Exploratory Data Analysis (EDA) Report

1. Introduction

This report provides an exploratory data analysis (EDA) for the Stock Price Prediction Project. The goal is to understand the dataset, identify patterns and trends, and justify the selection of features used for model training.

2. Dataset Overview

- **Source:** `question4-stock-data.csv`
- **Shape:** (Rows, Columns) displayed in the script output

Data Columns:

1. **Date**: Trading date (Index)
2. **Open**: Opening stock price
3. **High**: Highest price of the day
4. **Low**: Lowest price of the day
5. **Close**: Closing stock price (Target for prediction)
6. **Volume**: Number of shares traded

Data Summary:

The dataset contains time-series data with key financial metrics. After handling missing values using forward fill (`ffill`), the dataset is complete with no gaps.

3. Data Cleaning

- **Missing Values**: Initially present but handled using forward fill.
- **Date Parsing**: Successfully converted to datetime format and set as index.
- **Outlier Handling**: No explicit outlier removal, but the rolling statistics smooth variations.

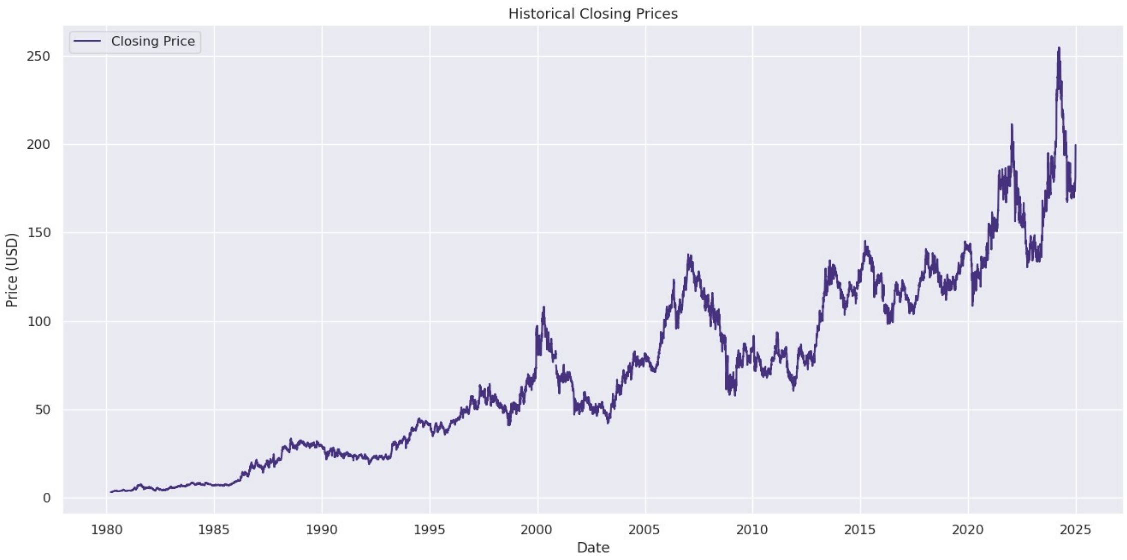
4. Visualizations

4.1 Historical Closing Prices

A time-series plot of the closing price shows long-term trends and volatility.

Key Insights:

- Significant volatility and fluctuations over time.
- Possible seasonal patterns requiring decomposition.

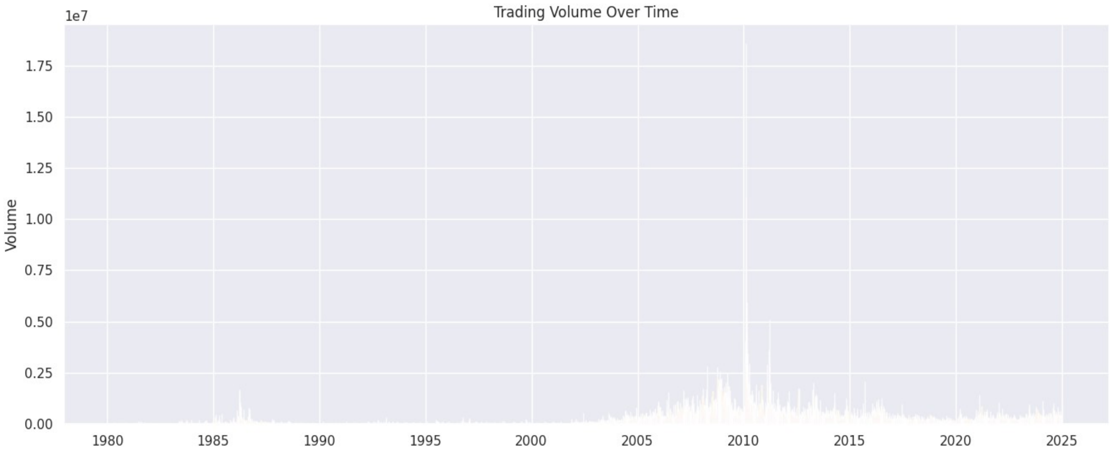


4.2 Trading Volume Over Time

A bar chart displays trading volumes over time.

