Chapter 4

Initial Findings

4.1 Overview

This chapter covered how sentiment analysis is integrated for forecasting stock markets with machine learning. The first step is doing exploratory data analysis (EDA) to see the trends of stock prices, how sentiment is distributed, and the correlations. The stock price dataset preprocessed; feature engineering applied by adding lagged prices, rolling statistics, and sentiment indicators based on sentiment. The analysis of financial news sentiment leads to the classification of the market sentiment as positive, neutral, or negative. Finally, a machine learning model is used to assess the cost of stock movements over sentiment: how news sentiment influences market trends.

4.2 Exploratory Data Analysis (EDA)

The initial stage of every data analysis project is to carry out exploratory data analysis. Exploratory Data Analysis is very important to do before the modeling stage. Exploratory Data Analysis (EDA) can be briefly interpreted as a process of understanding data to obtain as much information as possible. In addition, EDA can also be done to understand data patterns. In our project, the EDA is started by analyzing stock price and sentiment data to have a better understanding of the relationship between stock prices and sentiments, and ultimately prediction of price.

EDA generally comprises the following main activities:

Data Summary: The first activity was to summarize the datasets so we could analyze the structure, the types of features available, and understand about any missing/erroneous data. Averages,

medians, standard deviations, etc. were basic statistics considered in understanding the distribution of stock prices and sentiment scores.

Sentiment Distribution: How sentiment labels (positive, neutral, negative) were distributed across the dataset concerned to see whether there was any other source of bias in the sentiments used and how well the corresponding distribution of sentiments matched our expectations.

Stock Price Analysis: It involves checking the trends, patterns, or even seasonality in the stock prices of the company over time. How the stock price behave in different time period such as in a market crash or in an economic growth period also needs to be looked at.

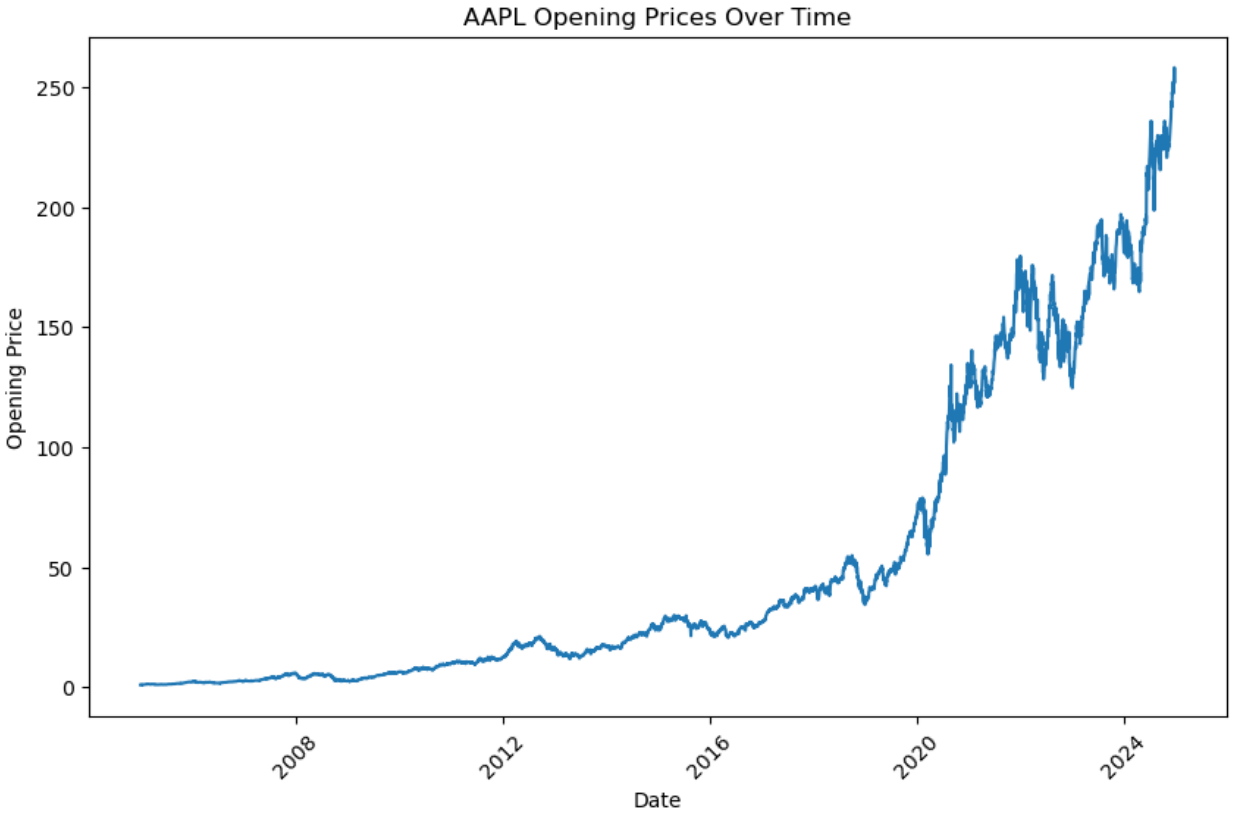
Sentiment Analysis: How sentiments are changing movements regarding stock price throughput-time was analyzed, as these could best reflect the market events that could migrate the sentiments and eventually send the stock prices tumbling or soaring like the earnings reports and political turmoil.

