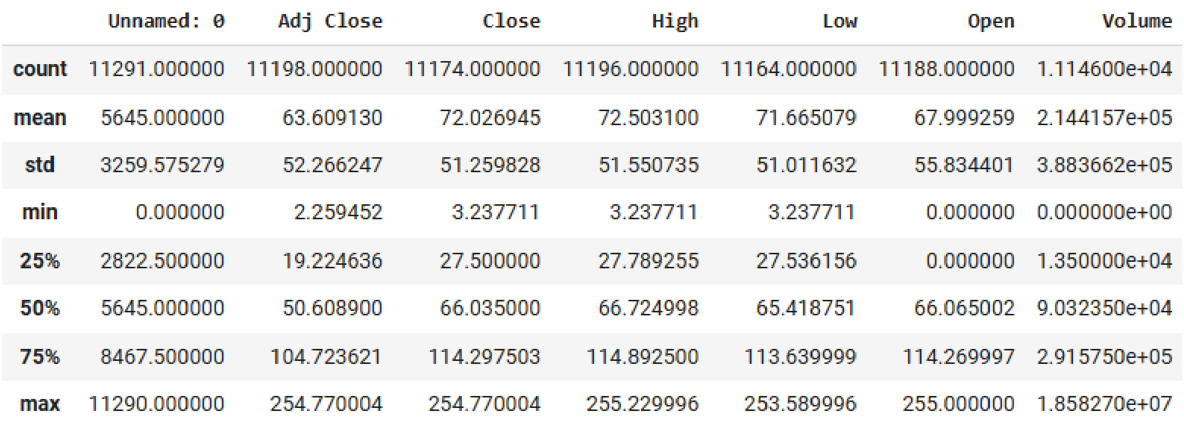
# *Task 04 - Stock Price Prediction Challenge*
# *Team Duo Dynamics*

## *Part 1: Stock Price Prediction Model*



| | Unnamed: 0 | Adj Close | Close | High | Low | Open | Volume |
|---|---|---|---|---|---|---|---|
| **count** | 11291.000000 | 11198.000000 | 11174.000000 | 11196.000000 | 11164.000000 | 11188.000000 | 1.114600e+04 |
| **mean** | 5645.000000 | 63.609130 | 72.026945 | 72.503100 | 71.665079 | 67.999259 | 2.144157e+05 |
| **std** | 3259.575279 | 52.266247 | 51.259828 | 51.550735 | 51.011632 | 55.834401 | 3.883662e+05 |
| **min** | 0.000000 | 2.259452 | 3.237711 | 3.237711 | 3.237711 | 0.000000 | 0.000000e+00 |
| **25%** | 2822.500000 | 19.224636 | 27.500000 | 27.789255 | 27.536156 | 0.000000 | 1.350000e+04 |
| **50%** | 5645.000000 | 50.608900 | 66.035000 | 66.724998 | 65.418751 | 66.065002 | 9.032350e+04 |
| **75%** | 8467.500000 | 104.723621 | 114.297503 | 114.892500 | 113.639999 | 114.269997 | 2.915750e+05 |
| **max** | 11290.000000 | 254.770004 | 254.770004 | 255.229996 | 253.589996 | 255.000000 | 1.858270e+07 |

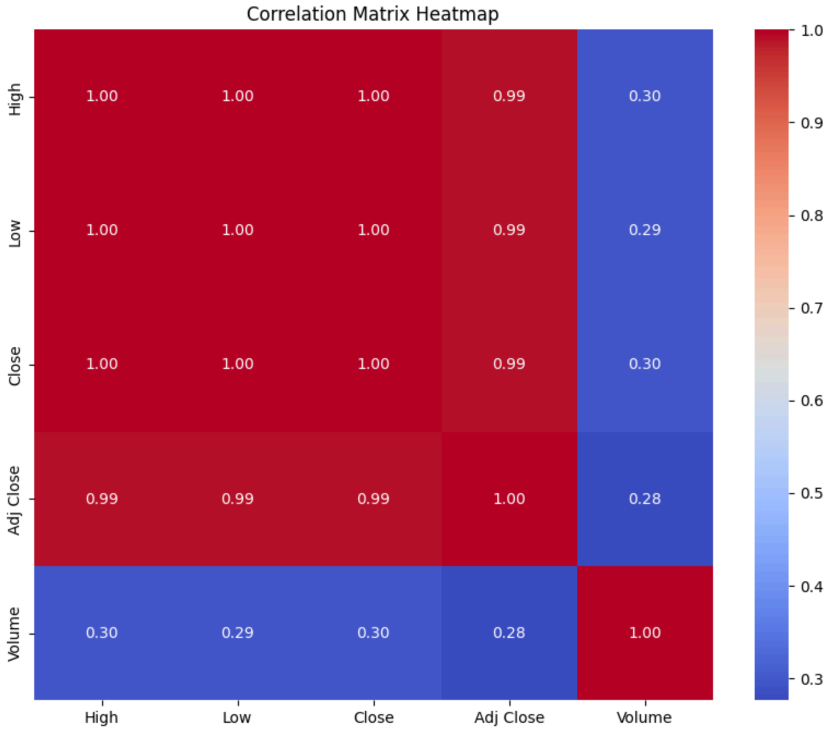
### ❖ *Handling missing values*

- Date (110): Cannot be imputed - Remove rows with missing dates if they are completely empty.
- Adj Close (93): Forward Fill - Stock prices are time-dependent; the last known value is a reasonable estimate.
- Close (117): Forward Fill  - Similar to Adj Close, forward fill works well.
- High (95): Forward Fill  - Since high values follow a trend, forward fill is better.
- Low (127): Forward Fill  -  Use forward fill as low prices change based on market trends.
- Open (103): Backward Fill - Opening prices often depend on recent trends but may use future values if missing.
- Volume (145): Median Imputation - Trading volume can be highly volatile; the median is more robust to outliers.

