How an 18-year-old earned millions by forecasting the stock market?

A Story of Predictive Modeling with the Nepal Stock Exchange (NEPSE) Data

Time Series Forecasting

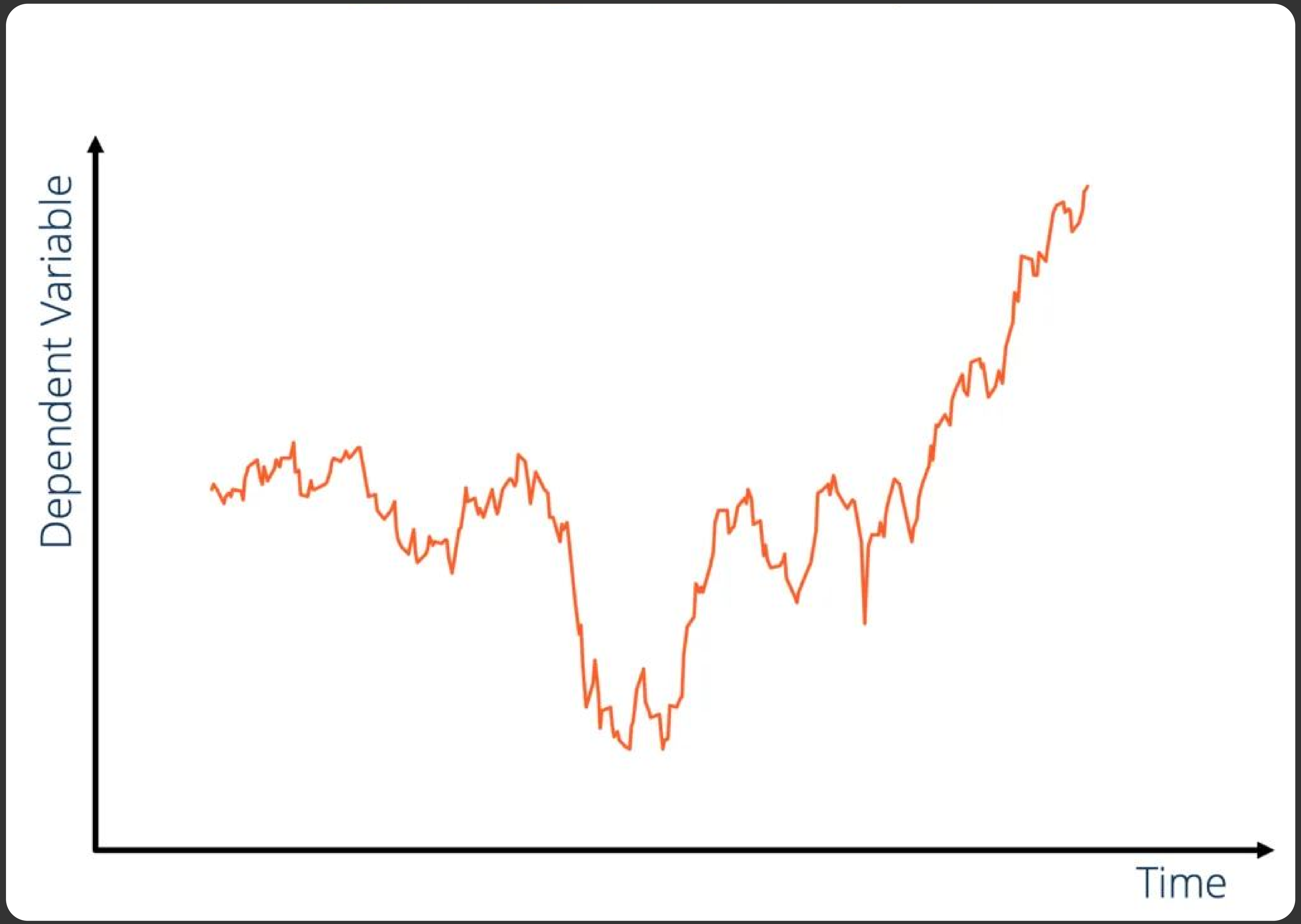
NEPSE

Birat Poudel

7th October, 2024

Table of Content

Let’s start with a brief introduction to Time Series



Introduction to Time Series Forecasting

What is Time Series Forecasting?

- 💡 Analysing historical data to predict future values
- 💡 Uses statistical techniques to identify patterns, trends, and seasonality for accurate predictions

Why is it important in the stock market?

- 💡 **Predictive Insights:** Helps anticipate stock prices for informed trading decisions
- 💡 **Strategic Planning:** Use forecasts to optimize strategies and timing

The Problem: Predicting Stock Prices

- 💡 Inherently difficult due to market volatility and external factors
- 💡 Successful predictions can lead to substantial rewards

Can we turn uncertainty into opportunity?

Data Pipeline

2022-01-01

2023-12-31

kaggle

2024-01-01

2024-09-20

NEPSE STOCK DATA

Data Overview

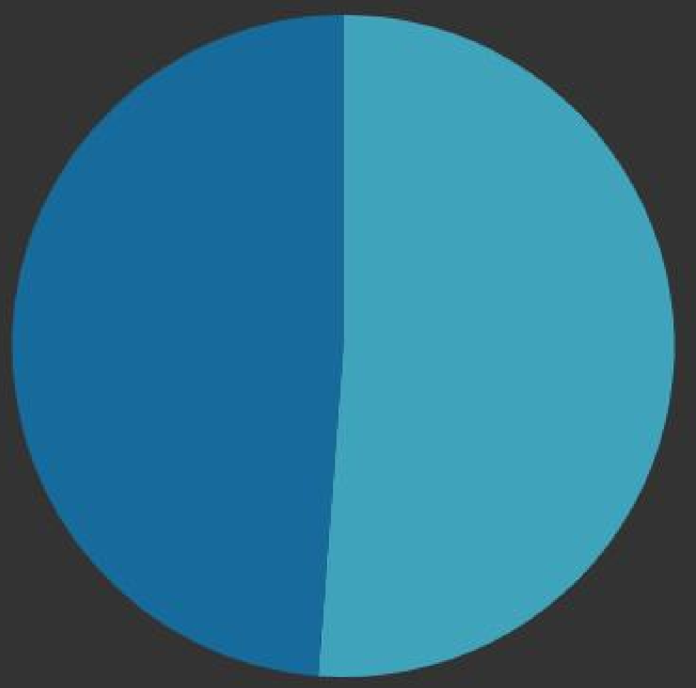
Introducing the Nepal Stock Exchange (NEPSE) Dataset

💡 Contains daily stock prices for 228 companies from 2022-01-01 to 2024-09-20

💡 Key variables include:

- **Open:** Opening price of the stock
- **Close:** Closing price of the stock
- **High:** Highest price during the day
- **Low:** Lowest price during the day
- **Volume:** Number of shares traded

This rich dataset forms the foundation for our forecasting analysis



Original Dataset

	Date	Close_ACLBSL	Close_ACLBSLP	Close_ADBL	Close_ADBLD83	Close_ADLB	Close_AHL	Close_AHPC	Close_AIL	Close_AKBSL	...	Vol_UPCL	Vol_UPPER
0	2022-01-01	1227.0	NaN	492.0	NaN	NaN	NaN	429.8	678.0	NaN	...	31544.0	57495.0
1	2022-01-02	1256.0	NaN	411.0	1071.0	NaN	NaN	442.0	682.0	NaN	...	37079.0	108017.0
2	2022-01-03	1302.0	NaN	415.0	NaN	NaN	NaN	435.0	700.0	NaN	...	36798.0	119110.0
3	2022-01-04	1390.0	NaN	422.0	1050.0	NaN	NaN	445.0	698.1	NaN	...	44737.0	160123.0
4	2022-01-05	1370.0	NaN	419.0	1050.0	NaN	NaN	450.0	686.1	NaN	...	80418.0	216456.0

