# Time Series Analytics Project

(Seasonal and non-seasonal dataset)

Author: Ankit Vikas Agrawal
Email: aagrawal1@stevens.edu / aavikas01@gmail.com
Project Supervisor: Dr. Hadi Safari Katesari

**Abstract**

The main objective of this paper is to explain two kinds of datasets—seasonal and non-seasonal—in the context of time series modelling. The goal is to determine the best time series model using various techniques which are discussed further in the report.

Time series analysis can be effectively performed only when the dataset is stationary. In this paper, I first preprocess the dataset, convert it into a univariate format, and set the datetime as the index. Subsequent steps include plotting the time series, ACF and PACF graphs which aid in identifying the appropriate model. I then test for stationarity using the ADF test. If the dataset is stationary, I can directly analyze the ACF and EACF bars to choose between AR or MA models. If it is not stationary, I apply techniques such as differencing and detrending to make it stationary.

Further, I apply ARIMA, SARIMA, and GARCH models and perform parameter estimation using criteria like AIC and BIC. The paper also covers various aspects of residual analysis including ACF plot, histogram, QQ plot, Shapiro-Wilk test, and Ljung-Box test. Finally, I perform forecasting on the original dataset to predict future values and evaluate the performance of the time series models.

**Keywords**: Seasonal and non-seasonal data, Stationarity, Data cleaning, Data Loading, Data Visualization, Differencing, Augmented Dickey–Fuller (ADF) test, Ljung-Box test, Shapiro-Wilk test AIC, ACF, PACF, Residual, ARIMA, SARIMA, GARCH, Forecasting.

**Introduction**

I gathered data from publicly available online sources. For the non-seasonal dataset, I used NVDA stock data spanning from 2023 to 2025, which contains six columns as shown in Figure 1. My analysis focused on the "Date" and "Close" columns. For the seasonal dataset, I used a large grocery store dataset obtained from Kaggle. Although the original dataset had more than ten columns, I focused on the "Date" and "Profit" columns, as shown in Figure 2. To simplify the analysis, I reduced the dataset to 201 rows.

| | Date | Open | High | Low | Close | Volume |
|---|---|---|---|---|---|---|
| 0 | 2025-04-07 | 87.46 | 101.75 | 86.62 | 97.64 | 611041250 |
| 1 | 2025-04-04 | 98.91 | 100.13 | 92.11 | 94.31 | 532273812 |
| 2 | 2025-04-03 | 103.51 | 105.63 | 101.60 | 101.80 | 338769406 |
| 3 | 2025-04-02 | 107.29 | 111.98 | 106.79 | 110.42 | 220601203 |
| 4 | 2025-04-01 | 108.52 | 110.20 | 106.47 | 110.15 | 222614000 |
| 5 | 2025-03-31 | 105.13 | 110.96 | 103.65 | 108.38 | 299212719 |
| 6 | 2025-03-28 | 111.49 | 112.87 | 109.07 | 109.67 | 229872500 |
| 7 | 2025-03-27 | 111.35 | 114.45 | 110.66 | 111.43 | 236902094 |
| 8 | 2025-03-26 | 118.73 | 118.84 | 112.71 | 113.76 | 296431719 |
| 9 | 2025-03-25 | 120.55 | 121.29 | 118.92 | 120.69 | 167447203 |

Fig: 1

| Order Date | Profit |
|---|---|
| 2014-01-31 | 31.015072 |
| 2014-02-28 | 18.745835 |
| 2014-03-31 | 3.176624 |
| 2014-04-30 | 25.843224 |
| 2014-05-31 | 22.448439 |
| 2014-06-30 | 36.863144 |
| 2014-07-31 | -5.884494 |
| 2014-08-31 | 34.758856 |
| 2014-09-30 | 31.074998 |
| 2014-10-31 | 21.687153 |
| 2014-11-30 | 29.220525 |
| 2014-12-31 | 32.315000 |
| 2015-01-31 | -56.569086 |
| 2015-02-28 | 43.966419 |

Fig: 2

**Data Visualization**

A. Non-seasonal dataset

a.

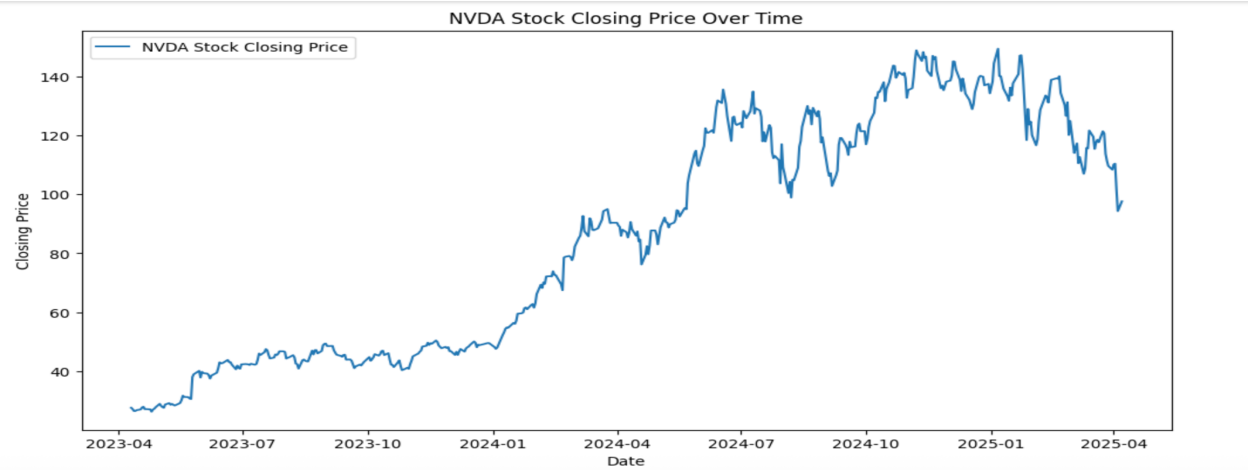



Fig: 3a

In Figure 3a, we can see that there is no clear pattern, which indicates that it is a non-stationary dataset.