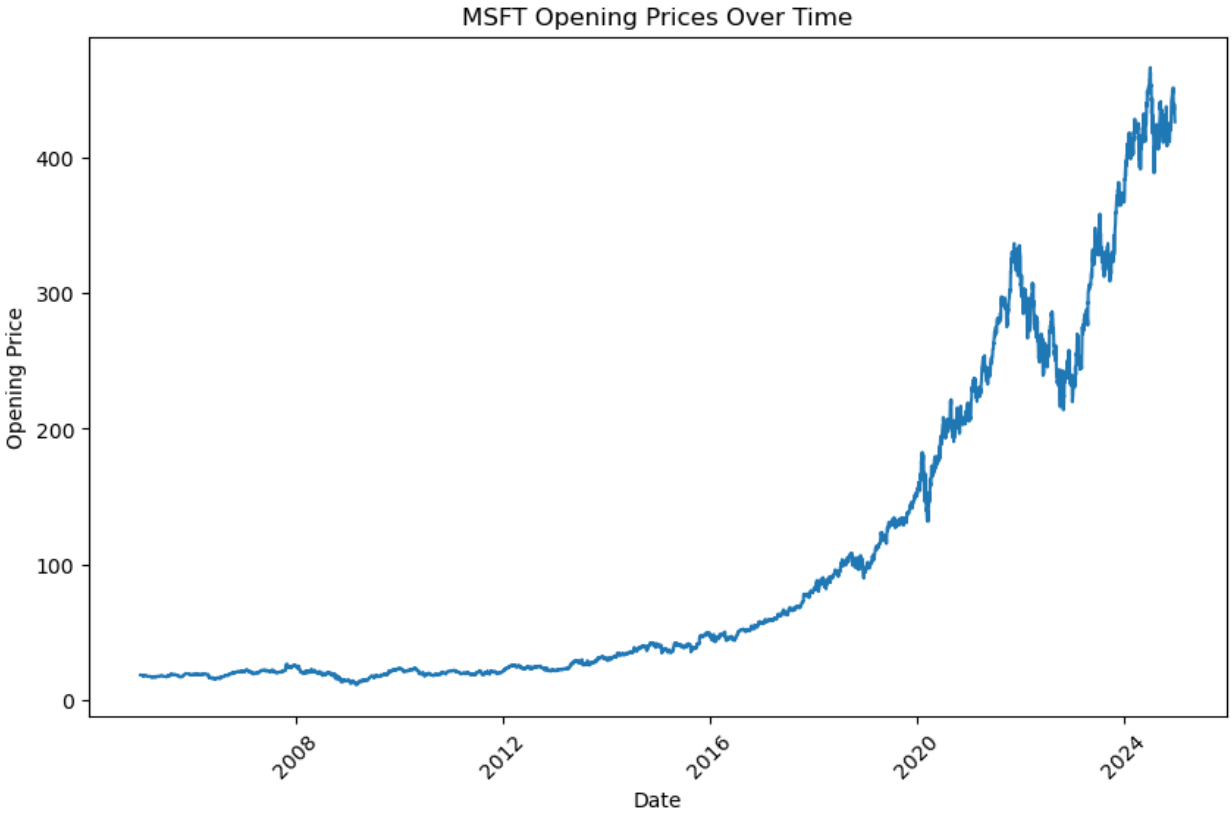
Correlation Analysis: We computed the correlations between sentiment data with stock price movements, particularly focusing on how sentiment influences stock volatility and price returns. It made it possible to determine if such data could act as a predictor for stock price trends.

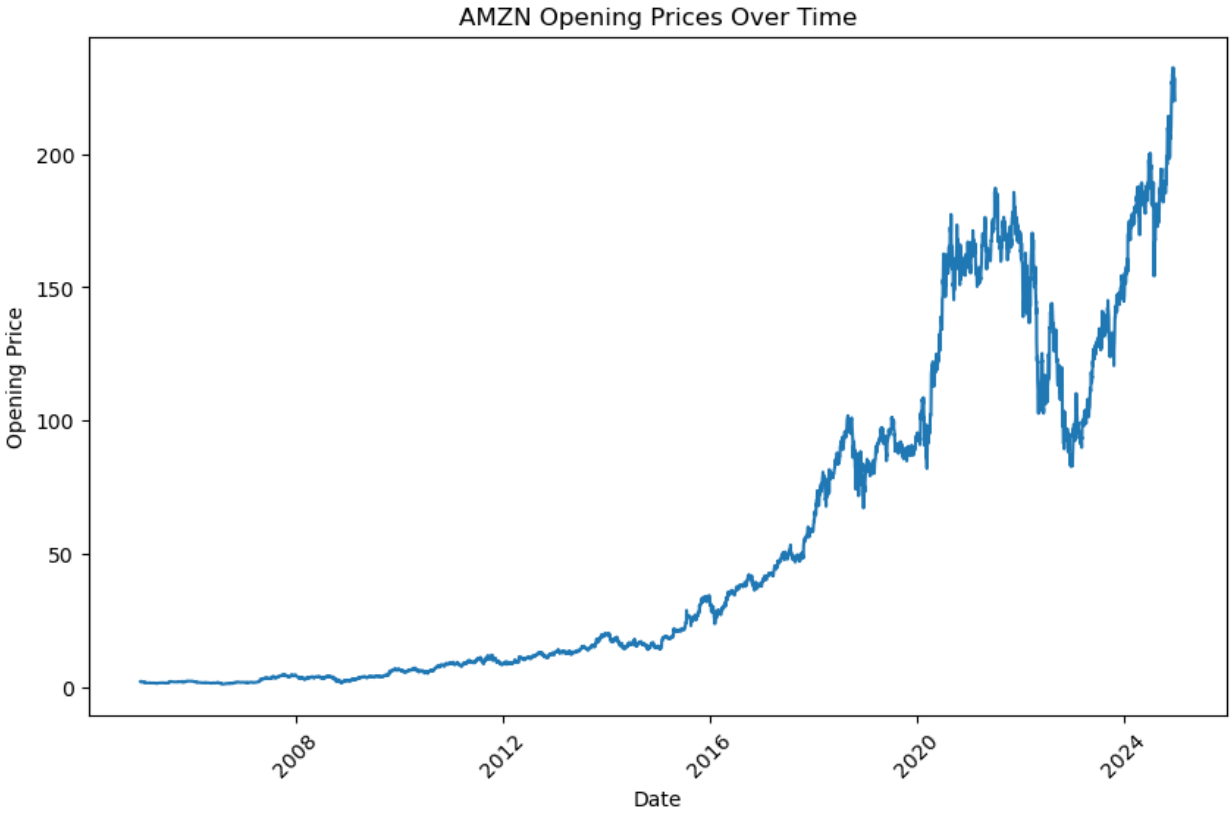
4.2.1 Understanding the stock price dataset

