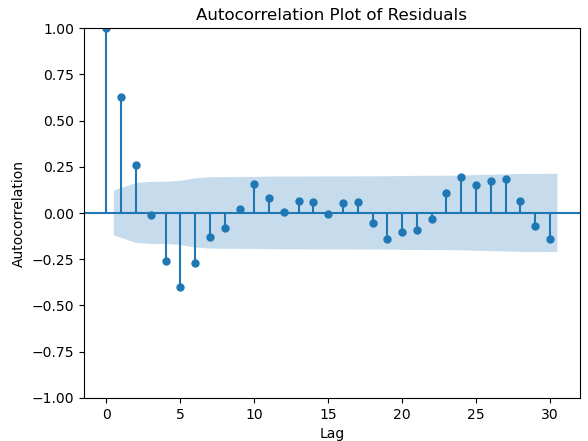
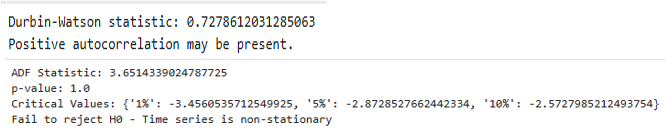
**Stationarity**: A crucial assumption for many time series models, as it ensures that the statistical properties of the series do not change over time. Significant spikes in Autocorrelation Plot for Residuals at various lags suggests that there may still be patterns in the residuals that are not captured by the trend and seasonal components hence non-stationarity.

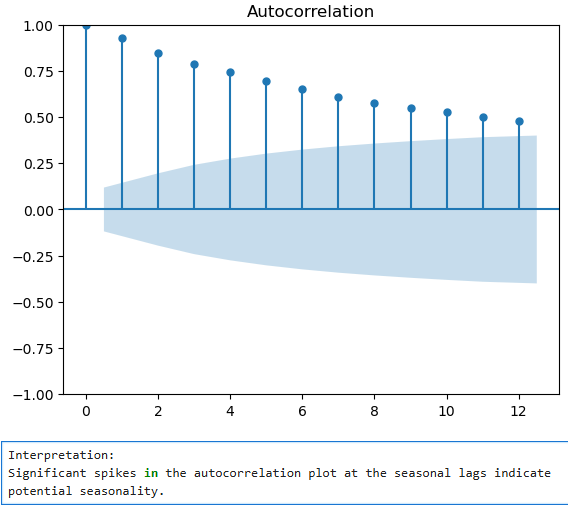
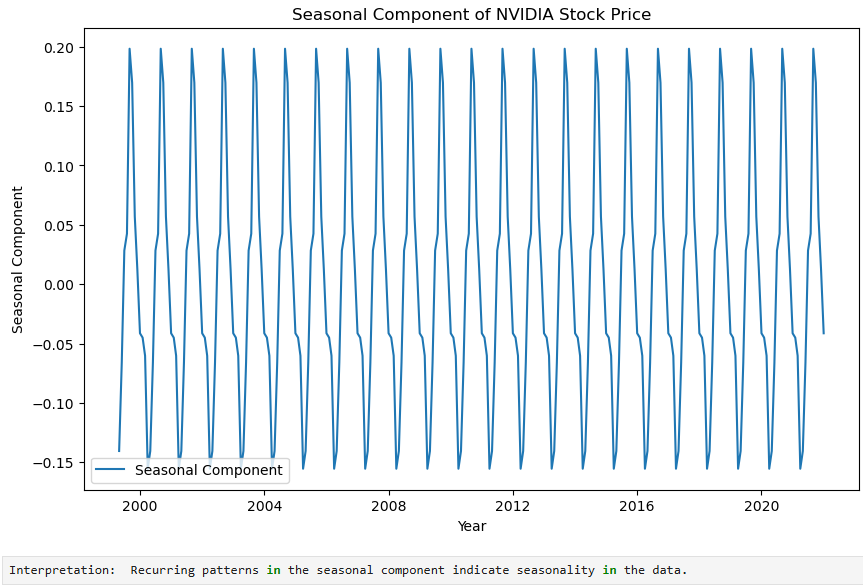
**Augmented Dickey-Fuller (ADF) test**: Used to determine whether a time series is stationary or not. If the test statistic is less than the critical value from the ADF distribution, we reject the null hypothesis (suggesting stationarity). However, if the test statistic is greater than the critical value, we fail to reject the null hypothesis (indicating non-stationarity).