💡 **Total Values:** 25,29,672

❌ **Missing Values:** 12,33,455 ✅ **Present Values:** 12,96,217

Let’s do some data preprocessing before extracting insights from it

Data Preprocessing

Handling Missing Values

💡 Taking top **50** companies that has less missing values percentage in **Open**, **Close**, **High**, **Low** and **Vol** fields

💡 Filling any **NaNs** in numeric columns with **forward fill** and then **backward fill** as a last resort

- **Forward Fill (ffill):** Fills missing values by taking the last valid observation and propagating it forward
- **Backward Fill (bfill):** Fills missing values by taking the next valid observation and propagating it backward

Preprocessed Data Overview

	Date	Close_ADBL	Close_AHPC	Close_AKJCL	Close_AKPL	Close_ALBSL	Close_ALICL	Close_API	Close_BARUN	Close_BFC	...	Vol_JFL	Vol_JOSHI	Vol_KKHC
0	2022-01-01	492.0	429.8	273.0	493.0	1415.0	1123.0	445.0	462.0	440.0	...	11922.0	53763.0	4942.0
1	2022-01-02	411.0	442.0	273.0	506.0	1438.0	1165.0	467.0	458.0	438.1	...	8566.0	39788.0	8084.0
2	2022-01-03	415.0	435.0	273.9	504.9	1518.7	1171.0	469.0	462.0	442.0	...	34660.0	48659.0	12504.0
3	2022-01-04	422.0	445.0	280.0	514.2	1508.0	1180.0	475.0	470.0	456.0	...	27374.0	72353.0	7265.0
4	2022-01-05	419.0	450.0	286.0	514.0	1460.0	1169.9	477.0	471.0	460.0	...	39191.0	145275.0	15918.0

💡 **Total Values:** 2,57,526

❌ **Missing Values:** 0 ✅ **Present Values:** 2,57,526

Let's select the 5 banking stocks ADBL, CZBIL, EBL, GBBL and MBL and visualize

- Line Plot of Closing Prices
- Correlation Matrix of Closing Prices
- Volume Traded over Time
- High and Low Prices over Time

ADBL

GBBL

CZBIL

MBL

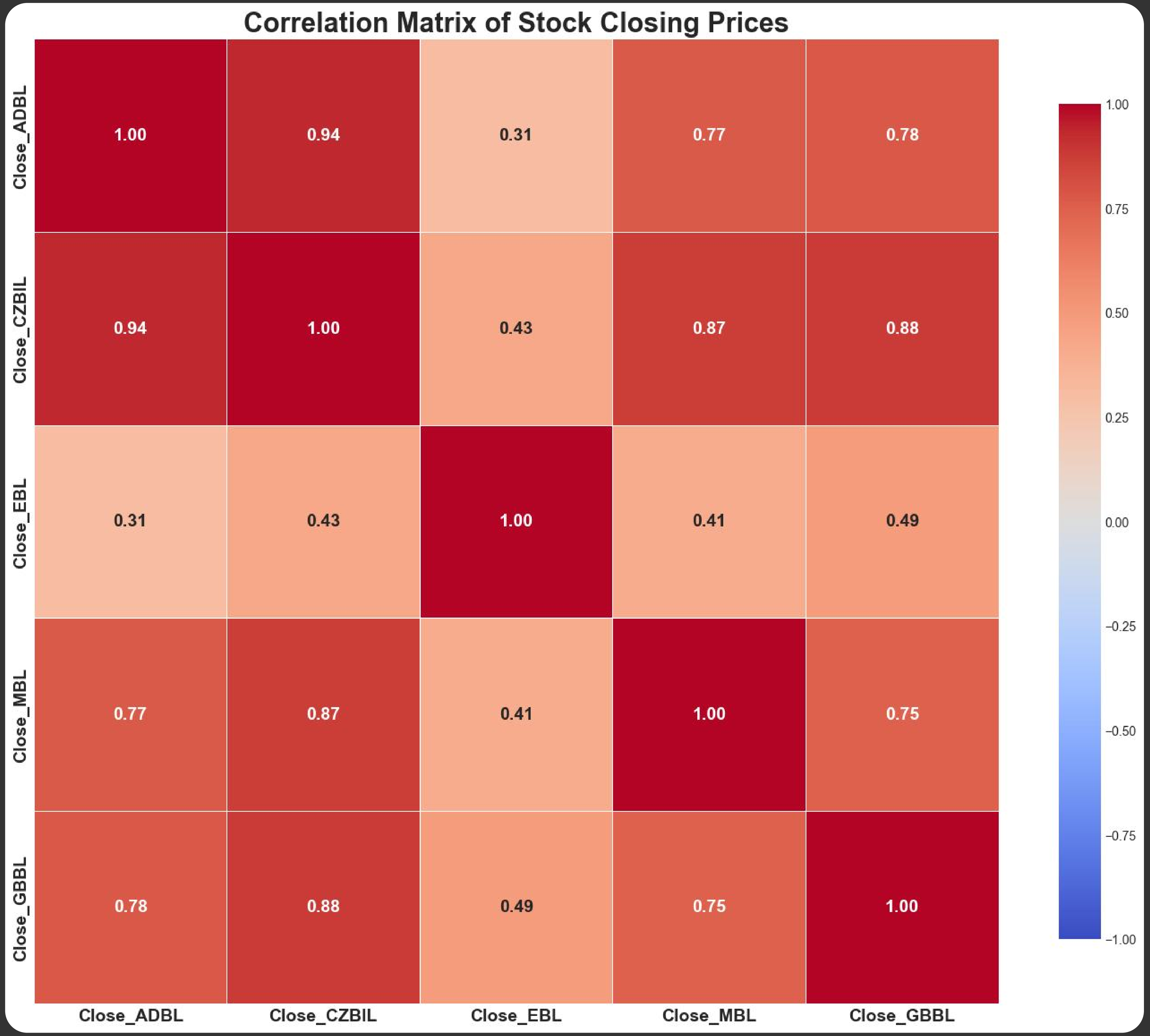
EBL

Line Plot of Closing Prices

- 💡 Shows the trend of the closing price over time for **ADBL**, **CZBIL**, **EBL**, **GBBL**, and **MBL**
- 💡 **EBL** consistently shows the highest closing price among selected stocks

