**Time Series Analysis and Modelling - Stock Market Predictions**



**Team Members:**

* Tobias Ngonga
* Sammy Sifuna
* Jael Akech
* Denis Kipkorir
* Hellen Mwangi
* Waruchu Kuria
* Alan Omondi

# **1. Business Understanding**

## **1.1 Overview**

Stock price prediction plays an important role in financial market and accurate forecasts can have significant implications for investors and financial institutions.

## **1.2 Problem Statement**

The major challenge of this project is to accurately predict the future closing value of a given stock across cross a given period of time in the future.

## **1.3 Business Objective**

## **1.3.1 General Objective**

This is to develop a robust stock price prediction model using recurrent neural networks (RNN).

## **1.3.2 Specific Objectives**

* + Build and implement RNN model for stock price prediction then improve the model with LSTM.
  + Evaluate the performance of the model using F1 score and accuracy of the models using MAE.
  + Use LSTM to forecast stock prices.
  + Create a user-friendly dashboard for stakeholders to access predictions.

## **1.4 Business Success Criteria**

* + Achieve a low MAE and MSE errors and accurately forecast stock prices.
  + Ability of the LSTM to capture long-term dependencies in stock price data.
  + Have the RNN model perform better than our baseline model which uses traditional time-series forecasting method (SARIMAX).
  + Usability of the models by investors and financial analysts.

# **2. Data Understanding & EDA**

## **2.1 Data Features**

During the exploration exercise, we checked for the value counts of each column, to understand how various parameters in the columns were distributed The column definitions are displayed below.

The following observations were made during the preliminary data exploration

**Columns/Features**

The data has the following features:

* + **Date**: date of the stock price observation.
  + **Open price**: opening price of the stock on the given date.
  + **High** **price**: highest price of the stock on the given date.
  + **Low** **price**: lowest price of the stock on the given date.
  + **Close** **price**: closing price of the stock on the given date.
  + **Adjusted** **Close** **price**: closing price after adjustments for all applicable splits and dividend distributions
  + **Volume**: number of shares of the stock traded on the given date.

**Shape**

* + It has 3,960 rows and 5 columns.

**Data Types**

* + All the data is numerical as expected.

**Missing Values**

* + There are no missing values.

**Checking for Duplicates**

* + There are no duplicates in the data.

**Time Series Comformity**

* The Date column is already in DateTime format and is the index.

# **3. Univariate Analysis**

## **3.1 Distributon of The Columns using Histplots**

## 

We made analysis in individual columns/variables to see how it’s distributed and get patterns from the columns.

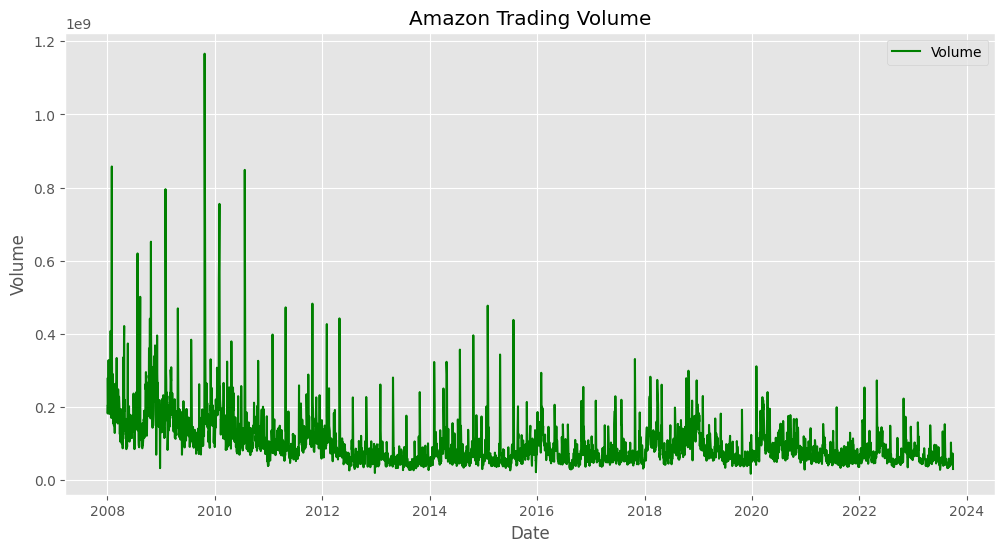
**Observations**

* The **Open, High, Low, Close, Adj Close** plots have similar distributions throughout the period under review (2008 to 2023).
  + They are trimodal (three peaks).
  + 0-25 dollars is the most frequent price.
* Volume seems to be heavily distributed around 100 million to 200 million.
* All are skewed to the right.
* No outliers are visible from these graphs.

#### **b. Time Series Plots for Open, High, Low, Close, and Adj Close Columns**

* The plots below visualize the historical price trends over the period under revie
* Similar seasonality and trend characteristics. This will be confirmed in later sections.

