Coca Cola Stock Analysis & Machine Learning

Exploratory Data Analysis | Regression & Classification | ML Model

Objective

- Perform exploratory data analysis (EDA) on historical stock price data to understand trends and patterns
- Engineer time-based and lagged features to capture temporal dependencies in stock price movements
- Build classification models to predict the direction of stock price movement (Up or Down)
- Develop an ARIMA time series model to forecast adjusted closing prices for future periods
- Evaluate classification models using metrics such as accuracy, precision, recall, and F1-scoreAssess ARIMA forecasting performance using RMSE and MAPE metrics
- Visualize data trends, decomposition components, model predictions, and future forecasts through comprehensive plots

Dataset Overview

Loaded using pandas

```
# Reading the dataset
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, StandardScaler
from sklearn.svm import SVR
from sklearn.linear_model import SGDRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
from keras.models import Sequential
from keras.layers import Dense, LSTM
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.stattools import adfuller, kpss
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.arima.model import ARIMA
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
sns.set_style("darkgrid")
df=pd.read_csv(r"C:\Users\hp\OneDrive - subho\Desktop\UNIFIED\Coca Cola Stock - Live and Updated  ( ML _ FA _ DA projects)\Coca-
print(df)
type(df)
```

Columns types

```
Date                datetime64[ns]
Open                       float64
High                       float64
Low                        float64
Close                      float64
Volume                       int64
Dividends                  float64
Stock Splits                 int64
adjusted_close             float64
dtype: object
```

Missing values checked

```
Date            0
Open            0
High            0
Low             0
Close           0
Volume          0
Dividends       0
Stock Splits    0
dtype: int64
```

Head (5)

```
df.head()
```

	Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
0	1962-01-02	0.050016	0.051378	0.050016	0.050016	806400	0.0	0
1	1962-01-03	0.049273	0.049273	0.048159	0.048902	1574400	0.0	0
2	1962-01-04	0.049026	0.049645	0.049026	0.049273	844800	0.0	0
3	1962-01-05	0.049273	0.049892	0.048035	0.048159	1420800	0.0	0
4	1962-01-08	0.047787	0.047787	0.046735	0.047664	2035200	0.0	0

Number of rows and column

```
df.shape
```

```
(15311, 8)
```

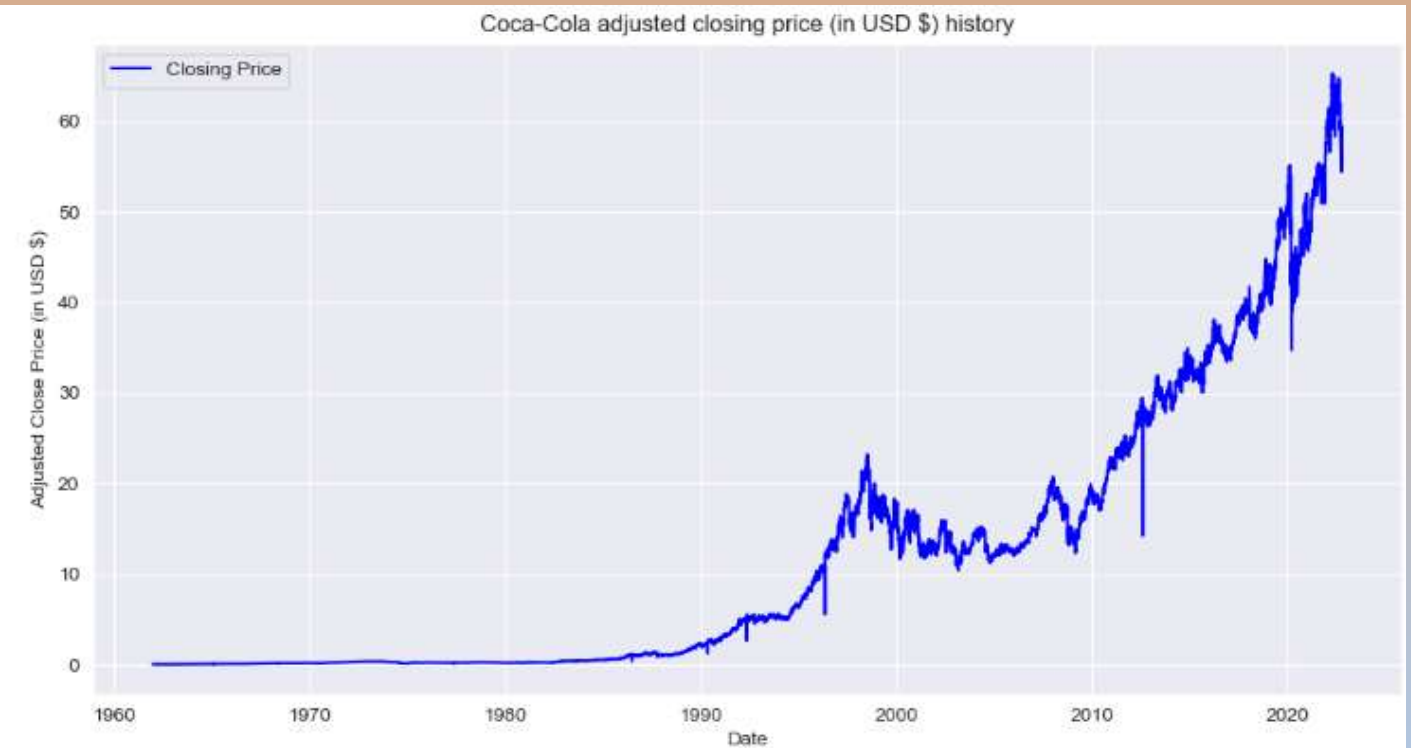
Unique values for each column

```
# Finging unique values for each column
df.nunique()
```

```
Date            15311
Open            14855
High            14547
Low             14580
Close           11105
Volume          10396
Dividends          64
Stock Splits        3
dtype: int64
```