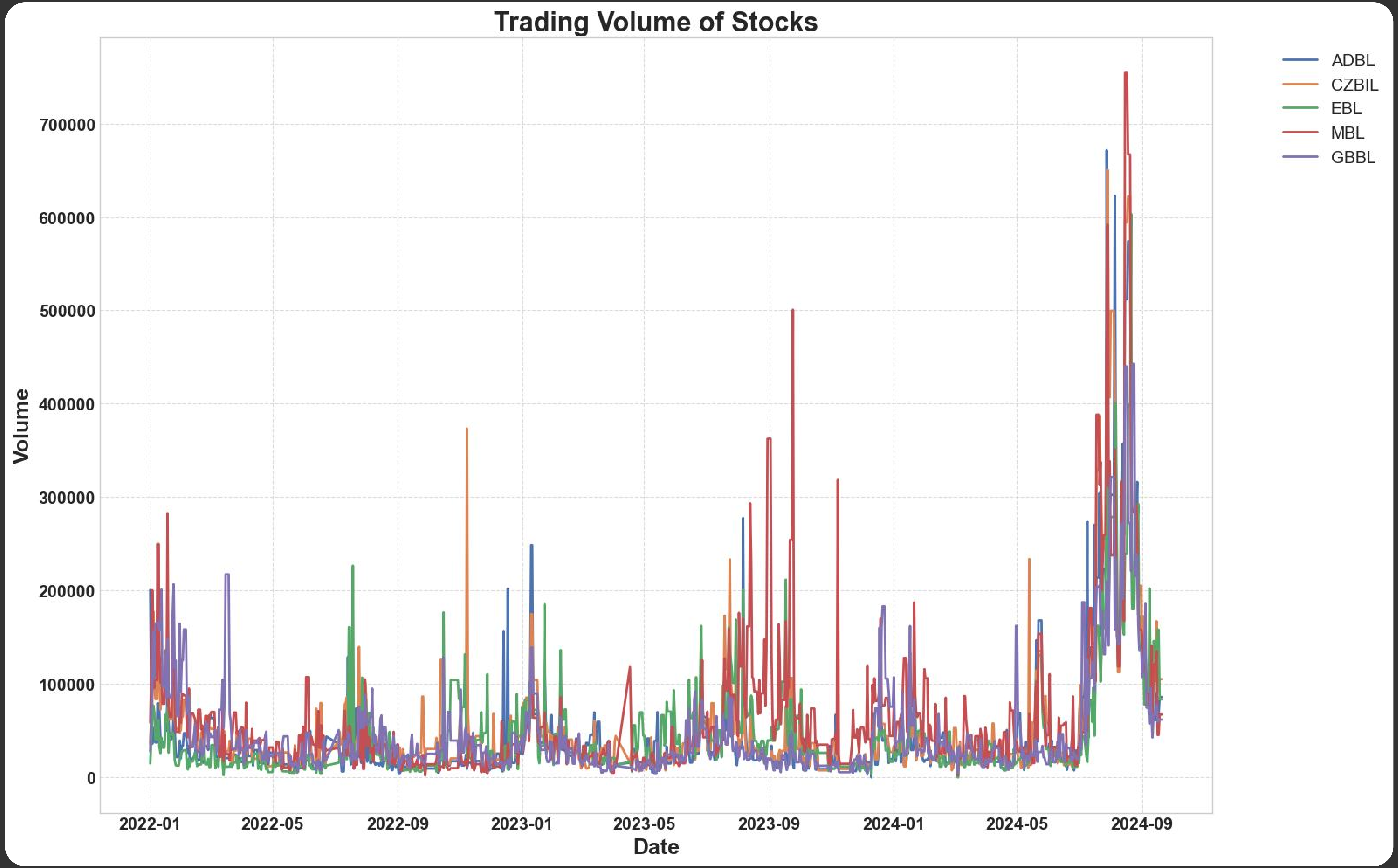
Correlation Matrix of Closing Prices

- 💡 If two stocks have a high positive correlation, they tend to move in the same direction
- 💡 This can indicate that they are influenced by similar market factors



Volume Traded over Time

- 💡 Visualizes the volume of stocks traded over time for **ADBL**, **CZBIL**, **EBL**, **GBBL**, and **MBL**
- 💡 **MBL** consistently maintains the highest trading volume throughout the period, especially in recent months



Initial Analysis

💡 From the initial analysis the following three banking stocks are recommended

1. ADBL (Agricultural Development Bank Limited)

💡 With an annualized volatility of 22.41%, ADBL offers a balance between potential returns and risk

- For most stocks, especially in the banking sector, a "good" volatility percentage might be considered in the moderate range, around 15% to 25%

2. CZBIL (Citizens Bank International Limited)

💡 CZBIL has a highest average daily **trading volume of 53263 shares** compared to selected 5 stocks

- **Suggesting good liquidity and ease of trading**

3. EBL (Everest Bank Limited)

💡 The absolute change in the closing price of EBL stock from the beginning to the end of the dataset is most positive compared to 5 selected stocks

- It indicates an upward trend
- The stock price has increased over the period

एभरेष्ट बैंक लिमिटेड
EVEREST BANK LIMITED
(A joint-Venture with punjab national bank, India)
Consistent, Strong & Dependable

Basics of Time Series Components

Trend

- 💡 The long-term direction in which the data is moving
- 💡 Ex: Steady increase in stock price over several months

Seasonality

- 💡 Regular patterns or fluctuations occurring at specific intervals (e.g., monthly, quarterly)
- 💡 Ex: Retail stock spikes during the holiday season each year

Cycles

- 💡 Irregular patterns that can span several years
- 💡 Ex: Stock price rises during economic expansions and falls in recessions

Noise

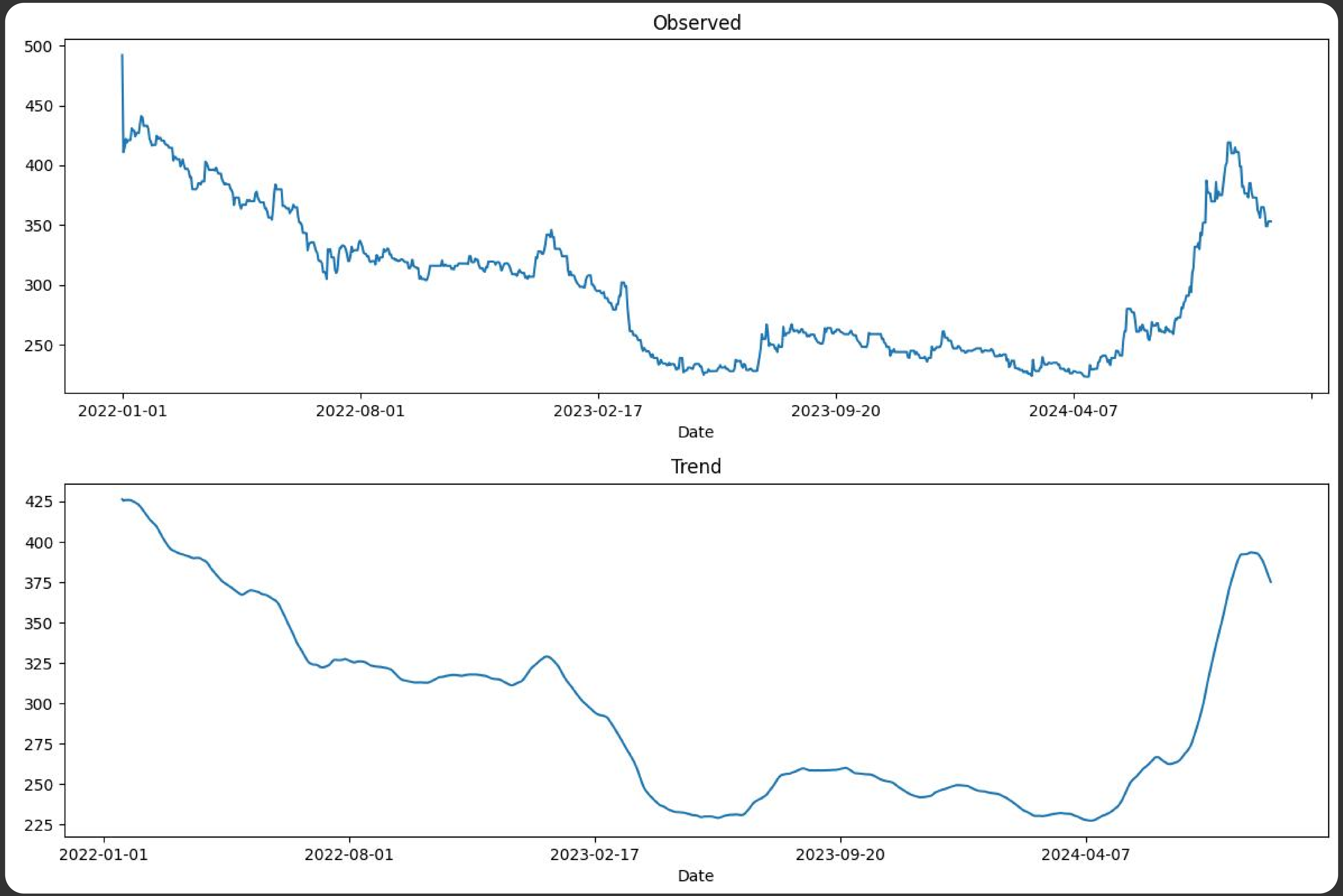
- 💡 Random variations in data that cannot be attributed to trend, seasonality or cycles
- 💡 Ex: Daily price fluctuations due to random news events

