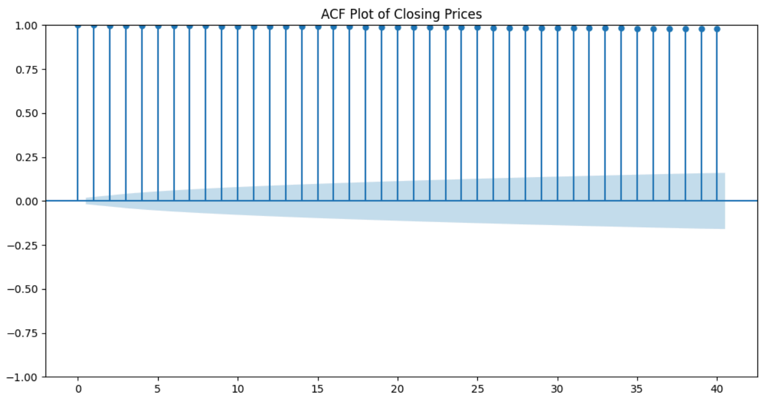
Task 04 : EDA Report

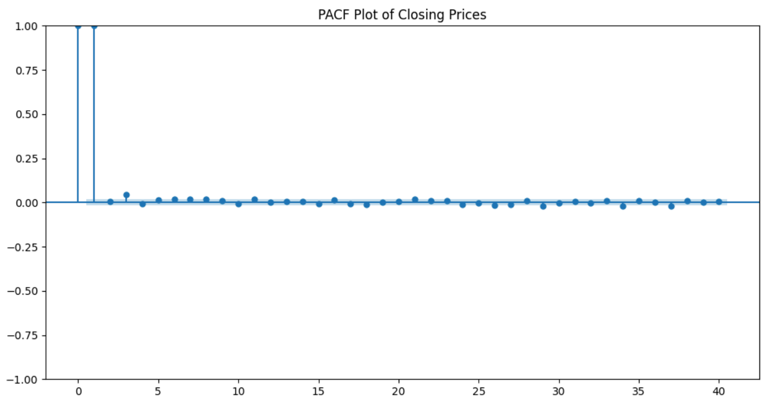
EDA for the given problem was done under 4 main sections

- **Time series analysis**
- **Distribution analysis**
- **Correlation analysis**
- **Technical Indicators**

Time series analysis

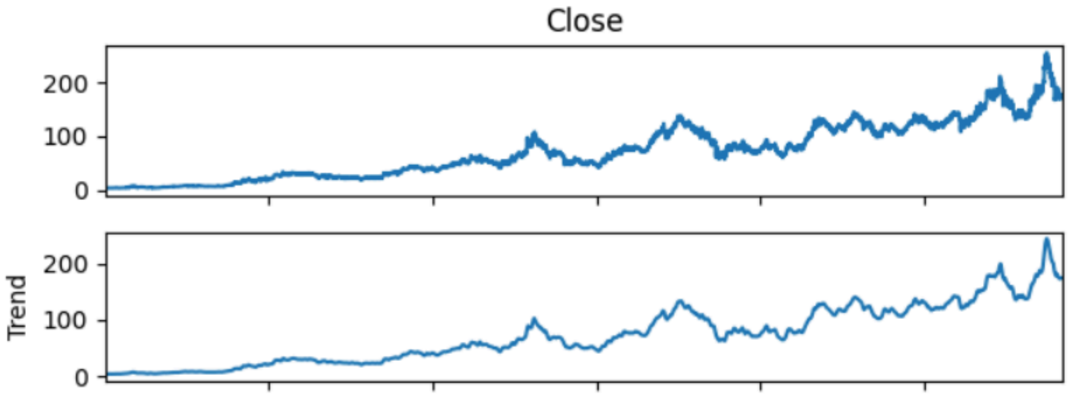
Auto correlation and partial autocorrelation plots were drawn to check for seasonality. We can infer that the model lacks clear stationarity from the look of those plots.





