**STOCK PRICE PREDICTION**

TEAM MEMBER

212621243027-Santhosh .T

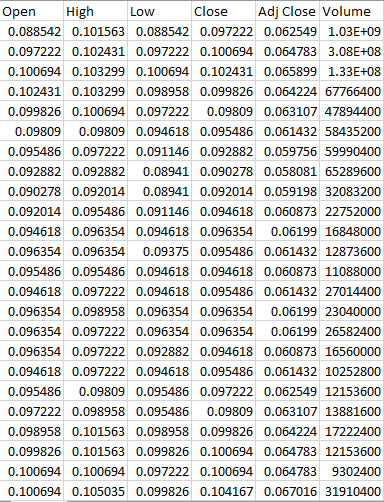
Phase-1 Document Submission

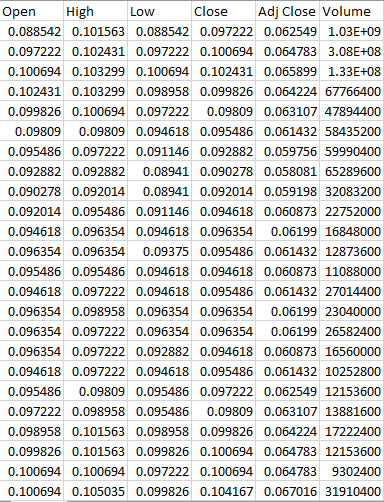
**Problem Definition:**

 The problem is to build a predictive model that forecasts stock prices based on historical market data. The goal is to create a tool that assists investors in making well-informed decisions and optimizing their investment strategies. This project involves data collection, data preprocessing, feature engineering, model selection, training, and evaluation.

**DESIGN THINKING**

1. **Data Collection:**

****

****

**FOR THIS SOURCE DATASET LINK**

[**(https://www.kaggle.com/datasets/prasoonkottarathil/microsoft-lifetime-stocks-dataset)**](file:///C:\Users\SMIT\Desktop\(https:\www.kaggle.com\datasets\prasoonkottarathil\microsoft-lifetime-stocks-dataset))

1. **Data Preprocessing:**

Data preprocessing is a crucial step when using Long Short-Term Memory (LSTM) networks for stock price predictions. Here's a general outline of the preprocessing steps:

* **Data Collection:**

Obtain historical stock price data, usually in the form of time series data, which includes date, open, high, low, close, and volume.

* **Data Cleaning:**

- Handle missing data: Interpolate, forward-fill, or backward-fill missing values.

- Remove outliers that can significantly affect your model's performance.

* **Feature Selection:**

- Decide which features to use. Common features include closing prices, volume, and technical indicators like moving averages, RSI, and MACD**.**

* **Normalization/Scaling:**

- Normalize or scale the data, often between 0 and 1, to ensure that all input features are on a similar scale. This helps the model converge faster and prevents the dominance of one feature over others.

**3. Feature Engineering:**

Feature engineering is a critical part of building accurate and effective LSTM models for various tasks, including stock price prediction. While LSTM networks can capture temporal patterns from raw data, feature engineering can help enhance the model's ability to learn and extract meaningful information. Here are some feature engineering techniques you can apply when working with LSTM models:

* **Lag Features:**

Create lag features by shifting your target variable (e.g., stock price) to previous time steps. These lag features can help the model learn from historical data and capture autocorrelations. For instance, you can use the closing price from the past 'n' days as features for predicting the future price.

* **Moving Averages:**

Compute and include various moving averages (e.g., simple moving average, exponential moving average) over different time windows. These moving averages can capture trends and smooth out noise in the data.

* **Technical Indicators:**

**-** Integrate popular technical indicators such as Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), Bollinger Bands, Stochastic Oscillator, and others. These indicators provide additional information about market sentiment and momentum.

* **Volatility Indicators:**

Include indicators that measure market volatility, like Average True Range (ATR) or historical volatility, to help the model adapt to changing market conditions.

* **Price Rate of Change:**

Calculate the rate of change of prices over a given time period. This can help the model understand the velocity of price movements.

* **Feature Scaling:**

Ensure that all engineered features are scaled or normalized to a similar range to prevent any one feature from dominating the model's learning process.

1. **Model Selection:**

Selecting the right LSTM model for your task involves experimenting with different architectural choices and hyperparameters. Here's a concise guide on how to approach model selection using LSTM:

* **Model Architecture:**

- Start with a basic LSTM model and gradually increase complexity if needed.

- Consider stacking multiple LSTM layers or using Bidirectional LSTMs for capturing more complex patterns.

- Experiment with other recurrent layers like GRU or combinations of LSTM and GRU.

* **Hyperparameter Tuning:**

- Adjust hyperparameters like the number of hidden units, dropout rates, and learning rate. Grid search or random search can be helpful.

* **Sequence Length:**

- Determine the optimal sequence length for input data. Longer sequences may capture more context but can lead to vanishing gradient problems.

* **Batch Size:**

- Experiment with batch sizes; smaller batches can lead to faster convergence, while larger batches can provide better generalization.

* **Learning Rate:**

- Fine-tune the learning rate. A smaller learning rate can help in fine-grained convergence, while a larger one can lead to faster initial training.

* **Regularization:**

- Implement regularization techniques like L1 or L2 regularization to prevent overfitting.

* **Initializations:**

- Try different weight initialization methods, such as Xavier or He initialization, to ensure a balanced learning process.

Remember that the optimal model and hyperparameters can vary depending on the specific problem you're solving and the characteristics of your data. It'sessential to experiment and iterate to find the best LSTM model for your use case.

