Level-1-Data Analytics and Predictive Modeling Project Documentation

Project Title:

Stock Market Prediction

Team Members:

- Madhavendra Singh Naruka

Project Overview:

The goal of this project is to predict stock market prices using historical data and advanced machine learning techniques. By leveraging predictive analytics, the project aims to provide insights that can assist in making informed investment decisions and developing trading strategies.

1. Introduction

- **Purpose**

The primary goal of this project is to gain insights into the stock market and create accurate predictions. These predictions will serve as the foundation for developing effective trading strategies.

- **Scope**

The project encompasses the following key stages:

1. Data Collection: Gather relevant data from various sources.
2. Pre-Processing: Clean, transform, and prepare the data for analysis.
3. Exploratory Data Analysis (EDA): Explore the dataset to identify patterns, anomalies, and potential features.
4. Feature Selection: Choose the most relevant features for modeling.
5. Model Development: Build predictive models using machine learning techniques.
6. Model Evaluation and Selection: Assess model performance and choose the best-performing one.
7. Prediction and Analysis: Apply the selected model to make stock price predictions.
8. Strategy Development: Based on the predictions, devise trading strategies to optimize returns.

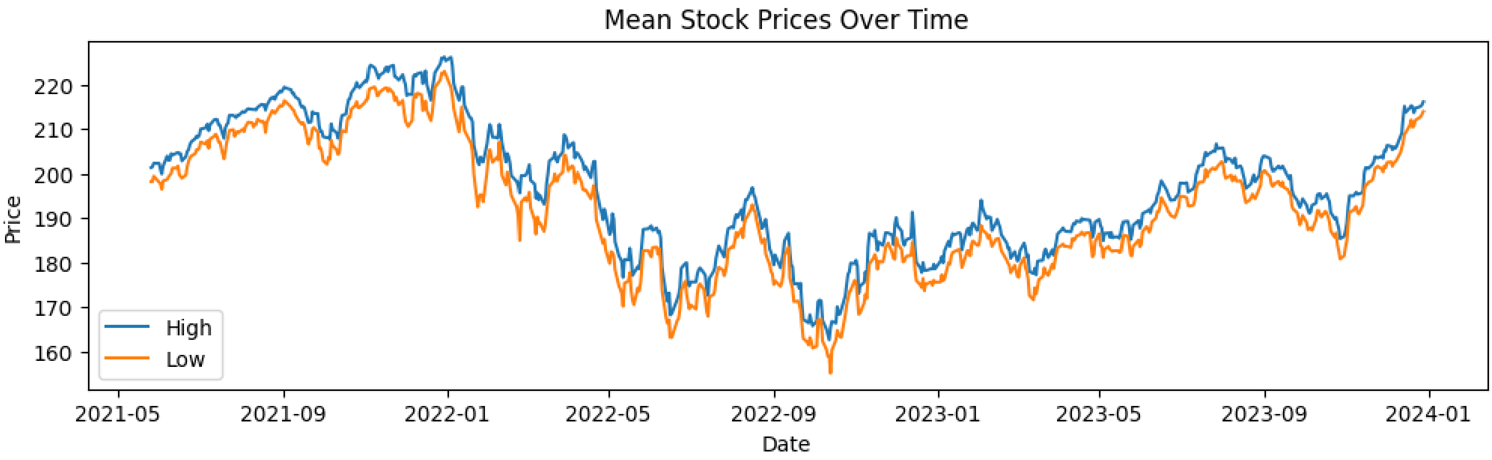
2. Data Collection

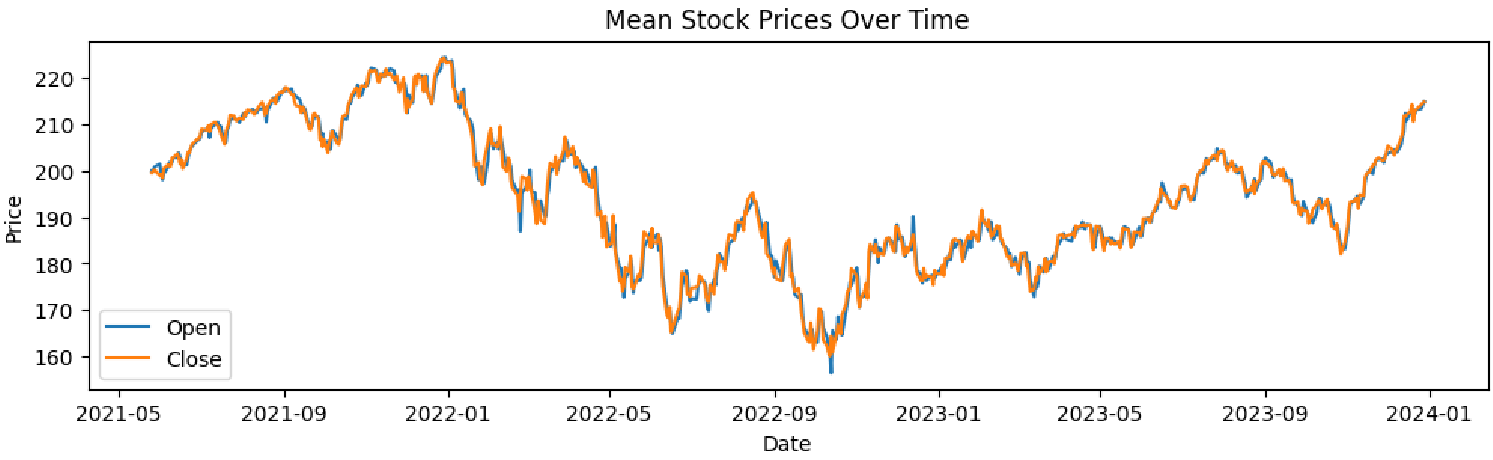
- **Data Sources**: For this project, I utilized the yfinance library to extract stock data from Yahoo Finance. Specifically, I collected historical stock data for 86 companies, spanning the period from January 1, 2021, to January 1, 2024.
- **Data Acquisition:** After extracting the data, I organized it into individual dataframes and stored them in a list for further analysis.

3. Data Preprocessing

- **Data Cleaning**: The Data was already cleaned but creating new features like SMA and EMA.
- **Data Transformation**: After conducting Exploratory Data Analysis (EDA), I scaled the data to prepare it for feature selection and model development.
- **Data Integration**: To streamline access and analysis, I consolidated the list of individual dataframes (originally extracted for different companies) into a single unified dataframe.
- **Data Export:**
  - For each of the 86 companies, I exported their stock data into separate CSV files.
  - Additionally, I consolidated all the individual company data into a single comprehensive CSV file.