Trend of Adjusted Closing Price History (Daily View)

- Observation: The adjusted closing price showed a steady increase from 1962 to 2022, starting around $0.037 and peaking at over $65. This reflects Coca-Cola's long-term growth and strong market performance over six decades.



Minimum 5 Observations:

	Date	adjusted_close
123	1962-06-27 04:00:00	0.037028
122	1962-06-26 04:00:00	0.037154
121	1962-06-25 04:00:00	0.037279
120	1962-06-22 04:00:00	0.037656
205	1962-10-23 04:00:00	0.037671

Maximum 5 Observations:

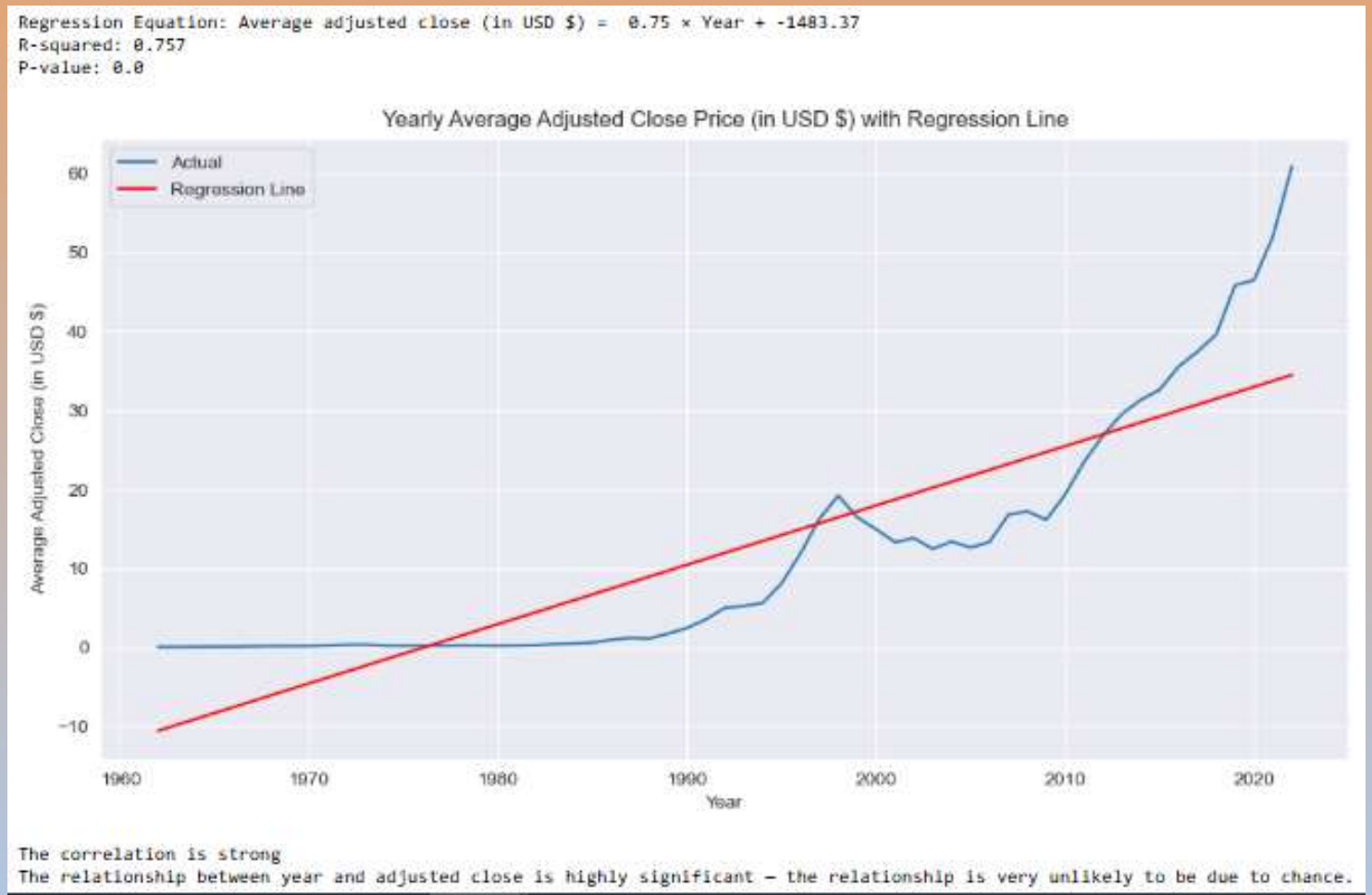
	Date	adjusted_close
15182	2022-04-25 04:00:00	64.993156
15179	2022-04-20 04:00:00	65.012871
15197	2022-05-16 04:00:00	65.012871
15185	2022-04-28 04:00:00	65.239563
15180	2022-04-21 04:00:00	65.259270