### ❖ *Correlation Analysis*

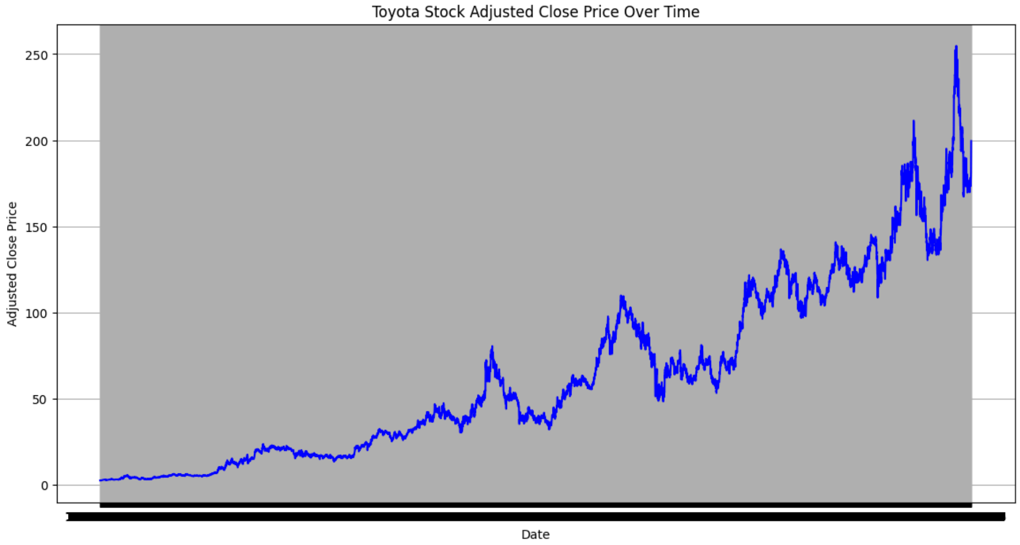
- "High," "Low," "Close," and "Adj Close" have very strong positive correlations (close to 1.00) with each other. This is expected, as these values are all related to the stock's price. When the high is high, the low, close, and adjusted close are likely to be high as well.
- "Volume" has a weak positive correlation (around 0.30) with the price metrics. This means that while there might be a slight tendency for volume to increase when prices are high, the relationship is not very strong.

- There are no strong negative correlations in the heatmap, which means that none of the variables move in opposite directions to a significant degree.



## *Exploratory Data Analysis ( EDA)*

### ★ *Time-Series Plot of Adjusted Close Price*

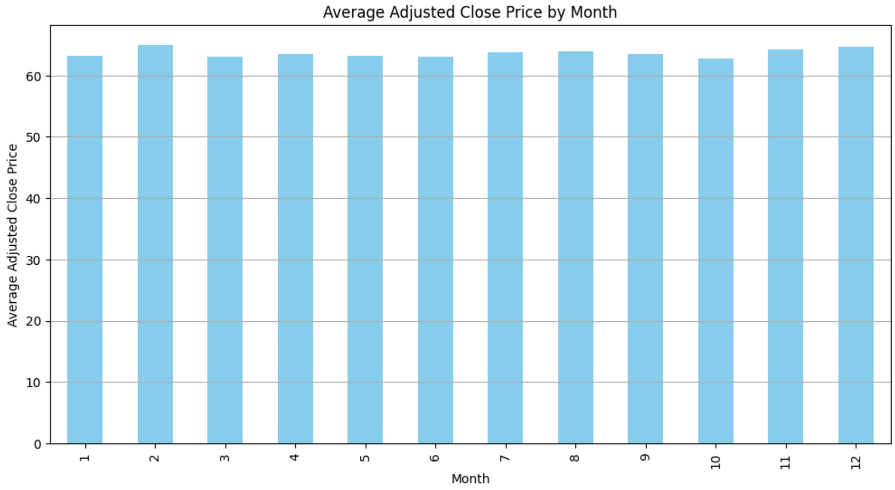


- The most striking feature is the clear long-term upward trend in the adjusted close price. This indicates that the stock has generally appreciated over the observed time period.

- While the trend is upward, there are significant fluctuations and periods of volatility. This suggests that the stock price is influenced by various factors, potentially including market conditions, company performance, and industry trends.