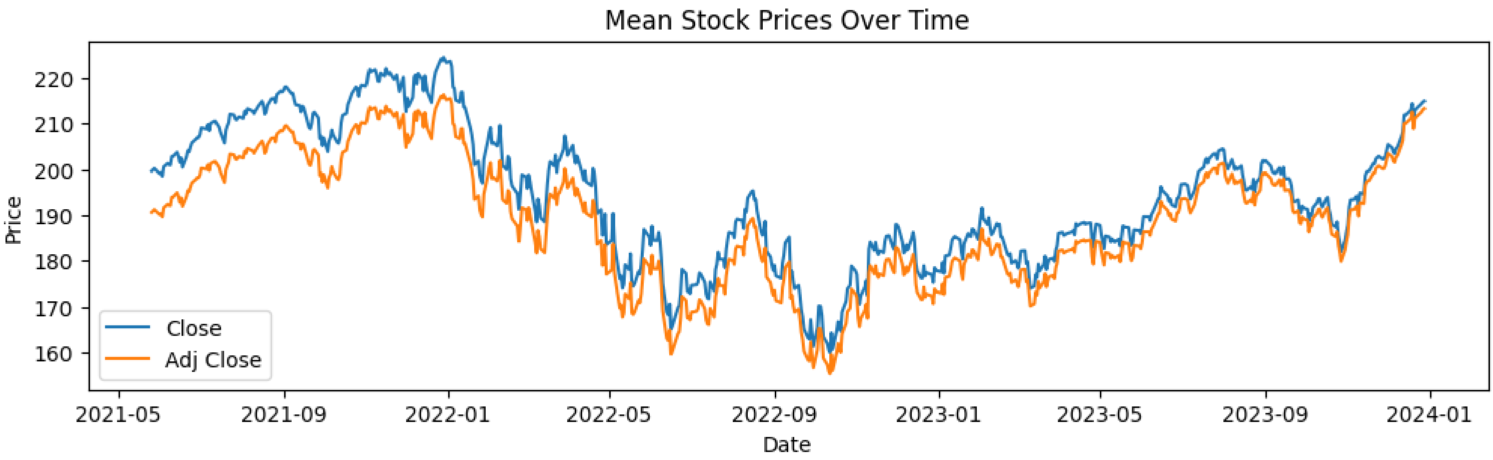
4. Exploratory Data Analysis (EDA)





Significant Findings:

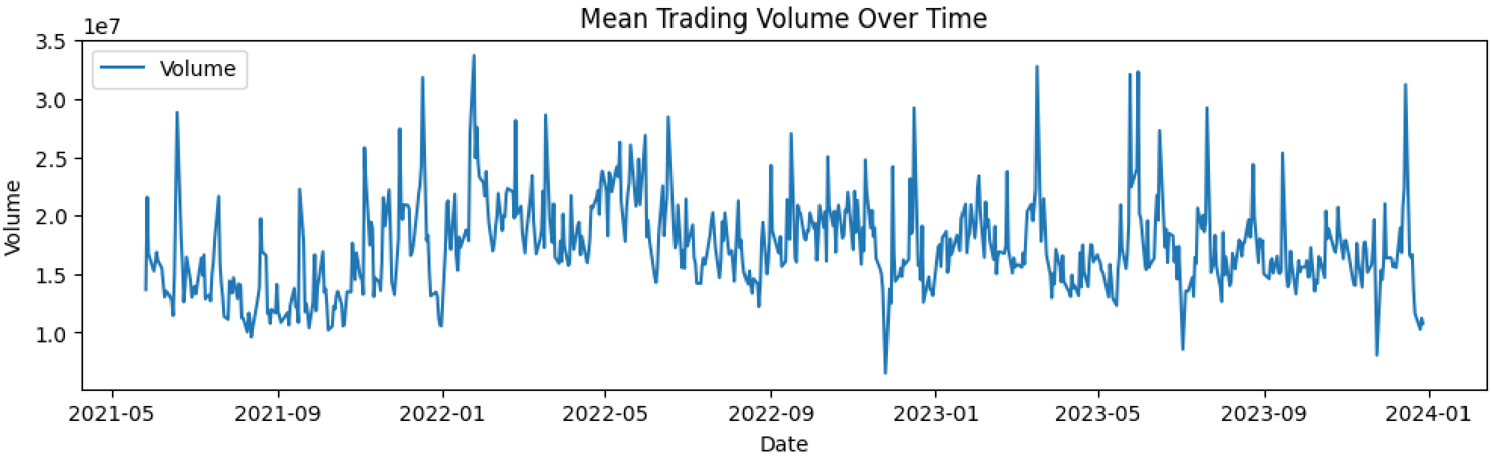
- **High and Low Price Volatility:** There are significant differences between the high and low prices during a trading day, indicating substantial price volatility.
- **Open and Close Price Stability:** The opening and closing prices are relatively similar, suggesting that despite intraday price fluctuations, the closing price does not differ significantly from the opening price.
- **Trading Strategies:** This data can be valuable for both day traders and long-term traders in developing their respective trading strategies.



Significant Findings :