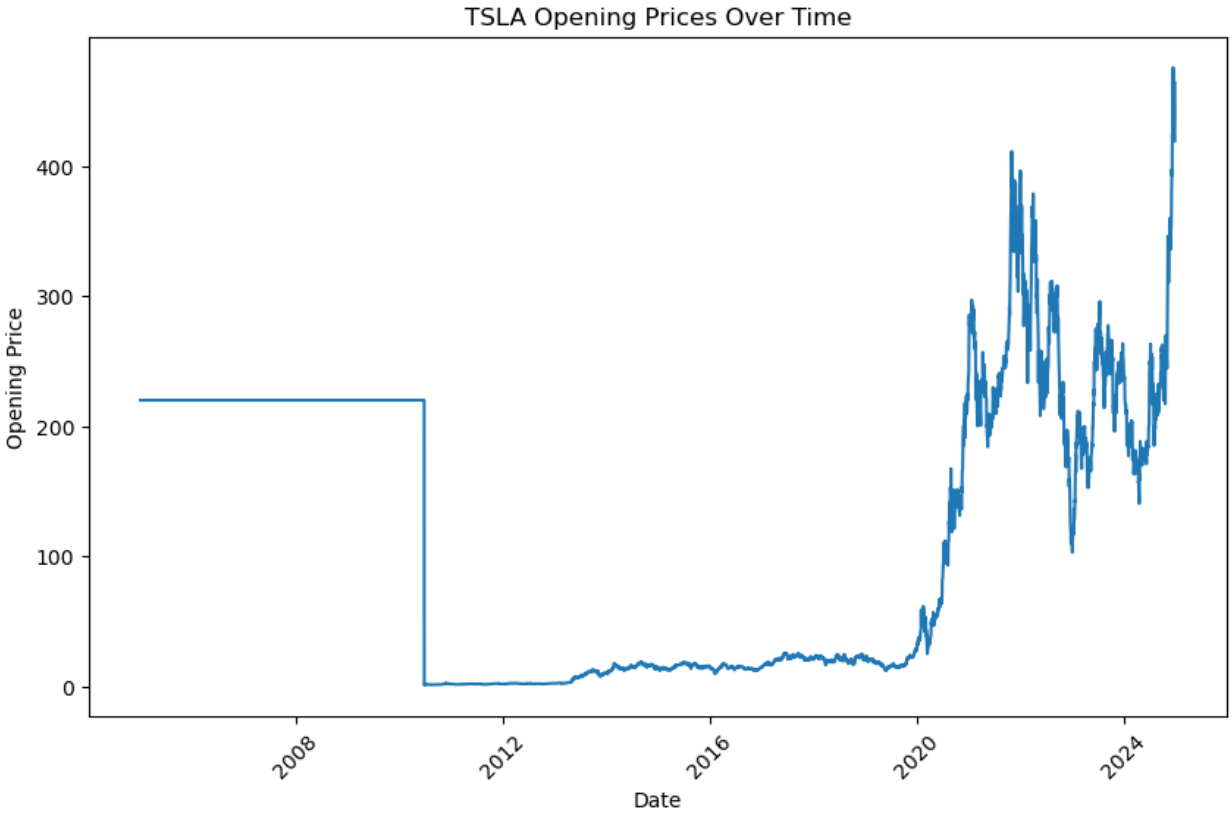
a. Line Plots for Opening Prices:

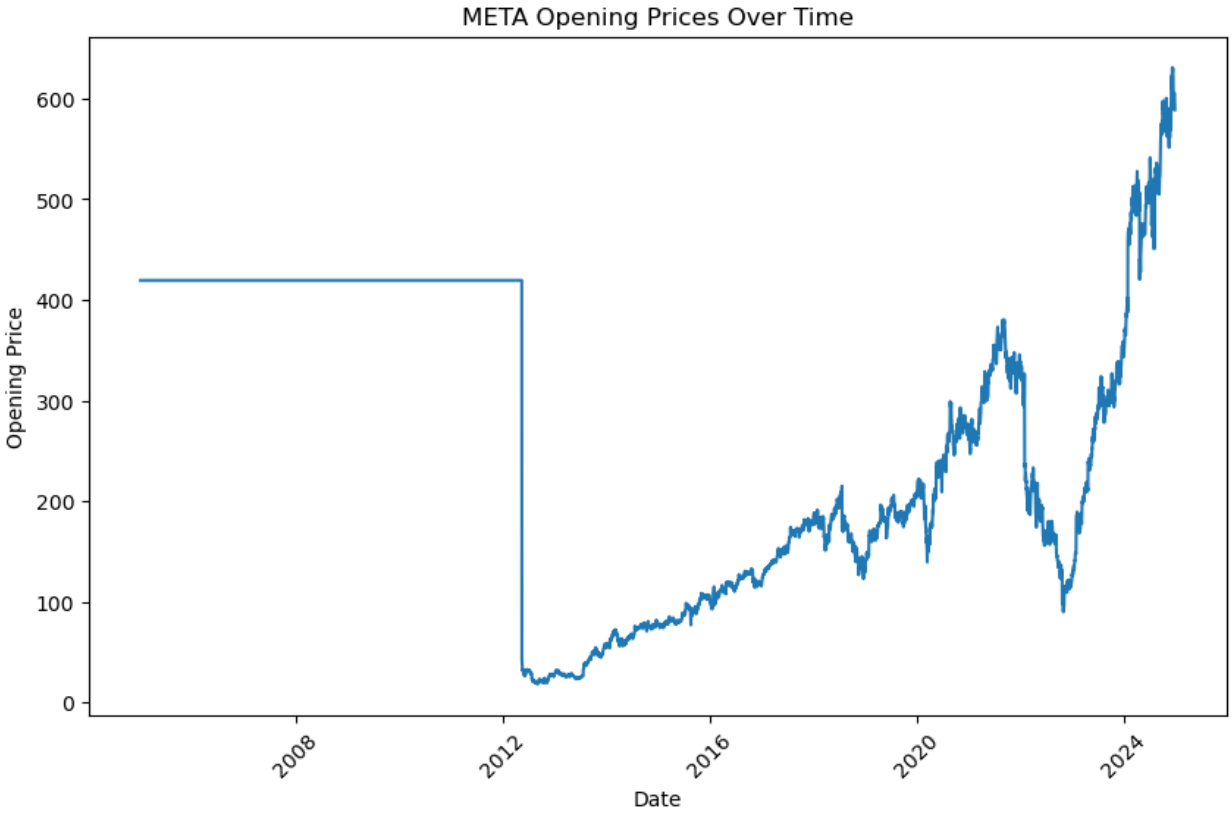
The objective of this line plot is to portray the opening price movements of all firms across a span of time. The data depict the changes in the opening prices over the years for each firm. Trends can tell how the company's stock performed based on years, thus coming with periods of growth, decline, or stability.