Let’s select the Close Price of ADBL and visualize time series components of it

- 💡 Trend
- 💡 Seasonality
- 💡 Residual

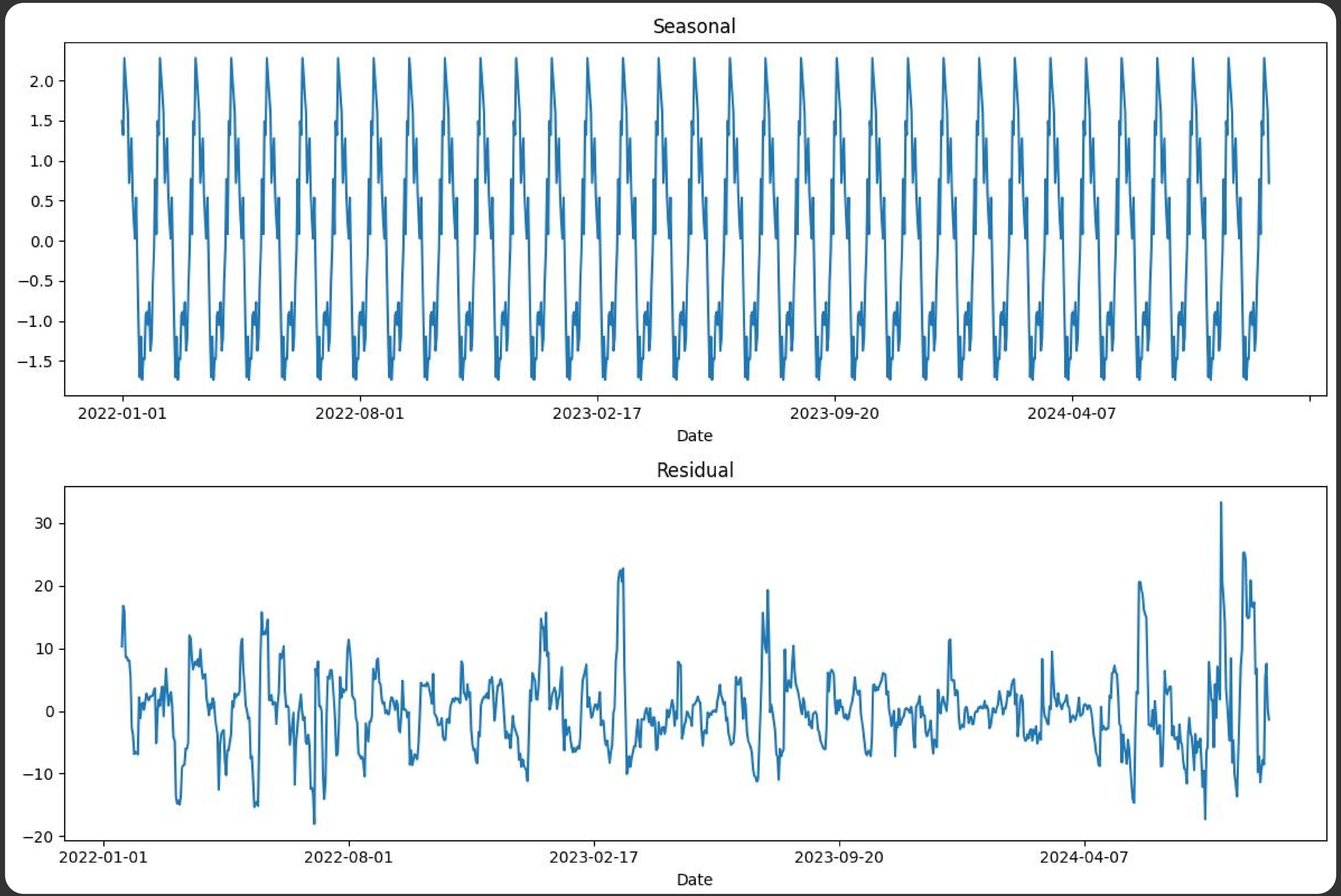
Close ADBL (Observed and Trend)

- 💡 The overall stock price shows a long-term decline from mid-2022 until early 2024
- 💡 Followed by a sharp upward movement in the last few months



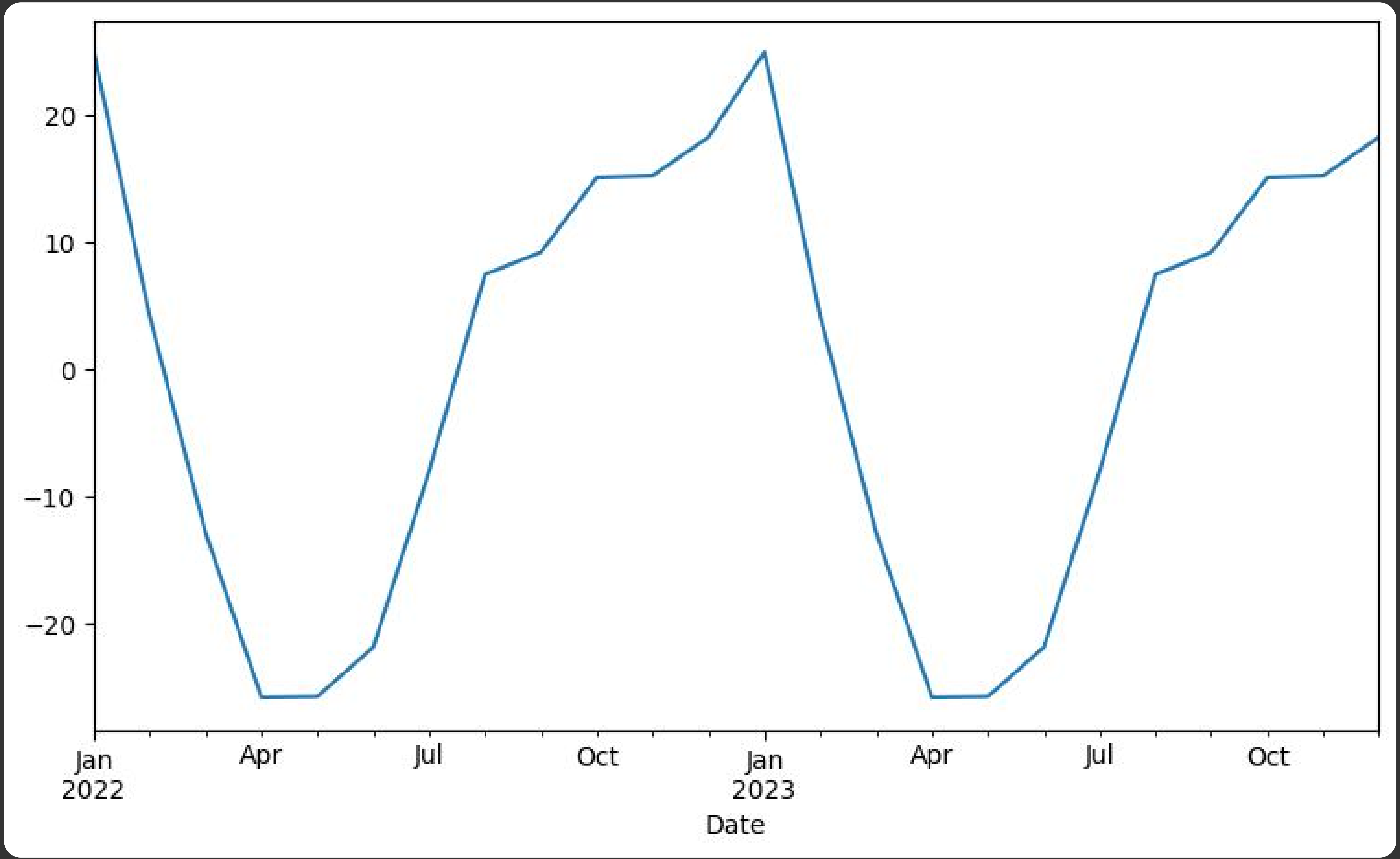
Close ADBL (Seasonality and Outlier)