**PROGRAM:**

pip install -q yfinance

pip install pandas\_datareader

**import** pandas **as** pd

**import** numpy **as** np

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

sns**.**set\_style('whitegrid')

plt**.**style**.**use("fivethirtyeight")

**%matplotlib** inline

*# For reading stock data from yahoo*

**from** pandas\_datareader.data **import** DataReader

**import** yfinance **as** yf

**from** pandas\_datareader **import** data **as** pdr

yf**.**pdr\_override()

*# For time stamps*

**from** datetime **import** datetime

*# The tech stocks we'll use for this analysis*

tech\_list **=** ['AAPL', 'GOOG', 'MSFT', 'AMZN']

*# Set up End and Start times for data grab*

tech\_list **=** ['AAPL', 'GOOG', 'MSFT', 'AMZN']

end **=** datetime**.**now()

start **=** datetime(end**.**year **-** 1, end**.**month, end**.**day)

**for** stock **in** tech\_list:

globals()[stock] **=** yf**.**download(stock, start, end)

company\_list **=** [AAPL, GOOG, MSFT, AMZN]

company\_name **=** ["APPLE", "GOOGLE", "MICROSOFT", "AMAZON"]

**for** company, com\_name **in** zip(company\_list, company\_name):

company["company\_name"] **=** com\_name

df **=** pd**.**concat(company\_list, axis**=**0)

df**.**tail(10)

*# Summary Stats*

AAPL**.**describe()

*# General info*

AAPL**.**info()

*# Let's see a historical view of the closing price*

plt**.**figure(figsize**=**(15, 10))

plt**.**subplots\_adjust(top**=**1.25, bottom**=**1.2)

**for** i, company **in** enumerate(company\_list, 1):

plt**.**subplot(2, 2, i)

company['Adj Close']**.**plot()

plt**.**ylabel('Adj Close')

plt**.**xlabel(**None**)

plt**.**title(f"Closing Price of {tech\_list[i **-** 1]}")

plt**.**tight\_layout()

*# Now let's plot the total volume of stock being traded each day*

plt**.**figure(figsize**=**(15, 10))

plt**.**subplots\_adjust(top**=**1.25, bottom**=**1.2)

**for** i, company **in** enumerate(company\_list, 1):

plt**.**subplot(2, 2, i)

company['Volume']**.**plot()

plt**.**ylabel('Volume')

plt**.**xlabel(**None**)

plt**.**title(f"Sales Volume for {tech\_list[i **-** 1]}")

plt**.**tight\_layout()

ma\_day **=** [10, 20, 50]

**for** ma **in** ma\_day:

**for** company **in** company\_list:

column\_name **=** f"MA for {ma} days"

company[column\_name] **=** company['Adj Close']**.**rolling(ma)**.**mean()

fig, axes **=** plt**.**subplots(nrows**=**2, ncols**=**2)

fig**.**set\_figheight(10)

fig**.**set\_figwidth(15)

AAPL[['Adj Close', 'MA for 10 days', 'MA for 20 days', 'MA for 50 days']]**.**plot(ax**=**axes[0,0])

axes[0,0]**.**set\_title('APPLE')

GOOG[['Adj Close', 'MA for 10 days', 'MA for 20 days', 'MA for 50 days']]**.**plot(ax**=**axes[0,1])

axes[0,1]**.**set\_title('GOOGLE')

MSFT[['Adj Close', 'MA for 10 days', 'MA for 20 days', 'MA for 50 days']]**.**plot(ax**=**axes[1,0])

axes[1,0]**.**set\_title('MICROSOFT')

AMZN[['Adj Close', 'MA for 10 days', 'MA for 20 days', 'MA for 50 days']]**.**plot(ax**=**axes[1,1])

axes[1,1]**.**set\_title('AMAZON')

fig**.**tight\_layout()

*# We'll use pct\_change to find the percent change for each day*

**for** company **in** company\_list:

company['Daily Return'] **=** company['Adj Close']**.**pct\_change()

*# Then we'll plot the daily return percentage*

fig, axes **=** plt**.**subplots(nrows**=**2, ncols**=**2)

fig**.**set\_figheight(10)

fig**.**set\_figwidth(15)

AAPL['Daily Return']**.**plot(ax**=**axes[0,0], legend**=True**, linestyle**=**'--', marker**=**'o')

axes[0,0]**.**set\_title('APPLE')

GOOG['Daily Return']**.**plot(ax**=**axes[0,1], legend**=True**, linestyle**=**'--', marker**=**'o')

axes[0,1]**.**set\_title('GOOGLE')

MSFT['Daily Return']**.**plot(ax**=**axes[1,0], legend**=True**, linestyle**=**'--', marker**=**'o')

axes[1,0]**.**set\_title('MICROSOFT')

AMZN['Daily Return']**.**plot(ax**=**axes[1,1], legend**=True**, linestyle**=**'--', marker**=**'o')

axes[1,1]**.**set\_title('AMAZON')

fig**.**tight\_layout()

plt**.**figure(figsize**=**(12, 9))

**for** i, company **in** enumerate(company\_list, 1):

plt**.**subplot(2, 2, i)

company['Daily Return']**.**hist(bins**=**50)

plt**.**xlabel('Daily Return')

plt**.**ylabel('Counts')

plt**.**title(f'{company\_name[i **-** 1]}')

plt**.**tight\_layout()

*# Grab all the closing prices for the tech stock list into one DataFrame*

closing\_df **=** pdr**.**get\_data\_yahoo(tech\_list, start**=**start, end**=**end)['Adj Close']

*# Make a new tech returns DataFrame*

tech\_rets **=** closing\_df**.**pct\_change()

tech\_rets**.**head()