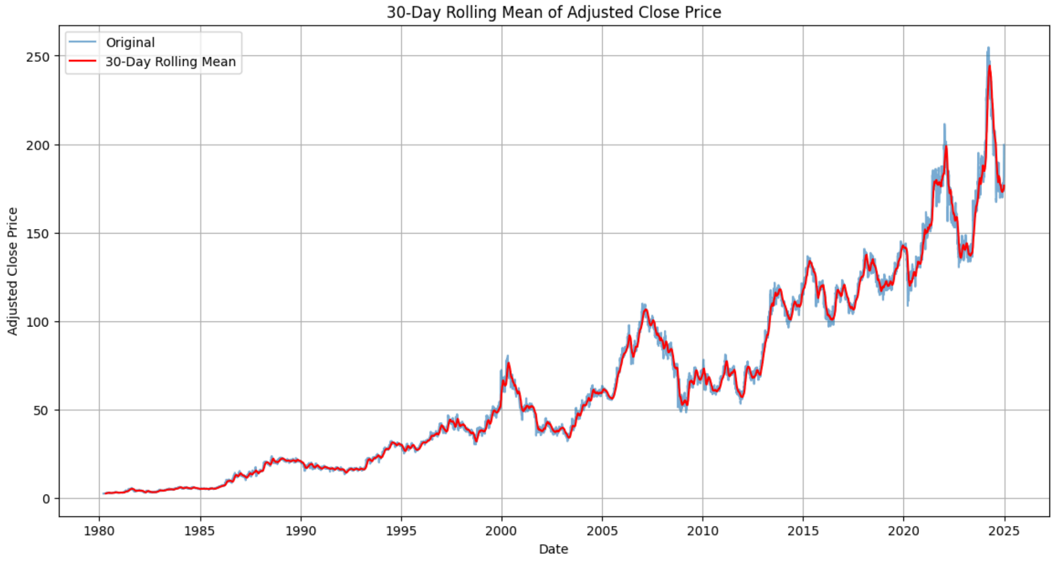
- We can potentially identify different phases in the stock's performance:
  1. Early Growth (Left): A gradual, relatively stable increase in price.
  2. Acceleration (Middle): A period of more rapid price appreciation.
  3. High Volatility (Right): Increased fluctuations and potential uncertainty

## ★ Monthly Average Prices



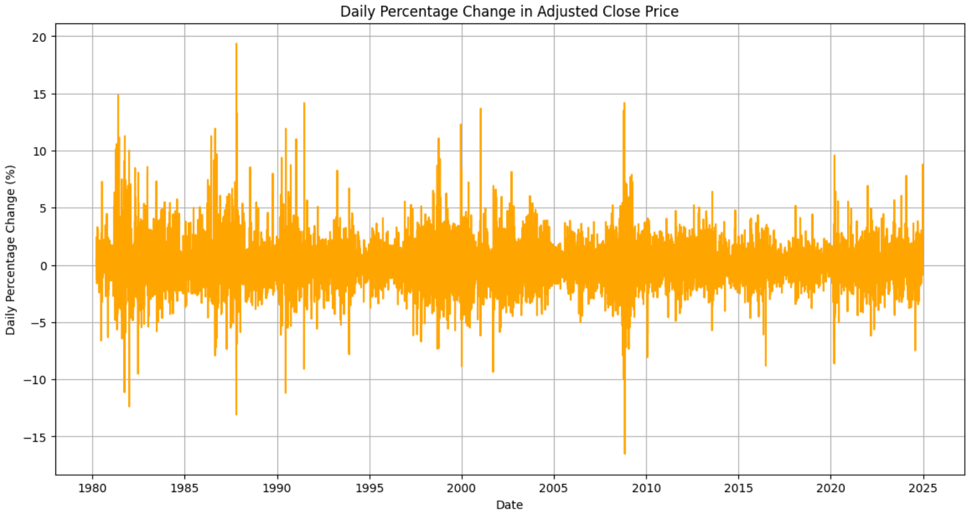
- Overall, the average adjusted close price appears relatively stable across the 12 months. There are no dramatic spikes or dips.
- While the averages are relatively close, there are subtle variations that suggest some degree of seasonality or monthly patterns.

## ★ Rolling Mean Analysis



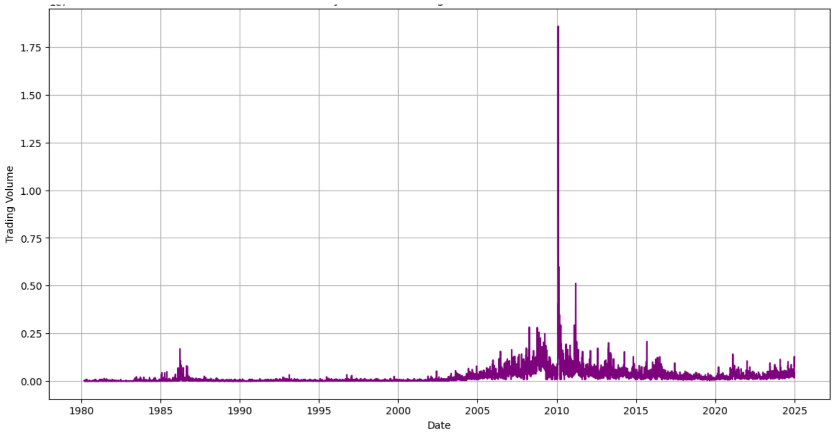
- A rolling mean, also known as a moving average, is a statistical technique used to analyze time-series data by creating a series of averages of different subsets of the full data set. It essentially smooths out short-term fluctuations and highlights longer-term trends or cycles.
- The red line representing the 30-day rolling mean clearly shows the overall upward trend in the adjusted close price. It smooths out the daily fluctuations (represented by the blue line) and makes the trend more apparent.
- The rolling mean line is smoother than the original price line, demonstrating its ability to reduce noise and highlight the underlying signal.
- The rolling mean confirms the long-term upward trend we observed in previous visualizations.