- 💡 Seasonality represent recurring fluctuations in stock prices
- 💡 Residuals are remaining random fluctuations after removing trend and seasonality



Drilling Down and Observing Seasonality

💡 Data peaks in early months (January) and dips in mid-year (July)



Decomposition of Time Series Components

💡 **Trend:**

- The overall stock price shows a long-term decline from mid-2022 until early 2024, followed by a sharp upward movement in the last few months

💡 **Seasonality:**

- Repeating patterns occur at regular intervals
- This could represent recurring monthly or quarterly fluctuations in stock prices

💡 **Residuals (Noise):**

- The remaining random fluctuations after removing trend and seasonality
- These residuals reflect unpredictable, short-term price variations caused by external factors or market anomalies

Understanding these components enable more accurate stock price predictions

Let's try to forecast the Close Price of ADBL using SARIMA

- 💡 Taking into consideration the past historical records of **Close Price** only
- 💡 Predicting the **Close Price** of ADBL

Basic Forecasting Method (SARIMA)

💡 Since the data has seasonality, we use **SARIMA**

💡 SARIMA (Seasonal Autoregressive Integrated Moving Average)

SARIMA Model Parameters

1. **(p, d, q) Non Seasonal Components**
   - p: No. of past observations used (AR)
   - d: No. of times differenced to make data stationary
   - q: No. of past forecast errors used (MA)
2. **(P, D, Q, s) Seasonal Components**
   - P: No. of past seasonal observations used
   - D: No. of seasonal differencing applied
   - Q: No. of past seasonal forecast errors used
   - s: Length of the seasonal cycle

Stationarity Test

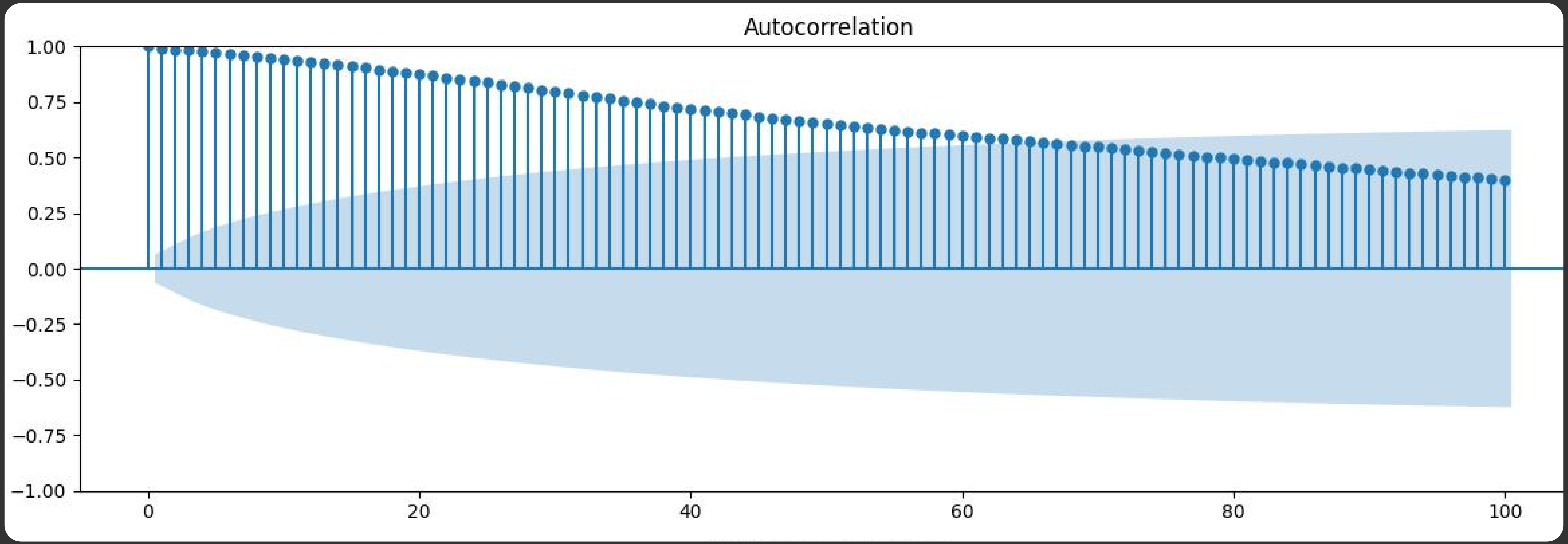
💡 Augmented Dickey-Fuller (ADF) test checks whether a time series is stationary or not by analysing if time series has a unit root (non stationary)

ADF Statistic	-2.16
p-value	0.22
Critical Values	1%: -3.661 5%: -2.961 10%: -2.619

Rejects Null Hypothesis of the ADF Test, The time series has a unit root (non-stationary)

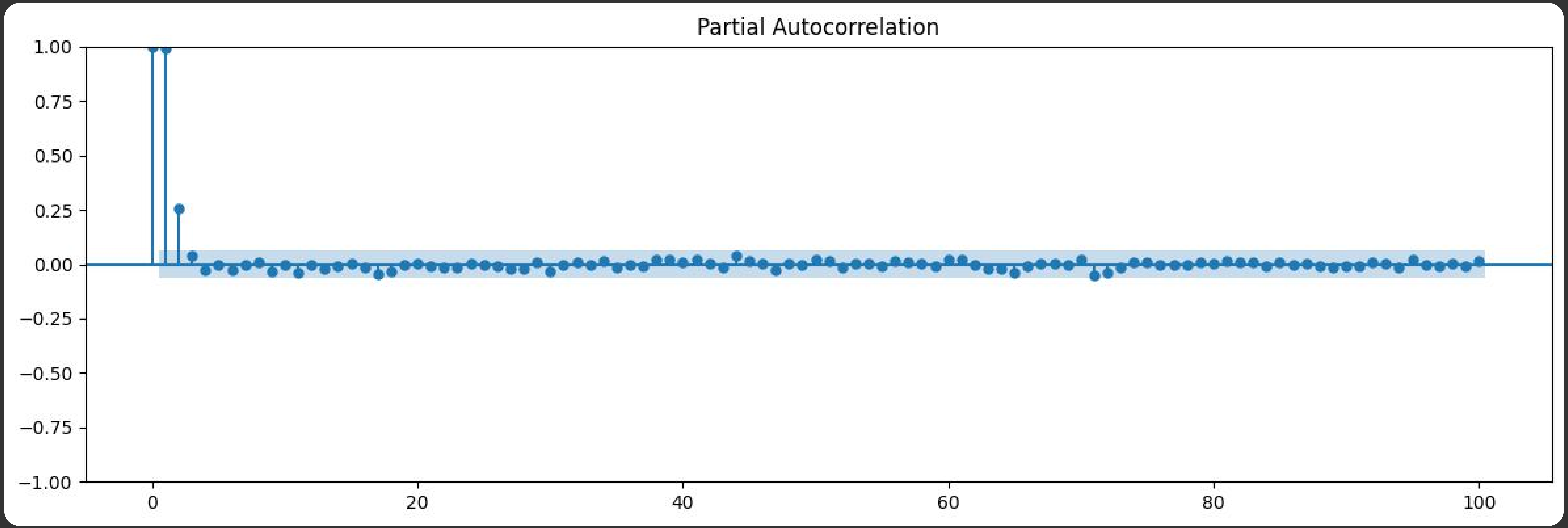
Autocorrelation and Partial Autocorrelation Analysis

💡 Plots for **ACF** and **PACF** with 100 lags, which help identify potential AR and MA terms for time series modeling



💡 The ACF plot shows a slow decay, suggesting non-stationarity, so d = 1

- Stock price trends are not stable over time



💡 The PACF plot shows a significant spike at lag 1, then it cuts off. This suggests p = 1

- **Recent stock price movements are highly influenced by immediate past values**