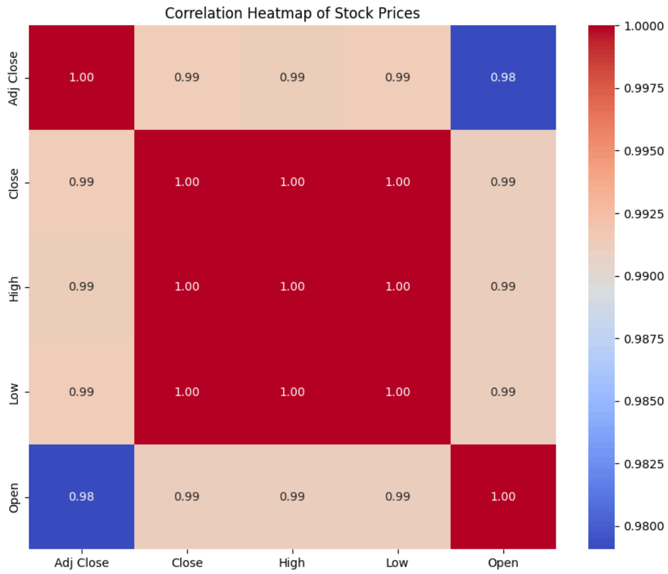
Also we did the ADF (Augemnted Dicky Fuller test too to confirm our findings)
Failed to reject null hypotheses

```
ADF Statistic: -0.837853
p-value: 0.807749
Critical Values:
        1%: -3.431
        5%: -2.862
        10%: -2.567
```



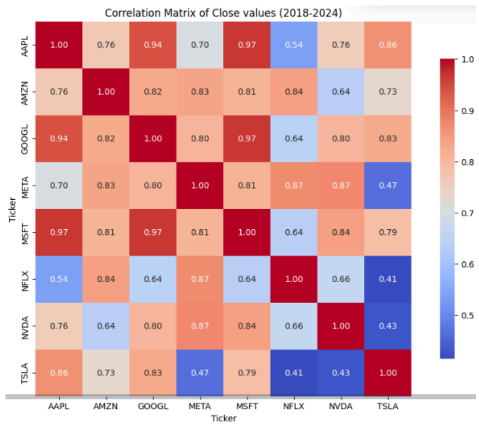
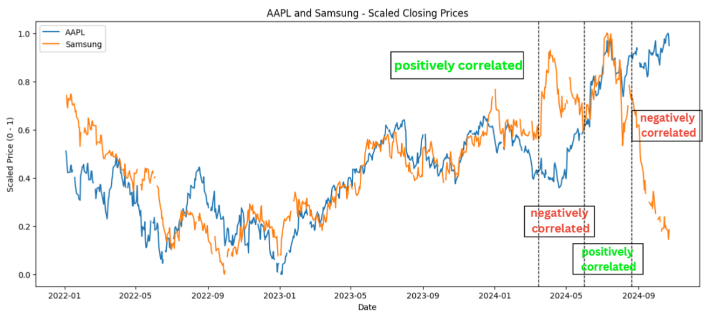
From Seasonal Trend decomposition it was clear that this does not have a clear seasonal trend, Insted it has a additive trend.

Correlation analysis



All of the given features of the stock price dataset are highly correlated with each other. Which makes them not that useful when used independently.