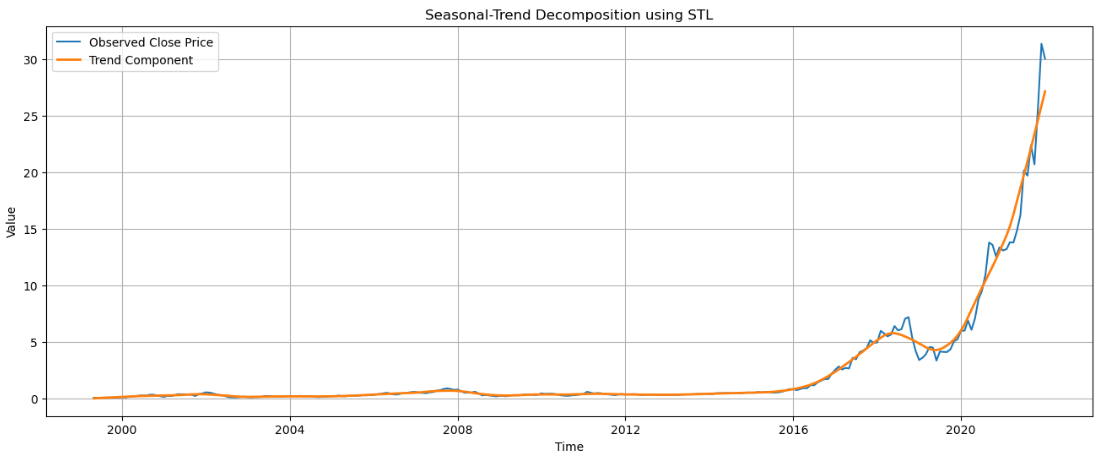
**Results:**

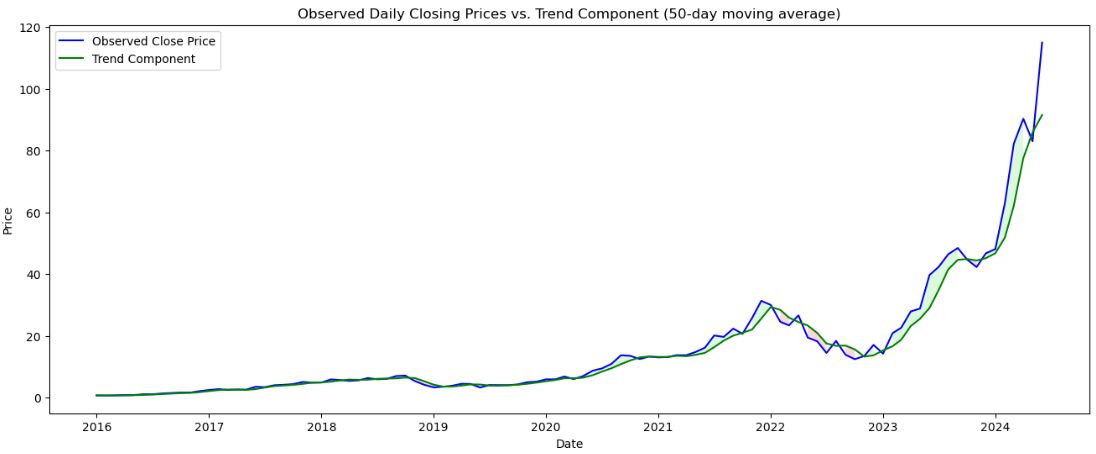
****

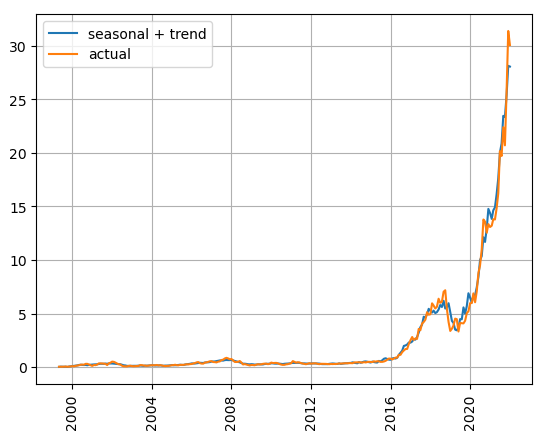
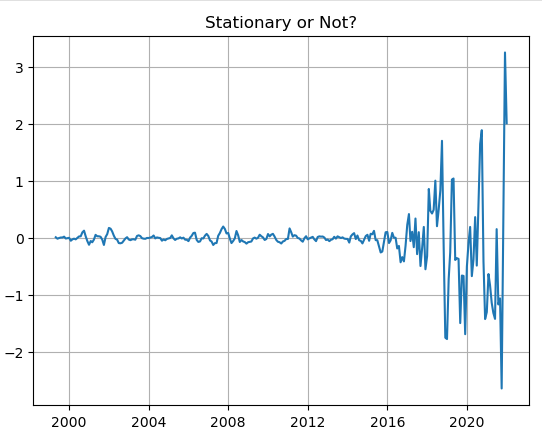
**Seasonality, Stationary and Autocorrelation assumptions**