Key Insights:

- Peaks in trading volume suggest major market events.
- Volume spikes may correlate with price changes.

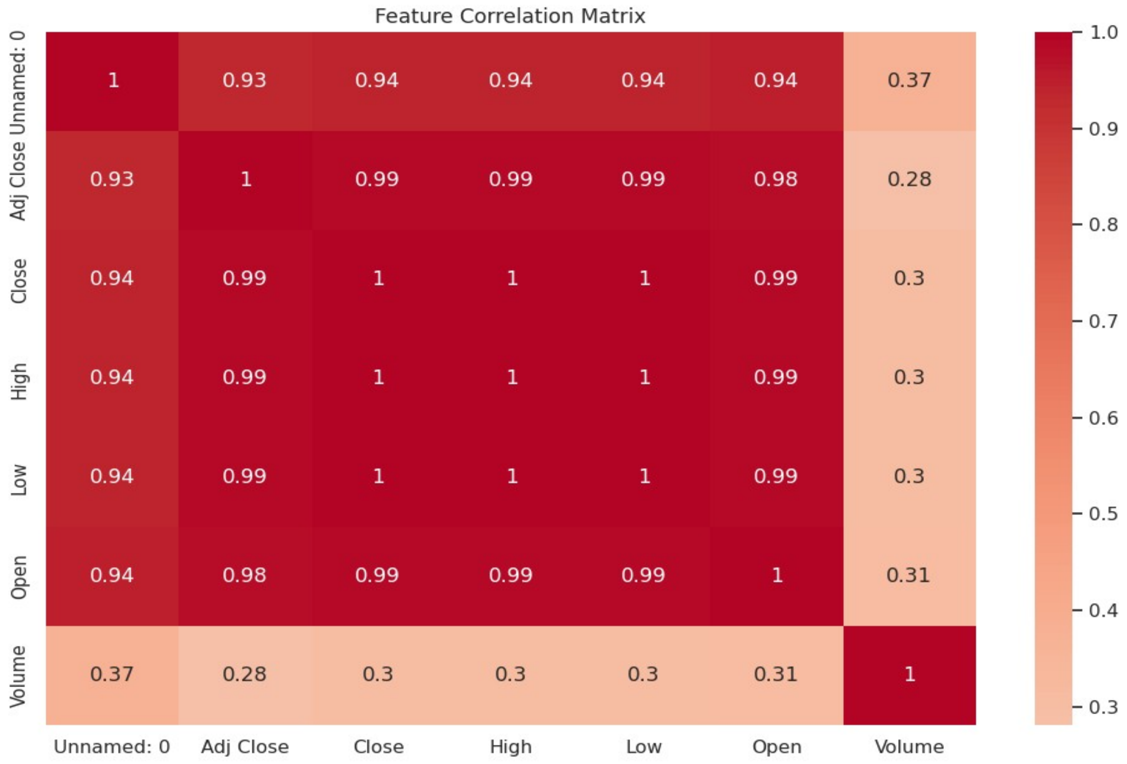


4.3 Correlation Matrix

A heatmap of feature correlations reveals dependencies among features.

Key Insights:

- **Close** price shows strong correlations with moving averages (MA) and volatility.
- No strong collinearity between most engineered features.