Additionally we've conducted more tests with external stock market data

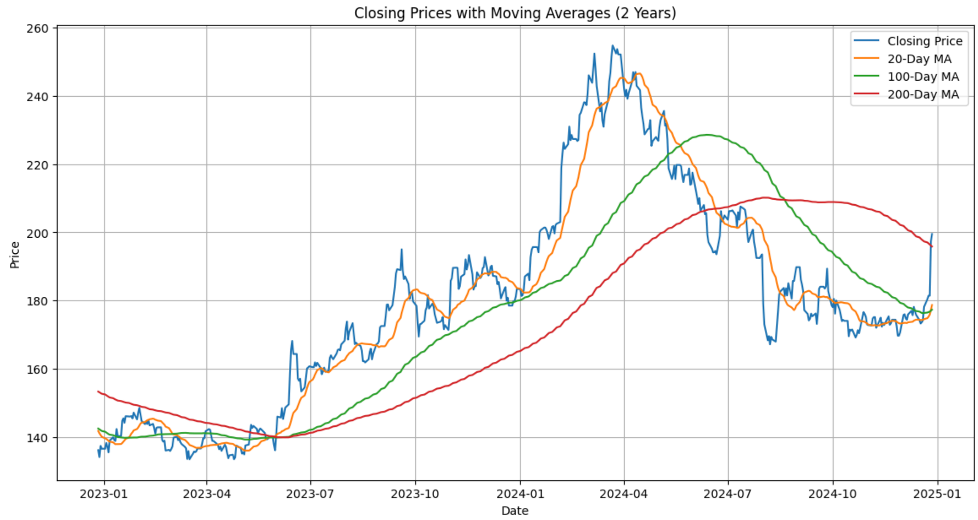
<table>
<tr>
<td>When examined there was no clear correlation between external stocks between each other.</td>
<td>
</td>
</tr>
<tr>
<td>Those correlations are complex. That means they change over time.</td>
<td>
</td>
</tr>
</table>

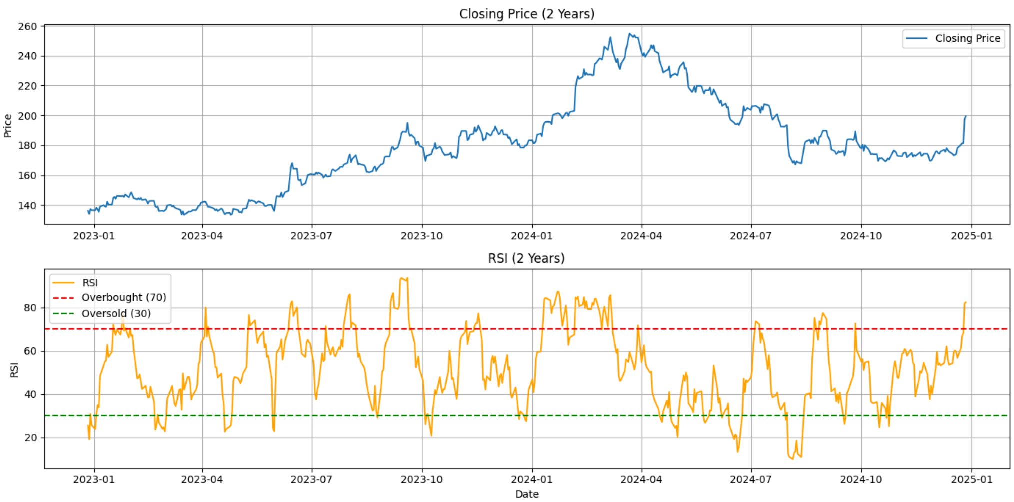
So we concluded that external stock prices will not directly correlate with our test set

Technical Indicators

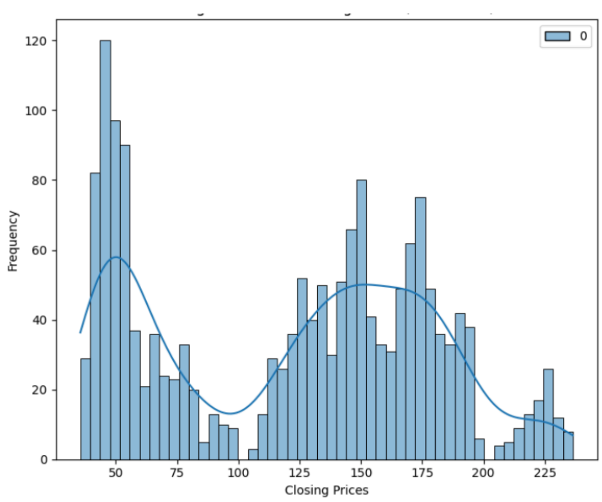
When it comes to the stock market we can not forget about the importance of technical indicators. We also tried incorporating some technical indicators related to stock market for our analysis