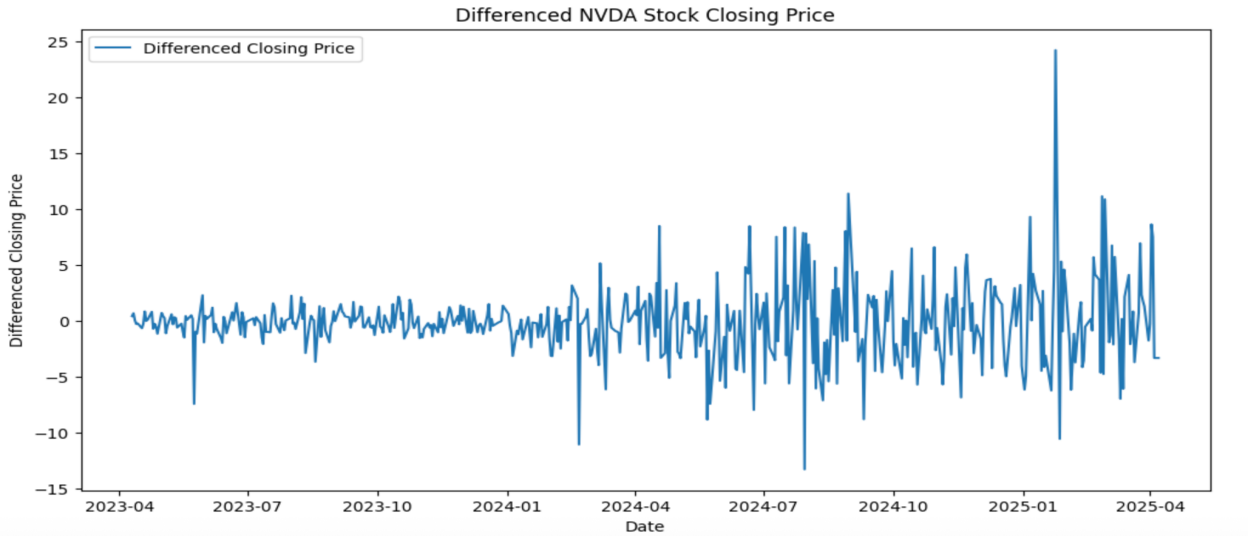



Fig: 3b

b. ACF plot

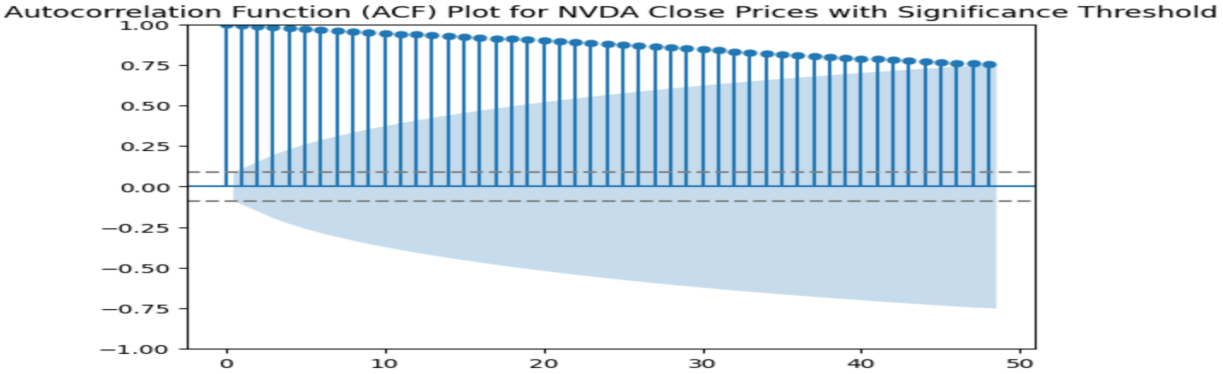



Fig: 4

The stock price shows strong positive autocorrelation at short lags, meaning yesterday's price heavily influences today's price, indicating a tight link between consecutive days. As the lag increases, the autocorrelation gradually decreases, suggesting that the impact of past prices on the current price weakens over time. However, significant autocorrelation persists even beyond short lags, revealing that the time series isn't entirely random—past prices retain some predictive power for future prices over an extended period. From figure 4 we can easily say it is a non-stationary.

c. ADF Test

```
result_original = adfuller(df['Close'])
print('ADF Statistic (Original):', result_original[0])
print('p-value (Original):', result_original[1])
print('Critical Values (Original):', result_original[4])
```

```
ADF Statistic (Original): -0.2988577938344866
p-value (Original): 0.9256659470708057
Critical Values (Original): {'1%': -3.4435761493506294, '5%': -2.867372960189225, '10%': -2.5698767442886696}
```

Fig: 5

In figure 5, The ADF statistic (-0.2989) is greater than all the critical values, and the p-value is significantly above 0.05, indicating that the dataset is non-stationary. Therefore, differencing is required to make it stationary.

B. Seasonal dataset

a.

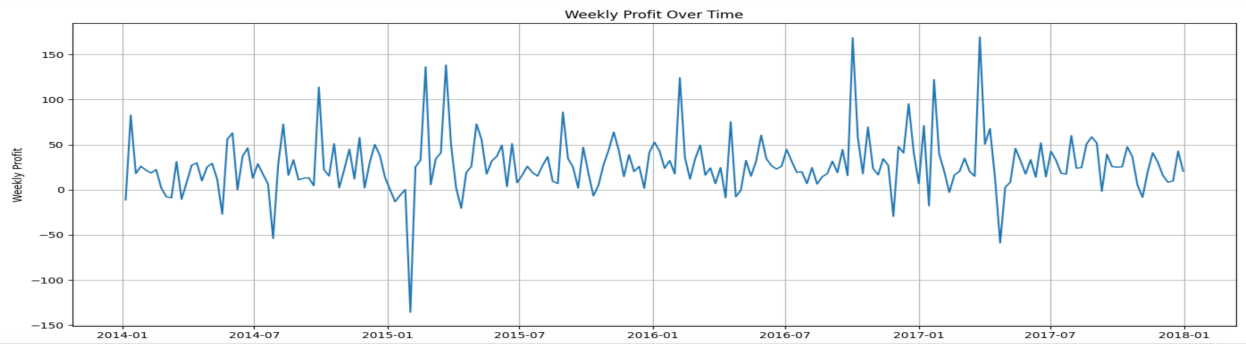



Fig: 6

```
ADF Statistic: -13.790688189145076
p-value: 8.938098275917793e-26
Critical Value (1%): -3.4621857592784546
Critical Value (5%): -2.875537986778846
Critical Value (10%): -2.574231080806213
The series is stationary.
```

Fig: 7

From Figure 6 and 7, we can see that there is a clear pattern and p-values, which indicates that it is a stationary dataset.