Trend of Adjusted Closing Price History (Monthly Average)

- Observation: The monthly average adjusted close price shows a steady rise from January to August, peaking at $12.22. Although there is a slight dip from September to November, the price remains relatively stable, indicating overall market resilience.

Yearly Average Adjusted Closing Price with Trend Line (Regression)

- Observation: There is a strong positive linear trend in the yearly average adjusted closing price, with an R^2 of 0.757.The regression indicates a statistically significant rise in price over time, confirming long-term value growth.



Maximum High Price Points in the Dataset

- Observation: The dataset shows peak high prices clustered around April–May 2022, with the highest value reaching $66.24. This indicates a significant surge in Coca-Cola's stock during this period, possibly due to market or company-specific events.

Maximum 5 Observations:

	Date	High
15181	2022-04-22 04:00:00	65.387402
15183	2022-04-26 04:00:00	65.416982
15197	2022-05-16 04:00:00	65.426838
15180	2022-04-21 04:00:00	66.037927
15182	2022-04-25 04:00:00	66.235058

Minimum Low Price Points in the Dataset

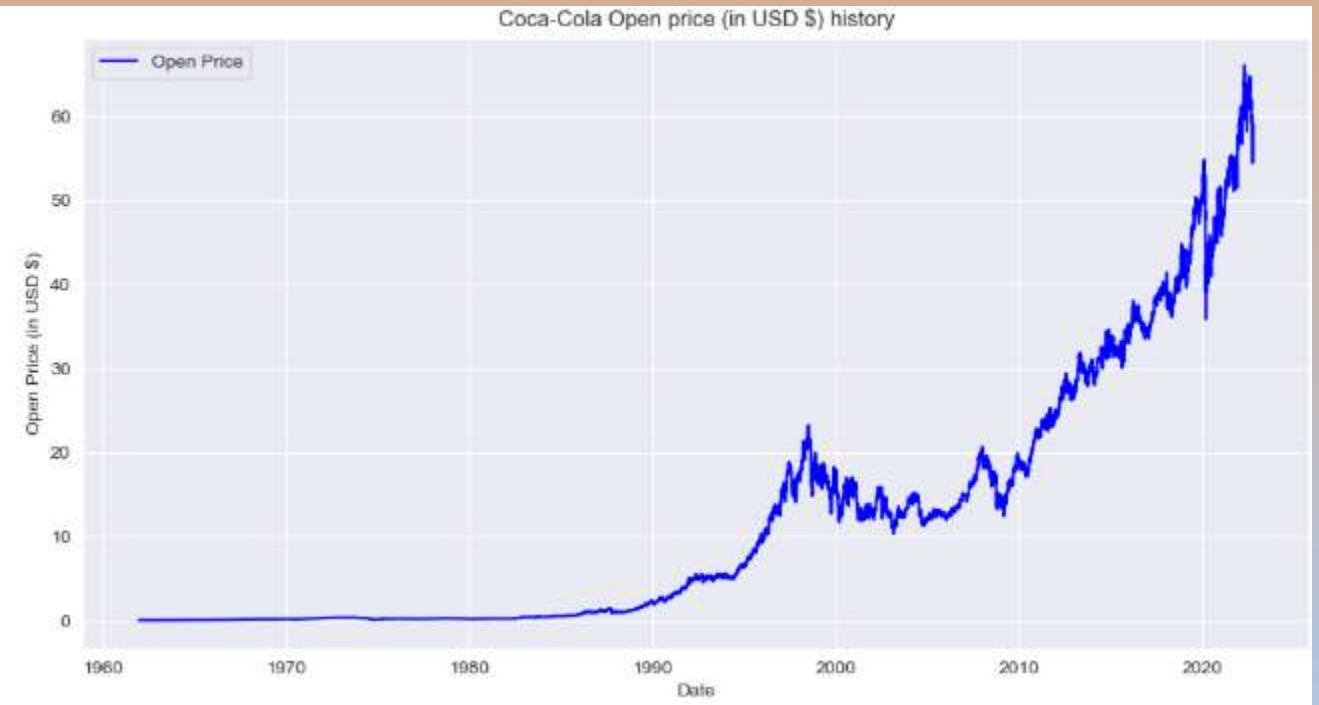
- Observation: The lowest recorded stock prices occurred in mid-1962, with the minimum low reaching $0.0349. This reflects the historical starting point of Coca-Cola's stock, highlighting its substantial growth over the decades.

Minimum 5 Observations:

	Date	Low
103	1962-05-29 04:00:00	0.034890
121	1962-06-25 04:00:00	0.036212
123	1962-06-27 04:00:00	0.036840
122	1962-06-26 04:00:00	0.036903
124	1962-06-28 04:00:00	0.037530

Trend of Adjusted Opening Price History (Daily View)

- Observation : The adjusted opening prices show a significant increase from the early 1960s, starting around $0.037. Recent data from 2022 indicates a peak opening price above $66, reflecting strong long-term growth.



Maximum 5 Observations:

	Date	Open
15186	2022-04-29 04:00:00	64.924157
15180	2022-04-21 04:00:00	65.032577
15181	2022-04-22 04:00:00	65.131141
15198	2022-05-17 04:00:00	65.150854
15182	2022-04-25 04:00:00	66.037933