*# Comparing Google to itself should show a perfectly linear relationship*

sns**.**jointplot(x**=**'GOOG', y**=**'GOOG', data**=**tech\_rets, kind**=**'scatter', color**=**'seagreen')

*# We can simply call pairplot on our DataFrame for an automatic visual analysis*

*# of all the comparisons*

sns**.**pairplot(tech\_rets, kind**=**'reg')

*# Set up our figure by naming it returns\_fig, call PairPLot on the DataFrame*

return\_fig **=** sns**.**PairGrid(tech\_rets**.**dropna())

*# Using map\_upper we can specify what the upper triangle will look like.*

return\_fig**.**map\_upper(plt**.**scatter, color**=**'purple')

*# We can also define the lower triangle in the figure, inclufing the plot type (kde)*

*# or the color map (BluePurple)*

return\_fig**.**map\_lower(sns**.**kdeplot, cmap**=**'cool\_d')

*# Finally we'll define the diagonal as a series of histogram plots of the daily return*

return\_fig**.**map\_diag(plt**.**hist, bins**=**30)

*# Set up our figure by naming it returns\_fig, call PairPLot on the DataFrame*

returns\_fig **=** sns**.**PairGrid(closing\_df)

*# Using map\_upper we can specify what the upper triangle will look like.*

returns\_fig**.**map\_upper(plt**.**scatter,color**=**'purple')

*# We can also define the lower triangle in the figure, inclufing the plot type (kde) or the color map (BluePurple)*

returns\_fig**.**map\_lower(sns**.**kdeplot,cmap**=**'cool\_d')

*# Finally we'll define the diagonal as a series of histogram plots of the daily return*

returns\_fig**.**map\_diag(plt**.**hist,bins**=**30)

plt**.**figure(figsize**=**(12, 10))

plt**.**subplot(2, 2, 1)

sns**.**heatmap(tech\_rets**.**corr(), annot**=True**, cmap**=**'summer')

plt**.**title('Correlation of stock return')

plt**.**subplot(2, 2, 2)

sns**.**heatmap(closing\_df**.**corr(), annot**=True**, cmap**=**'summer')

plt**.**title('Correlation of stock closing price')

rets **=** tech\_rets**.**dropna()

area **=** np**.**pi **\*** 20

plt**.**figure(figsize**=**(10, 8))

plt**.**scatter(rets**.**mean(), rets**.**std(), s**=**area)

plt**.**xlabel('Expected return')

plt**.**ylabel('Risk')

**for** label, x, y **in** zip(rets**.**columns, rets**.**mean(), rets**.**std()):

plt**.**annotate(label, xy**=**(x, y), xytext**=**(50, 50), textcoords**=**'offset points', ha**=**'right', va**=**'bottom',

arrowprops**=**dict(arrowstyle**=**'-', color**=**'blue', connectionstyle**=**'arc3,rad=-0.3'))

*# Get the stock quote*

df **=** pdr**.**get\_data\_yahoo('AAPL', start**=**'2012-01-01', end**=**datetime**.**now())

*# Show teh data*

Df

plt**.**figure(figsize**=**(16,6))

plt**.**title('Close Price History')

plt**.**plot(df['Close'])

plt**.**xlabel('Date', fontsize**=**18)

plt**.**ylabel('Close Price USD ($)', fontsize**=**18)

plt**.**show()

*# Create a new dataframe with only the 'Close column*

data **=** df**.**filter(['Close'])

*# Convert the dataframe to a numpy array*

dataset **=** data**.**values

*# Get the number of rows to train the model on*

training\_data\_len **=** int(np**.**ceil( len(dataset) **\*** .95 ))

training\_data\_len

*# Scale the data*

**from** sklearn.preprocessing **import** MinMaxScaler

scaler **=** MinMaxScaler(feature\_range**=**(0,1))

scaled\_data **=** scaler**.**fit\_transform(dataset)

scaled\_data

*# Create the training data set*

*# Create the scaled training data set*

train\_data **=** scaled\_data[0:int(training\_data\_len), :]

*# Split the data into x\_train and y\_train data sets*

x\_train **=** []

y\_train **=** []

**for** i **in** range(60, len(train\_data)):

x\_train**.**append(train\_data[i**-**60:i, 0])

y\_train**.**append(train\_data[i, 0])

**if** i**<=** 61:

print(x\_train)

print(y\_train)

print()

*# Convert the x\_train and y\_train to numpy arrays*

x\_train, y\_train **=** np**.**array(x\_train), np**.**array(y\_train)