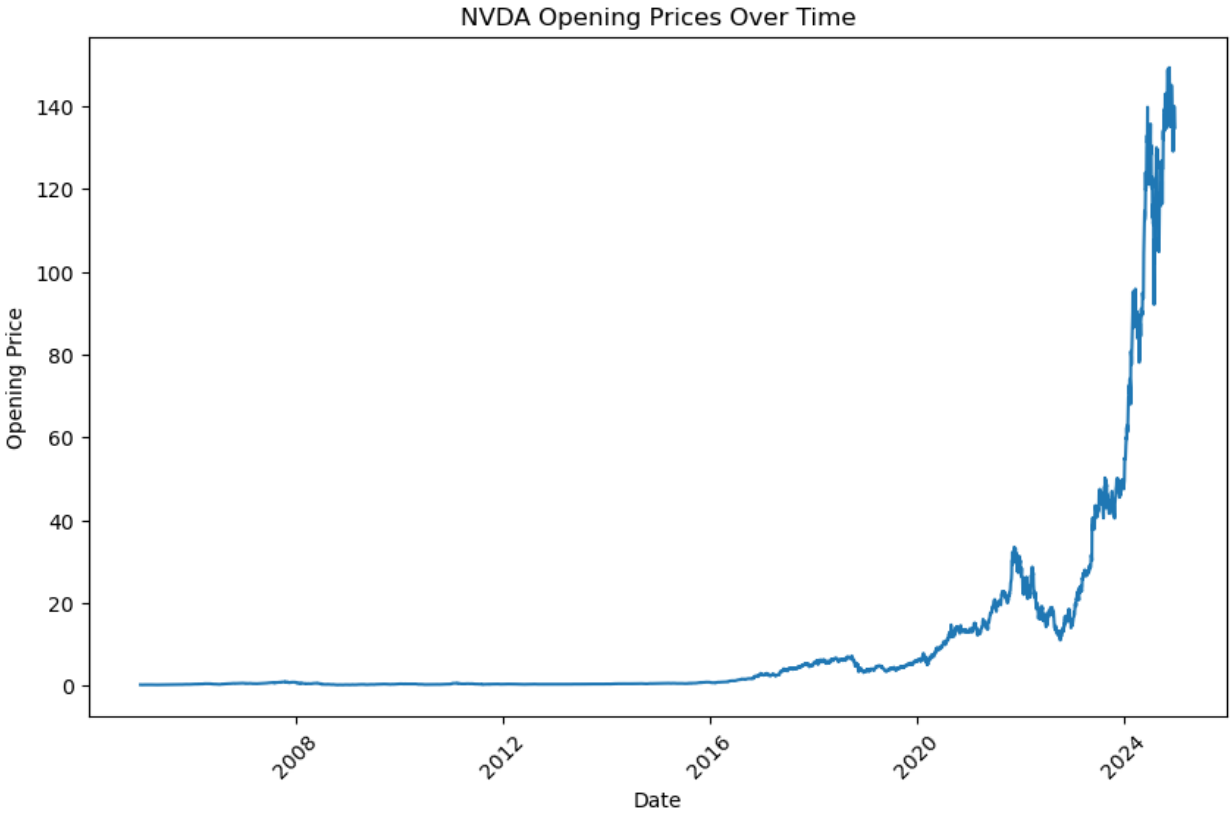


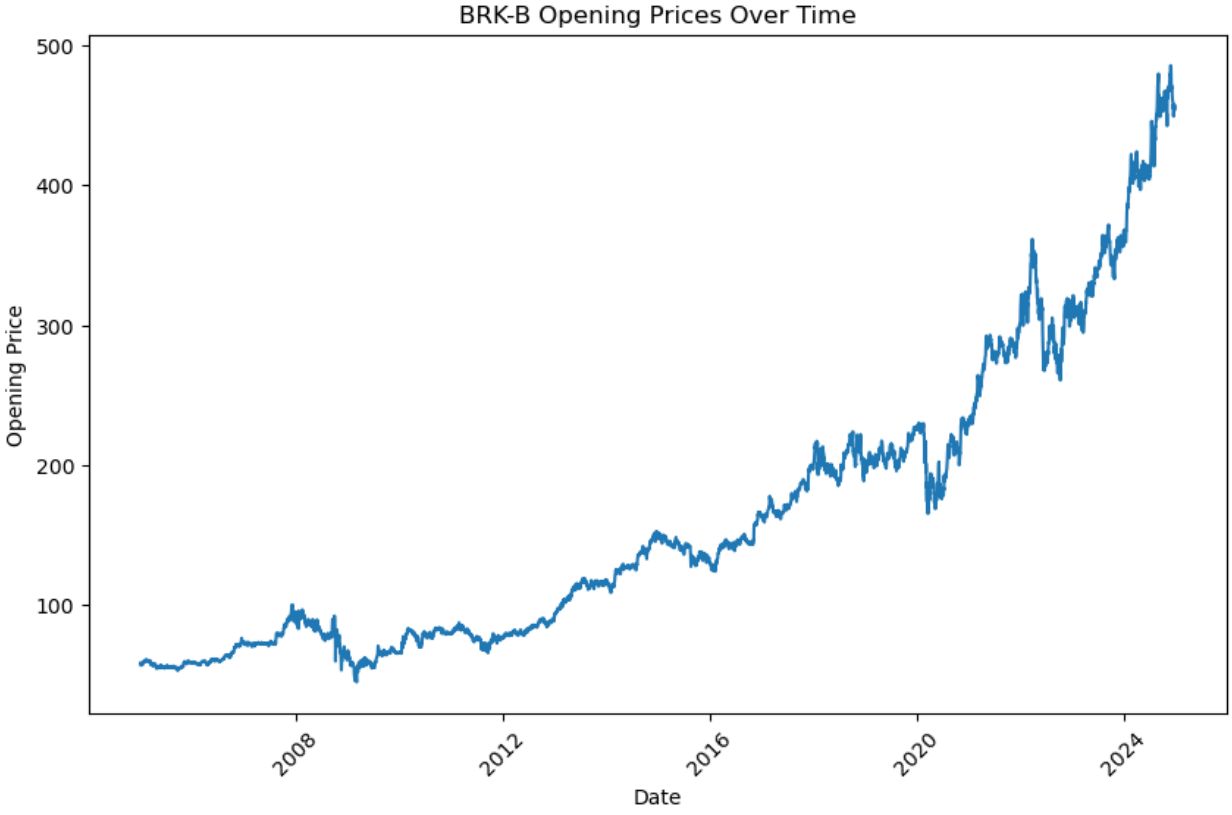

MSFT Opening Prices Over Time
Opening Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