💡 The ACF plot shows a significant spike at lag 1 and possibly lag 2, then it tails off. This suggests q = 1 or 2

- **Recent market behaviour can strongly influence near-term forecasts**

Non Seasonal (p, d, q)

- p (AR order): 1 (from PACF chart, showing significant lag at 1)
- d (Differencing order): 1 (to make the data stationary)
- q (MA order): 1 (from ACF chart, showing significant lag at 1)

Seasonal (P, D, Q, s)

- P (Seasonal AR order): 1 (significant seasonal lags in PACF)
- D (Seasonal differencing): 1 (required for seasonal stationarity)
- Q (Seasonal MA order): 1 (from seasonal lags in ACF)
- s (Seasonal period): 12 (monthly data, 12 months per year)

Final Model

SARIMA(1, 1, 2)(1, 0, 1, 12)

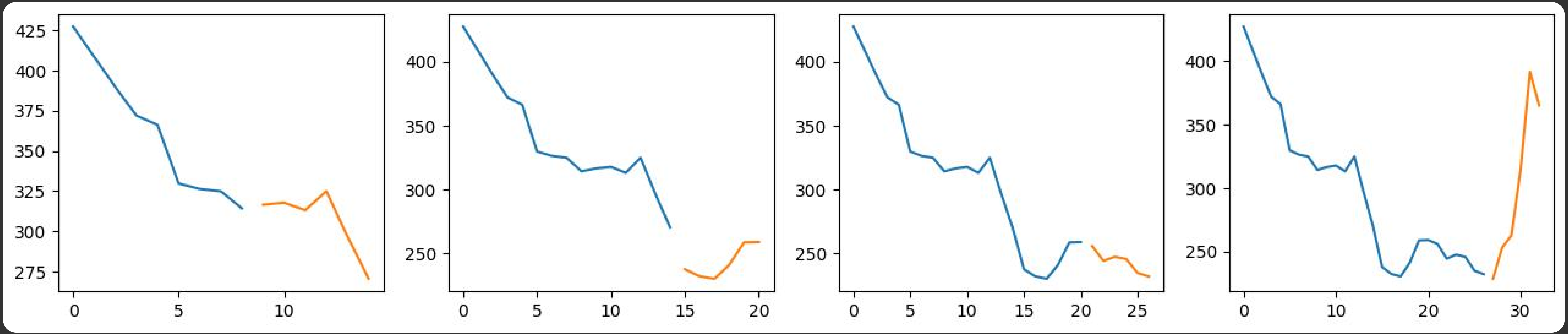
Hyperparameters Tuning

- 💡 Grid Search for Best Parameters
- 💡 The best model is determined by comparing the AIC (Akaike Information Criterion) values, with the model having the lowest AIC selected as the best

Model	AIC
SARIMA(1, 1, 2)(1, 0, 1, 12)	-12.495

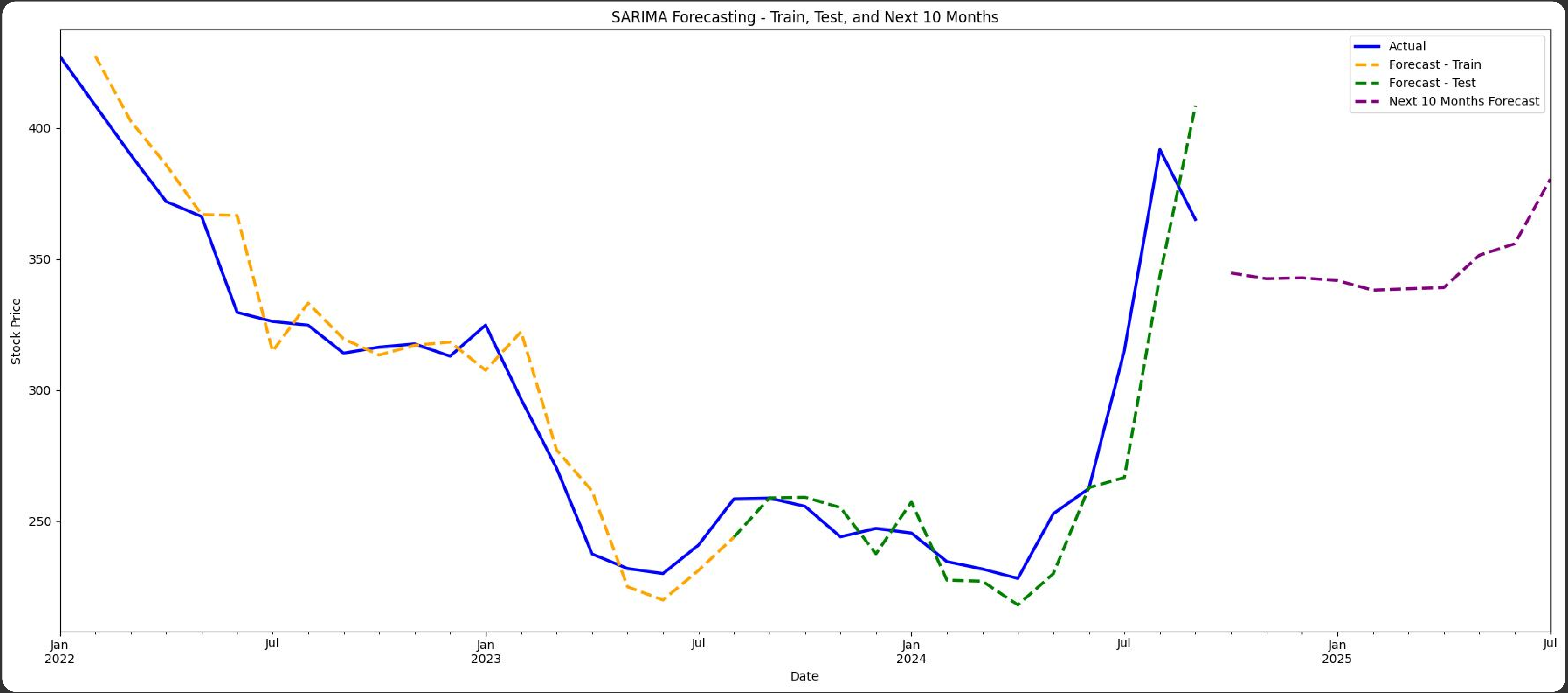
Back-Testing Training and Testing Data (Cross Validation)

- 💡 Creating multiple train-test splits keeping in mind the temporal order of our data during splits
- 💡 Each split uses a progressively larger portion of the data for training, while the remaining portion is used for testing



Train Test

SARIMA Forecasting



Next 7 Months ADBL Stock Close Price Prediction

Date	Close Price (NPR)
2024-10-31	344.73
2024-11-30	342.55
2024-12-31	342.88
2025-01-31	341.85
2025-02-28	338.19
2025-03-31	338.72
2025-04-30	339.16

Let’s try to forecast the Close Price of ADBL using LSTM

💡 Taking into consideration **Open**, **Close**, **High**, **Low** and **Vol** features of ADBL

We will do evaluation later on comparing the forecast accuracy between SARIMA and LSTM

Advanced Forecasting Method (LSTM)