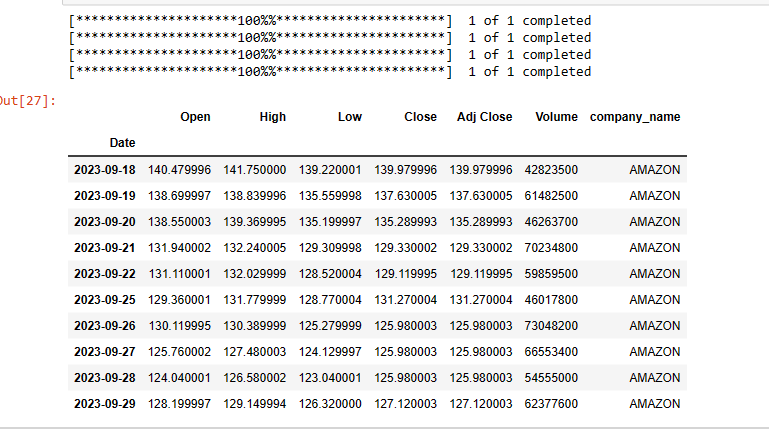
*# Reshape the data*

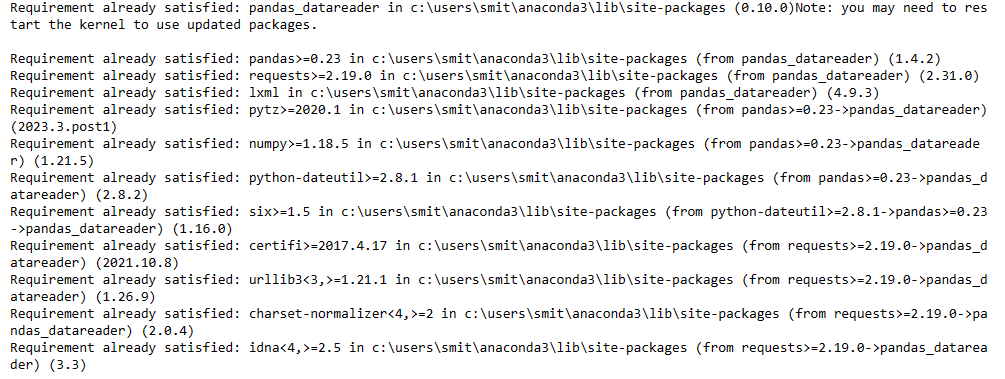
x\_train **=** np**.**reshape(x\_train, (x\_train**.**shape[0], x\_train**.**shape[1], 1))