## ★ Daily Percentage Change



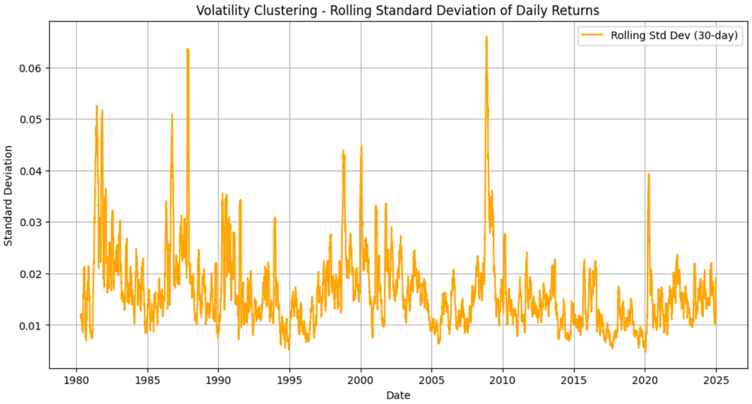
- This plot shows the percentage change in the adjusted closing price from one day to the next. A positive value indicates an increase in price, while a negative value indicates a decrease.
- The graph provides a clear visual representation of the stock's volatility. The wider the swings in the percentage change, the more volatile the stock is.

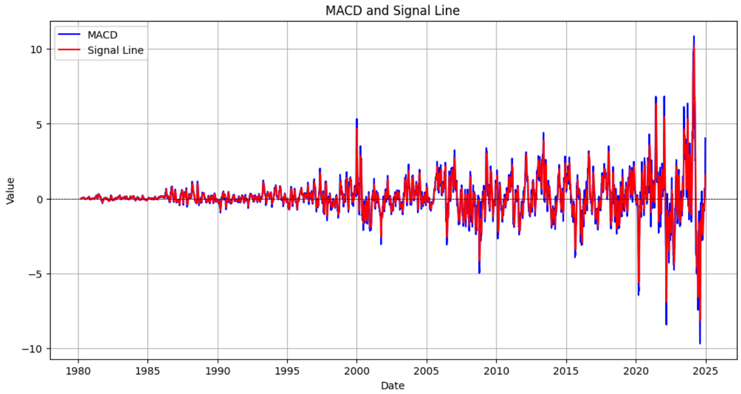
## ★ Volume Trends Analysis

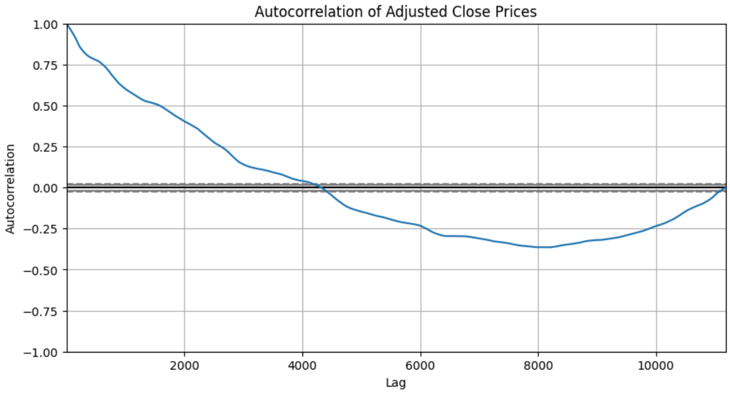


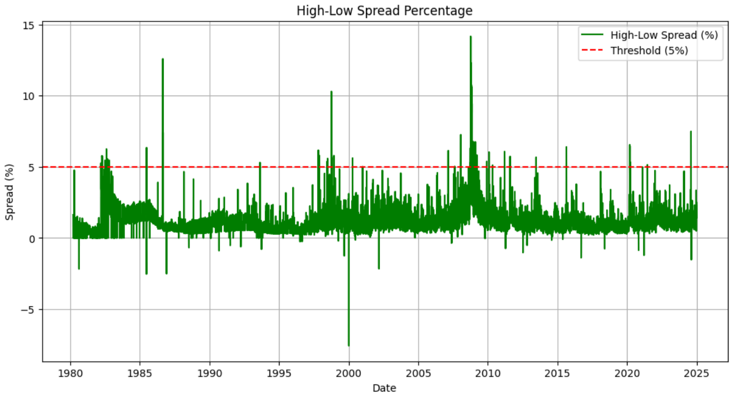
- For a large portion of the time period, the trading volume is relatively low. This suggests that the stock was not heavily traded during those years.
- There are several distinct spikes in volume, indicating periods of significantly higher trading activity. These spikes are particularly noticeable around 2008-2010 and in the mid-2010s.