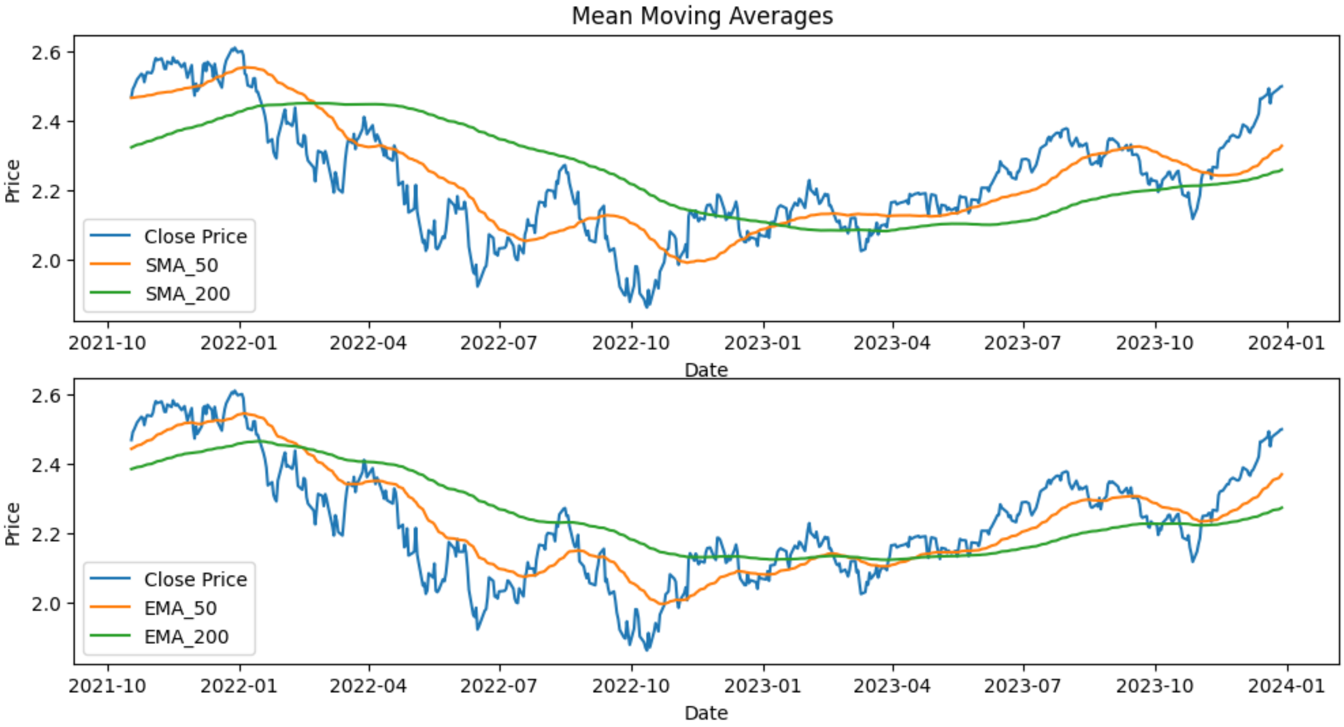
- The 'Adj Close' (Adjusted Close) prices are consistently lower than the 'Close' prices. This indicates adjustments made to the closing prices, possibly accounting for factors such as dividends, stock splits, and other corporate actions.

- The close correlation between 'Close' and 'Adj Close' prices throughout the period suggests that adjustments are consistent and follow the same overall trend as the unadjusted prices.



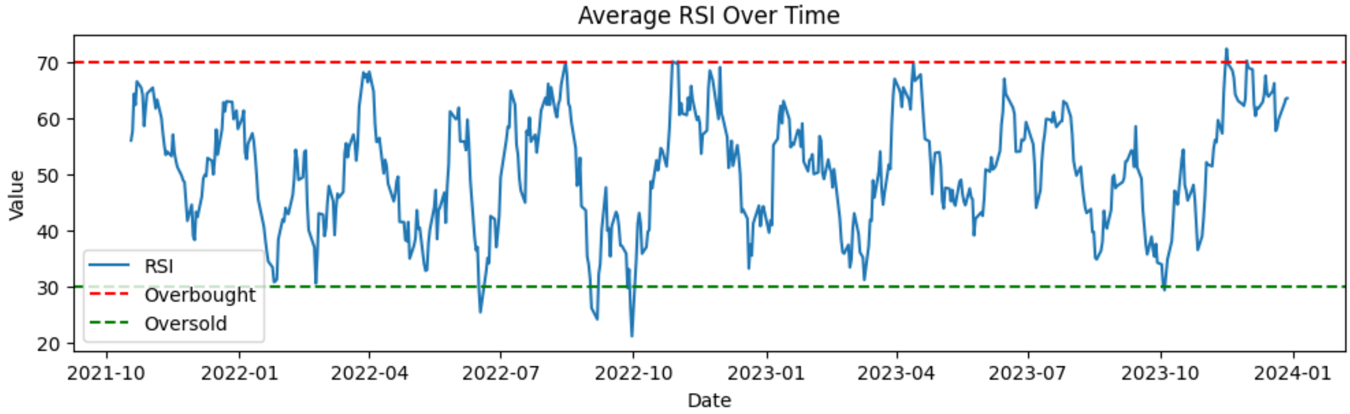
Significant Findings :

- The trading volume exhibits significant fluctuations throughout the observed period, with multiple peaks and troughs.
- The volume ranges mostly between 1 million to over 3 million trades, with the highest peak nearing 3.5 million trades. This indicates substantial trading activity during peak periods.
- The volume seems to increase around new year suggesting a time of peaked interest.



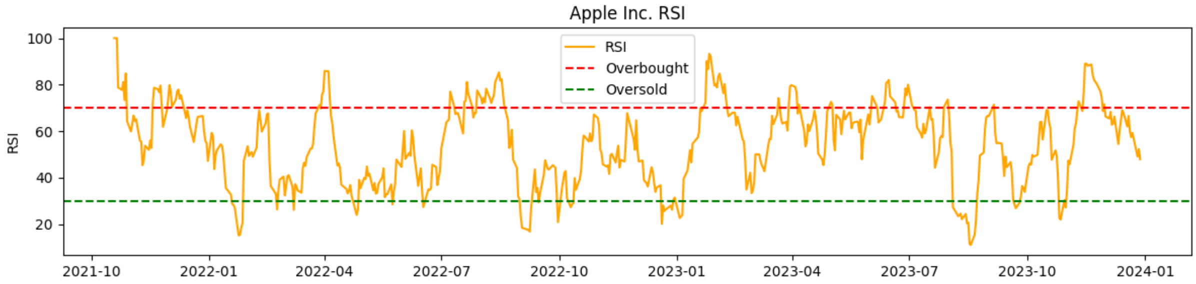
Significant Findings :