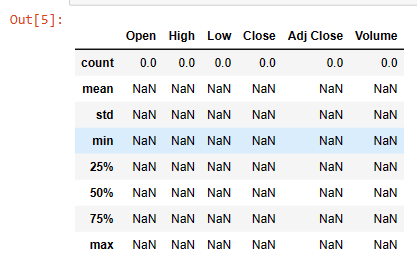
*# x\_train.shape*

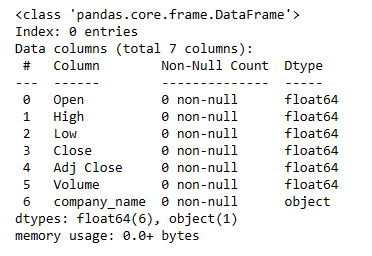
**Output:**

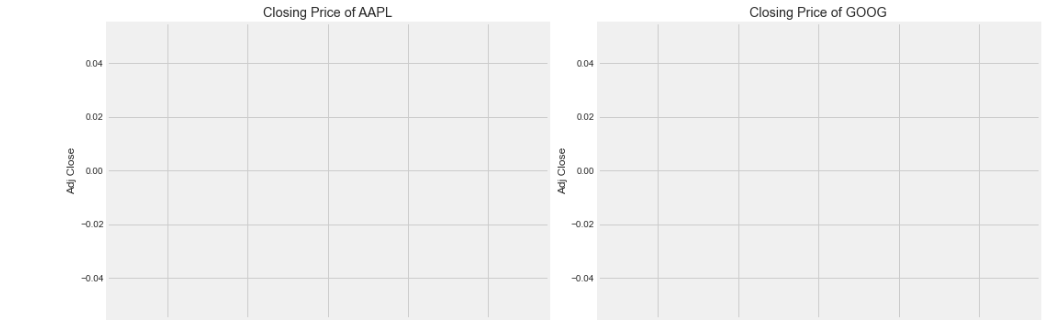
****

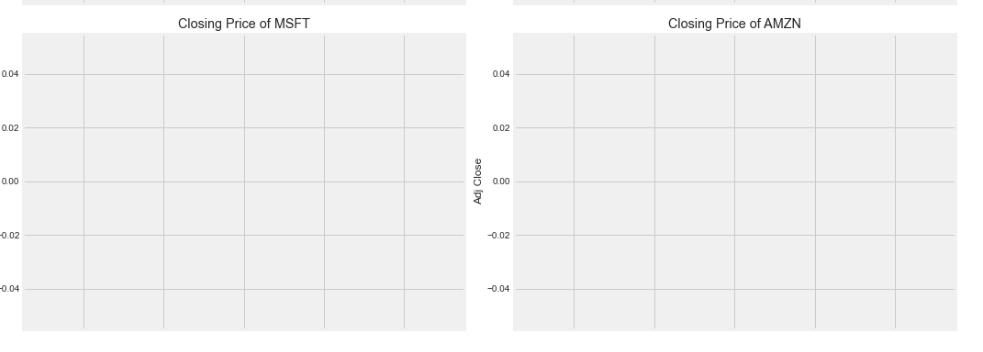
****

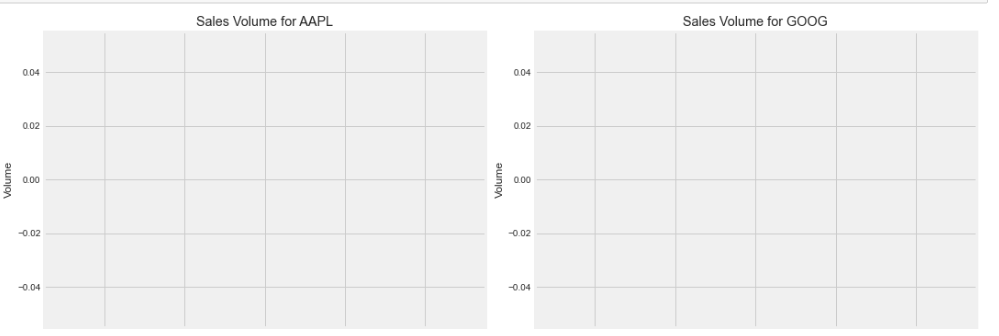
****

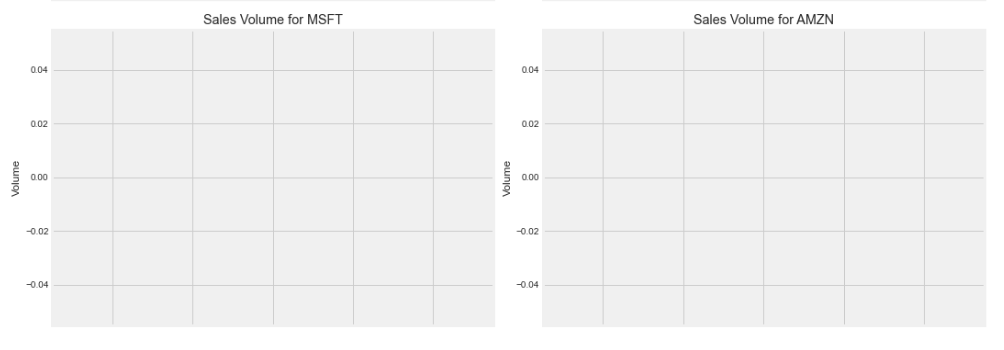
****

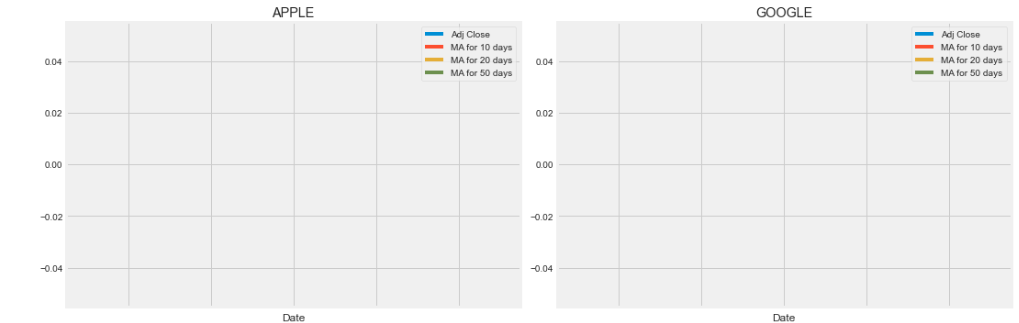
****

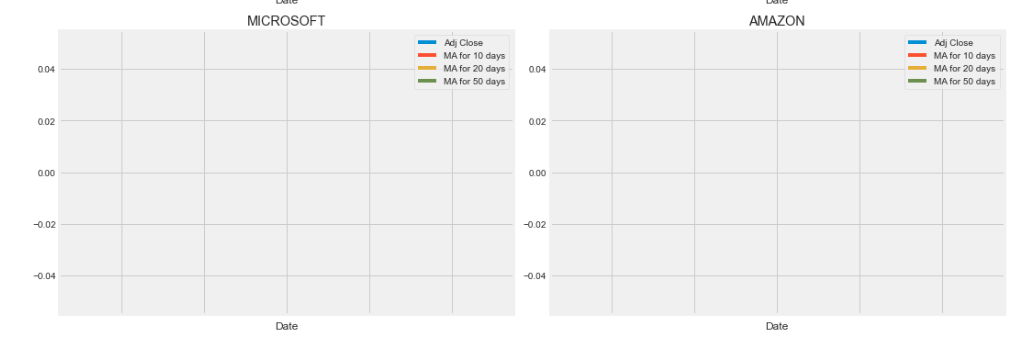
****

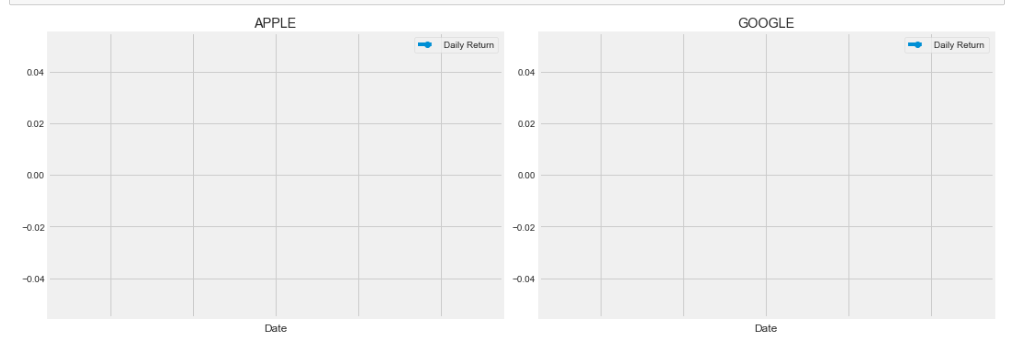
****

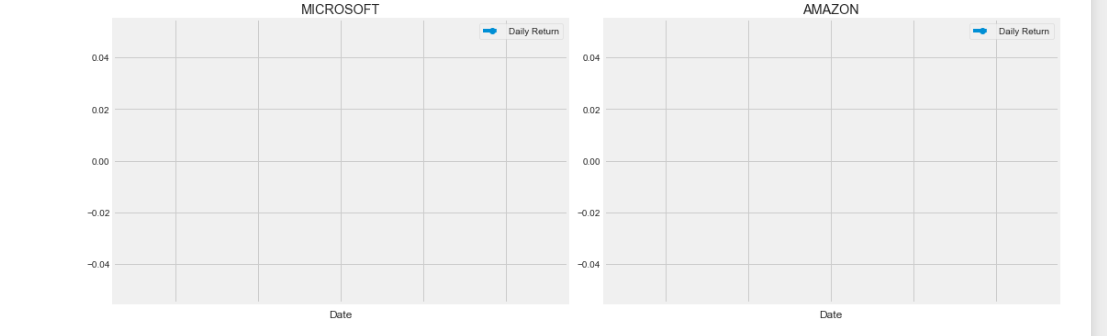