### ★ Detecting Anomalous Trading Volumes









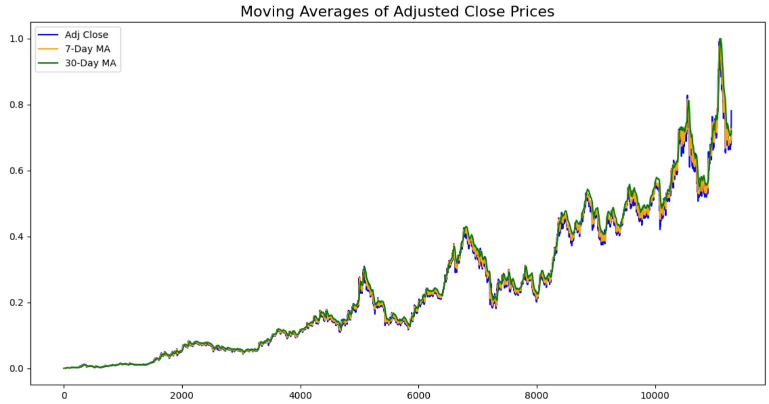
- Volatility Clustering: The analysis of rolling standard deviation reveals distinct periods of high and low volatility in stock, confirming the phenomenon of volatility clustering. These fluctuations, particularly prominent around major market events, highlight periods of increased risk and underscore the importance of adjusting trading strategies based on volatility levels.

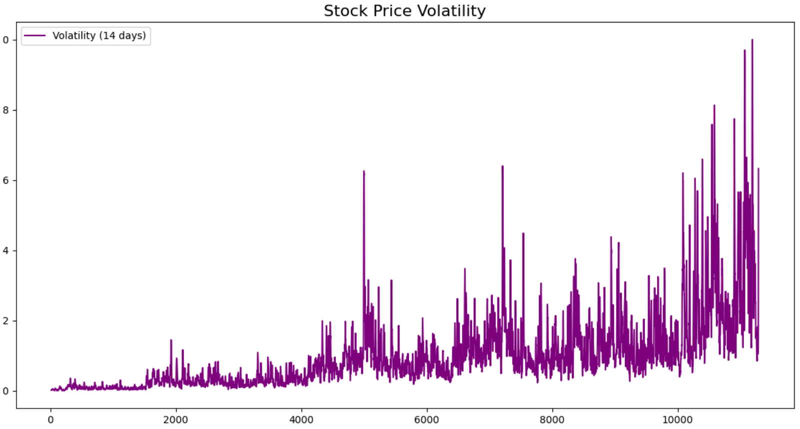
- MACD Momentum: The MACD indicator showcases frequent crossovers between the MACD and signal lines, suggesting dynamic momentum changes in stock. These crossovers provide potential buy/sell signals and hint at possible trend reversals, offering valuable insights for technical analysis and trading decisions.

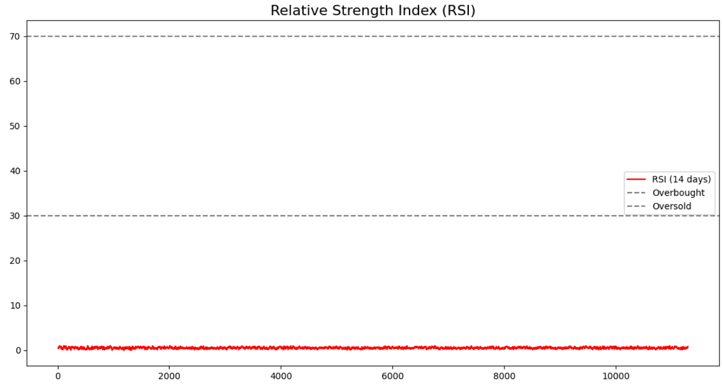
- Autocorrelation: The autocorrelation analysis demonstrates a strong short-term dependence in adjusted close prices, indicating that recent prices heavily influence current prices. This suggests trend persistence in the short term, while the slight negative autocorrelation at higher lags hints at potential long-term reversals, crucial information for time series modeling and forecasting.

- Price Spread and Volume: Examining the high-low price spread percentage reveals days with unusually high intraday volatility, often linked to significant market events. These spikes can signal trading opportunities or risks. Additionally, applying the z-score method to trading volume data helps detect anomalous trading activity, which could indicate unusual market behavior or data anomalies, crucial for data validation and identifying potential market anomalies.

## ❖ *Features Created*







1. Moving Averages of Adjusted Close Prices- This plot shows the adjusted close price along with its moving averages (MA) calculated over different periods (1-day, 5-day, and 10-day).

Moving averages smooth out short-term price fluctuations, making it easier to identify underlying trends.They help in identifying the direction and strength of the price trend.They can act as potential support or resistance levels.

  - ○ The 1-day MA (which is essentially the original adjusted close) shows the most volatility.
  - ○ The 5-day and 10-day MAs smooth out the noise, revealing the overall trend more clearly.
  - ○ Crossovers between the MAs and the price or between different MAs can be used as potential trading signals.

2. Stock Price Volatility - This plot shows the volatility of the stock price, likely calculated as the standard deviation of returns over a certain period (14 days as mentioned in the legend).

Volatility is a measure of risk. Higher volatility indicates greater price fluctuations and therefore higher risk. It helps in identifying periods of increased or decreased volatility.