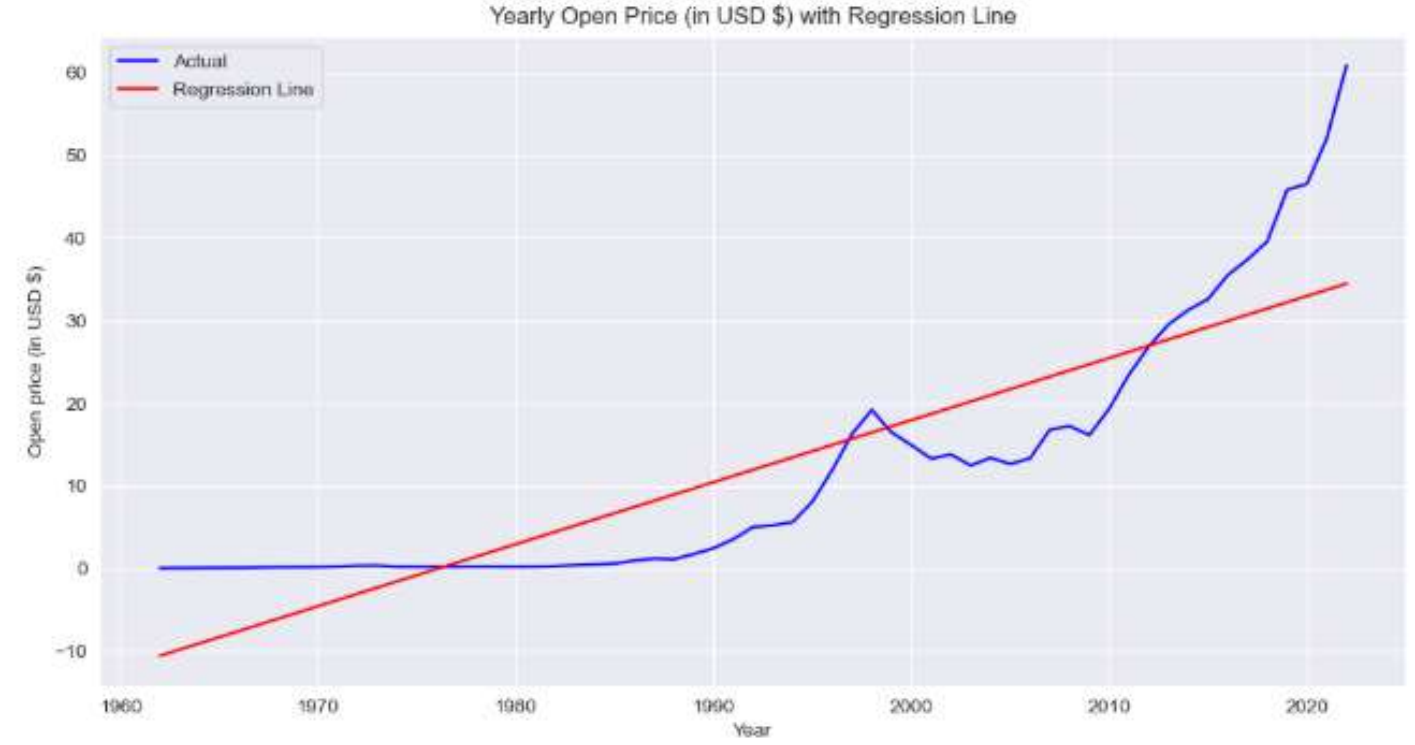
Minimum 5 Observations:

	Date	Open
123	1962-06-27 04:00:00	0.037154
122	1962-06-26 04:00:00	0.037279
124	1962-06-28 04:00:00	0.037530
121	1962-06-25 04:00:00	0.037656
206	1962-10-24 04:00:00	0.037671

Yearly Average Open Price with Trend Line (Regression)

- Observation: The yearly average opening price shows a strong upward trend over the years, supported by a high R-squared value of 0.757. The