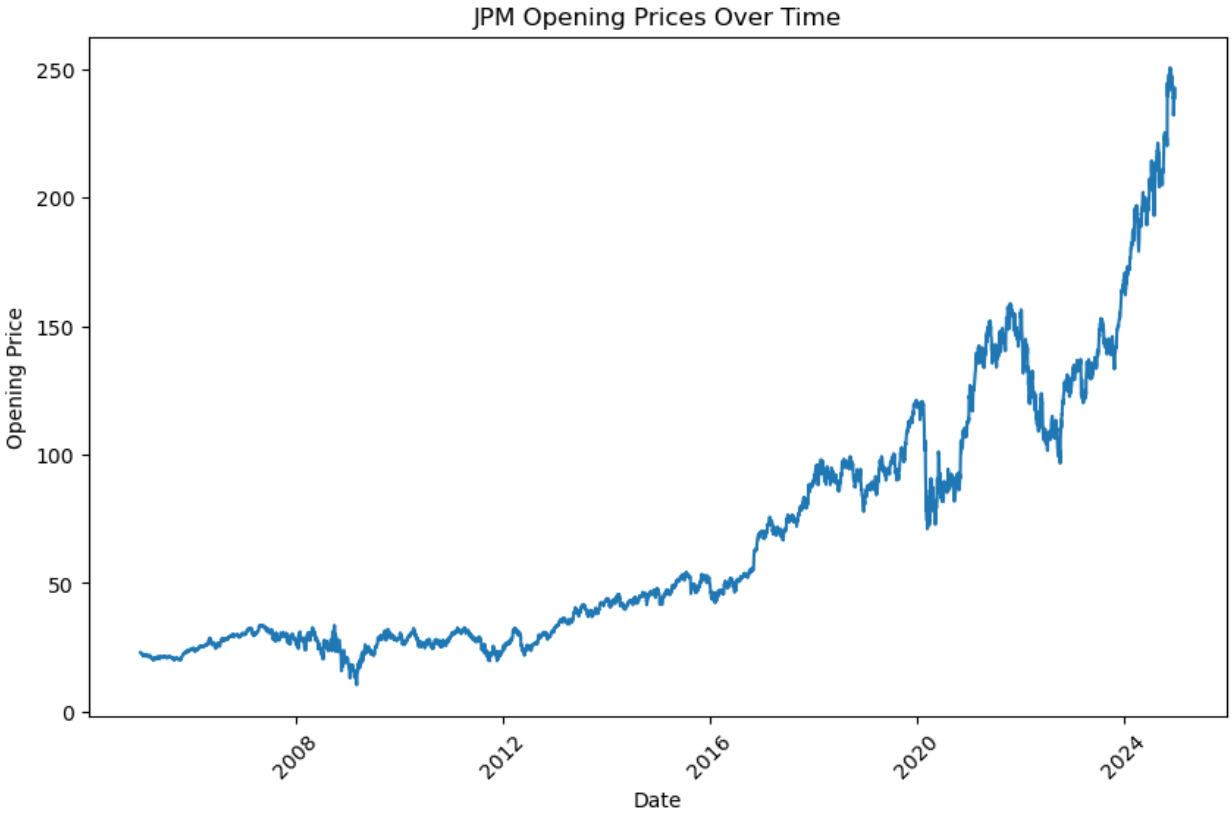

AMZN Opening Prices Over Time
Opening Price
0
50
100
150
200
2008
2012
2016
2020
2024
Date

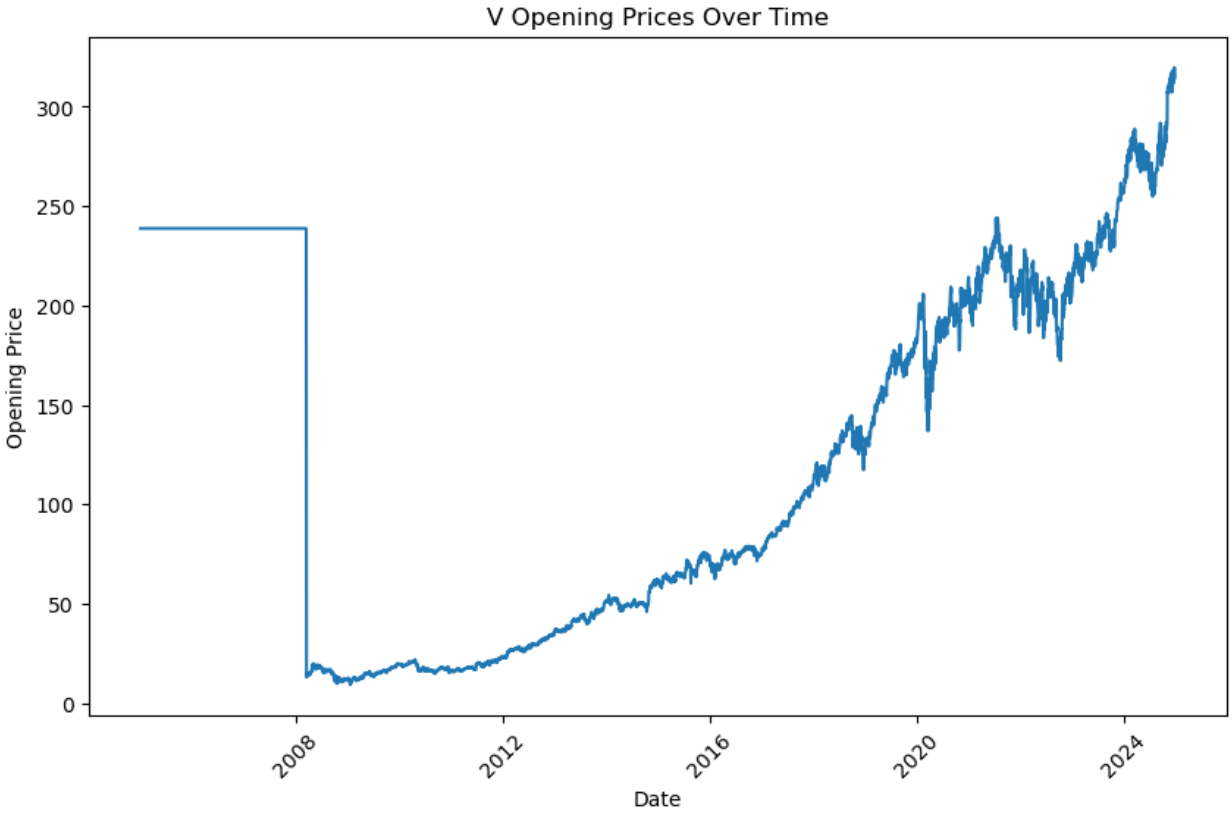

TSLA Opening Prices Over Time
Opening Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