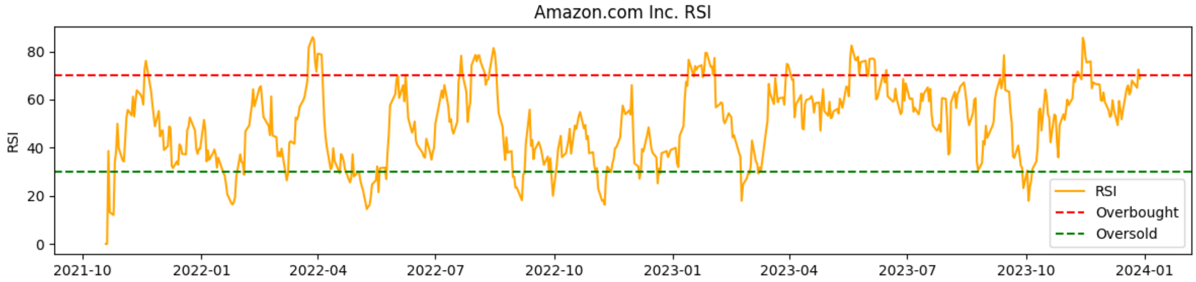
- **Trend Identification:**
  - The 200-day SMA and EMA provide a clear long-term trend, smoothing out short-term fluctuations.
- **Bullish and Bearish Phases:**
  - A bullish phase is observed when the 'Close Price' is above the 200-day SMA/EMA, particularly noticeable from mid-2023 to early 2024.
  - A bearish phase is seen when the 'Close Price' is below the 200-day SMA/EMA, evident from late 2021 to mid-2022.
- **Crossover Signals:**
  - Crossovers of the 50-day moving averages (SMA_50 and EMA_50) with the 200-day moving averages (SMA_200 and EMA_200) indicate significant trend reversals:
    - The 'Golden Cross' (50-day average crossing above 200-day average) around mid-2023 signals a shift to a bullish trend.
    - The 'Death Cross' (50-day average crossing below 200-day average) around early 2022 indicates a shift to a bearish trend.



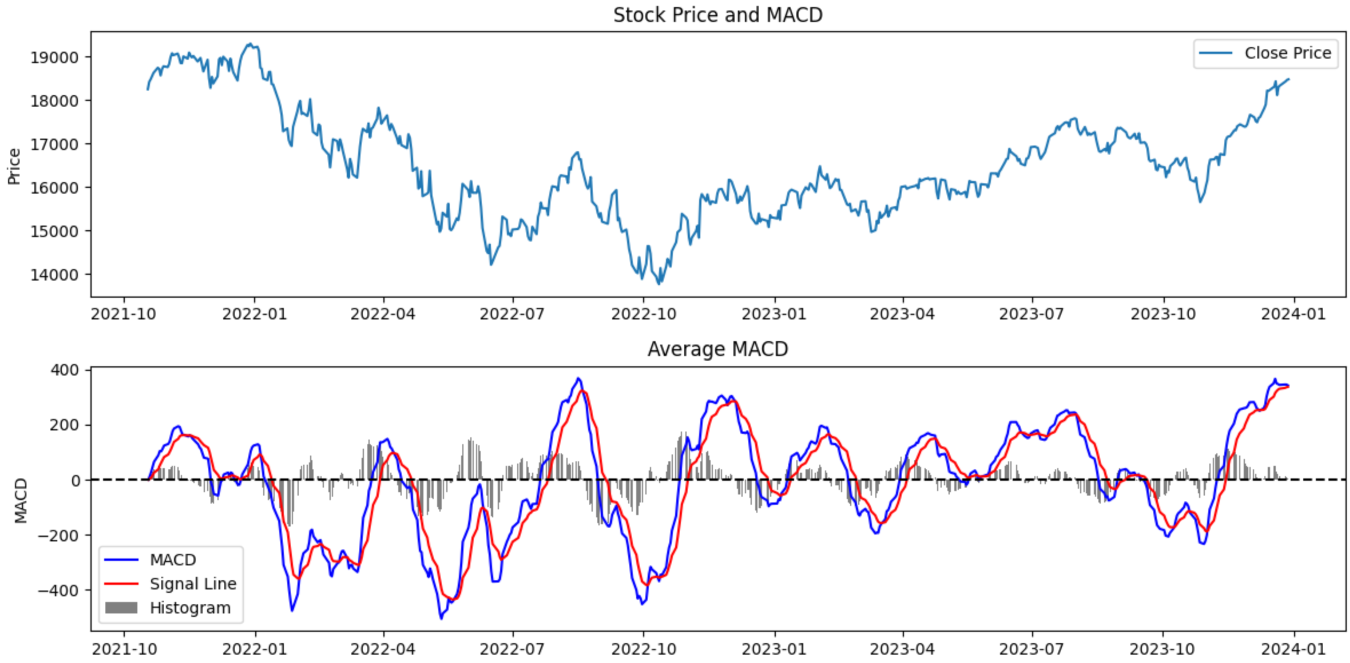
Significant Findings :

- The RSI chart shows multiple instances of reaching overbought conditions, particularly in late 2021, early 2022, and late 2023.
- The RSI also indicates several oversold conditions, especially in early 2022 and mid-2022, suggesting potential entry points for long positions i.e people buyinmg stock to hold them as a long term investment.
- RSI values between 30 and 70 are considered neutral. It implies that the stock is trading within a reasonable range, and there is no extreme momentum in either direction.
- We can see significant oversold stocks around October 2022.
- I also say significant fluctuations in individual stocks which are not visible in average RSI data. For example :





*** This suggests that individual stopck should be studied rather than market as a whole while looking for overbrough and oversold conditions.**



Significant Findings :

- Several MACD (blue line) and Signal Line (red line) crossovers are observed:
  - **Bullish Crossover**: When MACD crosses above the Signal Line, indicating potential buy signals. These occur around early 2022, mid-2022, and early 2023.
  - **Bearish Crossover**: When MACD crosses below the Signal Line, indicating potential sell signals. These occur around late 2021, mid-2022, and mid-2023.

- The **histogram** reflects the difference between the MACD and Signal Line:
  - **Positive values indicate bullish momentum**, observed in early 2022, late 2022, and late 2023.
  - **Negative values indicate bearish momentum**, observed in early 2022, mid-2022, and mid-2023.

The crossover of the MACD (Moving Average Convergence Divergence) and the signal line is a helpful signal for initiating buying or selling decisions. Additionally, the histogram assists in assessing the momentum of the stock.