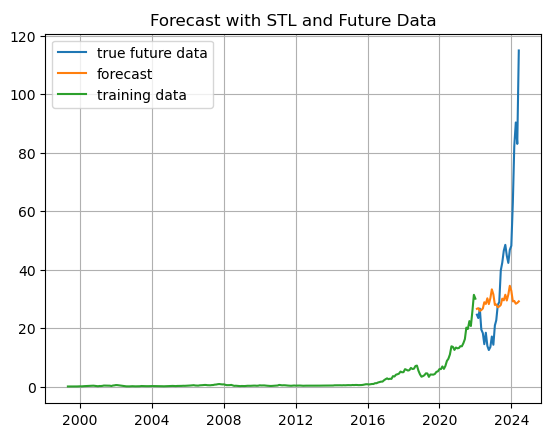
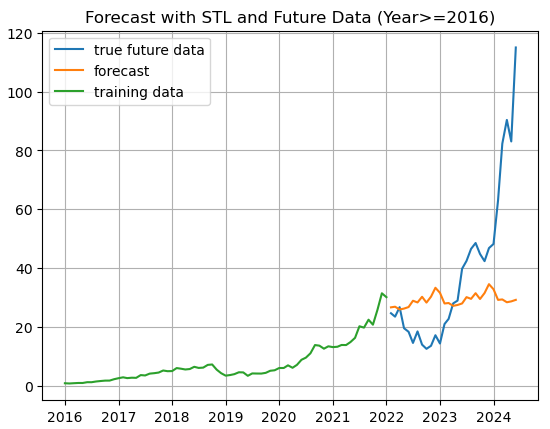
 

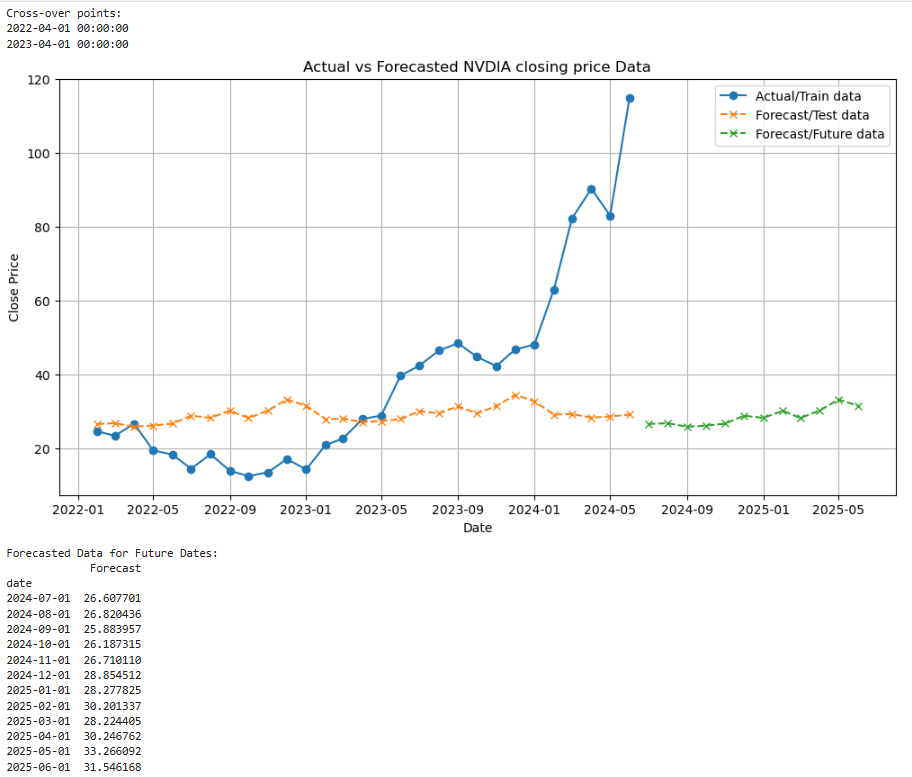


****





**Findings and Interpretation:**

**Trend and Seasonality**: There is a visible and significant up-trend as well as seasonality in the NVDIA stock prices data.