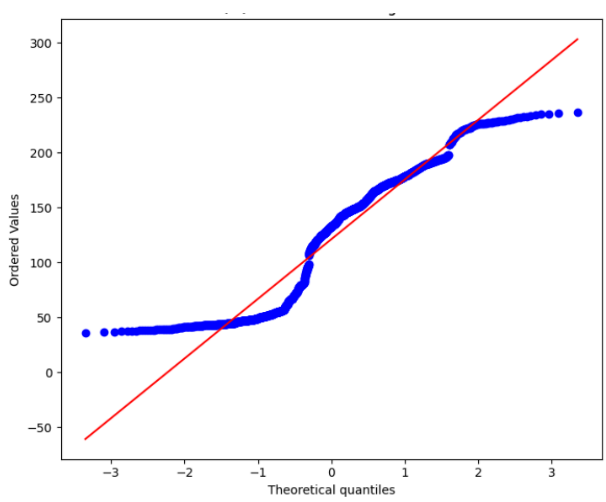
- Exponential Moving Average (EMA)
  - EMA is a type of moving average that gives more weight to recent prices, making it more responsive to new information compared to a simple moving average.
- Relative Strength Index (RSI)
  - RSI is a momentum oscillator that measures the speed and change of price movements, ranging from 0 to 100.





Distribution Analysis





Shapiro-Wilk Test

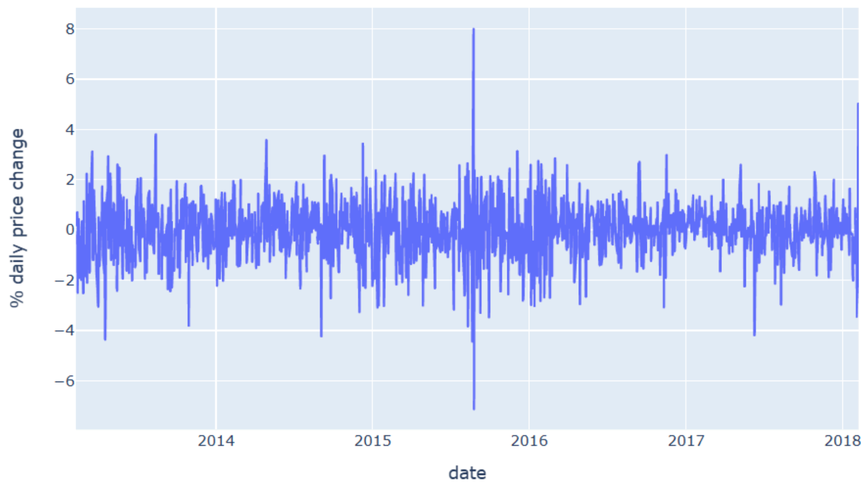
- Statistic: 0.9488
- p-value: $5.51 \times 10{-}24$

$p<0.05$ - returns do not follow a normal distribution

Kolmogorov-Smirnov Test

- Statistic: 0.0849
- p-value: $3.44 \times 10{-}11$

$p<0.05$ - returns do not follow a normal distribution

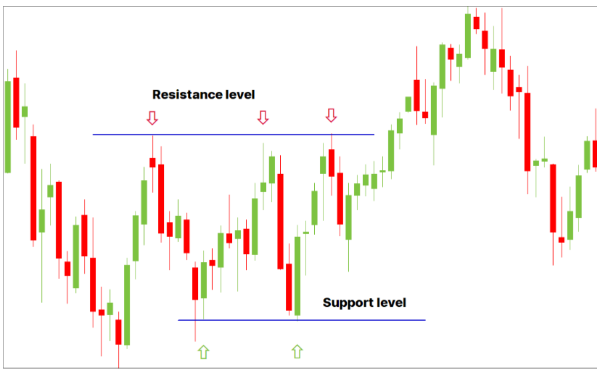


Daily price change is noisy and does not have a clear pattern

One of the problems we had was can we predict the trend?