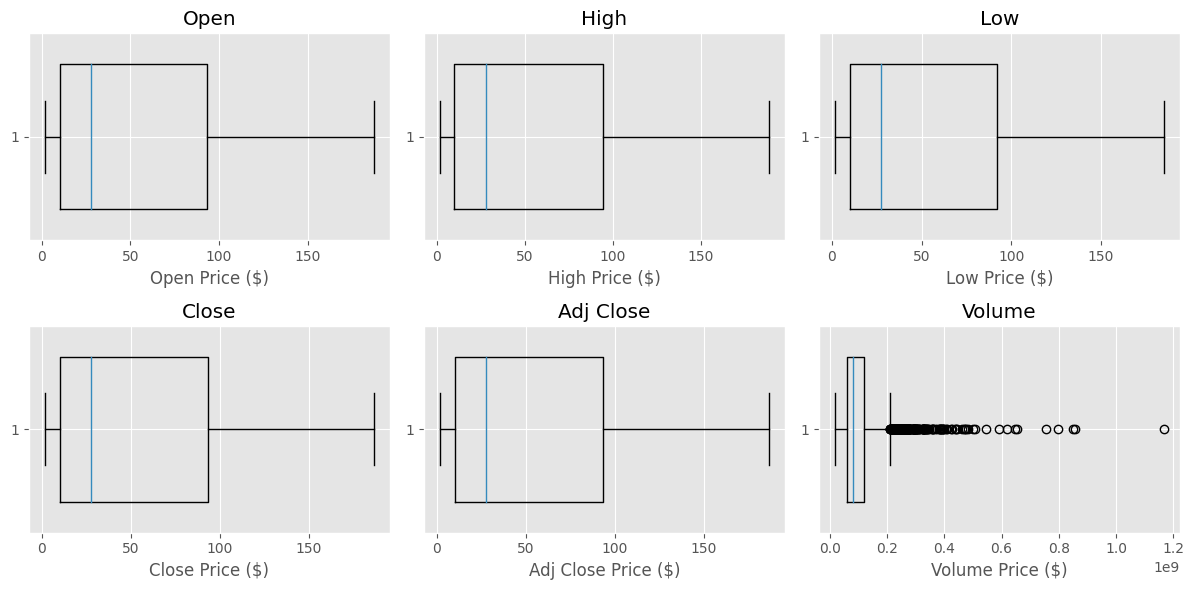
**Time Series Plot for Volume Column**

****

**Observation**

* The Volume column also looks seasonal and with trend.
* The volume decreases as time progresses with large share volumes bought in earlier years (2008 to 2010). This decreases sharply after 2010 and continues to drop to-date. (**why???\***)

### **2. Checking for Outliers or Anomalies using Box Plots**

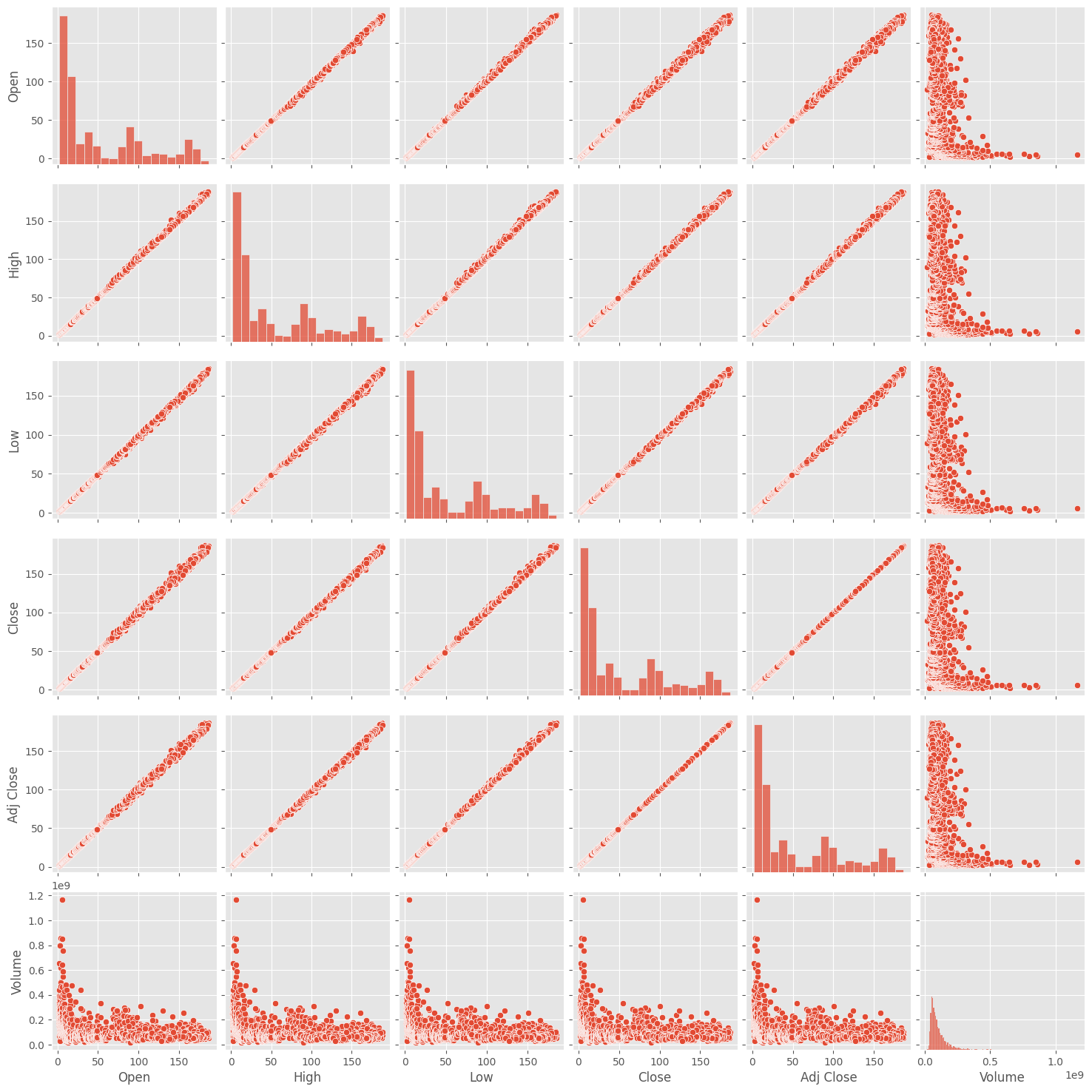


**Observations**

* The price is centered at approximately 10 and 90 dollars for all columns.
* In all features except Volume, there are no outliers detected.

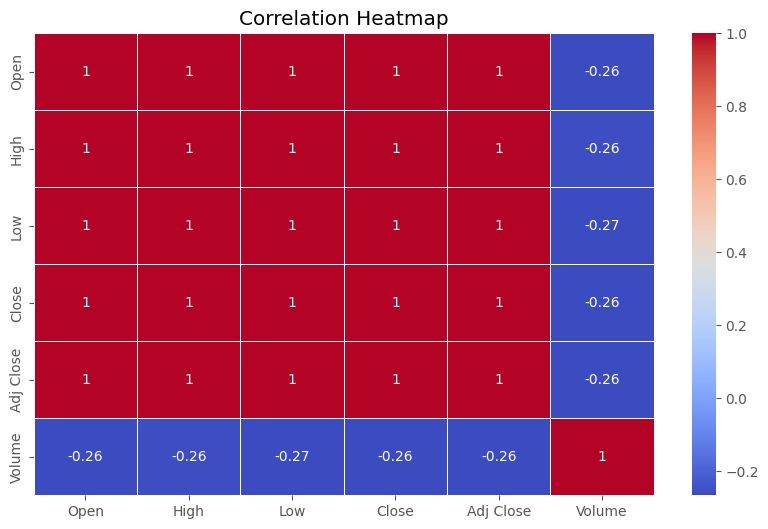
### **3. Bivariate and Multivariate Analysis**

#### **a. Pairplots and Heatmap to Show Correlation**



**Observations**

* The relationship among the variables is similar across the data.
* There is a linear relationship and very strong positive correlation among all features except with 'Volume'.
* There is a non-linear relationship and weak negative correlation with Volume.
* These observations are confirmed in the correlation heatmap below.