significant p-value (0.0) confirms that the increase in open price over time is statistically meaningful and unlikely due to random chance.

```
Regression Equation: Open price (in USD $) =  0.75 × Year + -1483.57
R-squared: 0.757
P-value: 0.0
```



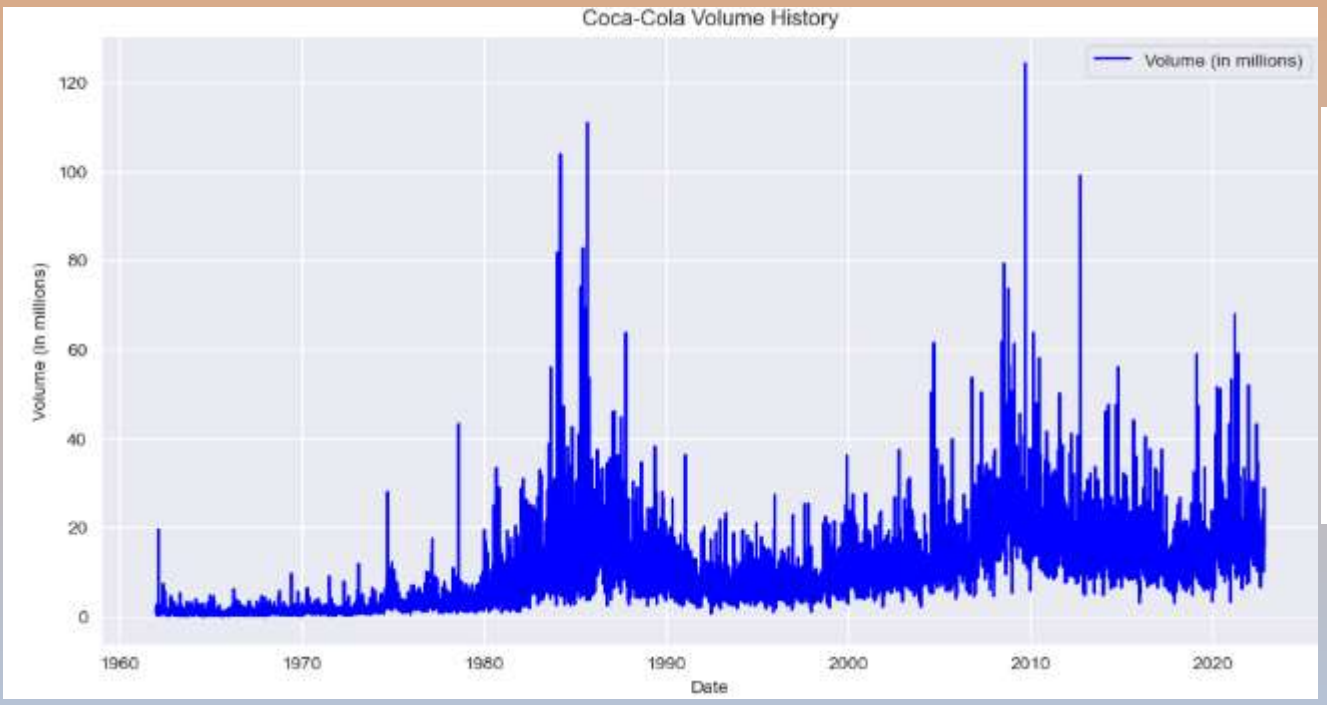

```
The correlation is strong
The relationship between year and open price is highly significant — the relationship is very unlikely to be due to chance.
```