**Autocorrelation**: Positive value of the Durbin – Watson test (0.73) indicates positive autocorrelation present suggesting that past values may influence future values in a positive manner.

**Stationarity**: Significant spikes were observed in the autocorrelation plot of residuals at various lags suggesting that there may still be patterns in the residuals that are not captured by the trend and seasonal components hence non-stationarity. This was confirmed by the Augmented Dickey-Fuller (ADF) test for stationarity (statistic = 3.65; p-value > 0.05).

**Model Evaluation/goodness of fit:** The Mean Absolute Error (MAE) of 18.74 indicates that, on average, the forecasts are off by approximately 18.74 units from the actual values. Since RMSE involves squaring the errors before taking the square root, it tends to penalize larger errors more heavily than MAE. The RMSE value of 27.29 means that, on average, the forecasts deviated from the actual values by approximately 27.29 units.

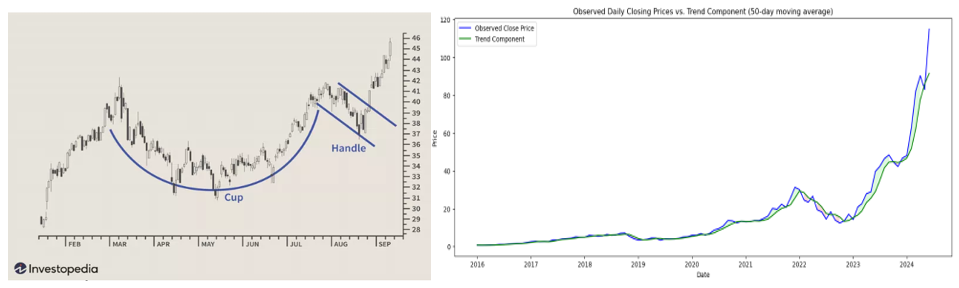
**What could have caused the sudden upward spike in the stock price?**

The explanation for the spike after crossing over the forecast line following the second attempt may be attributed to strong earnings reports (fundamental data), declining inflation (economic data), the CHIPS and Science Act signed into law by the president (news) that may have acted as a catalyst for the semi-conductor industrial group stocks. Also, the stock price’s break-out of the formed “cup” (consolidation phase) of the famous “Cup and Handle” technical pattern may have contributed to the spike. Furthermore, the markets tend to favor the stocks that lead every sector/industrial group and NVDA was the group leader for the semi-conductors even though the company did not directly benefit from the CHIPS and Science Act.

**The Cup and Handle pattern/ technical analysis strategy**

The Cup and Handle pattern, first described by William J. O'Neil in his 1988 book on technical analysis, ‘How to Make Money in Stocks’, is observed in the NVDIA data. It is just one of many technical analysis tools used by traders to identify potential buying opportunities in stocks. As with any technical pattern, it's important to combine it with other forms of analysis and risk management techniques to make informed trading decisions. Additionally, patterns in trading are probabilistic, meaning they can fail or succeed, so risk management is crucial.

If a stock has been in an uptrend, followed by a period of consolidation where it forms a cup shape. During this phase, the volume decreases as the price moves lower and then starts to round up forming the cup. After the cup formation, there's a brief pullback forming a handle with lower volume. The breakout occurs when the price moves above the high of the handle, often accompanied by increased volume, signaling a potential bullish continuation.



Markets are assumed to be efficient, hence The Efficient Market Hypothesis (EMH), but human behavior and cognitive biases can lead to market inefficiencies, that invalidate forecasts/statistical models, such as overreaction or underreaction to news which can be exploited by skilled investors/traders.

**Consequences of Model Errors:**

The errors in the model, reflected in the residuals, have implications for forecasting NVIDIA stock prices:

**Forecast Reliability**: High residual errors indicate potential challenges in accurately predicting future stock prices. This uncertainty can impact financial decisions and risk assessments based on the forecast.