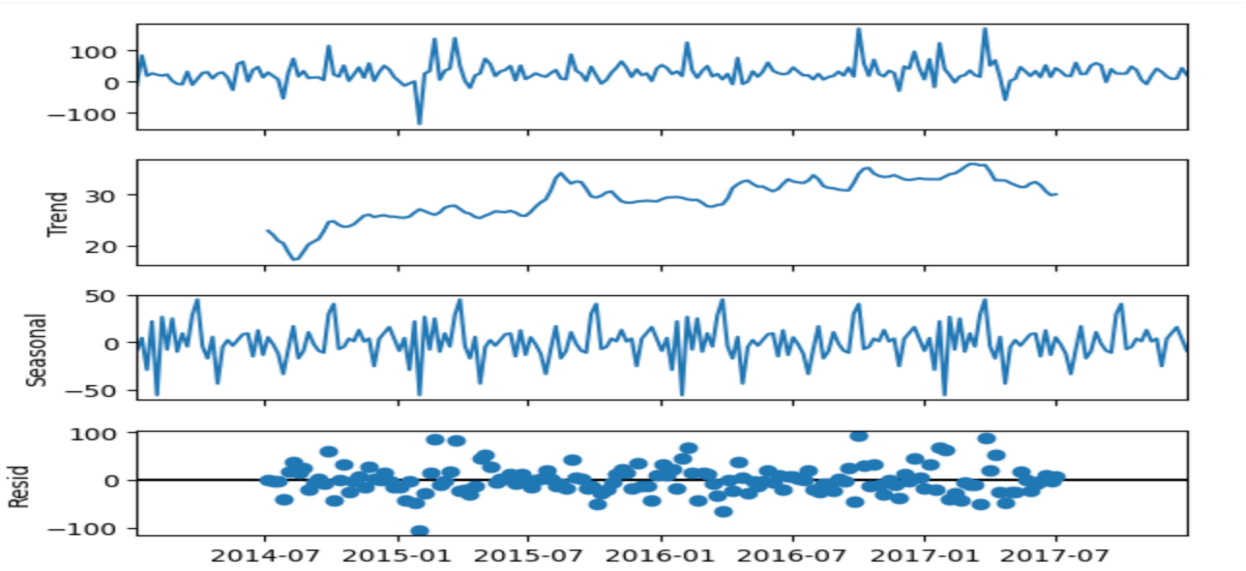



Fig: 8

**ACF and PACF plots**

A. Seasonal

a. ACF

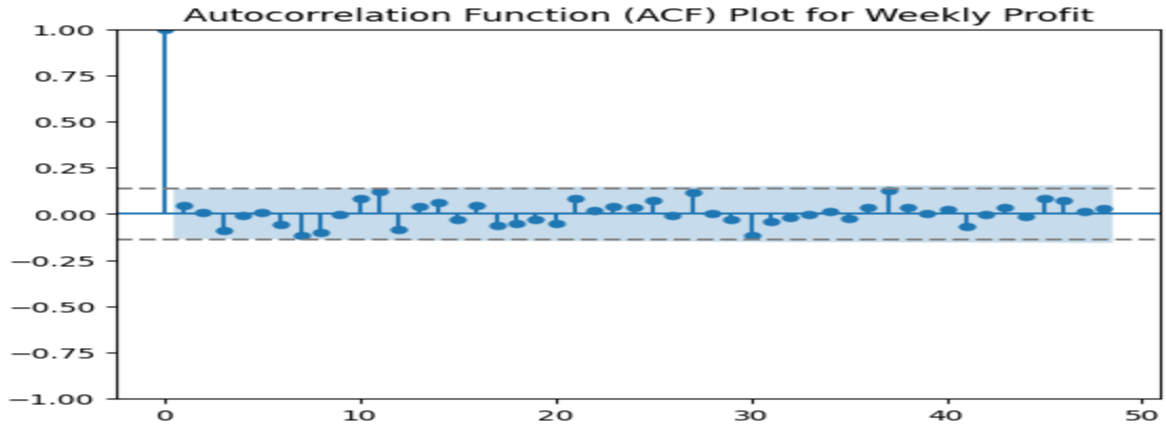



Fig: 9

b. PACF

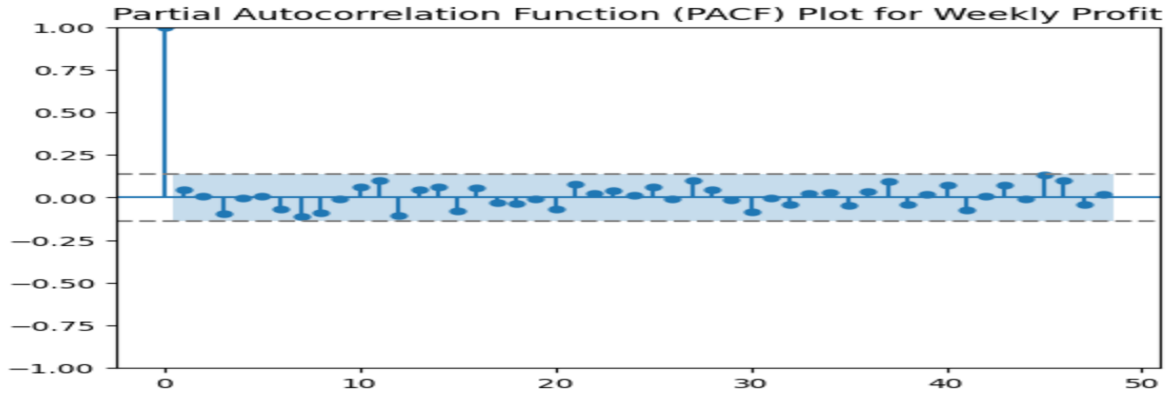



Fig: 10

B. Non-Seasonal

a. ACF

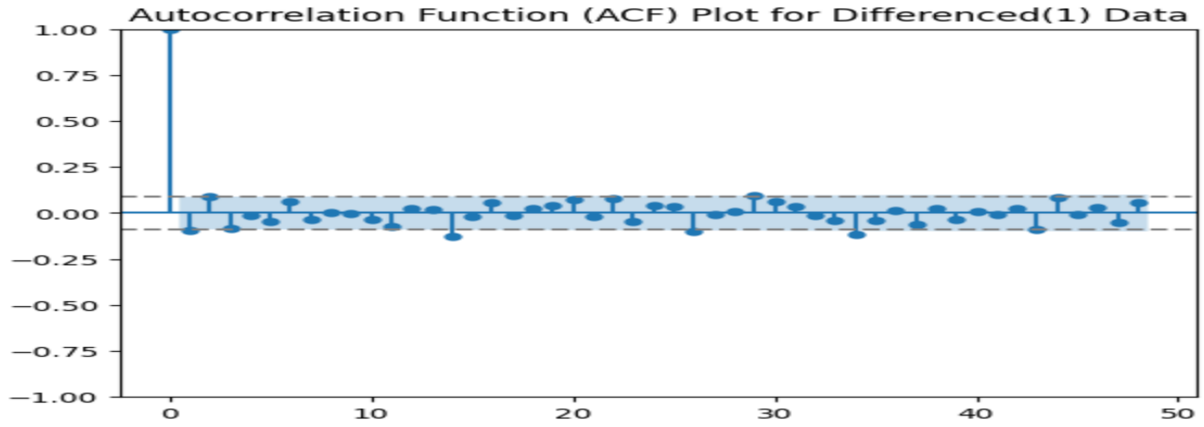



Fig: 11

b. PACF

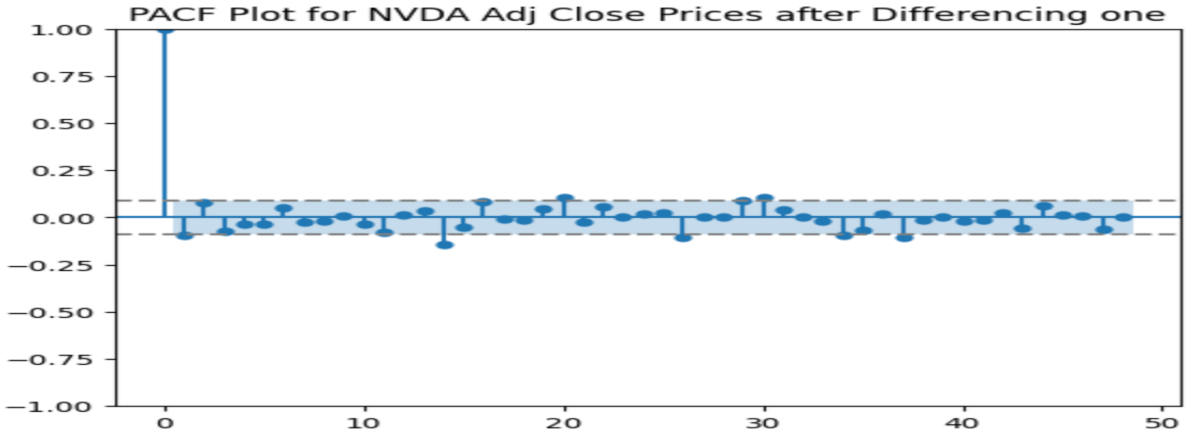



Fig: 12

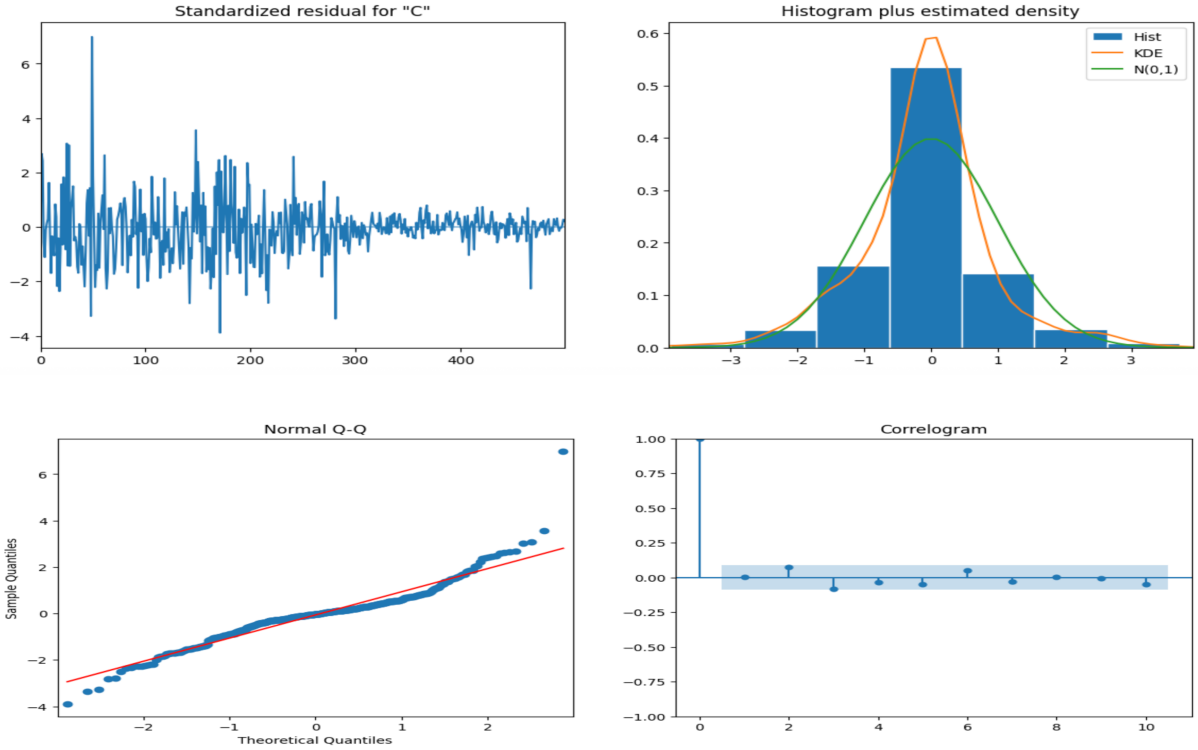
**Model of Time series**

For the seasonal dataset, I selected the parameters p=1, d=0, and q=1, which resulted in a better AIC value compared to other model combinations. Since the data exhibited seasonality, I used the SARIMA model

for forecasting. For the non-seasonal dataset, I experimented with various ARIMA models and found that the best AIC was achieved with p=1, d=1, and q=1. As this dataset consisted of financial data, I also applied the GARCH model, which yielded a better Ljung-Box test result compared to the ARIMA model, indicating improved residual behavior.

