Comprehensive EDA Report - Task 4

1. Data Overview

The dataset contains historical stock price data with the following key columns:

- **Date**: The trading date
- **Open**: Opening stock price
- **High**: Highest stock price during the trading session
- **Low**: Lowest stock price during the trading session
- **Close**: Closing stock price (target variable for prediction)
- **Volume**: Number of shares traded

Data Sample

A snapshot of the dataset:

	A	B	C	D	E	F	G	H
1		Date	Adj Close	Close	High	Low	Open	Volume
2	0	3/17/1980	2.296798	3.291227	3.344743	3.291227	0	41109
3	1	3/18/1980	2.306134	3.304606	3.358122	3.304606	0	9343
4	2	3/19/1980	2.306134	3.304606	3.304606	3.304606	3.304606	0
5	3	3/20/1980	2.306134	3.304606	3.358122	3.304606	0	10277
6	4	3/21/1980	2.362154	3.38488	3.438396	3.38488	0	8409
7	5	3/24/1980	2.37149	3.398259	3.451775	3.398259	0	2803
8	6	3/25/1980	2.352816	3.371501	3.425017	3.371501	0	27095
9	7	3/26/1980	2.352816	3.371501	3.425017	3.371501	0	1869
10	8	3/27/1980	2.324807	3.331364	3.38488	3.331364	0	14015
11	9	3/28/1980	2.31547	3.317985	3.371501	3.317985	0	36438

2. Data Cleaning and Preprocessing

- **Missing Values**: Checked for missing values and filled them using forward fill.
- **Date Conversion**: Converted 'Date' column to a datetime format and set it as the index.
- **Feature Scaling**: Applied MinMaxScaler (0,1) for normalizing data.

3. Data Visualization

3.1 Stock Price Trend Over Time

A line plot of the closing prices shows trends and seasonality.

3.2 Moving Averages

- **50-day Moving Average**: Captures mid-term trends.
- **200-day Moving Average**: Identifies long-term trends.

3.3 Volume Analysis

- Volume spikes correspond to sharp price movements, often signaling market activity.

4. Trend, Seasonality, and Anomalies

4.1 Trend Analysis

Using rolling averages, we identify an overall upward/downward trend in stock prices.

4.2 Seasonality Detection

Using decomposition techniques, we detect cyclic patterns in stock price behavior.

4.3 Anomaly Detection

By using boxplots and Z-score analysis, we identify outliers in price fluctuations.

5. Feature Selection Justification

Selected Features:

1. **Close Price (Target Variable)**
2. **Lagged Features**: Previous closing prices are used to model temporal dependencies.
3. **Rolling Statistics**: Moving averages (e.g., 10-day, 50-day) improve trend analysis.
4. **Trading Volume**: High volume often precedes significant price changes.

6. Data Preprocessing Decisions

- **Normalization**: Applied MinMax scaling to standardize input values.
- **Sequence Creation**: Transformed data into sequences of past 'n' days for LSTM input.
- **Train-Test Split**: 80% training, 20% testing.

Conclusion

The dataset exhibits strong trends, periodic seasonality, and volume-driven price changes. The selected features balance trend-following characteristics with anomaly detection.