5. Feature Engineering

- **Feature Creation**

I engineered several features using scikit-learn. Here's a brief overview of each feature and its purpose:

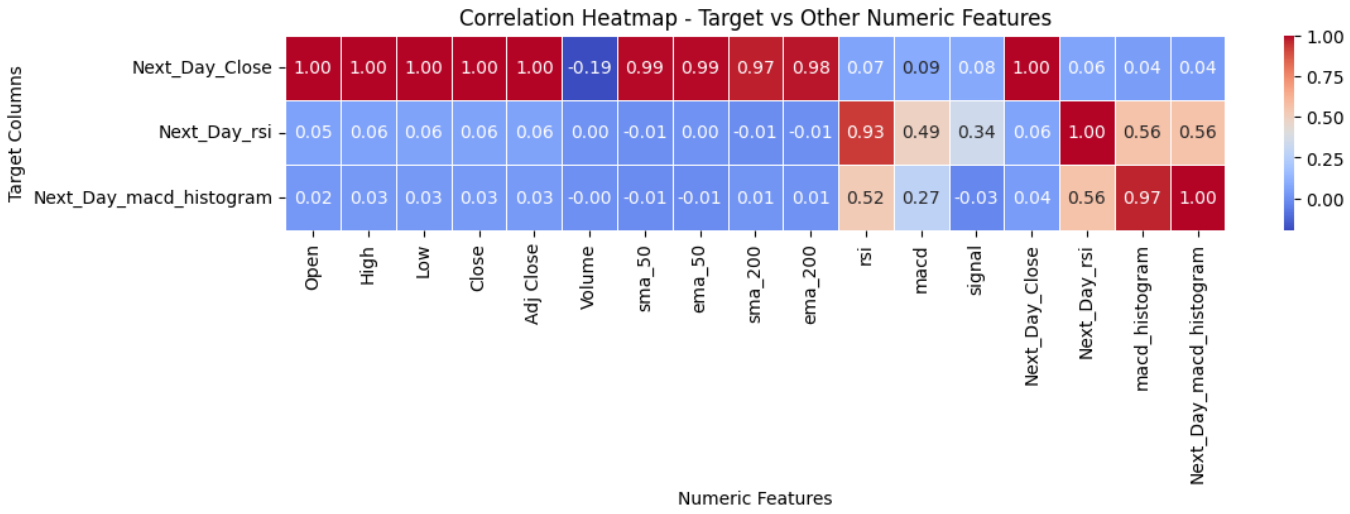
- **Simple Moving Average (SMA):**
  - *What it is*: SMA calculates the average closing price over a specified window (e.g., 50 days or 200 days).
  - *Why it's used*: SMA smoothes out price fluctuations, providing a trend-following indicator. It helps identify overall price direction.
- **Exponential Moving Average (EMA):**
  - *What it is*: EMA gives more weight to recent prices, making it sensitive to recent market movements.
  - *Why it's used*: EMA reacts faster to price changes, aiding in identifying short-term trends and potential reversals.
- **Relative Strength Index (RSI):**
  - *What it is*: RSI measures the strength and speed of price movements.
  - *Why it's used*: RSI helps assess whether a stock is overbought (above 70) or oversold (below 30), aiding in timing buy/sell decisions.
- **Moving Average Convergence Divergence (MACD):**
  - *What it is*: MACD combines two EMAs to create a trend-following momentum indicator.
  - *Why it's used*: MACD identifies trend changes and potential buy/sell signals. The crossover with the signal line is crucial.
- **Signal Line (for MACD):**
  - *What it is*: The signal line is a 9-day EMA of the MACD.
  - *Why it's used*: It provides additional confirmation for MACD crossovers and helps filter out noise.
- **MACD Histogram:**
  - *What it is*: The difference between the MACD and the signal line.
  - *Why it's used*: The histogram visualizes the strength of the trend. Positive values indicate bullish momentum, while negative values suggest bearish momentum.

- **Feature Selection**

In this step, I carefully chose features to enhance model performance. Here are the three target columns I selected:

1. **Next_Day_Close**:
   - **Purpose**: This feature aims to predict the closing price of the stock for the next trading day.
   - **Why it's important**: Accurate predictions of future prices help traders make informed decisions.
2. **Next_Day_rsi**:
   - **Purpose**: This feature predicts whether the stock will be overbought or oversold on the next trading day.
   - **Why it's relevant**: Identifying overbought or oversold conditions assists in timing buy/sell actions.
3. **Next_Day_macd_histogram**:
   - **Purpose**: This feature predicts the momentum of the stock for the following day.
   - **Why it matters**: Momentum analysis helps traders gauge the strength of price movements.

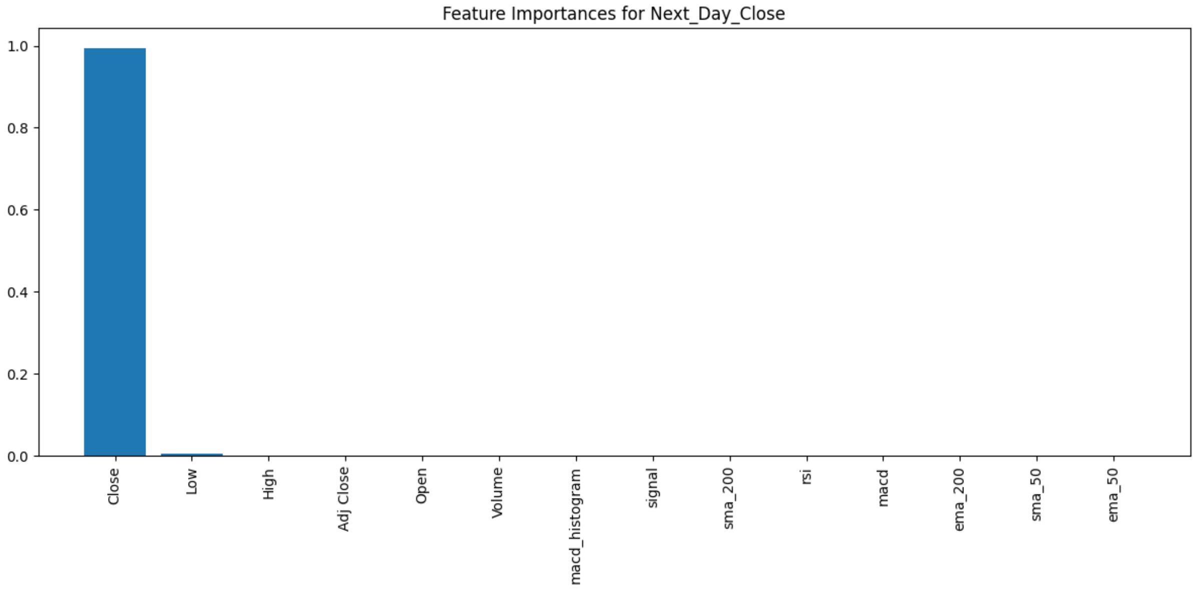
Correlation Analysis



- I have selected 3 target columns:
  - Next_Day_Close
  - Next_Day_rsi
  - Next_Day_macd_histogram
- For these targets, I plotted all other columns in my correlation matrix to create a correlation heatmap.
- **Findings:**
  - **Next_Day_Close:**
    - Has a strong positive correlation with several columns such as Open, High, Low, Close, Adj Close, Volume, SMA_50, EMA_50, SMA_200, and EMA_200.
  - **Next_Day_rsi:**
    - Has a strong positive correlation with RSI.
  - **Next_Day_macd_histogram:**
    - Has a strong positive correlation with MACD Histogram.