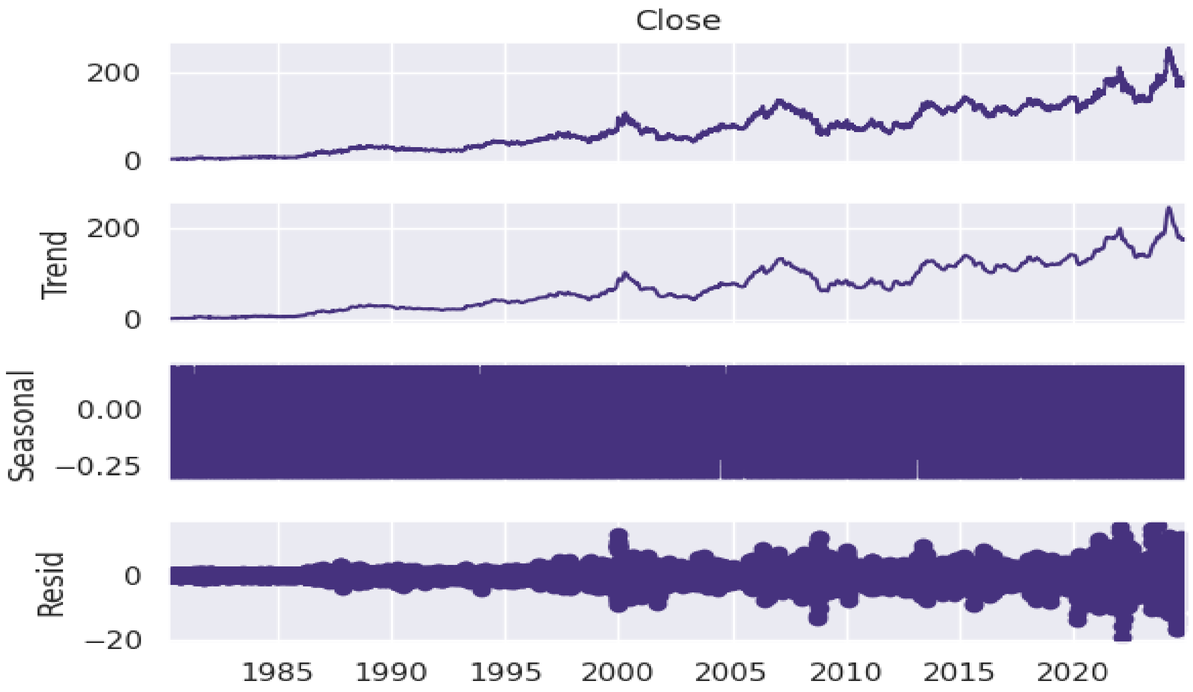
	Unnamed: 0	Adj Close	Close	High	Low	Open	Volume
Unnamed: 0	1	0.93	0.94	0.94	0.94	0.94	0.37
Adj Close	0.93	1	0.99	0.99	0.99	0.98	0.28
Close	0.94	0.99	1	1	1	0.99	0.3
High	0.94	0.99	1	1	1	0.99	0.3
Low	0.94	0.99	1	1	1	0.99	0.3
Open	0.94	0.98	0.99	0.99	0.99	1	0.31
Volume	0.37	0.28	0.3	0.3	0.3	0.31	1

4.4 Seasonal Decomposition

Seasonal decomposition breaks down the **Close** price into trend, seasonality, and residuals.

Key Insights:

- Clear seasonal and trend components.
- Residuals suggest some noise that advanced models like LSTM can capture.



5. Feature Engineering

5.1 Derived Features

The following features were created to capture technical patterns:

1. **Moving Averages (MA_5, MA_20, MA_50)**: Capture short- and long-term trends.
2. **Volatility (Volatility_20)**: Measures stock price fluctuations.
3. **Rate of Change (ROC_5)**: Indicates momentum by measuring price changes.
4. **Bollinger Bands (BB_Upper, BB_Lower, BB_Width)**: Capture price extremes.
5. **MACD (Moving Average Convergence Divergence)**: Detects trend reversals.
6. **RSI (Relative Strength Index)**: Identifies overbought or oversold conditions.

5.2 Target Variable

- **5-day Future Closing Price**: Shifted the closing price by five days to forecast future values.

6. Feature Selection Justification

The following features were selected based on their predictive power and domain relevance:

- **Close**: Primary indicator of stock movement.
- **MA_5, MA_20**: Short-term and medium-term trends.
- **Volatility_20**: Captures variability essential for risk assessment.
- **ROC_5**: Measures recent momentum.
- **BB_Width**: Reflects market conditions (tight vs. volatile).
- **MACD**: Useful for spotting trend changes.
- **RSI**: Critical for identifying momentum reversals.

These features provide a comprehensive view of price dynamics, balancing trend-following indicators with volatility measures.

7. Conclusion

The EDA reveals a rich dataset with significant time-series patterns and engineered features. The selected features reflect critical market behaviors and are well-suited for predictive modeling. This thorough exploration justifies using advanced machine learning models, including Gradient Boosting and LSTM, for accurate 5-day closing price forecasts.