LSTM Model Architecture

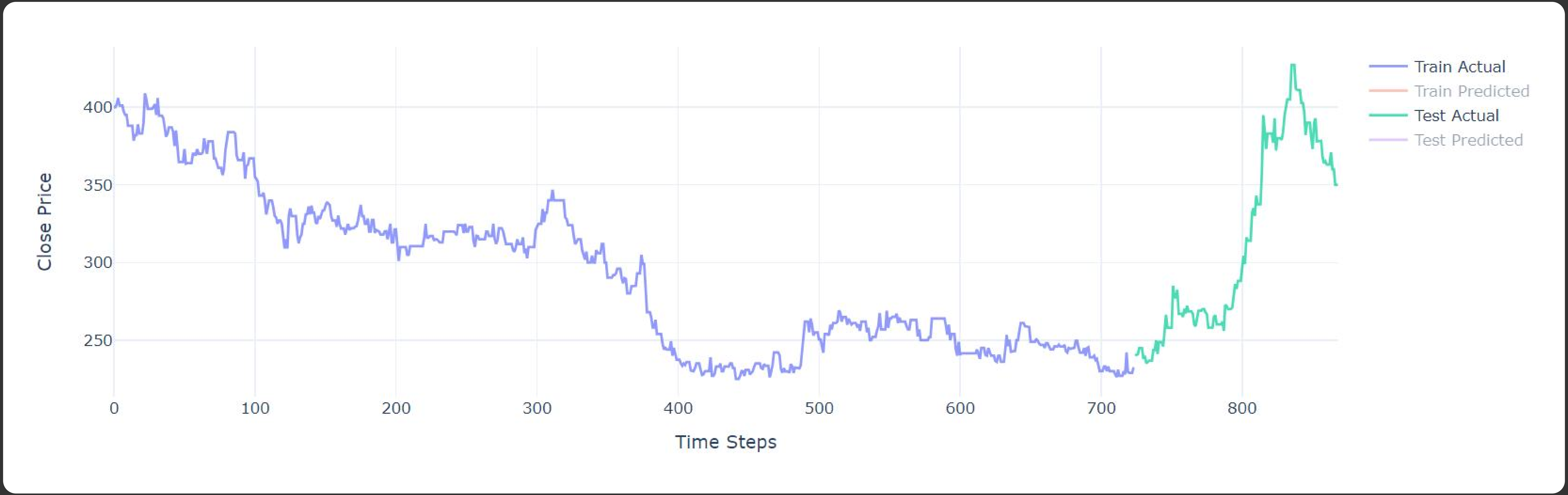
Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 50, 50)	11,200
dropout (Dropout)	(None, 50, 50)	0
lstm_1 (LSTM)	(None, 50, 50)	20,200
dropout_1 (Dropout)	(None, 50, 50)	0
lstm_2 (LSTM)	(None, 50)	20,200
dropout_2 (Dropout)	(None, 50)	0
dense (Dense)	(None, 5)	255

Construction of Input Sequences

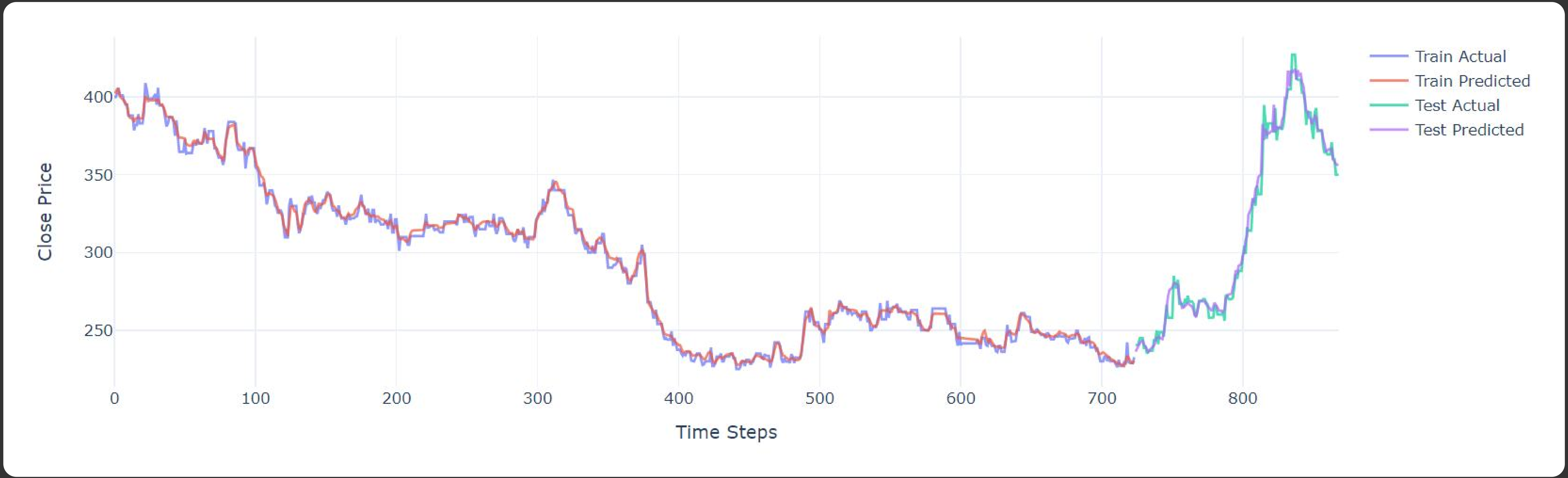
- 💡 Input sequences are constructed using a sliding window approach, where the model learns from the previous 50 time steps to predict the next value
- 💡 MinMaxScaler is used to normalize the data to a range of [0, 1], improving model performance

Train Test Split

💡 The dataset is split into approximately 80% for training and 20% for testing



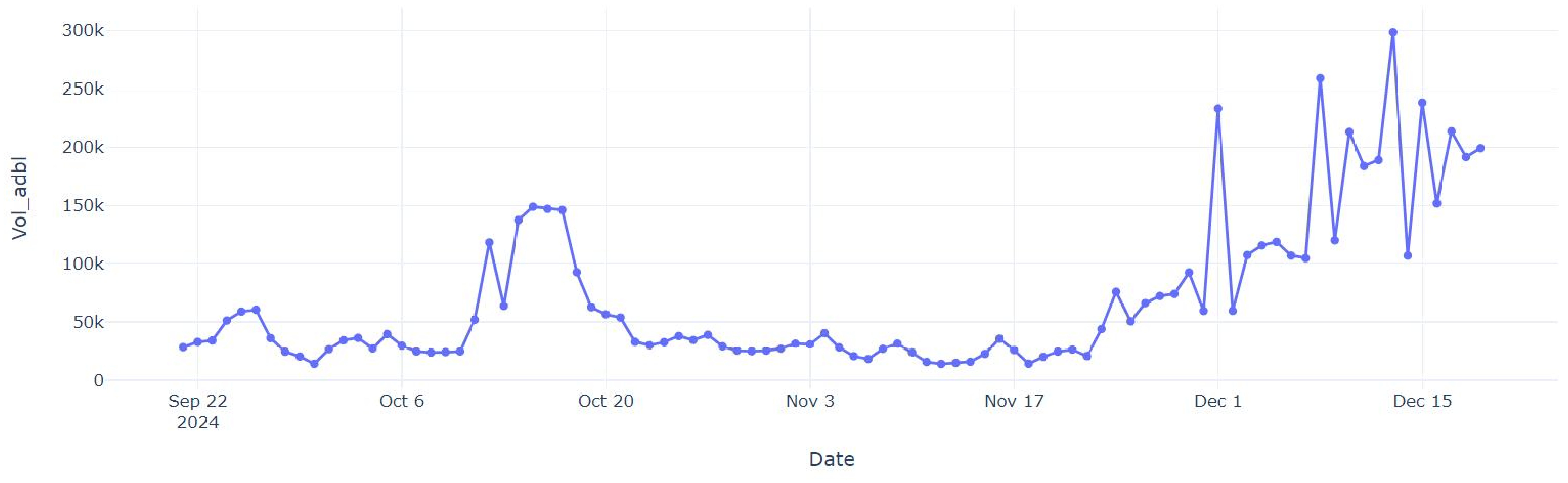
Train Test Split with Predictions



Last 50 Financial Days Prediction (LSTM)

Next 90 days forecast ADBL Close Price (LSTM)

Next 90 days forecast ADBL Share Volume Traded (LSTM)



Evaluation Metrics (Test Data)