**  
External factors that may have caused the positive correlations can be explained as follows:**

1. Amazon stock price is driven by the overall performance of the company. When the company is performing well, its stock price tends to go up.
2. It is correlated with the overall performance of the stock market. When the stock market is doing well, Amazon's stock price tends to go up. This is because investors are more likely to buy risky assets, such as stocks, when the market is doing well.

For the weak negative correlation with Volume:

1. The volume of trading in Amazon stock is higher when the stock price is volatile. This is because investors are more likely to trade stocks when they are experiencing large price swings.
2. Volume of trading in Amazon stock is higher when there is a lot of news about the company. This is because investors may be more likely to buy or sell shares of the company based on news about its performance, products, or competitive landscape.

The state of the economy, interest rates, inflation and overall market sentiment are the general external factors that may have influenced the patterns visualized above.

**Observations**

* The time series is seasonal. The price gradually increases as the years go with a sharp rise noted between 2018 and 2022.
* However, there is a decrease towards the year 2023.
* Another increase begins as 2023 comes to an end.

## **C. Data Preparation**

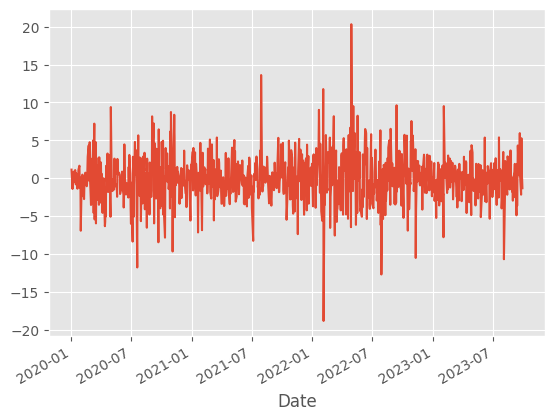
### **1. Feature Engineering**

In fature engineering we were identifying and selecting the most relevant and helpful data in the dataset we used.

#### **The date-related Features are as follows**

* Additional information from the date column will be extracted capture any potential patterns or seasonality in the stock prices:
  + **day\_of\_week**
  + **month**

#### **b. Returns**

* A new column \*\*Returns\*\* is added to calculate the difference in price for two consecutive days.
* As seen in the plot, the series is seasonal as the period progresses.

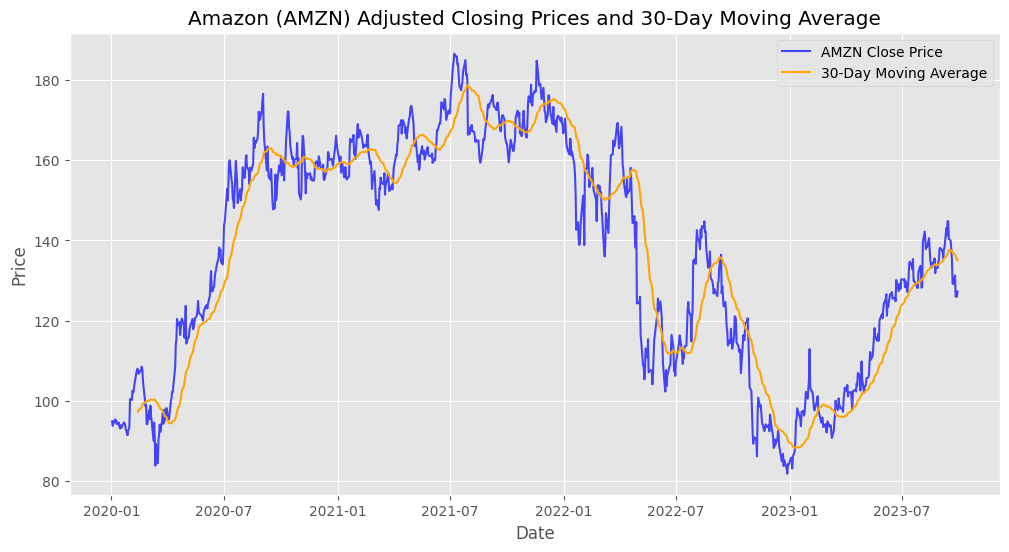
#### **c. Lag Features for Adj Close price**

* Creating lag features will capture the historical behavior of the stock prices.
* Lag features for the adjusted closing price have been calculated in the code below.

#### **d. Rolling Mean and Standard Deviation**

* To help capture trends and volatility.

#### **e. Moving Average**

* The 30-day moving average tracks the underlying trend in Amazon's closing stock prices over time, helping to smooth out short-term fluctuations and provide insights into longer-term price movements.

#### **f. Exponential Moving Average (EMA) Smoothing Factor**

* EMA can provide more weight to recent data points, which can be useful for capturing short-term trends.
* The small alpha value of 0.2 will result in a smoother EMA and higher sensitivity to recent price changes.

#### **g. Rate of Change (ROC) for Adj Close**

* Measures the percentage change in price over a specified period.
* Positive ROC values indicate upward momentum, while negative values indicate downward momentum.
* It helps traders and analysts identify potential trends or reversals in the price of an asset.

#### **h. Relative Strength Index (RSI)**

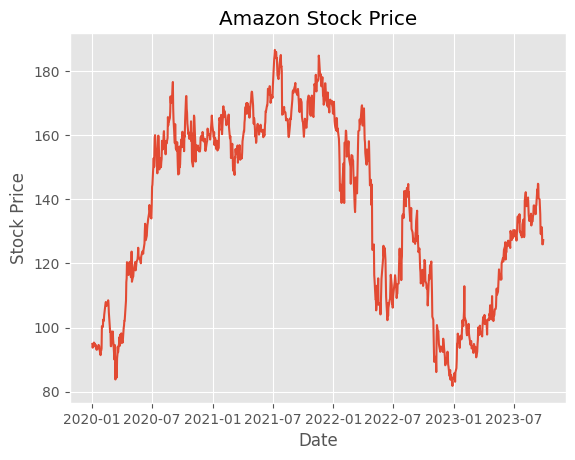
* It measures the speed and change of price movements.
* High RSI is normally placed as any value above 75% and a good indicator to sell.
* Low RSI is between 0 and 25%, a good indicator to buy.