Monthly Average Open Price Trend

- Observation: The monthly average open price shows a rising trend from January ($11.20) to a peak in August ($12.20), mirroring the adjusted close price trend. Post-August, a gradual decline is observed through November, before a slight recovery in December. This pattern reflects a typical mid-year peak and end-year stabilization in market behavior.

Trend of Volume Traded (Daily View)

- Observation : The daily trading volume shows substantial growth over the years, with peak volumes reaching over 124 million shares traded. Early years recorded very low volumes, often below 0.2 million, indicating much lower market activity in the initial decades.



Maximum 5 Observations (Volume in millions):

	Date	Volume
5888	1985-06-10 04:00:00	82.6560
12770	2012-09-21 04:00:00	98.9675
5576	1984-03-15 00:00:00	103.7760
5951	1985-09-09 04:00:00	110.7840
12011	2009-09-18 04:00:00	124.1690

Minimum 5 Observations (Volume in millions):

	Date	Volume
1007	1965-12-31 00:00:00	0.0768
452	1963-10-16 04:00:00	0.0768
1983	1969-12-26 00:00:00	0.1056
169	1962-08-31 04:00:00	0.1152
1285	1967-02-07 00:00:00	0.1152

Yearly Average Volume (in Millions) with Trend Line (Regression)

• Observation: The yearly average trading volume has shown a strong upward trend over time, with volumes increasing steadily each year. Statistical analysis confirms this relationship is highly significant and unlikely to be due to random chance.

Monthly Average Volume Trend (in Millions)

• Observation: The monthly average trading volume exhibits noticeable fluctuations throughout the year. A relatively lower volume is seen in the mid months, with a gradual increase peaking around early-year (typically Feb–Mar). Following this, a declining trend is observed towards the end of the year, indicating reduced market activity, possibly due to seasonal slowdowns or investor caution during year-end periods.

Machine Learning – regression model

• Observation: Created key time-related features (day, month, weekday) and encoded categorical data to capture seasonality and trends. Added lagged adjusted close prices to incorporate past price influence, enhancing model predictive power.

```
RMSE: 18.048915943651 53
MSE: 325.7633667409984
R²: -1.8176615127492 34
```