**Structural Insights**: Persistent patterns or non-random behavior in residuals suggest that additional factors or variables may influence NVIDIA stock prices, beyond those accounted for in the current model. Exploring and incorporating such factors could improve forecast accuracy.

**Conclusion:**

In conclusion, the seasonal and trend decomposition model provided valuable insights into the historical trends and patterns in NVIDIA stock prices. However, residual analysis indicates remaining uncertainty and potential for further model refinement. Continued monitoring and adjustment based on new data and insights will enhance the accuracy and reliability of future forecasts.

[Cup and Handle Pattern: How to Trade and Target with an Example (investopedia.com)](https://www.investopedia.com/terms/c/cupandhandle.asp)

[Efficient Market Hypothesis (EMH): Definition and Critique (investopedia.com)](https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp#:~:text=The%20efficient%20market%20hypothesis%20%28EMH%29%20or%20theory%20states,benefit%20from%20investing%20in%20a%20low-cost%2C%20passive%20portfolio.)

**PART 2: Fitting an ARIMA and SARIMAX models on NVIDIA stock prices data.**

**(Data range: 2016-01-01 to 2024-06-28)**

ARIMA (AutoRegressive Integrated Moving Average) and SARIMAX (Seasonal ARIMA with Exogenous variables) models are widely used for time series analysis and forecasting. These models assume that the future behavior of the time series will follow past patterns captured by the model. Uncertainty in forecasts should be considered, especially when extrapolating beyond the observed data. Prior to fitting ARIMA models, the assumptions are crucial to ensure the reliability and interpretability of the models.

**Assumptions of ARIMA Model:**

1. **Stationarity**: The time series data should be stationary or transformed to achieve stationarity. This typically includes having a constant mean, constant variance, and a covariance (or autocovariance) that does not depend on time.
2. **Linear Relationship**: ARIMA models assume that the relationship between the current value of the time series and its lagged values (autoregressive terms) and past forecast errors (moving average terms) is linear.
3. **No Perfect Multicollinearity**: The predictors (lagged values and exogenous variables in SARIMAX) should not be perfectly correlated. Perfect multicollinearity can lead to unreliable parameter estimates.
4. **Residuals Assumptions**:
   * **Mean of Residuals**: The residuals (errors) of the model should have a mean close to zero.
   * **Independence of Residuals**: The residuals should be independent of each other, meaning that the errors at one time point should not be correlated with errors at other time points.
   * **Constant Variance of Residuals**: The variance of the residuals should be constant over time (homoscedasticity).
5. **Normality of Residuals** (for inference and hypothesis testing):
   * While ARIMA models do not explicitly require residuals to be normally distributed, normality facilitates the use of standard statistical tests and confidence intervals for parameter estimates.

**Assumptions of SARIMAX Model (in addition to ARIMA assumptions):**

1. **Exogenous Variables**:
   * SARIMAX models can incorporate exogenous variables (external predictors not influenced by the model), which should be appropriately specified and integrated into the model.
2. **Seasonality**:
   * For seasonal time series data, SARIMAX models include seasonal components (seasonal AR and MA terms) to account for seasonal patterns.