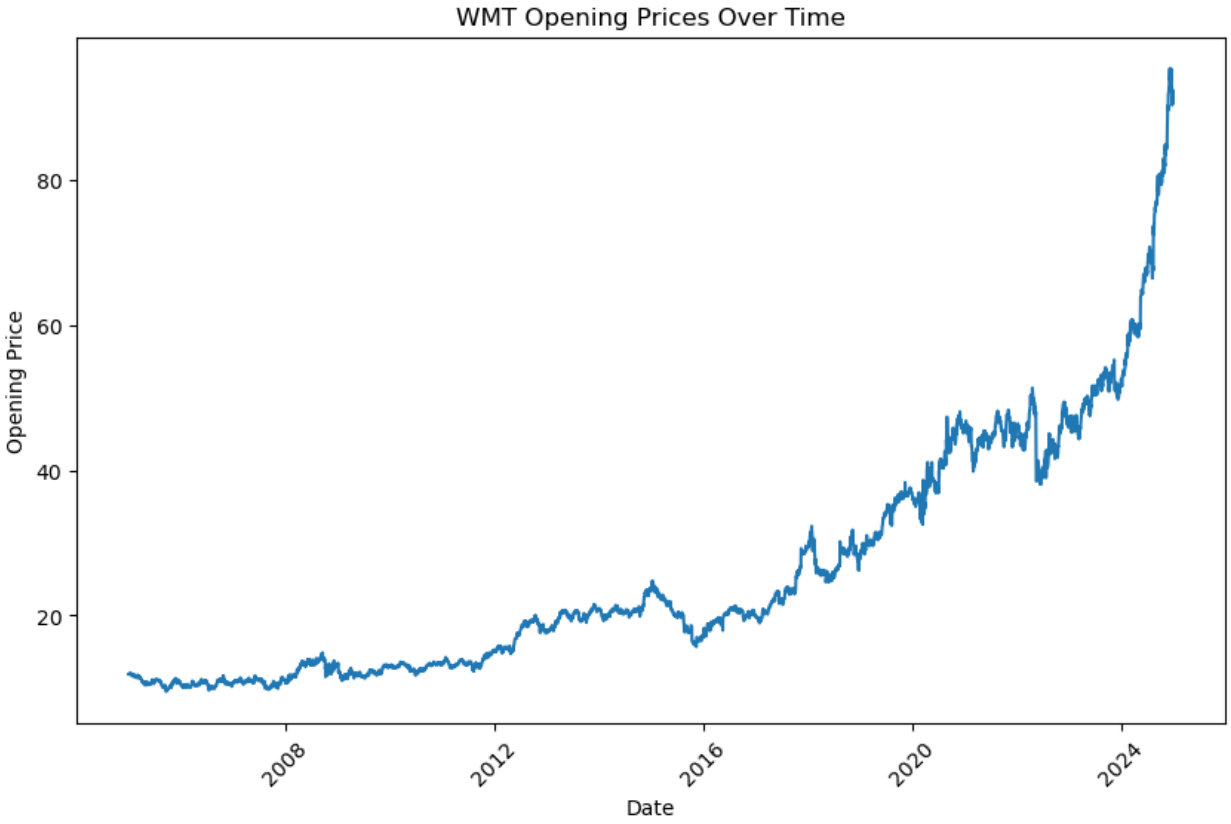

META Opening Prices Over Time
Opening Price
0
100
200
300
400
500
600
2008
2012
2016
2020
2024
Date

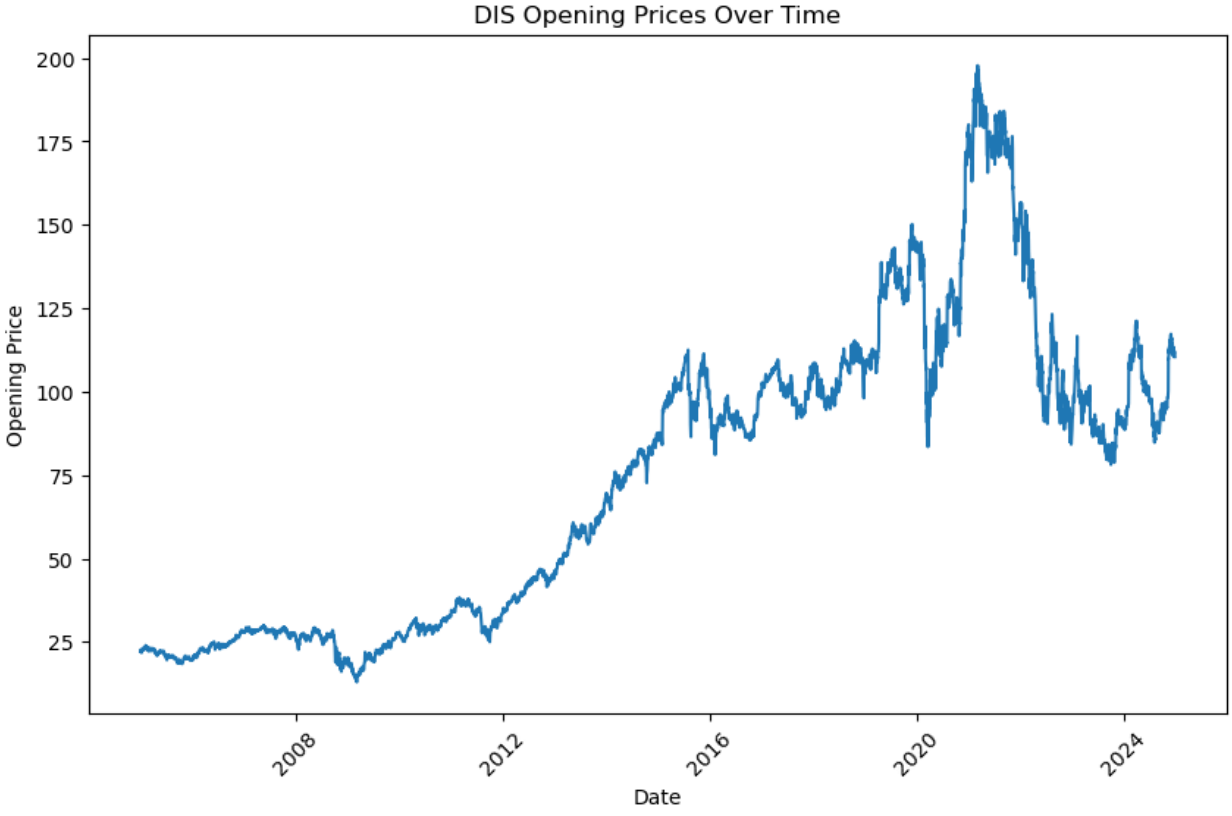

NVDA Opening Prices Over Time
Opening Price
0
20
40
60
80
100
120
140
2008
2012
2016
2020
2024
Date