#### **i. Average True Range (ATR)**

* A volatility indicator.
* It measures the average range between the high and low prices over a specified period.
* It helps traders and analysts assess the level of volatility in the price movements of an asset.

#### **a. Checking for Seasonality**

* First, we will check if there is seasonality and if it is statistically significant for each column in the time series.
* Remove seasonality using the seasonal\_decompose function.
* Twelve (12) periods are specified to represent a year, the duration of a season in the data. **to check later**
* After extracting the trend, seasonality and residuals, it plots each of these components in separate subplots.
* It is important to remove seasonality and trend because if they are part of the time series, there will be effects in the forecast value



Seasonality: Close is Not Seasonal

Seasonality: Adj Close is Not Seasonal

Seasonality: Volume is Seasonal

Seasonality: day\_of\_week is Not Seasonal

Seasonality: Returns is Not Seasonal

Seasonality: Adj\_Close\_Lag\_1 is Not Seasonal

Seasonality: Adj\_Close\_Lag\_2 is Not Seasonal

Seasonality: Adj\_Close\_Lag\_3 is Not Seasonal

Seasonality: Rolling\_Mean is Not Seasonal

Seasonality: Rolling\_Std is Not Seasonal

Seasonality: 30-Day MA is Not Seasonal

Seasonality: EMA is Not Seasonal

Seasonality: ROC is Not Seasonal

Seasonality: RSI is Seasonal

Seasonality: ATR is Not Seasonal

* Only Volume is seasonal. (**may change after adjusting threshold. small seasonal variations may not exceed this threshold**)

#### **b. Checking for Stationarity**

* The **Augmented Dickey-Fuller (ADF) test** will be used to determine the stationarity of the time series.
* The ADF test helps us assess whether a time series is stationary by comparing it to the null hypothesis that it has a unit root (meaning it's non-stationary).
* The test involves estimating the model's **coefficients**, calculating a **test statistic**, and comparing it to **critical values** to determine whether the null hypothesis can be rejected.
* The **p-value** resulting from the test indicates the strength of evidence against the null hypothesis.
  + If the p-value is less than or equal to the significance level of 0.05, there's evidence to reject the null hypothesis, suggesting the series is stationary.
  + If the p-value is greater than the significance level of 0.05, there's weak evidence to reject the null hypothesis, indicating the series is likely non-stationary.

High p-value : 0.26425987348123803

High is non-stationary.

Low p-value : 0.28981128316091487

Low is non-stationary.

Close p-value : 0.27950680519689974

Close is non-stationary.

Adj Close p-value : 0.27950680519689974

Adj Close is non-stationary.

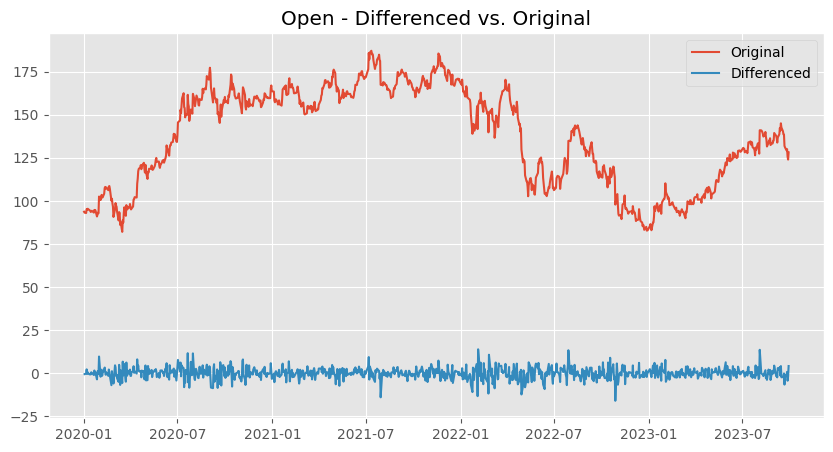
Volume p-value : 5.208366769489285e-06

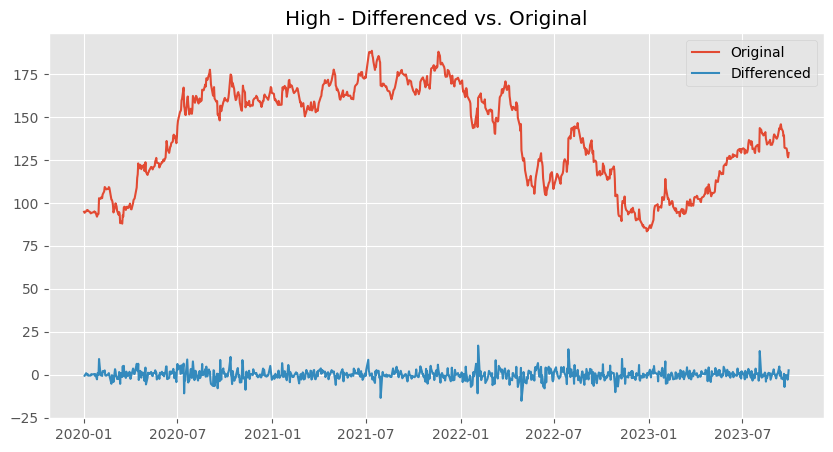
Volume is stationary.

day\_of\_week p-value : 1.7318352906631962e-09

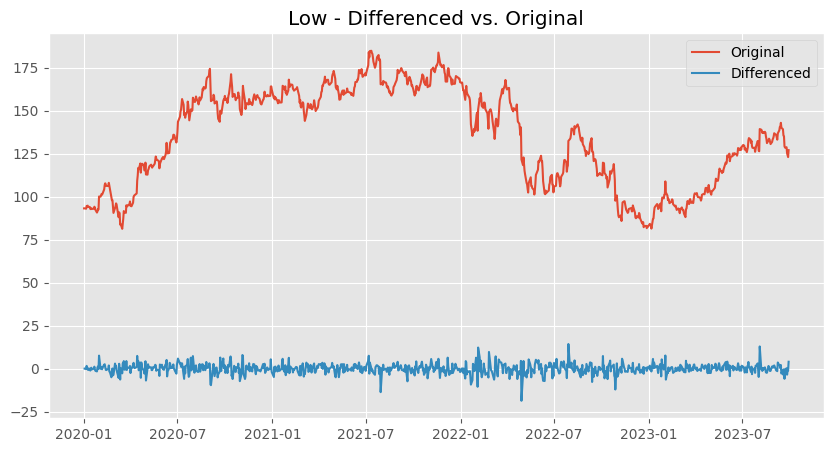
**Remove Stationarity Through Differencing**