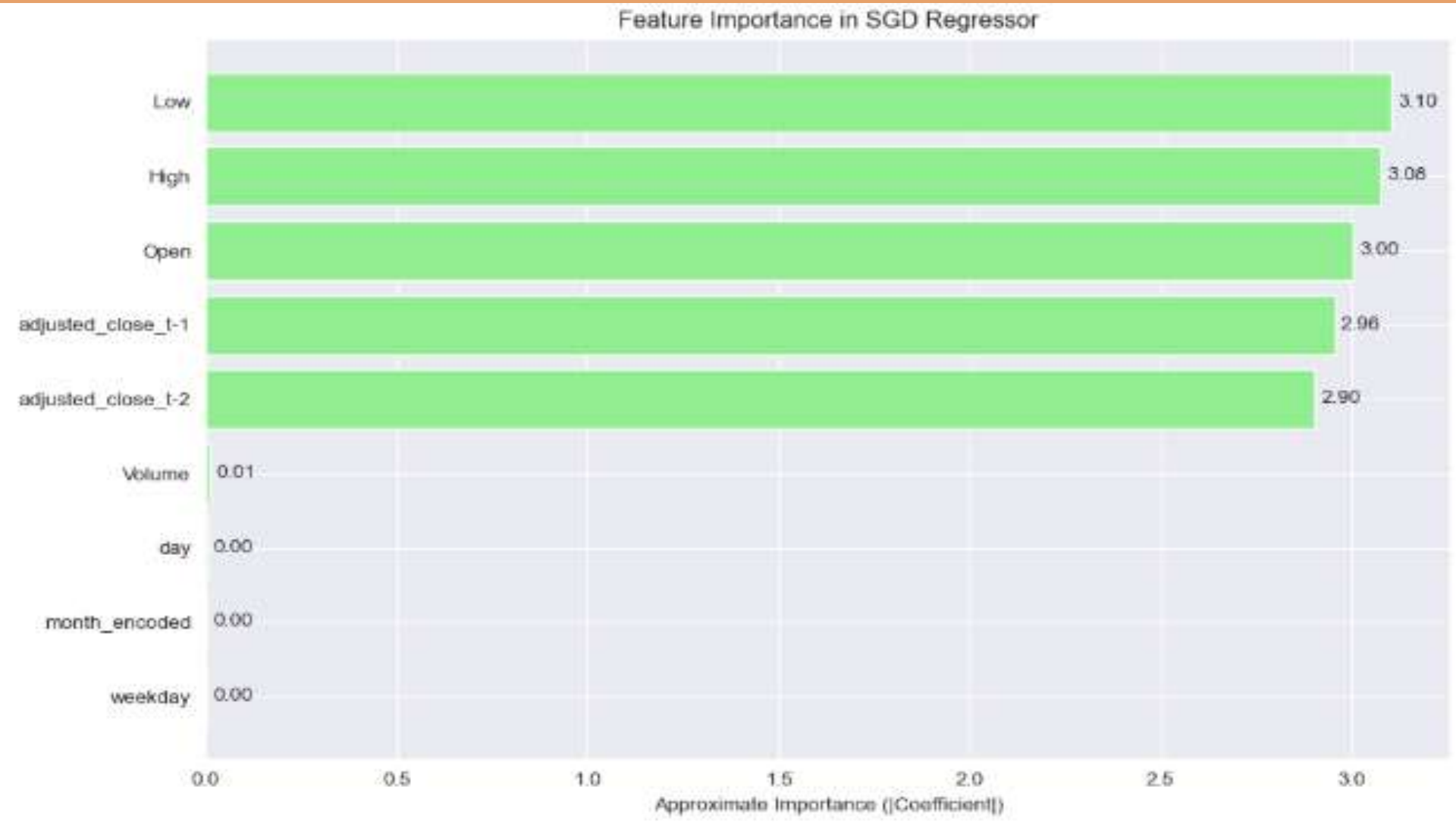
```
RMSE: 28.362361617297502
MSE: 804.4235565103505
R²: -5.962173033226879
```

```
RMSE: 0.42948414486526415
MSE: 0.18445663069064722
R²: 0.998402510951 8025
```

```
RMSE: 2.4158172911013285
MSE: 5.836173183984162
R²: 0.969224258117 2675
```

Feature importance in SGD

- Observation : The most important features influencing the SGD model are Low, High and Open. These features play a key role in predicting coca cola stock effectively.



Machine Learning – classification model

• Observation: Created time-related features such as day, month, weekday, and lagged adjusted close values to capture trends and seasonality. Encoded categorical data (month names) and defined a binary target for price direction (up/down) to enhance model performance.