**Why SARIMAX Model?**

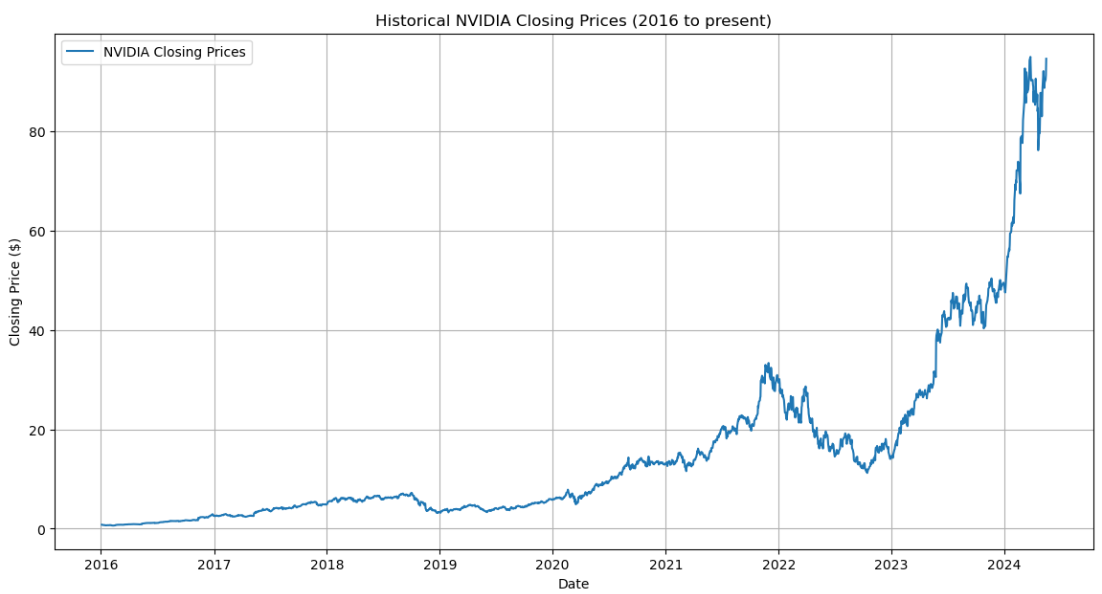
SARIMAX is advantageous when dealing with time series data that exhibits seasonal patterns and/or when incorporating external variables can enhance forecasting accuracy. It offers more flexibility and capabilities compared to ARIMA, making it a preferred choice in many real-world forecasting scenarios where these features are beneficial.

**Data transformations:**

1. Data for modeling was restricted to records later than 2016-01-01.
2. To achieve stationarity, the closing prices where transformed by applying first-order differencing.
3. Data was split leaving the last 30 records as the testing/future set.

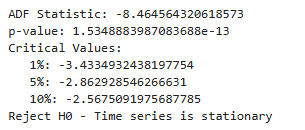
**Model building:**

**Step 1: Visualizing the time series data.**

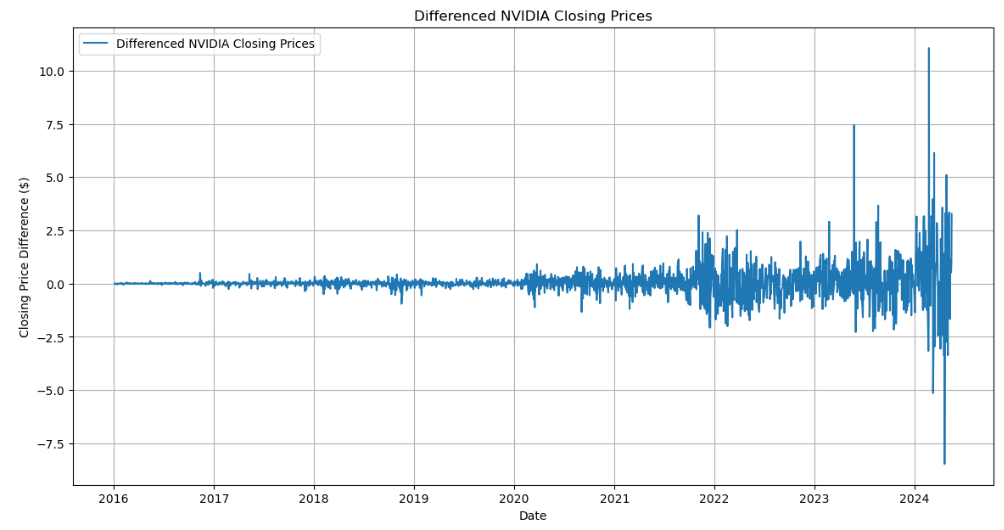


A positive upward trend is observed in the data.