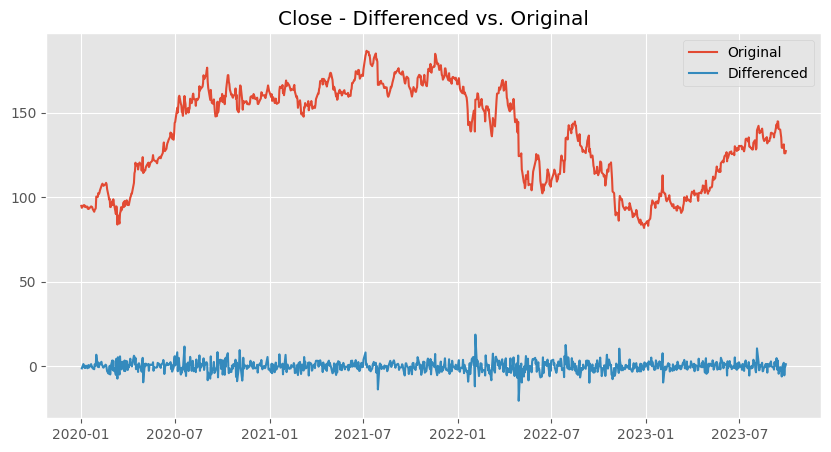
* Some of the series is stationary:
  + Volume
  + Returns
  + day\_of\_week
  + ROC
  + RSI
* The rest are non-stationary.
* Next, differencing will be applied to remove stationarity and make the model-building process accurate and reliable.
* The data's statistical properties will remain consistent, allowing for meaningful insights and accurate predictions.

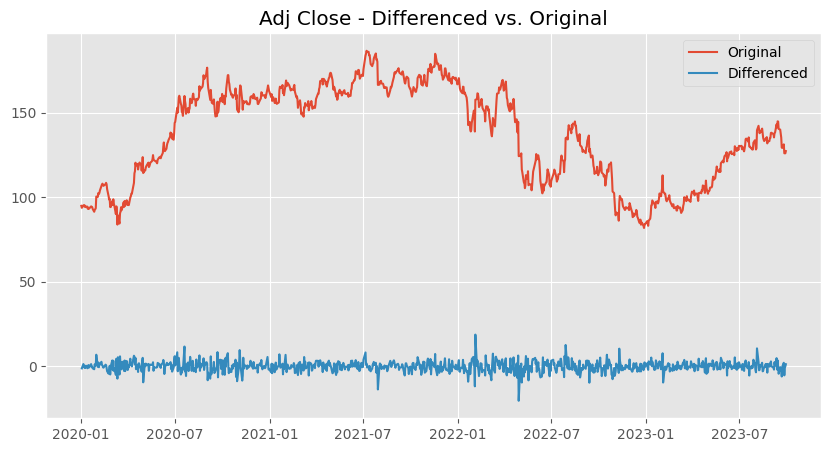


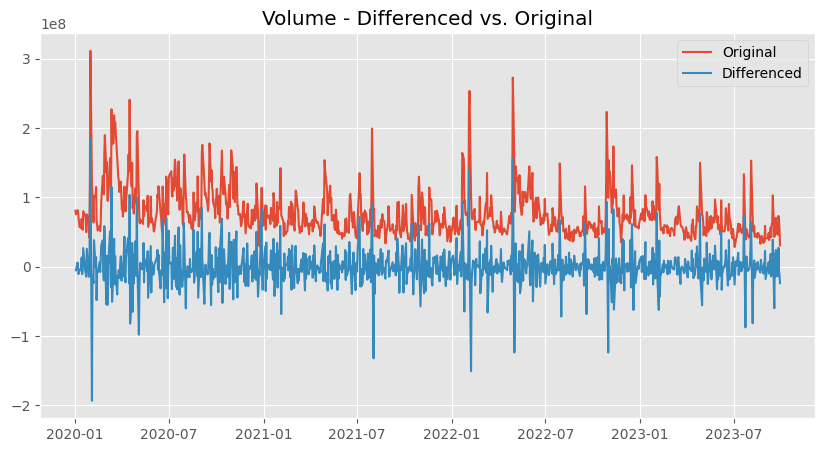


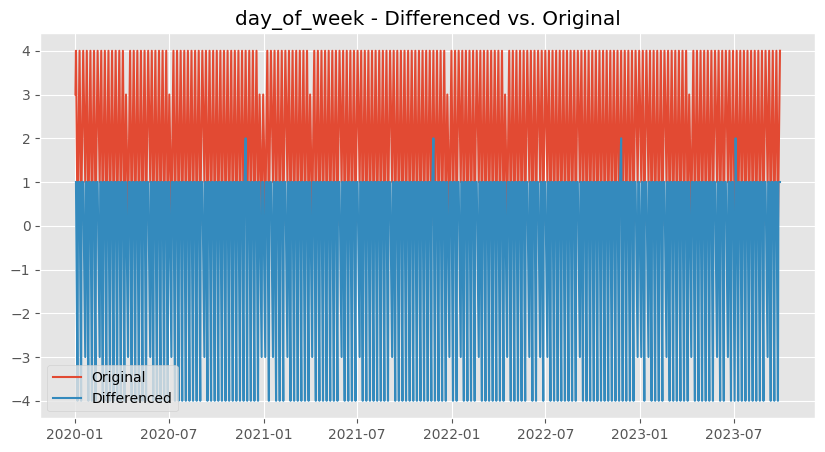
### **3. Autocorrelation and Partial Correlation of Differenced Series**

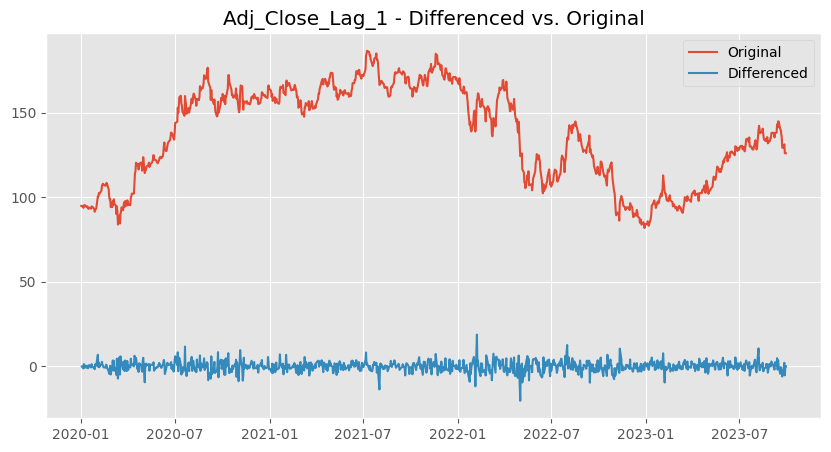
* Autocorrelation measures the linear relationship between lagged values of a time series.