- Daily Trend Prediction is Challenging
  - Market volatility and external factors make next-day movement forecasting difficult.

- Patterns Provide Predictability
  - Recognizable patterns in historical data, like head-and-shoulders or support/resistance, improve model prediction accuracy.



Problems discovered from the Analysis

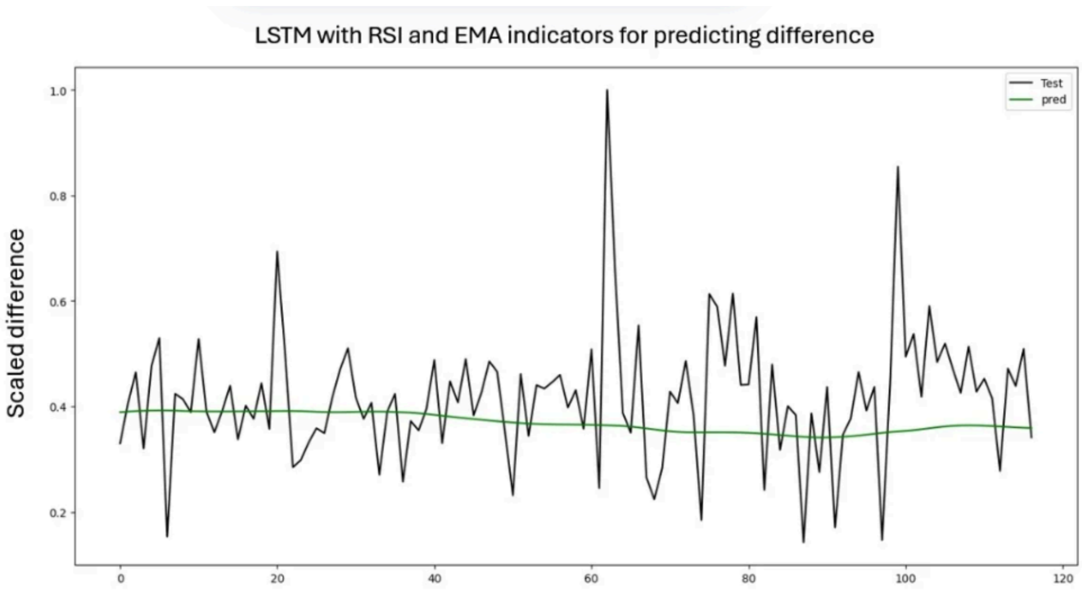
What to predict in the stock market?

1. Stock Closing Prices
   - Forecast the exact closing price of a stock on a future date.

2. Stock Trends / Difference
   - Predict the direction (upward or downward) or momentum over a time period (trend analysis)

Predicting Stock Trends / Difference

- Predict the Difference Between Tomorrow's and Today's Closing Values.
- Predict the Direction of Price Change (Up or Down)



As shown in the above plot it was not easy to predict differences even by using the best models.

Data Preprocessing and Feature Selection

Handling Missing values

```
# Check for missing values
df.isnull().sum()
```

	0
Unnamed: 0	0
Date	110
Adj Close	93
Close	117
High	95
Low	127
Open	103
Volume	145

Since we have rows where the date is also null, we cannot proceed further with those.

- The stock market does not operate every day (e.g., weekends, holidays).
- Missing values should be handled while considering only valid trading days.