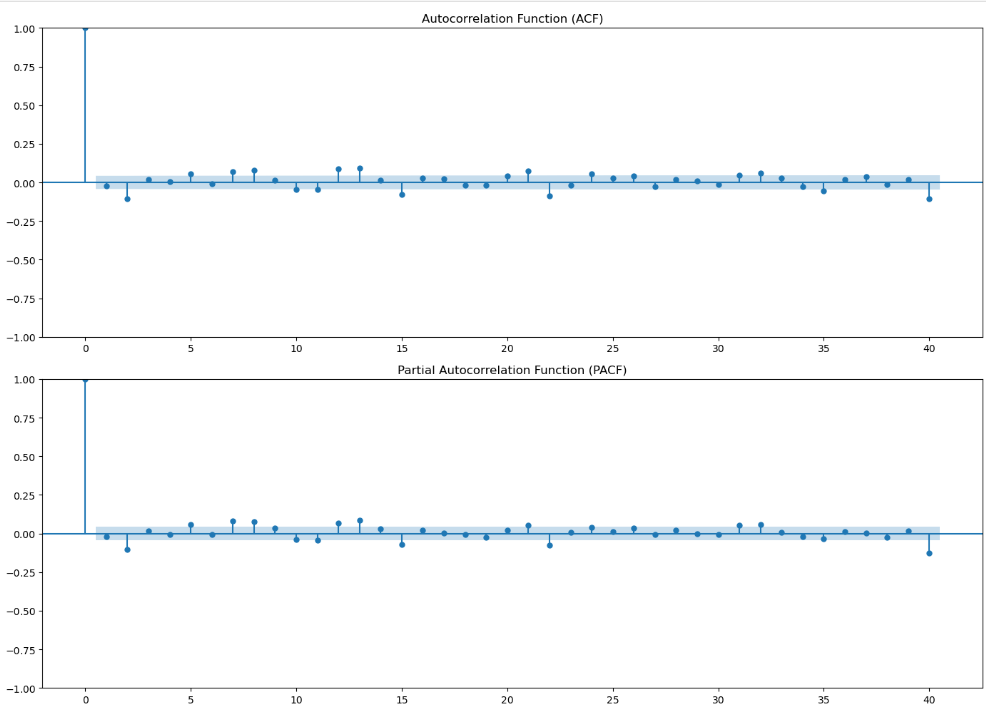
**Step 2**: Stationarity check using the ADF test.



After applying first order differencing, stationary was achieved.



**Step 3**: ACF and PACF plots to identify Autoregressive (AR) order p and Moving Average (MA) order q respectively.



Both the ACF and PACF plots show a significant spike at lag 0 beyond which the values drop off closer to zero hence we are choosing 𝑞=1 and 𝑝=1

**Step 4**: Instantiate and fit the model using statsmodels. Print the summary of the fitted model for coefficient details and statistical tests.

**ARIMA Model:**

SARIMAX Results

==============================================================================

Dep. Variable: Close No. Observations: 2105

Model: **ARIMA**(1, 0, 1) Log Likelihood -2342.644

Date: Sun, 30 Jun 2024 AIC 4693.288

Time: 15:56:10 BIC 4715.896

Sample: 0 HQIC 4701.567

- 2105

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

const 0.0444 0.015 2.911 0.004 0.015 0.074

ar.L1 0.4481 0.125 3.573 0.000 0.202 0.694

ma.L1 -0.4990 0.121 -4.132 0.000 -0.736 -0.262

sigma2 0.5422 0.004 137.351 0.000 0.534 0.550

===================================================================================

Ljung-Box (L1) (Q): 1.32 Jarque-Bera (JB): 164285.93

Prob(Q): 0.25 Prob(JB): 0.00

Heteroskedasticity (H): 137.15 Skew: 1.99

Prob(H) (two-sided): 0.00 Kurtosis: 46.10

===================================================================================

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

MSE on test data: 16.1824 MAE: 3.0492 RMSE: 4.0227

**SARIMAX Model:**

SARIMAX Results

==========================================================================================

Dep. Variable: Close No. Observations: 2105

Model: SARIMAX(1, 0, 1)x(1, 0, 1, 12) Log Likelihood -2334.860

Date: Sun, 30 Jun 2024 AIC 4679.719

Time: 16:25:12 BIC 4707.980

Sample: 0 HQIC 4690.069

- 2105

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 0.4405 0.120 3.665 0.000 0.205 0.676

ma.L1 -0.4920 0.116 -4.231 0.000 -0.720 -0.264

ar.S.L12 0.4395 0.082 5.342 0.000 0.278 0.601

ma.S.L12 -0.3376 0.086 -3.935 0.000 -0.506 -0.169

sigma2 0.5382 0.004 133.463 0.000 0.530 0.546

===================================================================================

Ljung-Box (L1) (Q): 0.78 Jarque-Bera (JB): 146718.17

Prob(Q): 0.38 Prob(JB): 0.00

Heteroskedasticity (H): 157.72 Skew: 1.82

Prob(H) (two-sided): 0.00 Kurtosis: 43.74

===================================================================================