  - ○ The plot shows periods of high and low volatility, indicating that the stock's volatility is not constant.
  - ○ Spikes in volatility often coincide with market events or news.
  - ○ Understanding volatility can help investors adjust their trading strategies and manage risk.

3. Relative Strength Index (RSI) - This plot shows the Relative Strength Index (RSI), a momentum indicator that measures the speed and change of price movements.

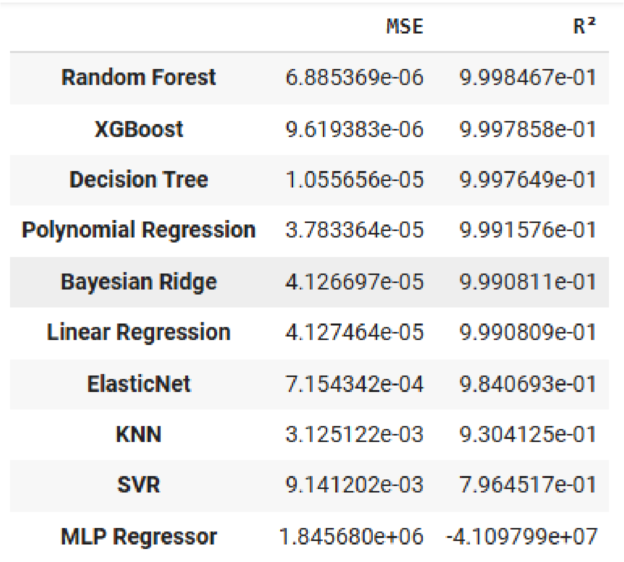
It helps in identifying overbought (RSI above 70) or oversold (RSI below 30) conditions.It indicates the strength of the price momentum. It can signal potential price reversals.

  - ○ In the provided image, the RSI is consistently very low. This suggests that the stock might be in an oversold condition for the entire observed period.
  - ○ However, with such consistently low values, it might also indicate an error in the calculation or a need to re-evaluate the parameters used for calculating the RSI.

### ❖ ***Model Selection***

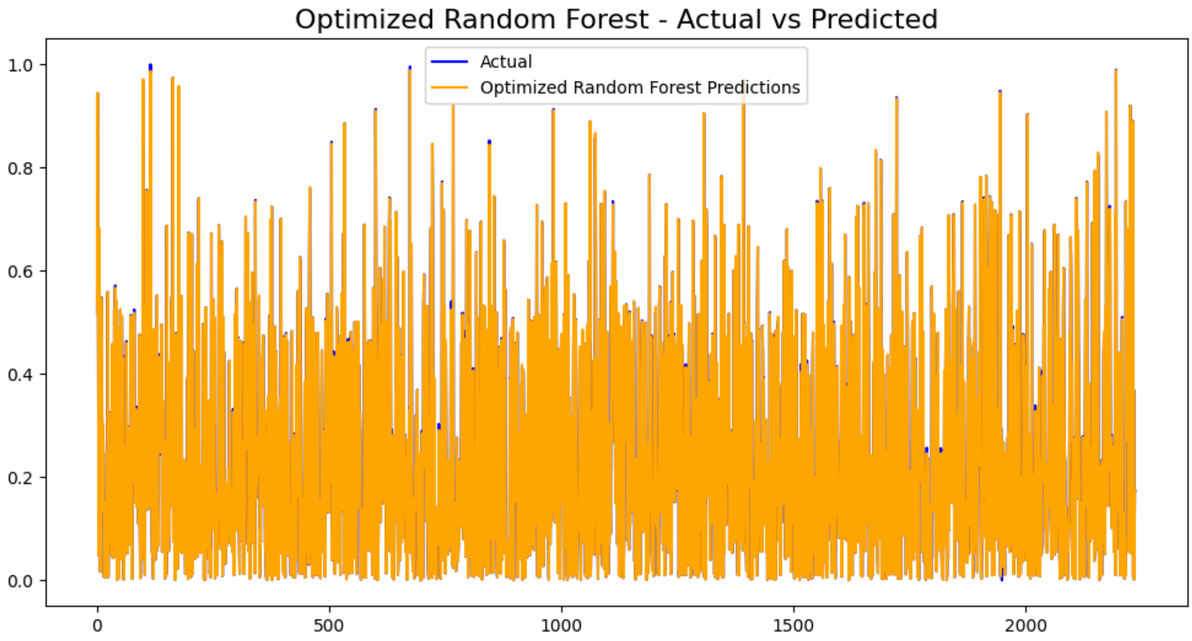
Metrics used

- Mean Squared Error (MSE): This measures the average squared difference between the predicted and actual values. Lower MSE indicates better accuracy.
- R-squared (R^2): This represents the proportion of variance in the dependent variable that is explained by the independent variables. Higher R^2 indicates a better fit.



| | MSE | R² |
|---|---|---|
| **Random Forest** | 6.885369e-06 | 9.998467e-01 |
| **XGBoost** | 9.619383e-06 | 9.997858e-01 |
| **Decision Tree** | 1.055656e-05 | 9.997649e-01 |
| **Polynomial Regression** | 3.783364e-05 | 9.991576e-01 |
| **Bayesian Ridge** | 4.126697e-05 | 9.990811e-01 |
| **Linear Regression** | 4.127464e-05 | 9.990809e-01 |
| **ElasticNet** | 7.154342e-04 | 9.840693e-01 |
| **KNN** | 3.125122e-03 | 9.304125e-01 |
| **SVR** | 9.141202e-03 | 7.964517e-01 |
| **MLP Regressor** | 1.845680e+06 | -4.109799e+07 |

After training multiple models and evaluation there score we realized that random forests is the best model as it has the lowest MSE among them.
And then we used grid search with cross validation to optimized the RF
Model.