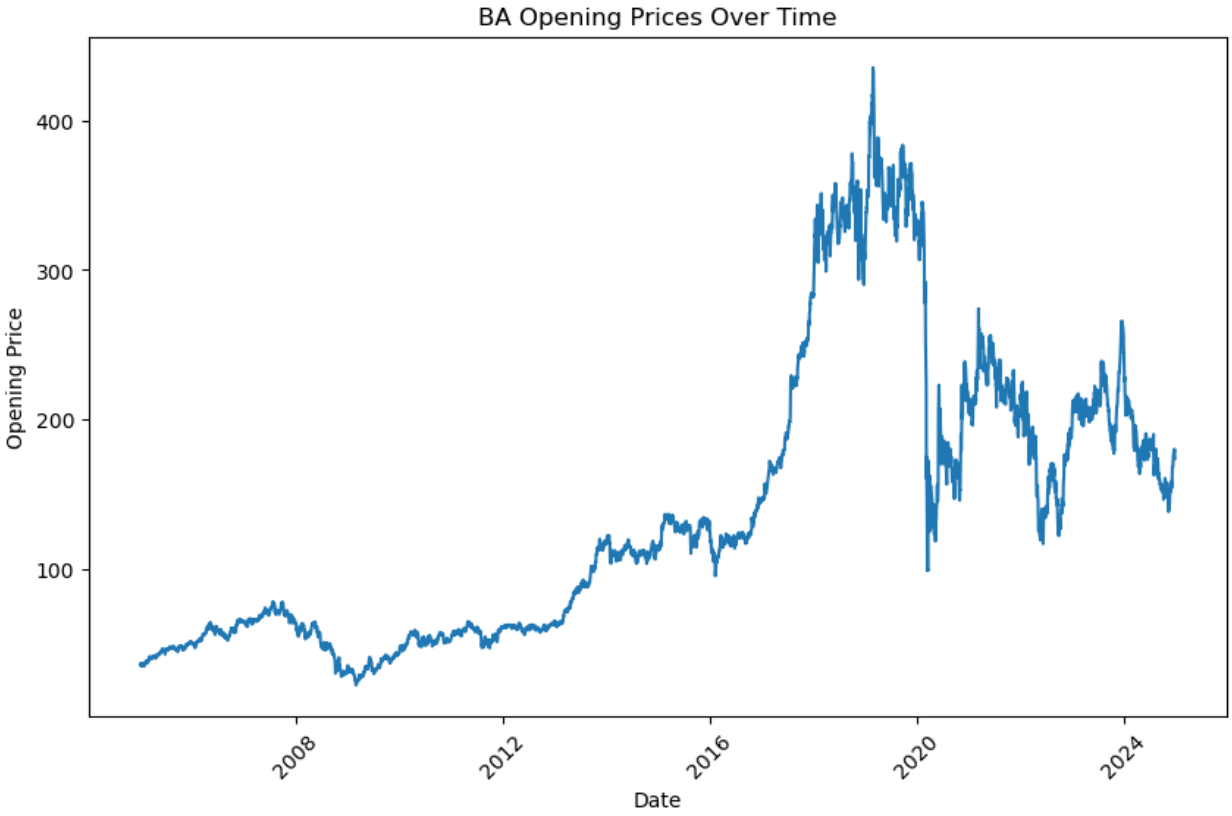

BRK-B Opening Prices Over Time
Opening Price
100
200
300
400
500
2008
2012
2016
2020
2024
Date

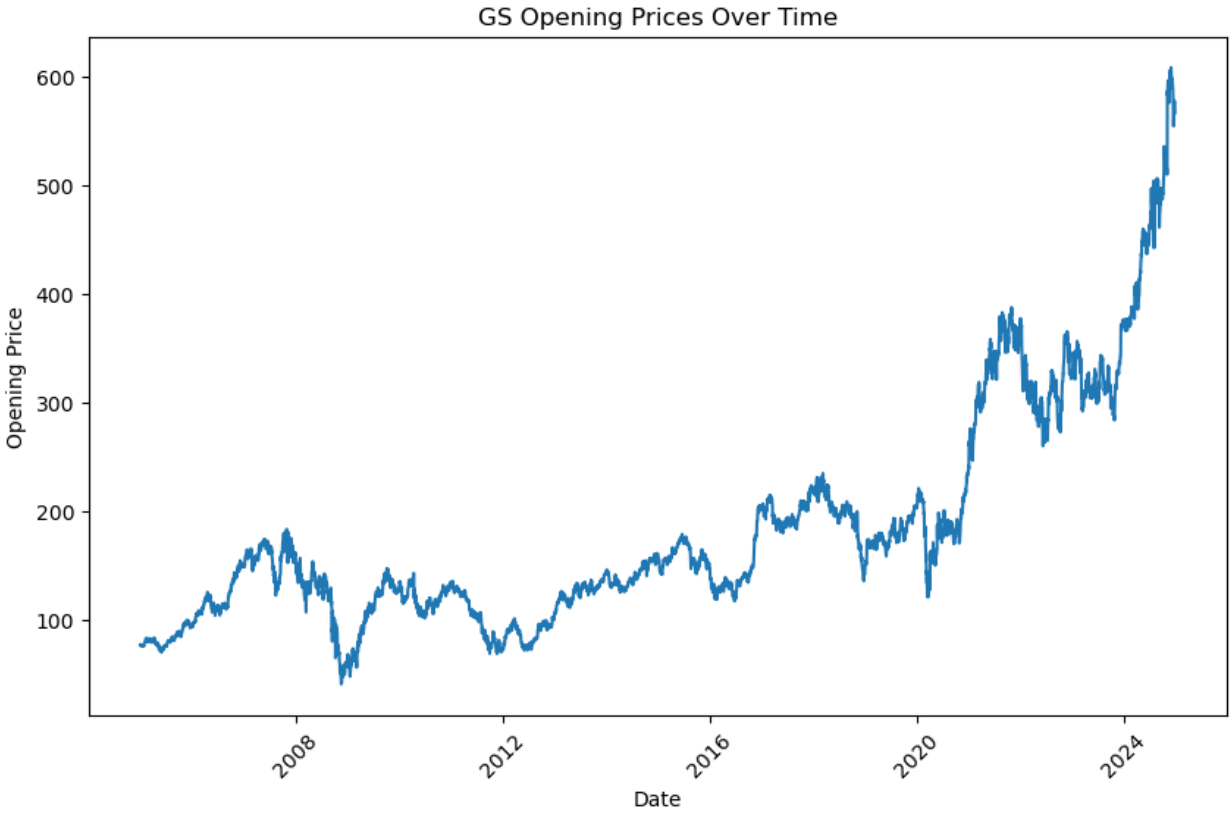

JPM Opening Prices Over Time
Opening Price
0
50
100
150
200
250
2008
2012
2016
2020
2024
Date