**Residual Analysis**

A. Non-seasonal



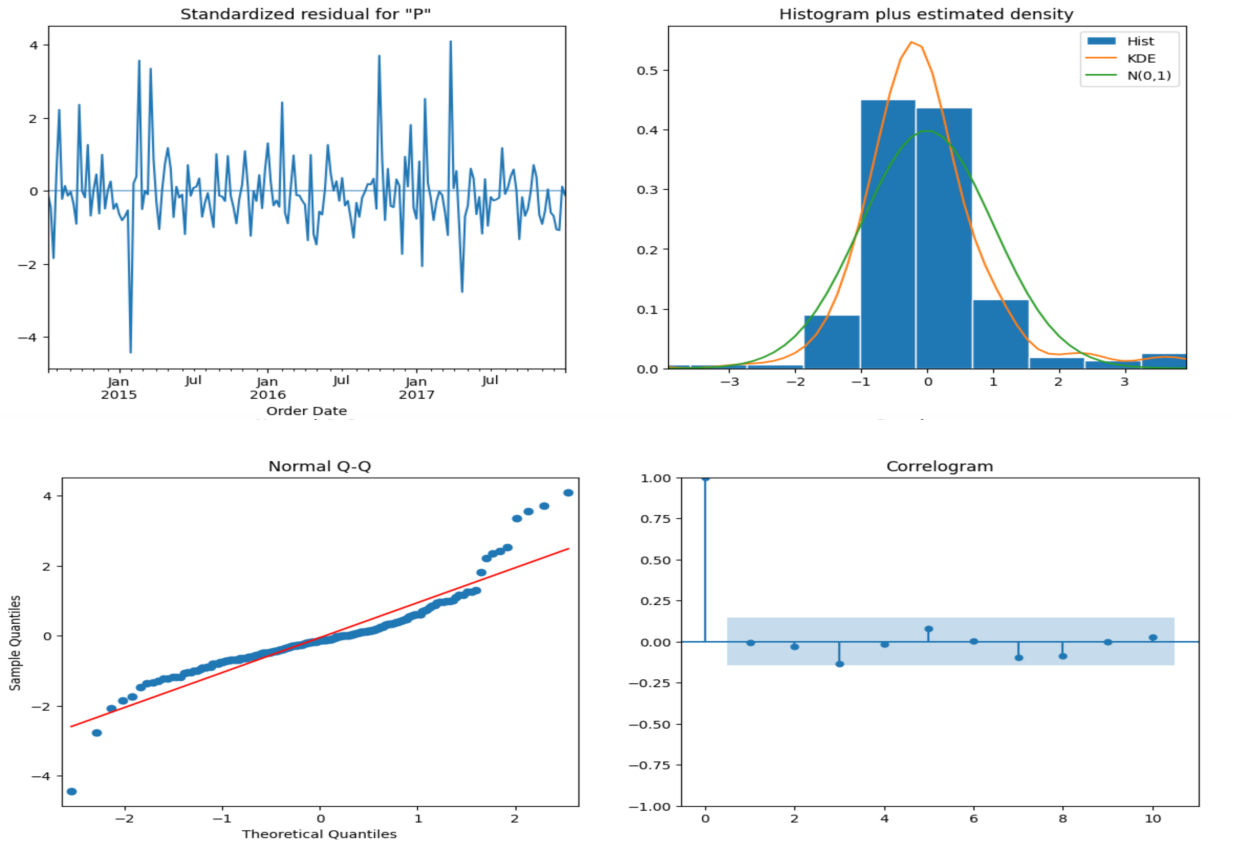

From the standardized residuals graph, we can observe that fluctuations around zero decrease in the later part of the series, which is a positive indication. The histogram displays a bell-shaped curve, suggesting that the residuals are approximately normally distributed. In the correlogram, all lags beyond the first fall within the confidence interval, indicating the presence of white noise.

```
stat, p = shapiro(data)

# Print results
print("Shapiro-Wilk Test Statistic:", stat)
print("p-value:", p)

if p > 0.05:
    print("Data appears to be normally distributed (fail to reject H0).")
else:
    print("Data does not appear to be normally distributed (reject H0).")
```

```
Shapiro-Wilk Test Statistic: 0.9219184111044073
p-value: 1.8984881534724404e-15
Data does not appear to be normally distributed (reject H0).
```

The Shapiro-Wilk test indicates that the data is not normally distributed, which is expected and common for financial datasets.