Mean Squared Error on test data: 16.401449120157398

Mean Absolute Error (MAE) on test data: 3.0506824569766673

Root Mean Squared Error (RMSE) on test data: 4.049870259669734

ARIMA(1, 0, 1) Model Equation: y\_t = 0.4405 \* y\_t-1 + -0.4920 \* e\_t-1 + e\_t

SARIMA(1, 0, 1)(1, 0, 1, 12) Model Equation: y\_t = 0.4405 \* y\_t-1 + -0.4920 \* e\_t-1 + 0.4395 \* y\_t-12 + -0.3376 \* e\_t-12 + e\_t

**Interpretation:**

A positive AR coefficient (ar.L1) value of 0.4405 indicates that the current value of the series is positively correlated with its previous value lagged by 1 period while the negative MA coefficient value (ma.L1) of -0.4920 indicates that the current error term (residual) is negatively correlated with the previous error term lagged by 1 period.

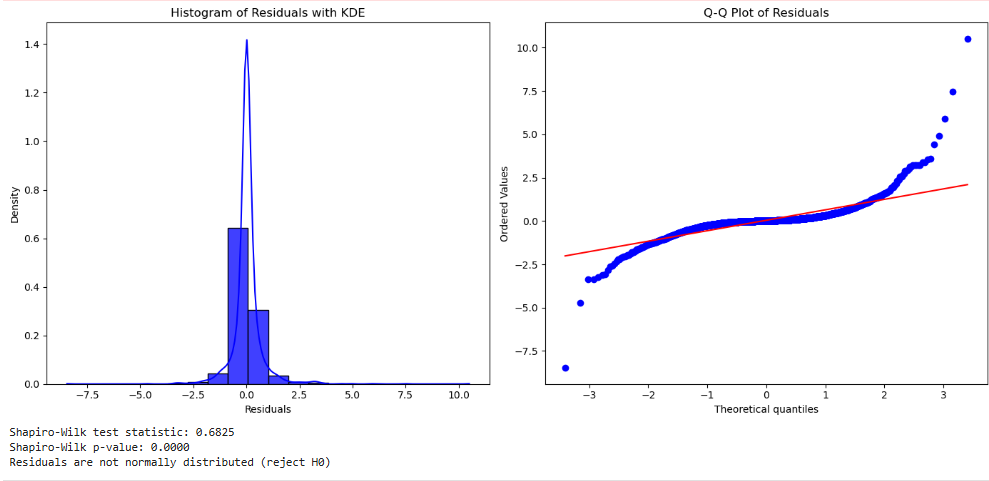
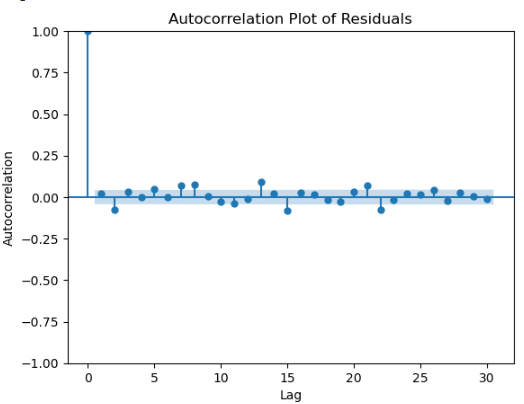
The seasonal AR coefficient (ar.S.L12) of 0.4395 suggests a positive correlation between the current value and its value 12 periods ago (monthly seasonality) while the Seasonal MA coefficient (ma.S.L12) of -0.3376 suggests a negative correlation between the current error term and the error term 12 periods ago (seasonal lag of 12 months).

Ljung-Box test statistic and p-value of 0.38 suggests that the residuals are uncorrelated while the Jarque-Bera test statistic and p-value indicate non-normality of residuals.

**Conclusion**

The SARIMAX model fitted shows significant coefficients for both non-seasonal and seasonal components, suggesting that both the current value and seasonal patterns from previous periods influence the forecasted values. The model's performance metrics indicate reasonably good fit on the test data, with low errors and well-behaved residuals according to diagnostic tests. However, further validation and refinement may be necessary depending on the specific application and forecasting accuracy requirements.

**Step 5**: Diagnostic tests based on the residuals

The autocorrelation plot of residuals shows some significant lags beyond the confidence band/intervals suggesting that there is still structure in the residuals that the model has not captured. This could indicate that the model might need further adjustment or that additional explanatory variables should be included.

**Appendix**