CLASSIFICATION RANDOM FOREST MODEL

Accuracy: 0.4786017641293695

Classification Report:

	precision	recall	f1-score	support
0	0.48	0.98	0.64	1451
1	0.58	0.03	0.06	1610
accuracy			0.48	3061
macro avg	0.53	0.50	0.35	3061
weighted avg	0.53	0.48	0.33	3061

CLASSIFICATION SVC MODEL

Accuracy: 0.4996732026143791

Classification Report:

	precision	recall	f1-score	support
0	0.49	0.98	0.65	1451
1	0.79	0.07	0.12	1609
accuracy			0.50	3060
macro avg	0.64	0.52	0.39	3060
weighted avg	0.64	0.50	0.37	3060

CLASSIFICATION SGD MODEL

Accuracy: 0.6294117647058823

Classification Report:

	precision	recall	f1-score	support
0	0.89	0.25	0.39	1451
1	0.59	0.97	0.73	1609
accuracy			0.63	3060
macro avg	0.74	0.61	0.56	3060
weighted avg	0.73	0.63	0.57	3060

CLASSIFICATION MODEL LSTM

LSTM Accuracy: 0.4766884531590414

LSTM Classification Report:

	precision	recall	f1-score	support
0.0	0.48	0.99	0.64	2178
1.0	0.61	0.01	0.02	2412
accuracy			0.48	4590
macro avg	0.54	0.50	0.33	4590
weighted avg	0.55	0.48	0.32	4590

Models

- Regression and classification models were applied to predict and classify stock price movement direction using features such as open, high, low prices, volume, date components, and lagged adjusted close values.

- Regression models – Regression models were trained using extensive feature engineering including date components (day, month, weekday), encoded month names, and lagged adjusted close prices to capture temporal patterns. Among the models, the LSTM achieved the best performance with an RMSE of 2.42 and R^2 of 0.97, demonstrating excellent predictive accuracy for stock prices. The Random Forest Regressor showed moderate results with an RMSE of 18.05 and a negative R^2, indicating poor fit, while the SGD Regressor surprisingly achieved strong results with an RMSE of 0.43 and R^2 of 0.998, likely due to effective feature scaling and linear assumptions. The SVR model performed poorly, with high RMSE (28.36) and negative R^2, indicating limited predictive capability. Overall, deep learning and well-scaled linear models outperformed traditional ensemble and kernel-based regressors in this task.

- Classification models – Among the classification models tested for predicting stock price direction, the SGD Classifier achieved the highest accuracy of 62.94% and demonstrated a strong recall of 0.97 for upward price movement, indicating effective identification of positive trends. The Support Vector Classifier (SVC) and Random Forest Classifier yielded similar moderate accuracies around 48-50%, but both struggled with recall on the minority class, reflecting challenges with class imbalance. The LSTM model also showed comparable accuracy (~47%) and high recall for the majority class, yet poor detection of downward movement. Overall, the classification models exhibited moderate predictive ability, with linear models like SGD outperforming more complex ensemble and deep learning approaches in this scenario, though class imbalance impacted performance, particularly in minority class recall and F1 scores.

Time Series (ARIMA MODEL)

• Observation: The ARIMA model effectively captures the long-term growth trend in Coca-Cola's stock prices. Its smooth forecast provides a reliable estimate for future performance, especially for strategic long-term planning.

Forecasting for next 6 months

• Observation: The 6-month forecast for Coca-Cola's adjusted close price shows a steady upward trend, indicating consistent positive growth. This suggests strong market confidence and potential long-term value for investors.

Final Observations & Storyline

- The adjusted closing price of Coca-Cola stock rose dramatically from $0.037 in 1962 to a peak of $66.24 in April–May 2022. This represents over 1,700-fold growth, highlighting strong long-term shareholder value and market confidence over six decades. Monthly averages show a rise from $11.20 in January to $12.22 in August, followed by a slight decline, indicating seasonal market trends.

- Yearly average adjusted closing prices follow a statistically significant upward trend with an R^2 of 0.757 and p-value < 0.001. This confirms consistent annual price growth, reflecting Coca-Cola's expanding market presence and sustained brand strength. The upward trend is mirrored in adjusted opening prices, which also rose from $0.037 to above $66 over the same period.

- Monthly average opening prices peak at $12.20 in August, then decline slightly before stabilizing in December, closely following the adjusted close price trends. Trading volume increased significantly from under 0.2 million daily shares in early years to peaks over 124 million shares recently, indicating greater market liquidity and investor participation.

- Yearly average trading volume shows a strong upward trend, statistically significant and aligned with Coca-Cola's growing market capitalization. Monthly volumes peak early in the year (February–March) and taper off toward year-end, reflecting typical seasonal trading patterns and investor behavior. The historic low price of $0.0349 in 1962 contrasts sharply with recent highs, emphasizing massive long-term appreciation.

- Overall, Coca-Cola stock demonstrates resilience with strong price appreciation, increasing liquidity, and predictable seasonal volume patterns. These trends underscore the company's status as a stable, blue-chip investment with a proven track record of delivering value over time.