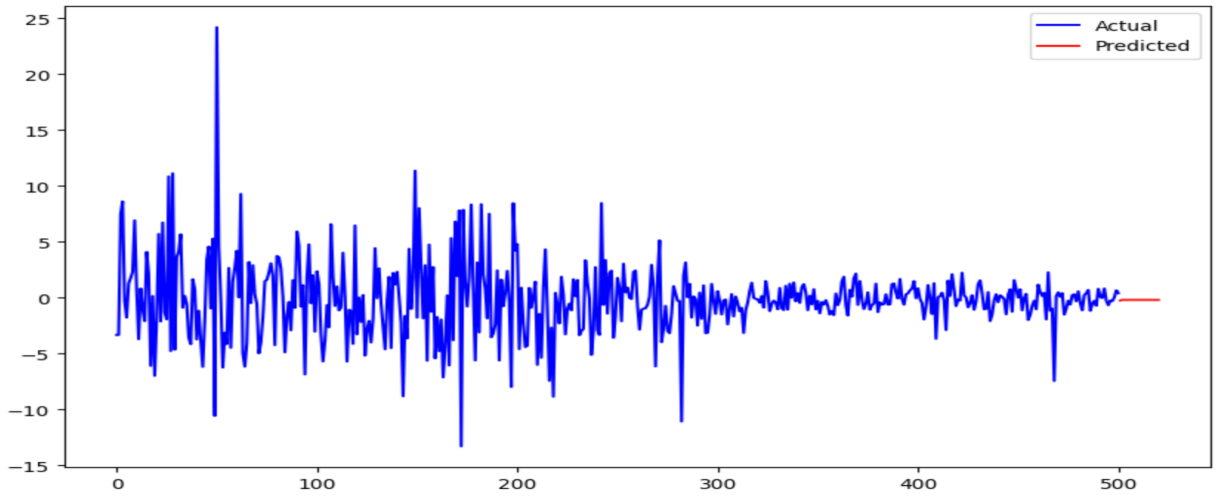
B. Seasonal



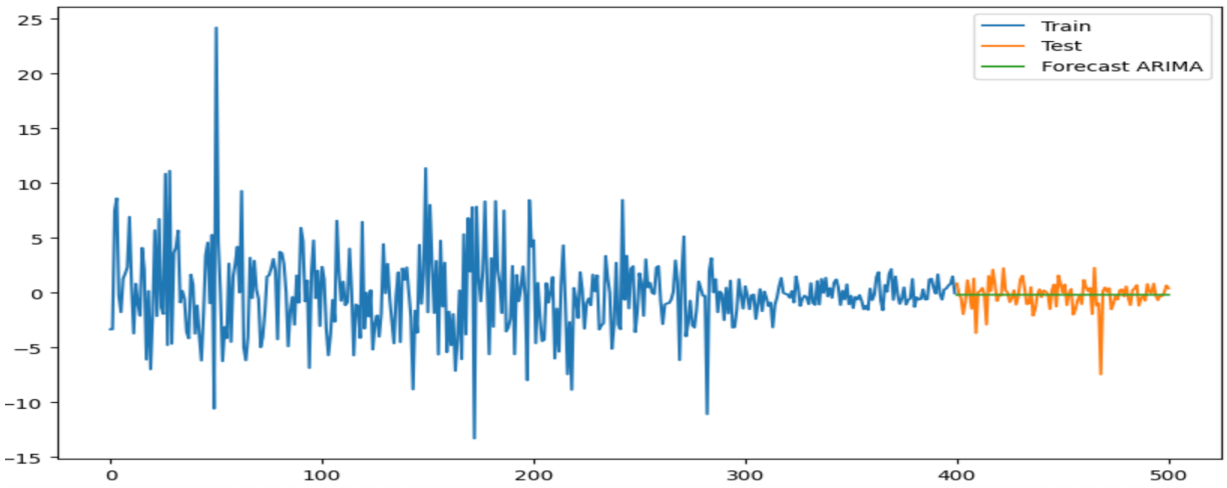

**Forecasting**

A. Non-seasonal

a. Normal forecasting

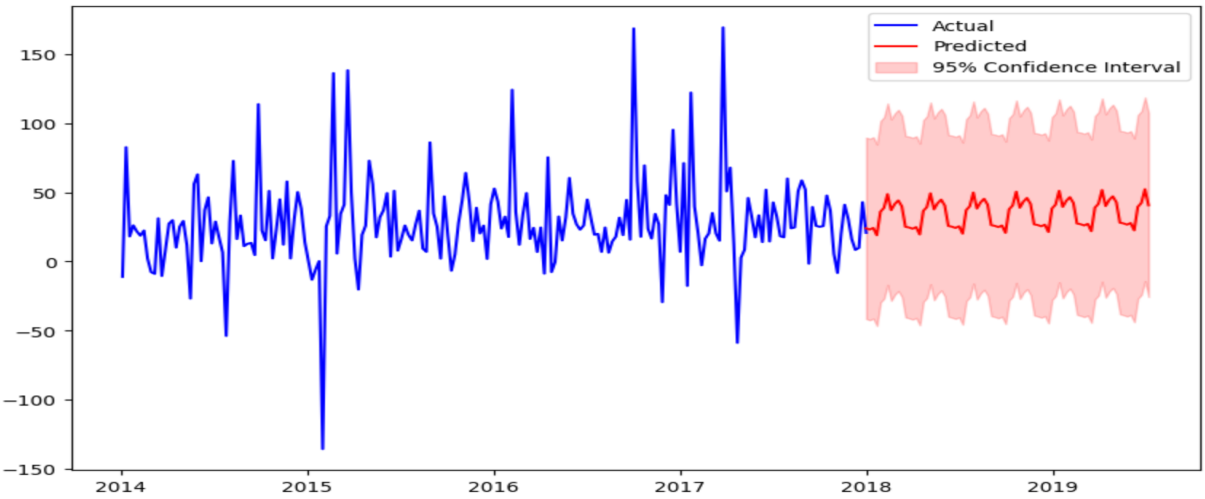


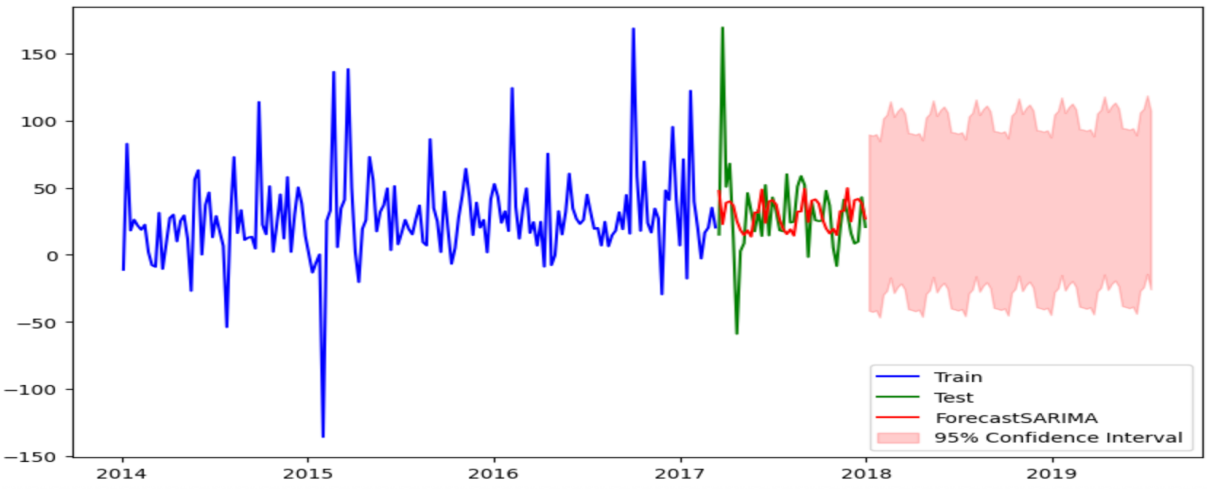

Here, I attempted to predict the next few days of NVDA's closing stock price, which is shown in orange. Due to the nature of the data and the use of the mean close value, the forecast appears as a straight line.