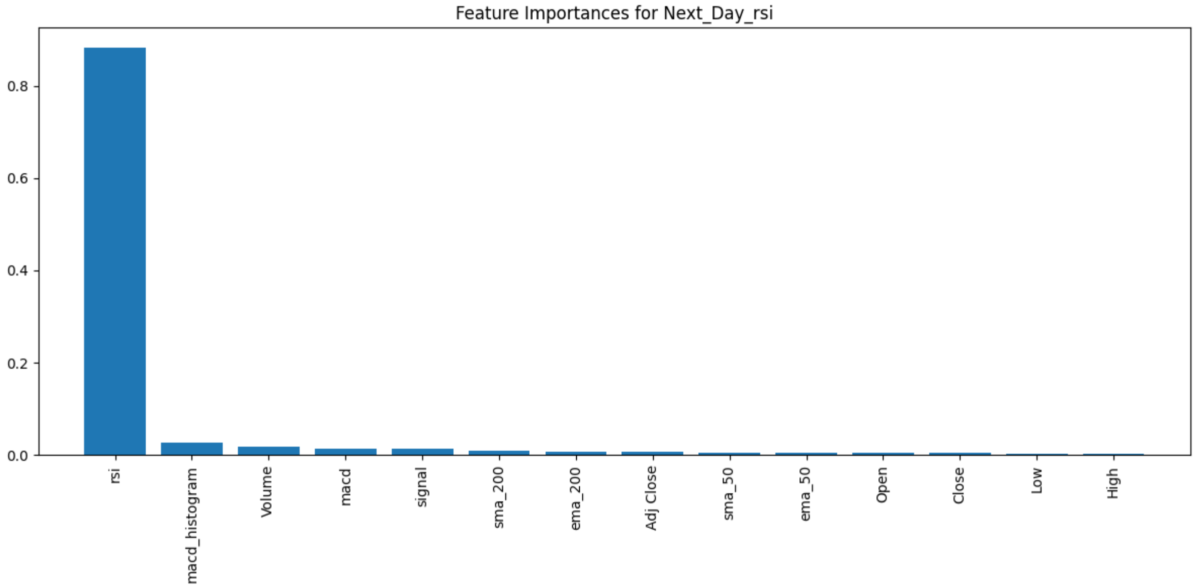
Feature Importance Ranking

- I have selected 3 target columns:
  - Next_Day_Close
  - Next_Day_rsi
  - Next_Day_macd_histogram
- Findings:
  - **Next_Day_Close**



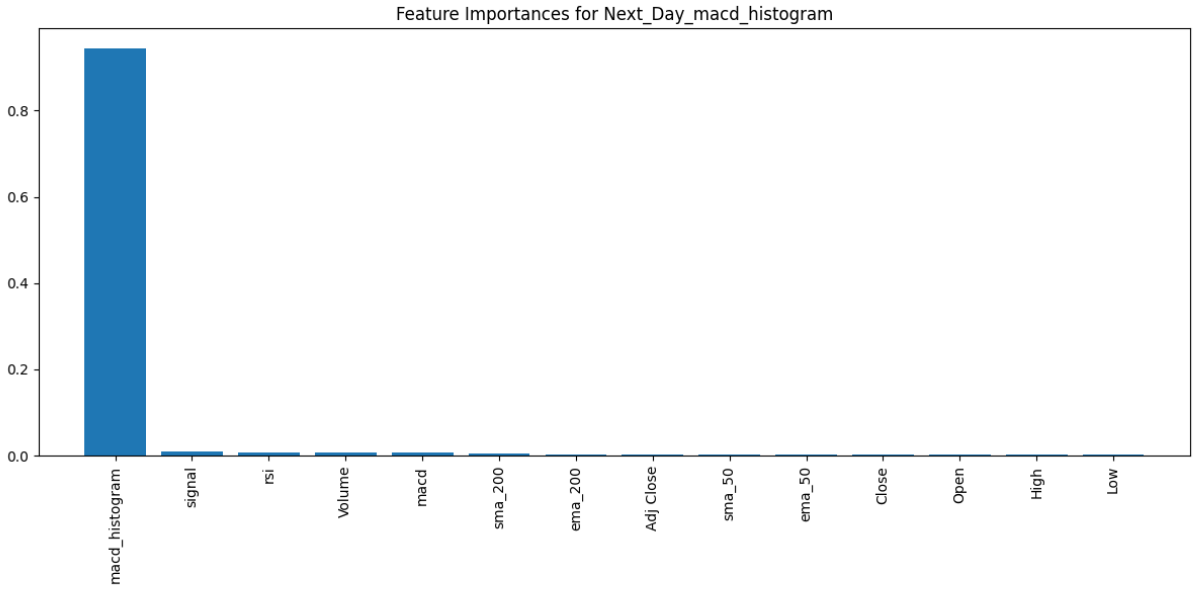
Close is suggested to be most important feature.

  - Next_Day_rsi



RSI is suggested to be most important feature.

- Next_Day_macd_histogram



MACD_histogram is suggested to be most important feature.