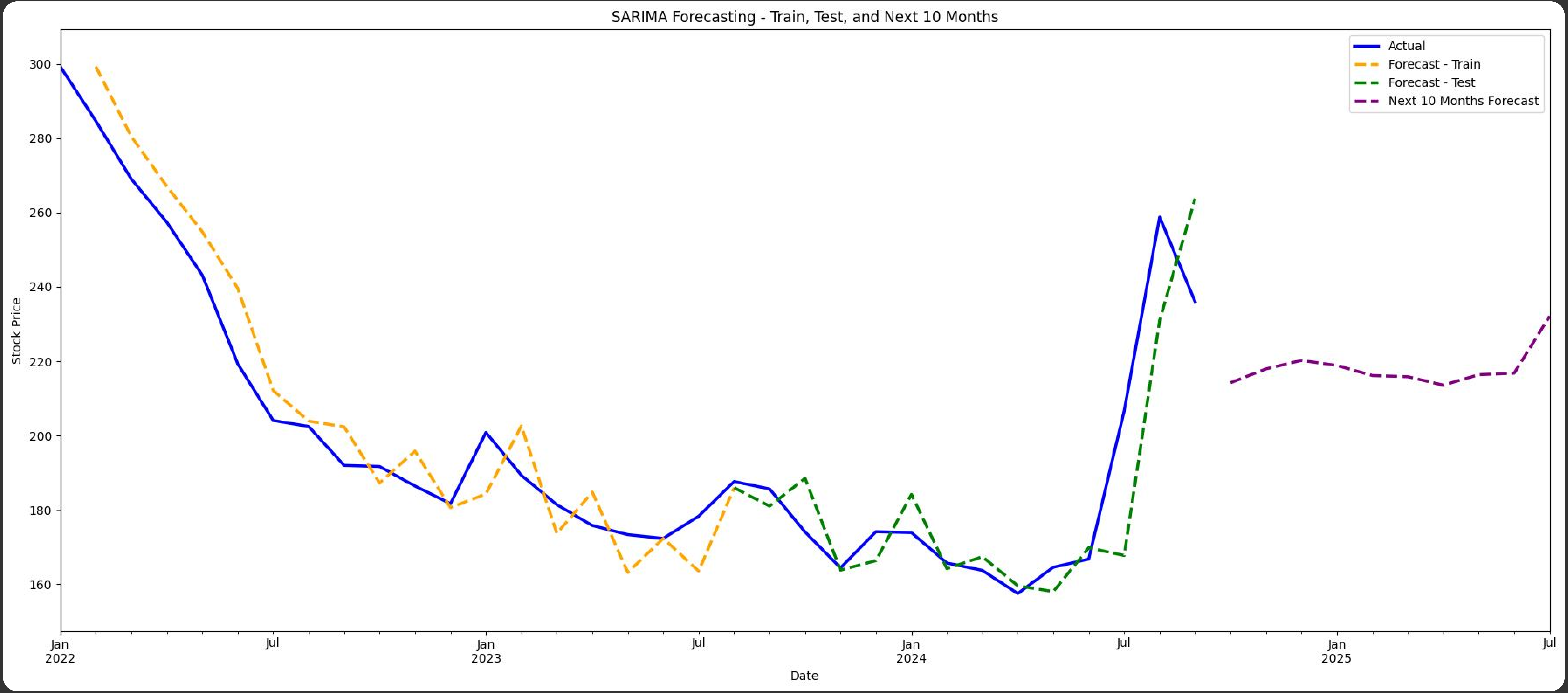
METRIC	SARIMA*	LSTM
MAE	41.91	2.45
MSE	3360.58	9.03
RMSE	53.28	3.00
MAPE	14.63%	0.85%

**In SARIMA we have done monthly resampling of daily data*

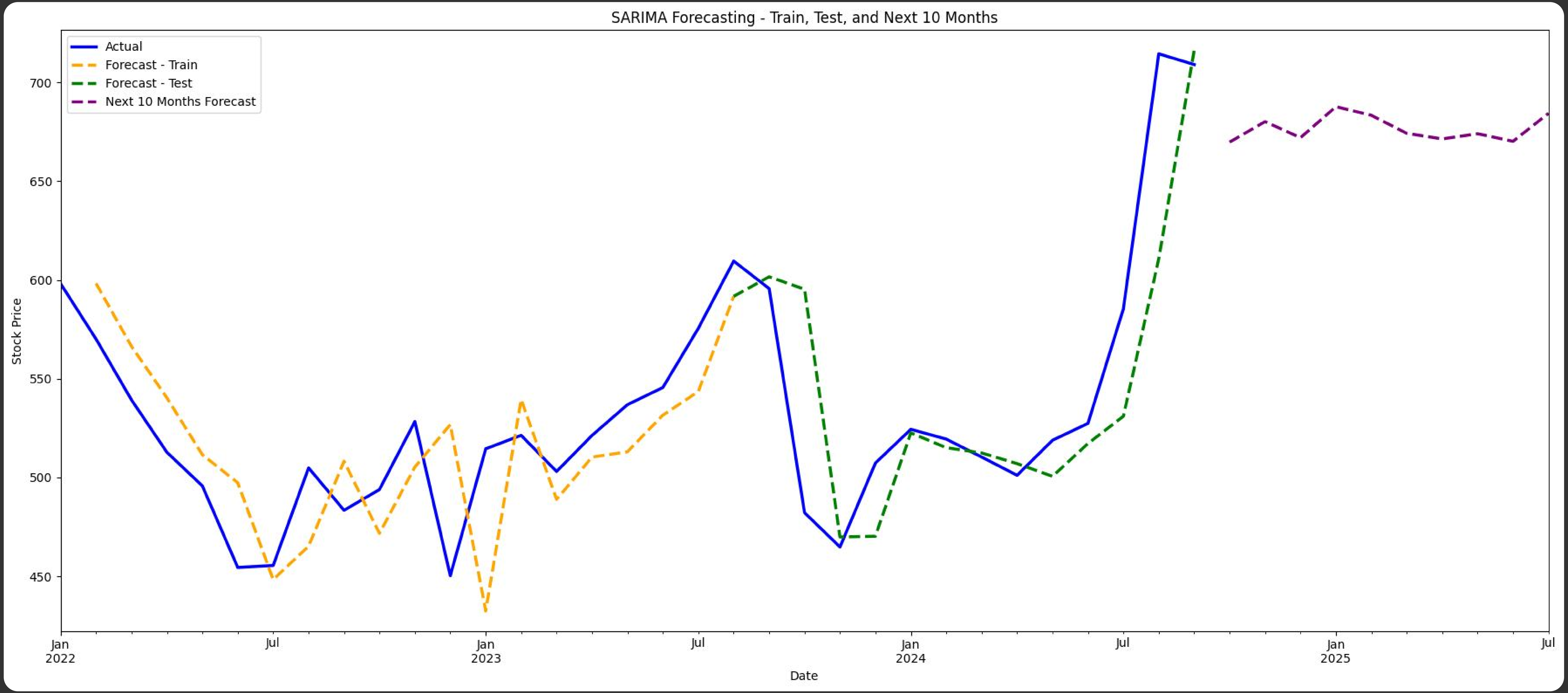
Similarly let’s forecast the close price of the stocks

- 💡 Citizens Bank International Limited (CZBIL)
- 💡 Everest Bank Limited (EBL)

CZBIL Stock’s Close Price



EBL Stock's Close Price



2024 Market Dynamics

- 💡 All selected stocks exhibit an upward trend in 2024, with **ADBL** showing the most significant stable price increases
- 💡 A sharp price spike occurs for all three stocks around mid-2024, followed by a slight decline
- 💡 Based on the brief analysis:
  - 🏆 **ADBL** is recommended to buy
  - ✨ It shows significant price increases in 2024 with stable movements, making it a solid choice
  - ✨ Consider purchasing around 250-300 shares to balance potential returns and risk

Future Plan

- 💡 Extending the dataset to a **5-year** period
- 💡 Incorporating new features, such as **stock sentiment** derived from news articles
    - **We can leverage the stock sentiment feature to enhance the analysis**

Thank You!