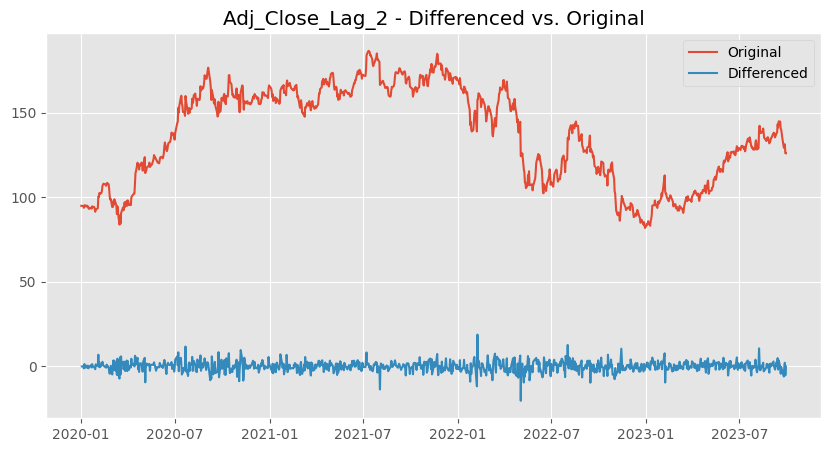


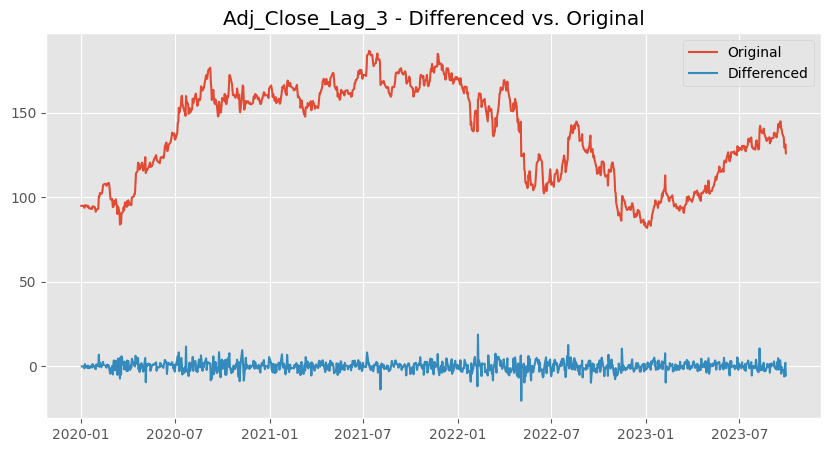


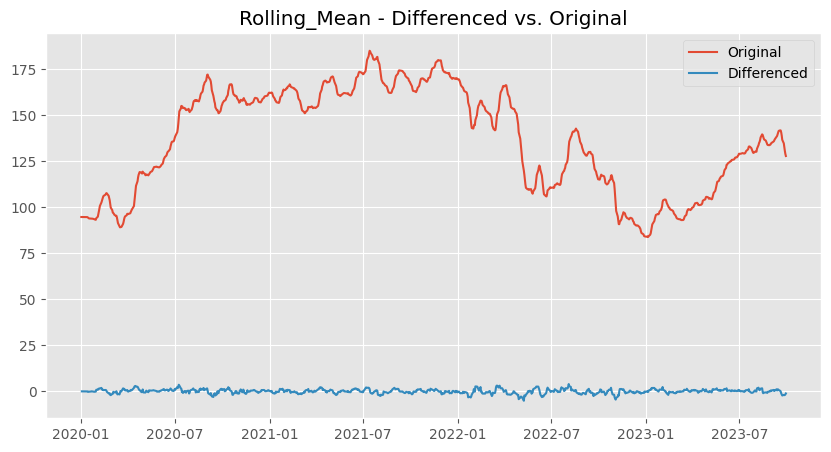


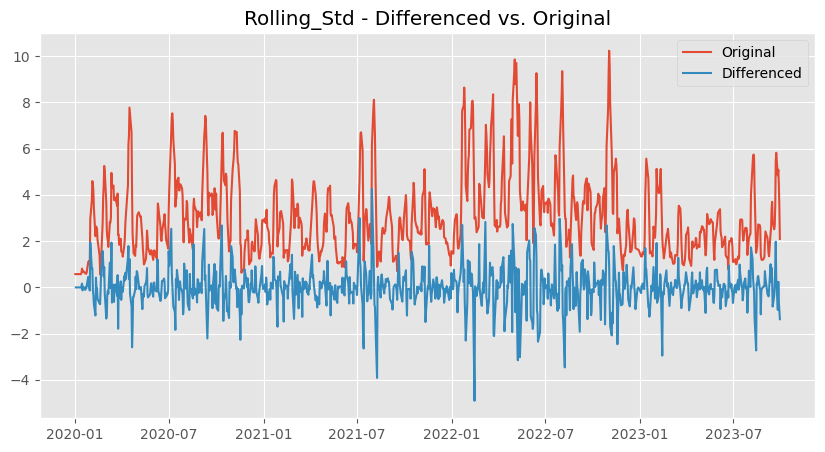


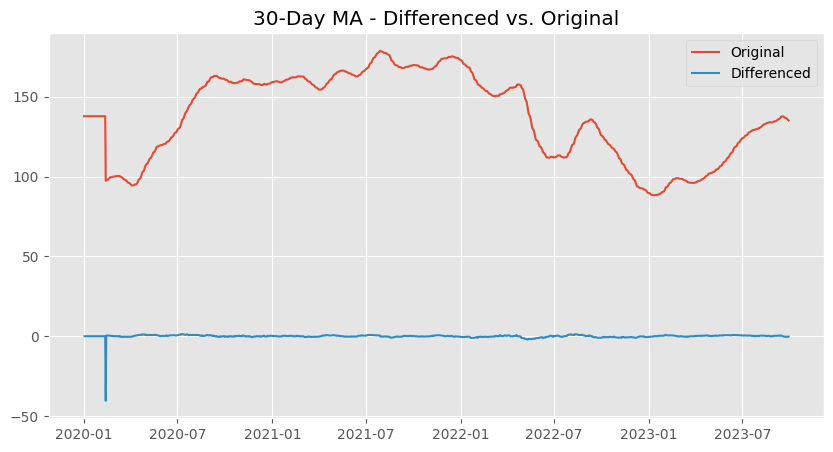


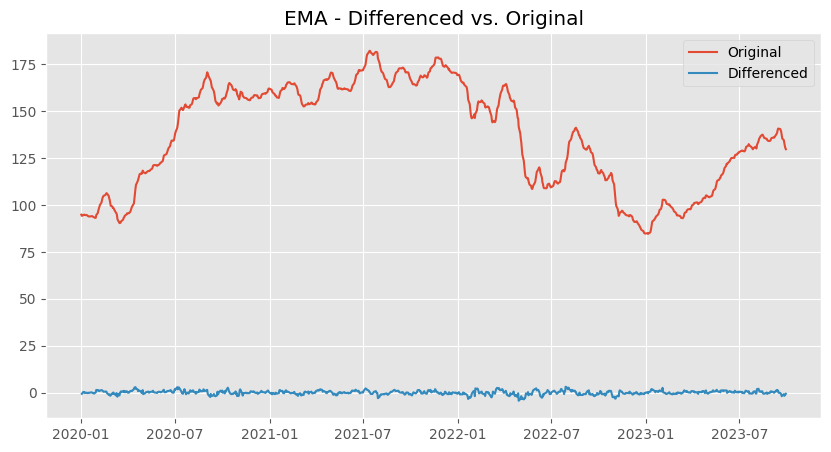


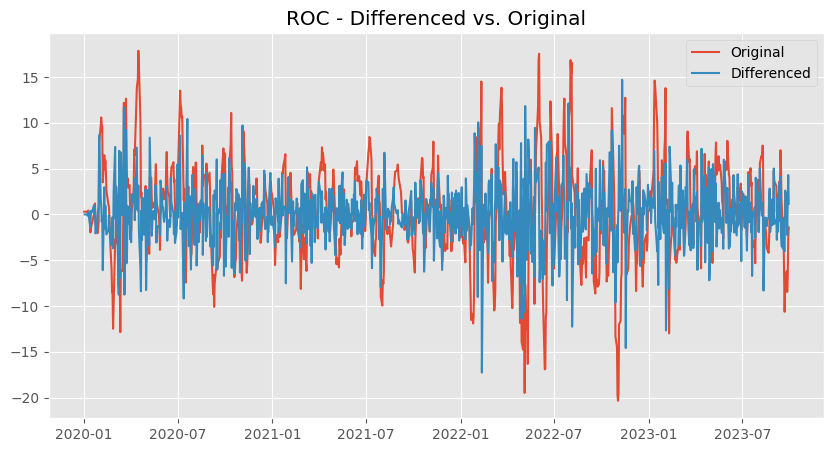


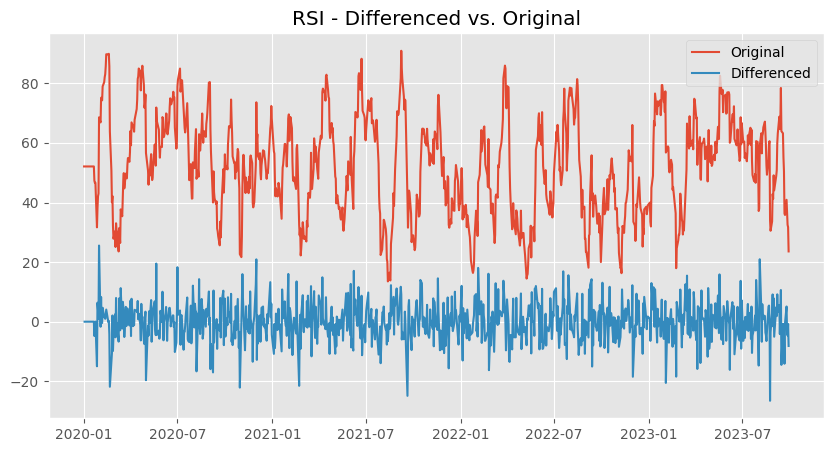


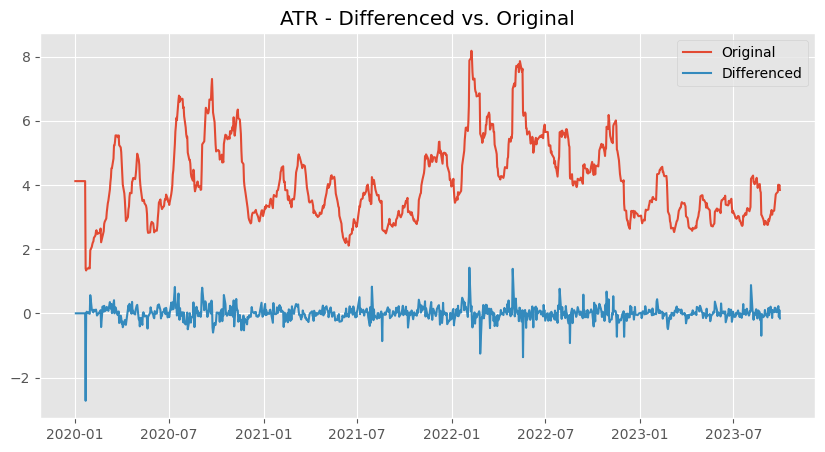






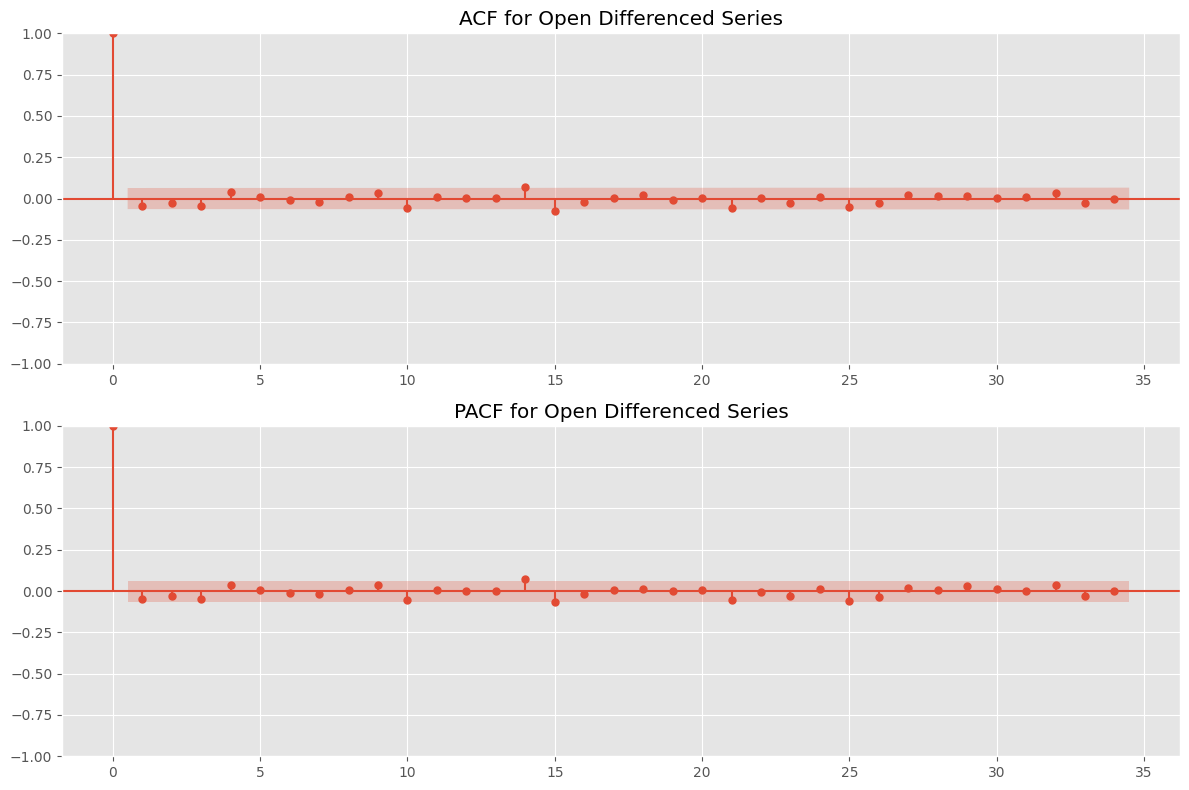






### **3. Autocorrelation and Partial Correlation of Differenced Series**

* Autocorrelation measures the linear relationship between lagged values of a time series.
* Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) will be used to identify the appropriate orders of autoregressive (AR) and moving average (MA) terms for modeling.
* The tools will guide in understanding the underlying structure of the time series, guiding the selection of model parameters, and ensuring that the model captures the relevant patterns in the data.
* They will help in building accurate and effective models.



**Observations**

* The differencing was mostly successful in removing stationarity. (**to edit**)

### **4. Performing Train-Test Split**

### **5. Scaling**

## **D. Modelling**

### **1. Baseline Model**

* The project uses the SARIMA model as the baseline model.
* Bedore proceeding to SARIMA model, first we conduct the Augmented Dickey-Fuller Test to ensure that the data is stationary