****

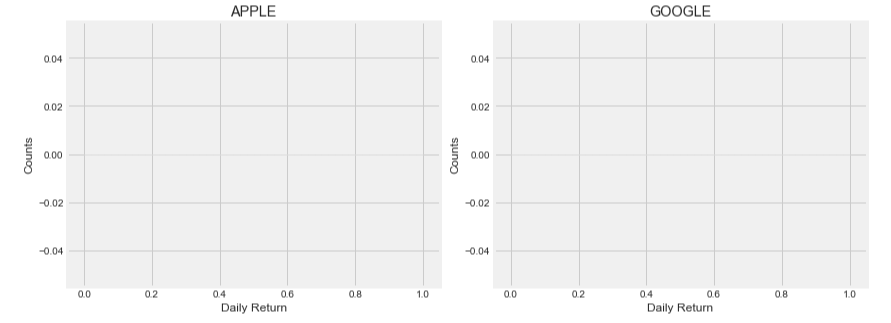
****

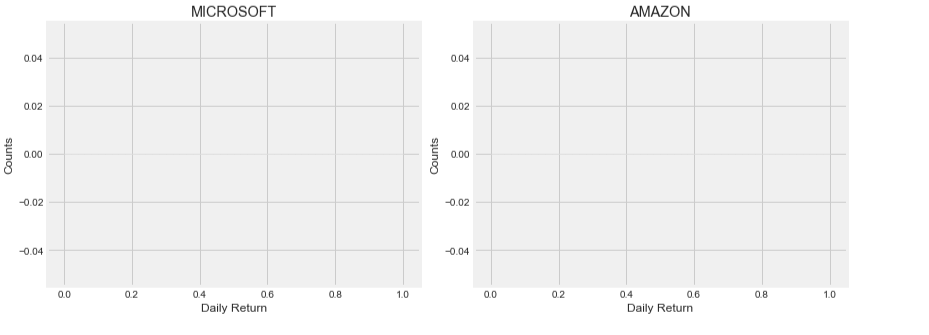
****

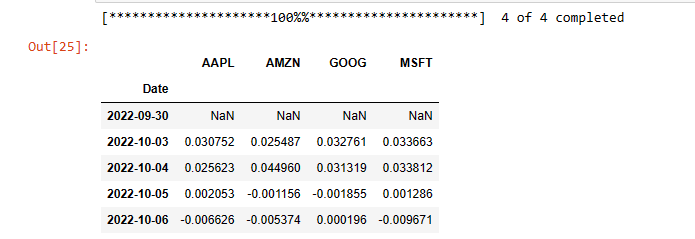
****

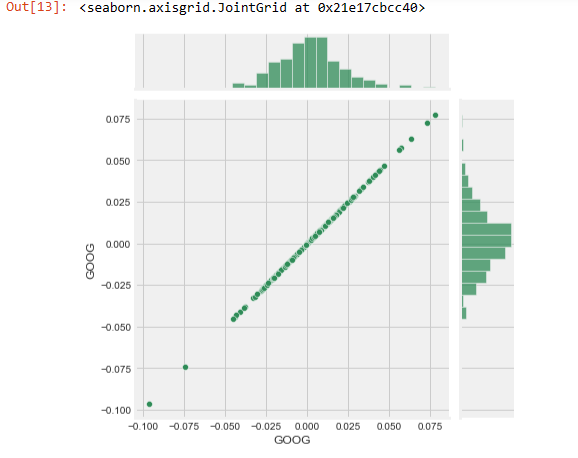
****

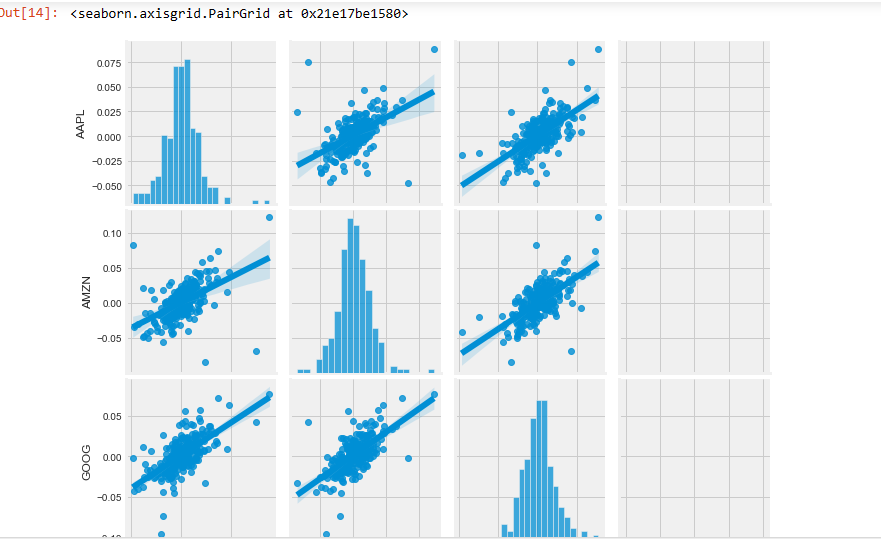
****

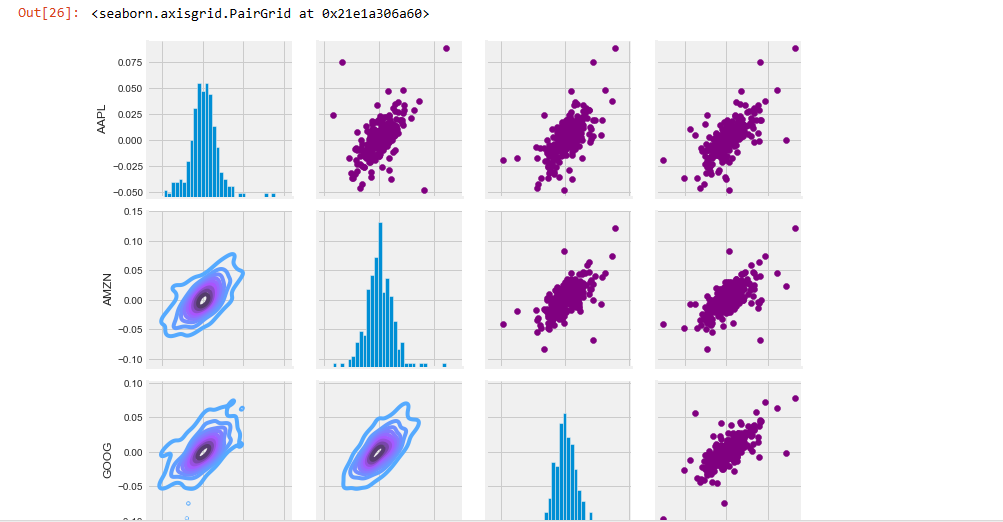
****

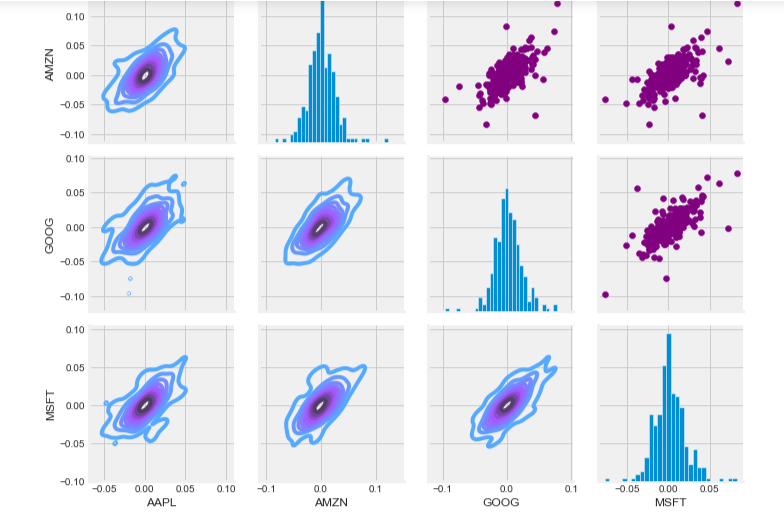
****

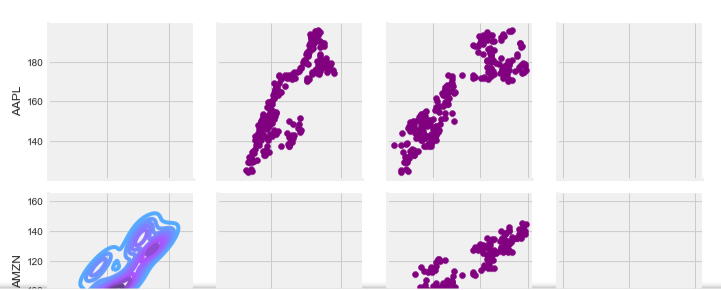
****

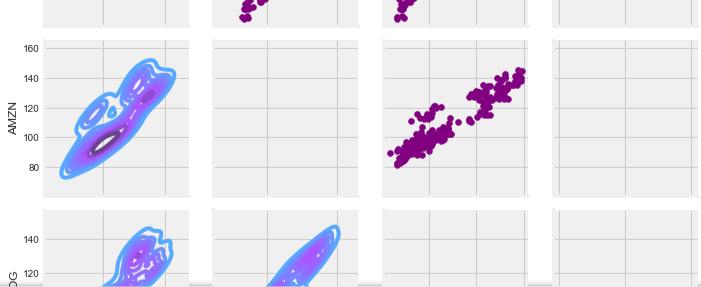
****

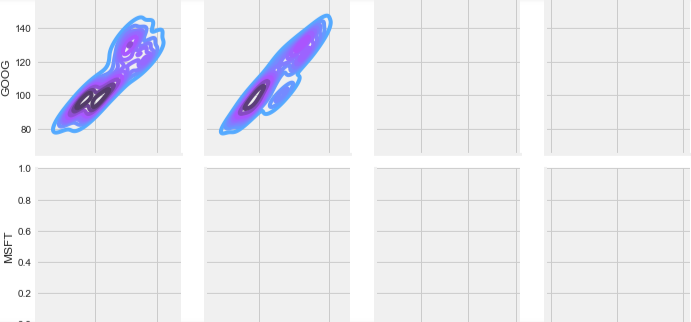
****

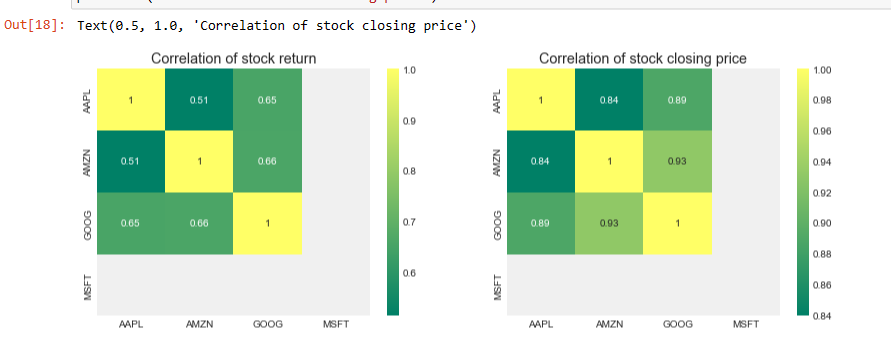