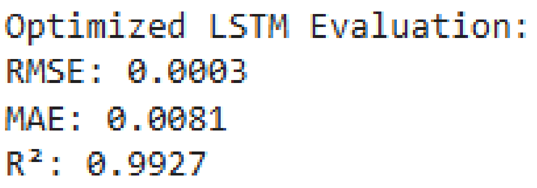


Optimized Random Forest - Actual vs Predicted

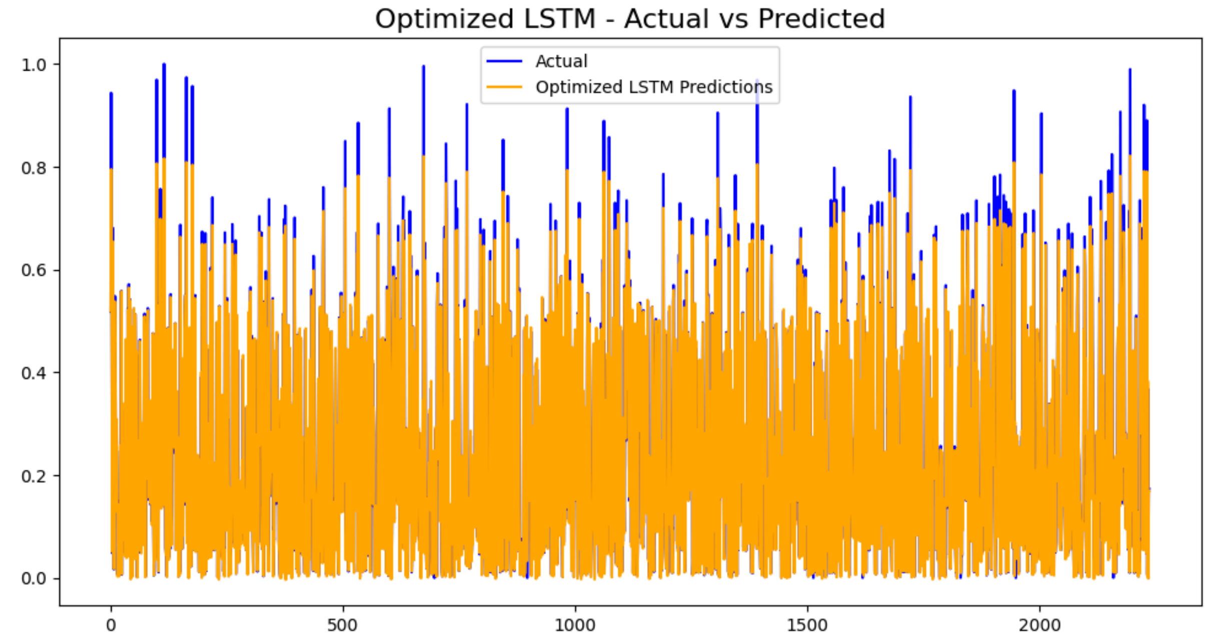
## Limitations of Random Forest Model

- Random Forest does not inherently understand time dependencies.
- It works well for short-term predictions but lacks forecasting power for 10+ days.
- The quality of predictions relies heavily on well-crafted features.

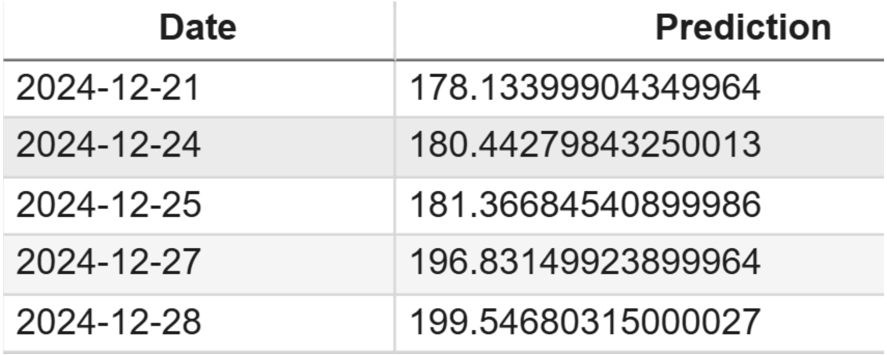
## For better sequential awareness we trained a LSTM Model as well



```
Optimized LSTM Evaluation:
RMSE: 0.0003
MAE: 0.0081
R²: 0.9927
```