### **a. Augmented Dickey-Fuller Test**

### **b. Differencing to Make the Training Data Stationary**

* Since the result from the Augmented Dickley-Fuller Test showed that the data is non-stationary, differencing is done to make the data stationary.

### **b. Hyperparameters Tuning Using GridSearch CV**

* Hyperparameter tuning using GridSearch CV is done to find the optimal values of hyperparameters (p,d,q) to be used during modeling
* For this project, we focus on the Adjusted Close prices which provides a more accurate representation of the stock values over time. The column of the Adjusted Close is therefore used in modeling.

### **c. SARIMA**

Covariance Type: opg

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

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

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

ar.L1 -0.0190 0.034 -0.555 0.579 -0.086 0.048

ma.L1 -0.9981 0.038 -26.420 0.000 -1.072 -0.924

ar.S.L12 -0.0312 0.038 -0.820 0.412 -0.106 0.043

ma.S.L12 -0.9994 0.790 -1.265 0.206 -2.547 0.549

sigma2 0.0010 0.001 1.293 0.196 -0.001 0.002

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

Ljung-Box (L1) (Q): 0.00 Jarque-Bera (JB): 441.51

Prob(Q): 0.99 Prob(JB): 0.00

Heteroskedasticity (H): 1.64 Skew: -0.02

Prob(H) (two-sided): 0.00 Kurtosis: 6.78

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

Warnings:

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

* The SARIMAX result shows lower AIC of -2928.914 and BIC of -2905.888 indicating a good performance of the model.
* The higher Log Likelihood of 1469.457 shows that the model has a good fit on the data used.