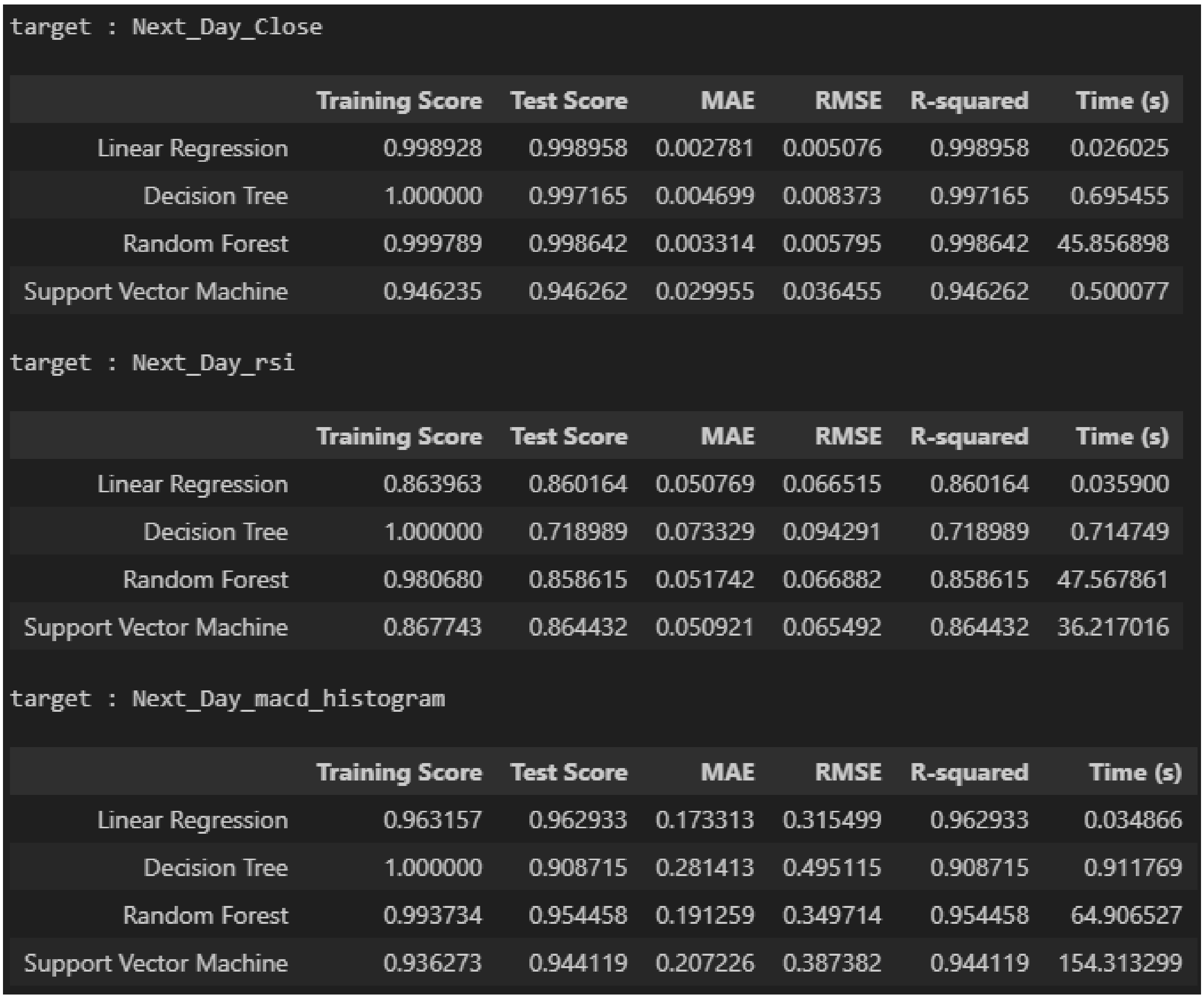
6. Model Development

- **Model Selection**:

  I selected various machine learning models, including Linear Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), to predict stock prices. This diverse approach ensures a comprehensive analysis, leveraging each model's strengths to capture different patterns and relationships in the data, ultimately aiming for more accurate and robust predictions.

- **Model Training**: I used PCA with 10 components to effectively capture relationships in the data and a train-test divide of 80-20%.

- **Model Evaluation**:



```
target : Next_Day_Close
```

	Training Score	Test Score	MAE	RMSE	R-squared	Time (s)
Linear Regression	0.998928	0.998958	0.002781	0.005076	0.998958	0.026025
Decision Tree	1.000000	0.997165	0.004699	0.008373	0.997165	0.695455
Random Forest	0.999789	0.998642	0.003314	0.005795	0.998642	45.856898
Support Vector Machine	0.946235	0.946262	0.029955	0.036455	0.946262	0.500077

```
target : Next_Day_rsi
```

	Training Score	Test Score	MAE	RMSE	R-squared	Time (s)
Linear Regression	0.863963	0.860164	0.050769	0.066515	0.860164	0.035900
Decision Tree	1.000000	0.718989	0.073329	0.094291	0.718989	0.714749
Random Forest	0.980680	0.858615	0.051742	0.066882	0.858615	47.567861
Support Vector Machine	0.867743	0.864432	0.050921	0.065492	0.864432	36.217016

```
target : Next_Day_macd_histogram
```

	Training Score	Test Score	MAE	RMSE	R-squared	Time (s)
Linear Regression	0.963157	0.962933	0.173313	0.315499	0.962933	0.034866
Decision Tree	1.000000	0.908715	0.281413	0.495115	0.908715	0.911769
Random Forest	0.993734	0.954458	0.191259	0.349714	0.954458	64.906527
Support Vector Machine	0.936273	0.944119	0.207226	0.387382	0.944119	154.313299

***Linear Regression has yielded highest results for all targets**

7. Model Interpretation

- **Feature Importance**:

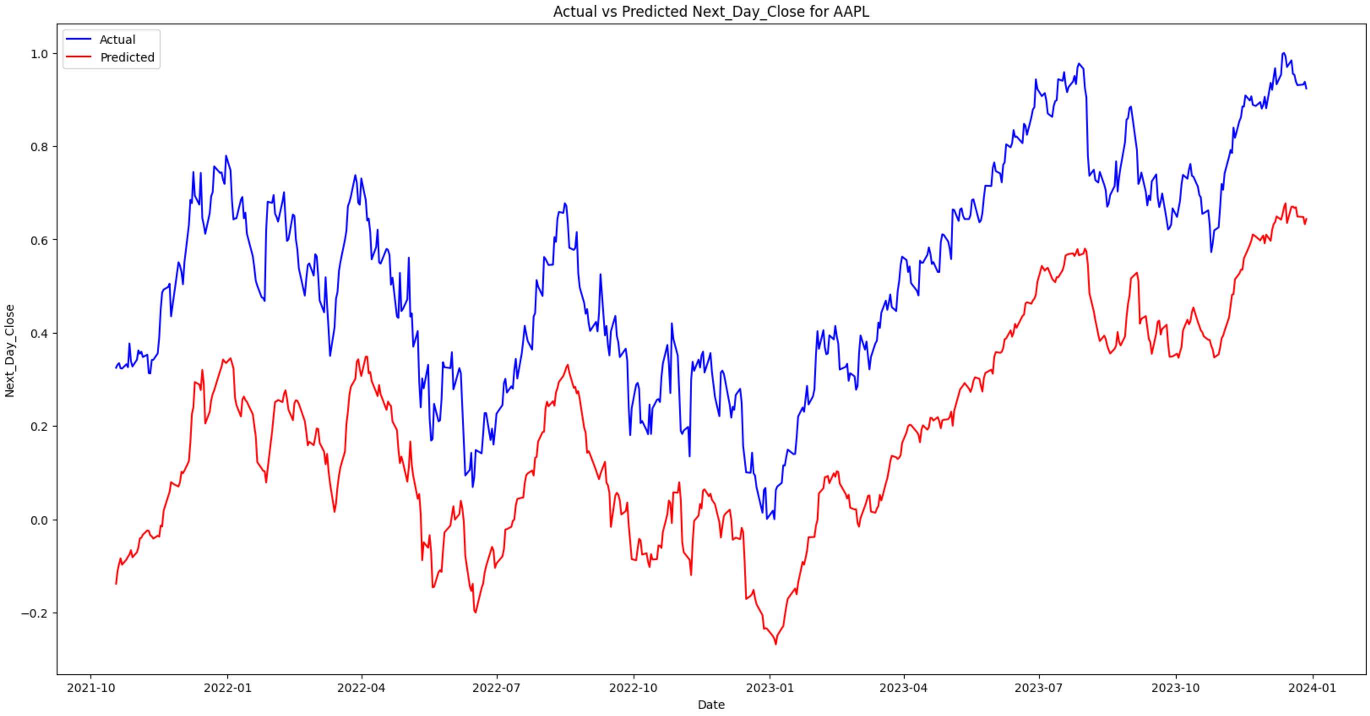
  I selected Principal Component Analysis (PCA) which is a dimensionality reduction technique that transforms features into a set of uncorrelated components, capturing the most variance in the data. This simplifies the dataset, reduces noise, and enhances model performance by focusing on the most important features. I selected PCA to improve feature importance and model accuracy.

- **Model Insights**:

  The highest accuracy of Linear Regression in predicting stock prices indicates strong linear relationships in the data, effective feature selection, and low noise levels. This model's simplicity helps avoid overfitting but may miss complex patterns.

8. Model Deployment

- **Deployment Plan**: I deployed my model in Jupyter Notebook for demonstration purposes, the models successfully captured the trends but accuracy can be enhanced.



- **Monitoring and Maintenance:**
  - **Data Usage and User Feedback:** Monitoring involves using data analytics to track model performance and user feedback to assess sentiment accuracy and user satisfaction.

- **Model Enhancement:**

  To enhance the model:

  - I propose training it on a larger dataset using robust computing resources.
  - Regularly retraining the model with new data will ensure its accuracy and relevance over time.

9. Conclusion

Summary:

- I realized the dynamic and sophisticated nature of Data Science and Machine Learning.
- I also recognized that having a field expert can significantly improve the efficiency and accuracy of large-scale data science projects.

Challenges:

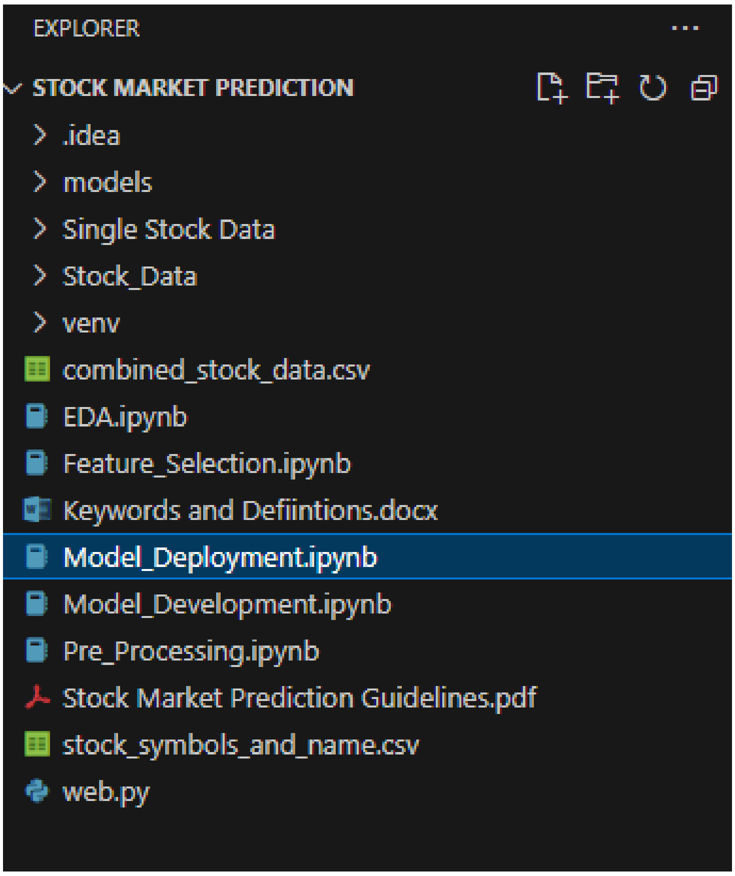
- I had to restart the project multiple times due to my limited knowledge; one incorrect assumption can critically affect the entire project.

Future Work:

- While the stock market follows some general trends, it is greatly affected by current affairs, market reputation, and government decisions. For example, the great fluctuations witnessed during the election results were attributed to mass panic, resulting in an extreme bearish trend and then gradually stabilizing. This provided a short window for investors to make a fortune. To account for such mass sentiment, ChatGPT APIs or other large model capabilities could be used to make the most out of these opportunities.

10. Appendices

- **Additional Visualizations**:



```
STOCK MARKET PREDICTION
  > .idea
  > models
  > Single Stock Data
  > Stock_Data
  > venv
  combined_stock_data.csv
  EDA.ipynb
  Feature_Selection.ipynb
  Keywords and Defiintions.docx
  Model_Deployment.ipynb
  Model_Development.ipynb
  Pre_Processing.ipynb
  Stock Market Prediction Guidelines.pdf
  stock_symbols_and_name.csv
  web.py
```

  - File Structure

- **References**: List any references or resources used in the project.
  - ChatGPT
  - BlackBox
  - YouTube
  - Co-Pilot