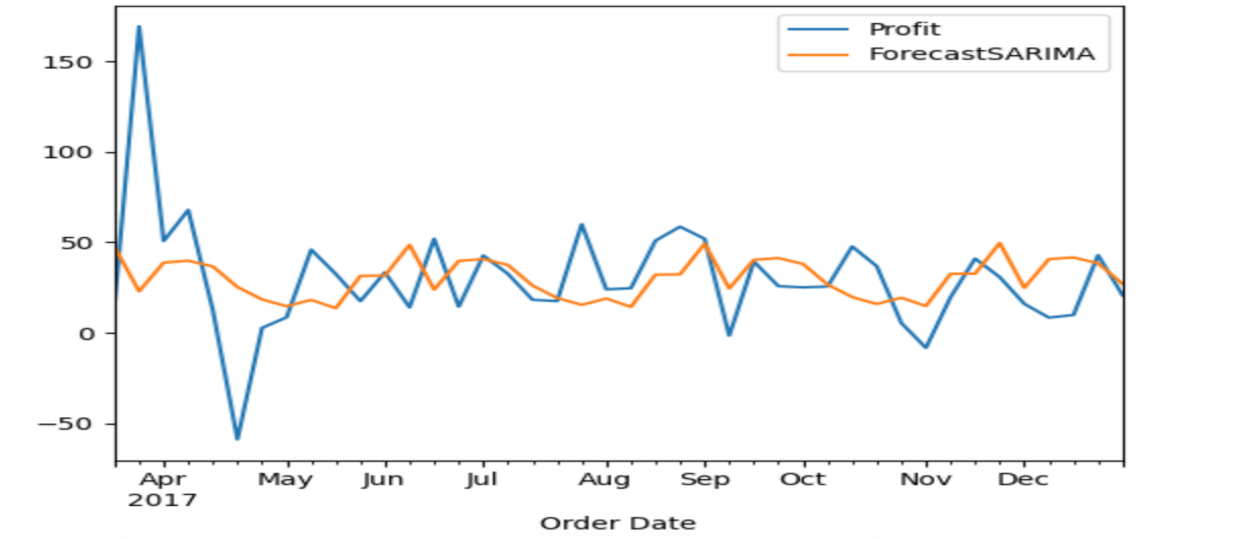

Here I divided the data into train and test data to predict the close stock.

B. Seasonal



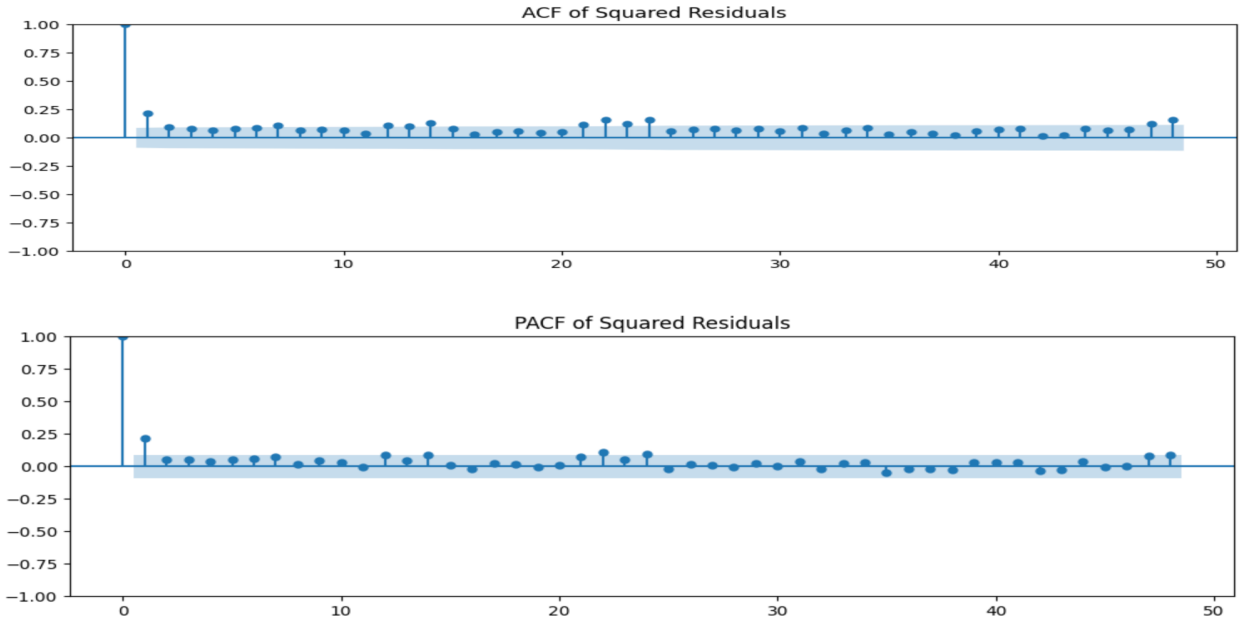

Here as well, I attempted to predict the profit using a standard approach by splitting the data into training and testing sets, as illustrated in the graph.




In this figure, there is better visual representation of predicted value using SARIMA model.