****

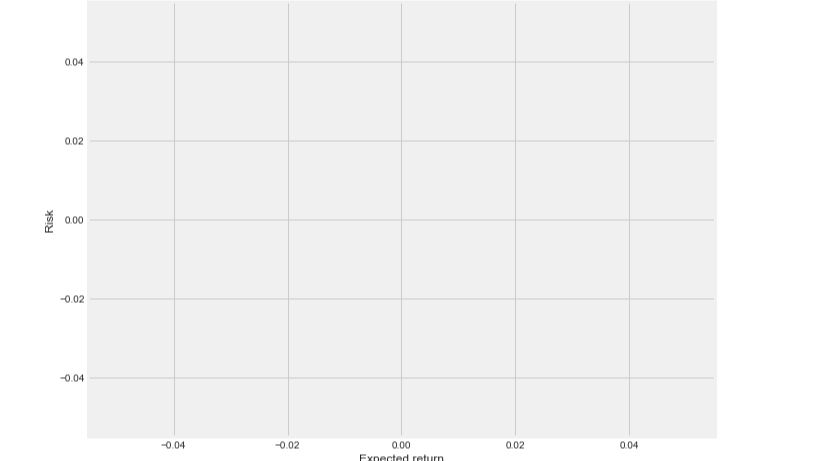
****

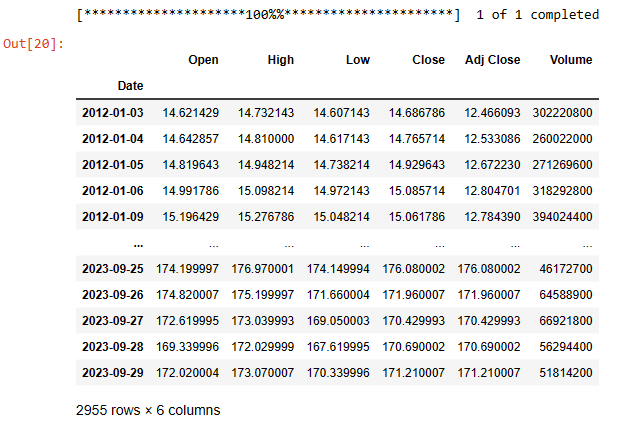
****

****

****

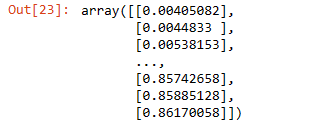
****

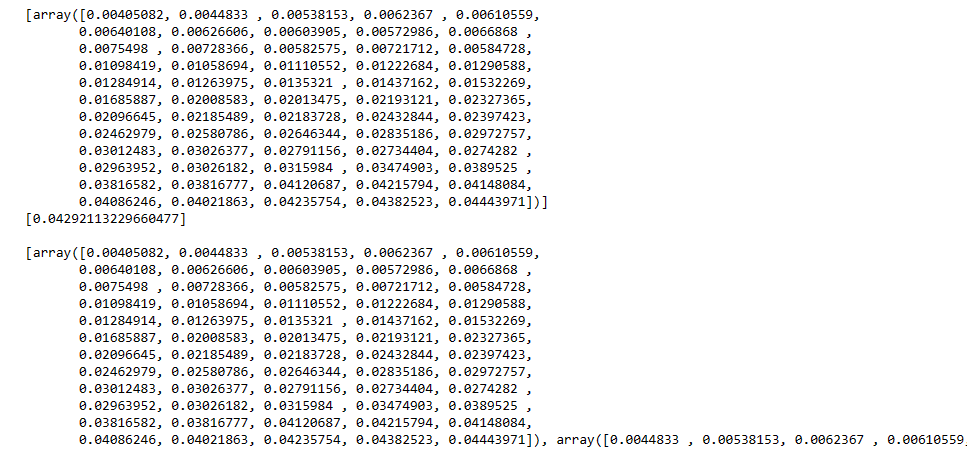
****

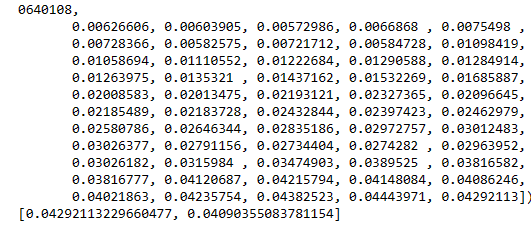
****

****

****

****

****

****

1. **Model Training:**

Train your LSTM model with the preprocessed data. Experiment with different architectures, hyperparameters, and optimization techniques.

1. **Evaluation:**

Evaluate the model's performance using appropriate metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and visualizations like price predictions vs. actual prices.\

1. **CONCLUSION:**

In Phase 1, we have established a clear understanding of our goal: STOCK PRICE PREDICTION USING LSTM We outlined a structured approach that includes data source selection, data preprocessing, feature selection, model selection, model training, and evaluation. This sets the stage for our project's successful execution in subsequent phases