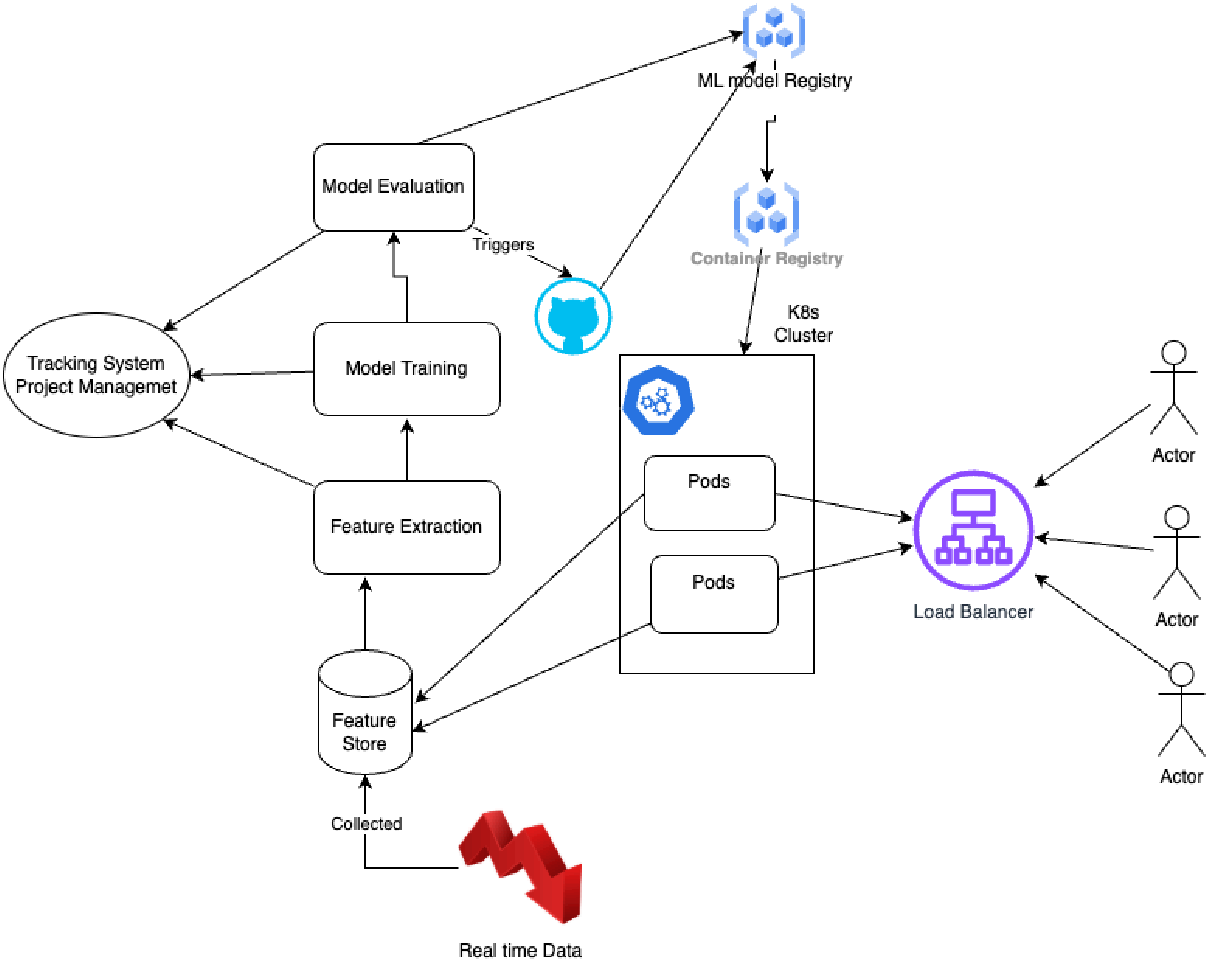
**Next 5 days predictions**



| Date | Prediction |
|---|---|
| 2024-12-21 | 178.13399904349964 |
| 2024-12-24 | 180.44279843250013 |
| 2024-12-25 | 181.36684540899986 |
| 2024-12-27 | 196.83149923899964 |
| 2024-12-28 | 199.54680315000027 |

| Component | Technology/Approach | Justification | Trade-offs |
|---|---|---|---|
| Data Collection & Ingestion | Alpha Vantage, Yahoo Finance, Quandl API, Kafka (for streaming) | Ensures access to both real-time and historical data. | API rate limits may require caching or alternative sources. |
| Data Processing & Feature Engineering | Pandas, Dask, Spark, AWS Lambda | Handles large datasets efficiently for feature extraction. | Need to balance speed vs. complexity of engineered features. |
| Model Training & Deployment | TensorFlow/PyTorch for deep learning, XGBoost for traditional ML, Docker + Kubernetes for deployment | Supports scalable, flexible model training & inference. | High training cost for deep learning models. |
| Prediction & Insights Delivery | FastAPI/Flask for APIs, Streamlit/Plotly for dashboards | Provides real-time insights for traders/analysts. | Web-based visualization may not suit all users. |
| Model Monitoring & Retraining | MLflow, Prometheus, Airflow for scheduling | Tracks model drift and retrains automatically. | Retraining frequency needs to be optimized. |

## End To End System Design


ML model Registry
Container Registry
Model Evaluation
Triggers
Tracking System
Project Managemet
Model Training
K8s
Cluster
Pods
Pods
Actor
Actor
Actor
Load Balancer
Feature Extraction
Feature
Store
Collected
Real time Data