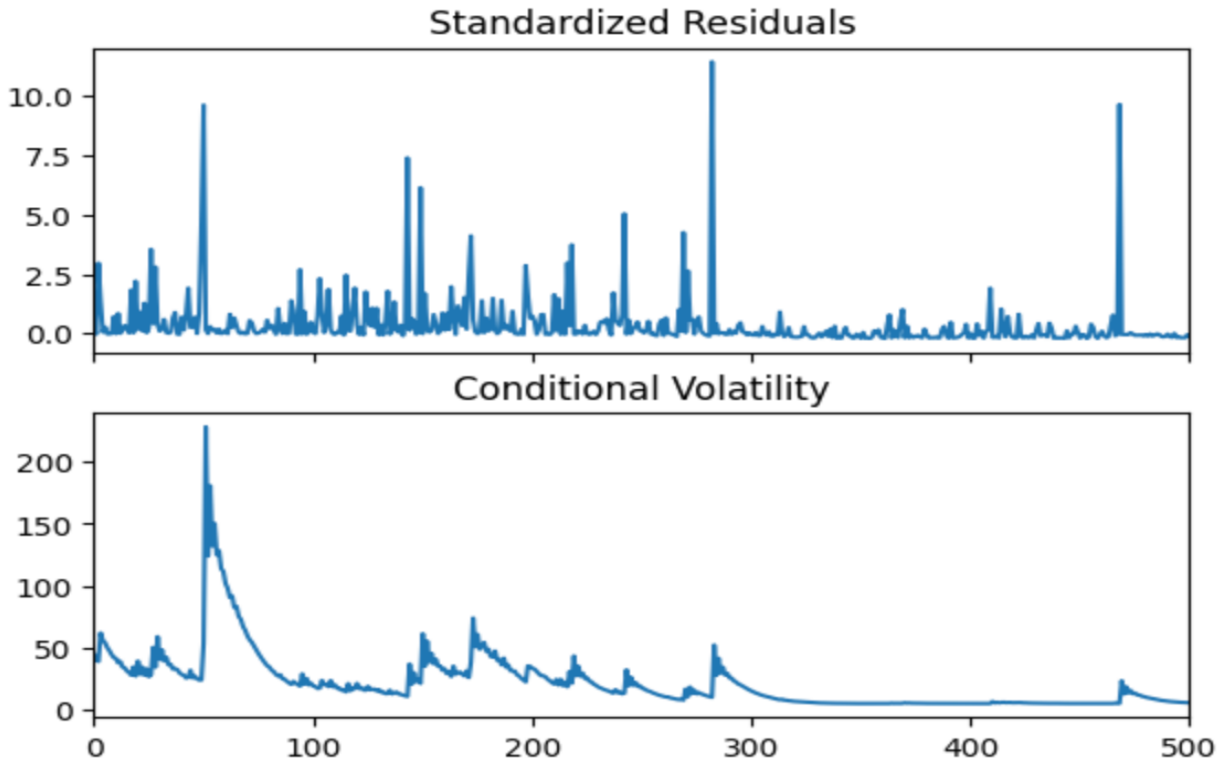
**GARCH Model**

Non-seasonal




For the non-seasonal data, I implemented the GARCH model. From the ACF graph, we observe that two lags fall outside the confidence interval (excluding the first and later lags), suggesting that the value of q could be 1 or 2. The PACF graph shows only one significant lag outside the confidence interval, indicating p as 1. Based on these observations, I tested various combinations, and the best result was obtained with p = 1 and q = 2.

**Diagnostics**



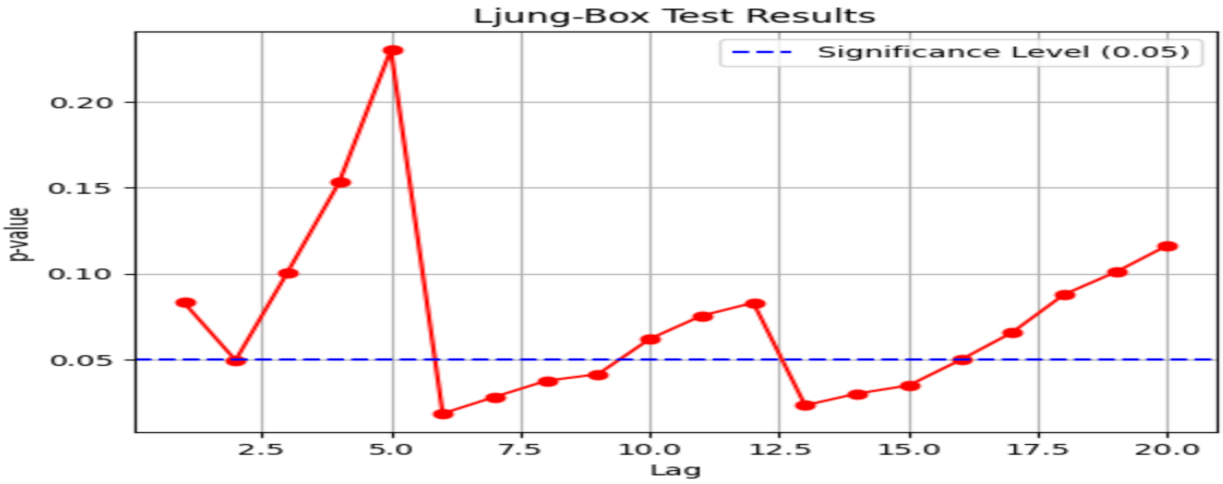

Standardized Residuals: This plot shows the model's errors, adjusted for the estimated volatility. Ideally, these residuals should look like white noise, meaning they have no discernible patterns. Large spikes or trends in this plot indicate that the GARCH model isn't fully capturing the volatility dynamics.

Conditional Volatility: This plot illustrates how the GARCH model estimates the volatility of the time series over time.

**Residual Analysis of GARCH model**

Ljung-Box Test Results:

| | lb_stat | lb_pvalue |
|---|---|---|
| 1 | 3.001522 | 0.083186 |
| 2 | 6.009235 | 0.049558 |
| 3 | 6.239033 | 0.100543 |
| 4 | 6.693593 | 0.152994 |
| 5 | 6.874028 | 0.230179 |
| 6 | 15.229046 | 0.018548 |
| 7 | 15.680577 | 0.028200 |
| 8 | 16.352810 | 0.037600 |
| 9 | 17.499147 | 0.041450 |
| 10 | 17.606637 | 0.061973 |
| 11 | 18.273491 | 0.075448 |
| 12 | 19.236631 | 0.082976 |
| 13 | 24.962605 | 0.023346 |
| 14 | 25.476277 | 0.030146 |
| 15 | 26.309785 | 0.034890 |
| 16 | 26.317864 | 0.049717 |
| 17 | 26.503731 | 0.065759 |
| 18 | 26.551016 | 0.087810 |
| 19 | 27.164798 | 0.100873 |
| 20 | 27.712404 | 0.116384 |




The Ljung-Box test results for the GARCH model show that many p-values are greater than 0.05, which is a positive indication and can be observed in the graph where several points lie above the significance level. However, some points fall below the significance threshold, suggesting that there is still room for improvement in the GARCH model.

**Conclusion**

This time series analysis project provided a comprehensive exploration of forecasting methodologies applied to both seasonal and non-seasonal datasets. Under the guidance of my professor, I followed a structured, step-by-step approach to analyze and model the data effectively. For the non-seasonal dataset (NVDA stock prices), I implemented ARIMA and GARCH models to capture volatility and trends, while for the seasonal dataset, I employed SARIMA to account for periodic patterns.

Throughout the project, I gained hands-on experience in data visualization, cleaning, model selection, and residual diagnostics. Key statistical tests such as the ADF test (for stationarity), Ljung-Box test (for autocorrelation), and Shapiro test (for normality) were performed to validate model assumptions. Additionally, I conducted forecasting using both full-dataset training and train-test splits, evaluating model performance through error metrics.

While the project demonstrates a solid understanding of time series techniques, there remain areas for improvement. Some models exhibited limitations in capturing extreme volatility or complex seasonal patterns, suggesting that advanced methods like machine learning-based forecasting (e.g., LSTM, Prophet) or hybrid models could enhance predictive accuracy. Further refinement in hyperparameter tuning and feature engineering could also strengthen results.

Overall, this project significantly deepened my expertise in time series analysis, equipping me with essential skills for real-world forecasting challenges. Moving forward, I aim to explore more sophisticated models and larger datasets to further refine my analytical capabilities.

**Acknowledgments**
I sincerely thank my professor for the clear roadmap and guidance, which were instrumental in successfully completing this project.

**Links**

https://www.marketwatch.com/investing/stock/nvda/download-data?startDate=04/08/2023&endDate=4/8/2024

https://www.kaggle.com/datasets/vivek468/superstore-dataset-final