**Observations**

* In the Histogram, the blue KDI line follows closely with the N(0,1) line showing a standard notation for a normal distribution with mean of 0 and standard deviation of 1. This indicates that the residuals are normally distributed.
* The Normal Q-Q plotshows the ordered distribution of Residuals along the linear trend of the sample of a normal distribution with N(0,1) indicating that the residuals are normally distributed.
* The Correlogram shows that the time series residuals have low correlation with the lagged version of itself.
* It is concluded that the SARIMAX model provides a good fit that can help in forecasting future values.

**Prediction on the Training Data and Test data**

* Perform a prediction on both the training dataset and Test dataset to see the the performance of the model on both the Training dataset and Test dataset.

### **d. FBProphet**

* The first step was to prepare the data for the model.
* The pipeline was built to create a DataFrame from a copy of the working DataFrame, monthly\_data. Here, the ds column was defined
* It then calculates some metrics that will be used as part of diagnostics.
* The prophet\_model function changes the Adj Close column to y, creates a subset, resets index, fits, forecasts and predicts using Facebook Prophet library.

prophet\_subset\_reset = prophet\_subset.reset\_index(drop=True)

### **2. RNN**

### **3. LSTM**

### **4. Evaluation of model**

### **5. Deployment**

# **6. Conclusion**

# **7. Limitations**

# **8. Recommendations**

# **9.Future Work**

### 