Strategy for missing values

- Convert the Date column to datetime.
- Identify the full range of trading dates (excluding weekends and holidays).
- Fill missing dates and forward-fill their corresponding values.

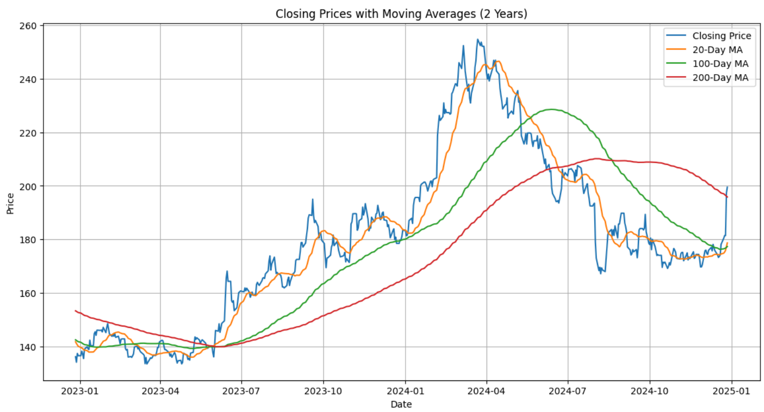
Filling Other Missing Values

- Forward Fill (ffill) for price-related columns (Adj Close, Close, High, Low, Open) since stock prices don't jump arbitrarily.
- Interpolate for Volume, as it's a fluctuating metric and forward-fill is not the best choice.

With this technique we were able to get a continuous time series dataset with no gaps

Feature Engineering

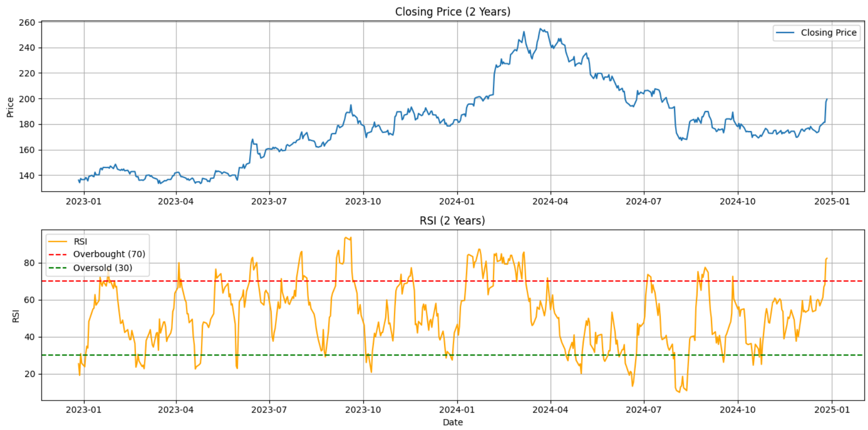
From our analysis it was clear that some of the technical indicators used in the stock market can also be used as good features for our models.



So we decided to derive multiple statistical features and use them in our experimentations.

Lag Features	Close_Lag_1, Close_Lag_5	Used to predict future price movement based on recent historical values. Helps capture autoregressive behavior in stock prices.
Moving Averages	MA_5, MA_10, MA_20	Short-term averages (e.g., MA_5) help in detecting quick trends. Long-term averages (e.g., MA_20) help smooth noise and indicate overall market direction.

Volatility	Volatility_5, Volatility_10	High volatility means higher uncertainty and potential sharp price swings. Helps in risk assessment and predicting large price movements.
Momentum	Momentum_5	A positive value suggests an upward trend, while a negative value indicates a downward trend. Helps traders identify breakout points where a stock might start rising or falling.



Also we tried utilizing indicators like RSI

- RSI is a momentum oscillator that measures the speed and change of price movements, ranging from 0 to 100.
- **Typically, values above 70 indicate overbought conditions, while values below 30 suggest oversold conditions**.