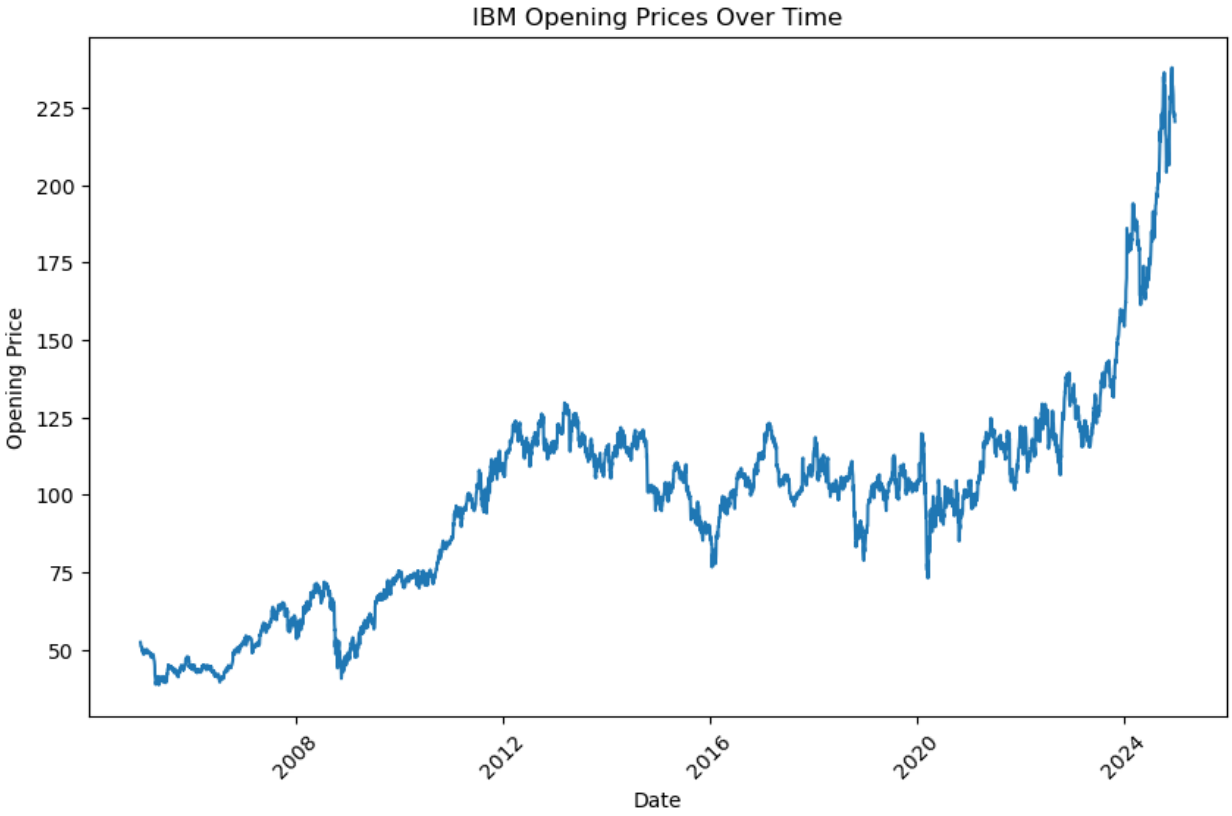

V Opening Prices Over Time
Opening Price
0
50
100
150
200
250
300
2008
2012
2016
2020
2024
Date

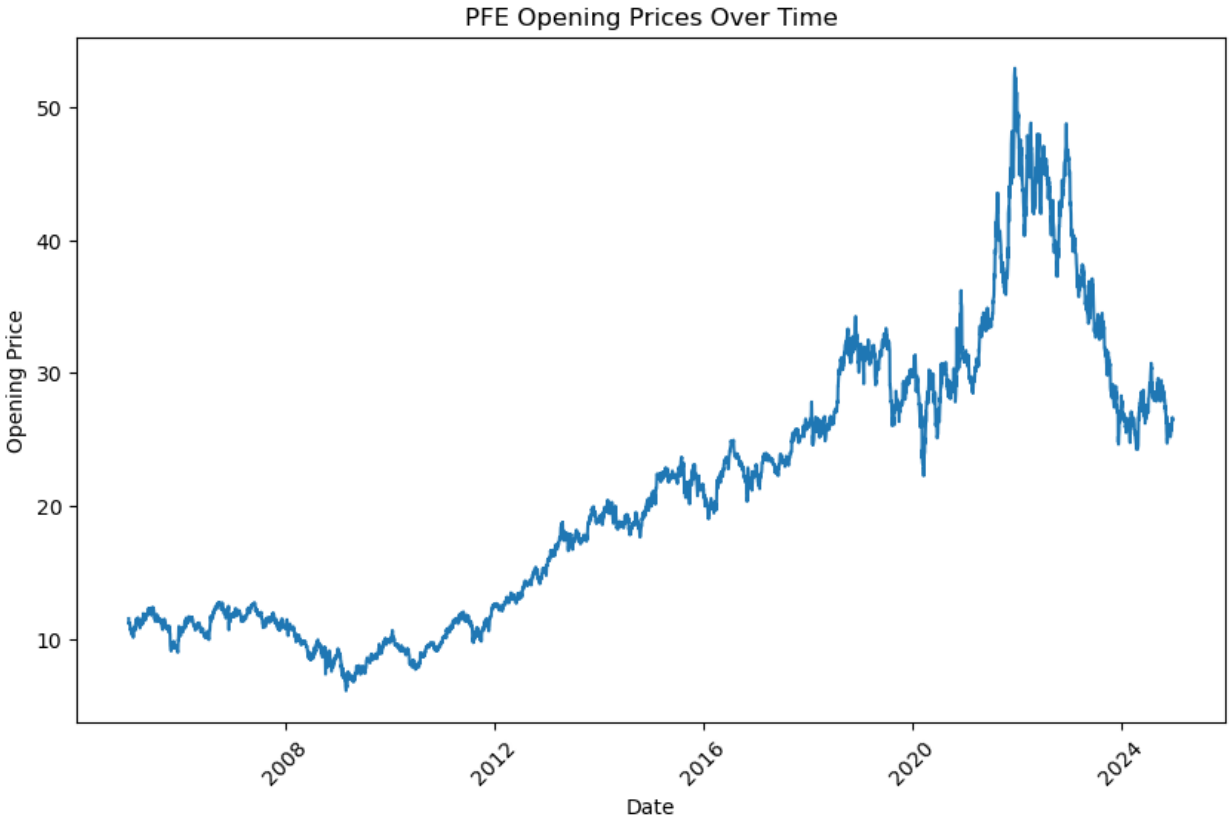

WMT Opening Prices Over Time
Opening Price
80
60
40
20
2008
2012
2016
2020
2024
Date


DIS Opening Prices Over Time
Opening Price
200
175
150
125
100
75
50
25
2008
2012
2016
2020
2024
Date


BA Opening Prices Over Time
Opening Price
400
300
200
100
2008
2012
2016
2020
2024
Date

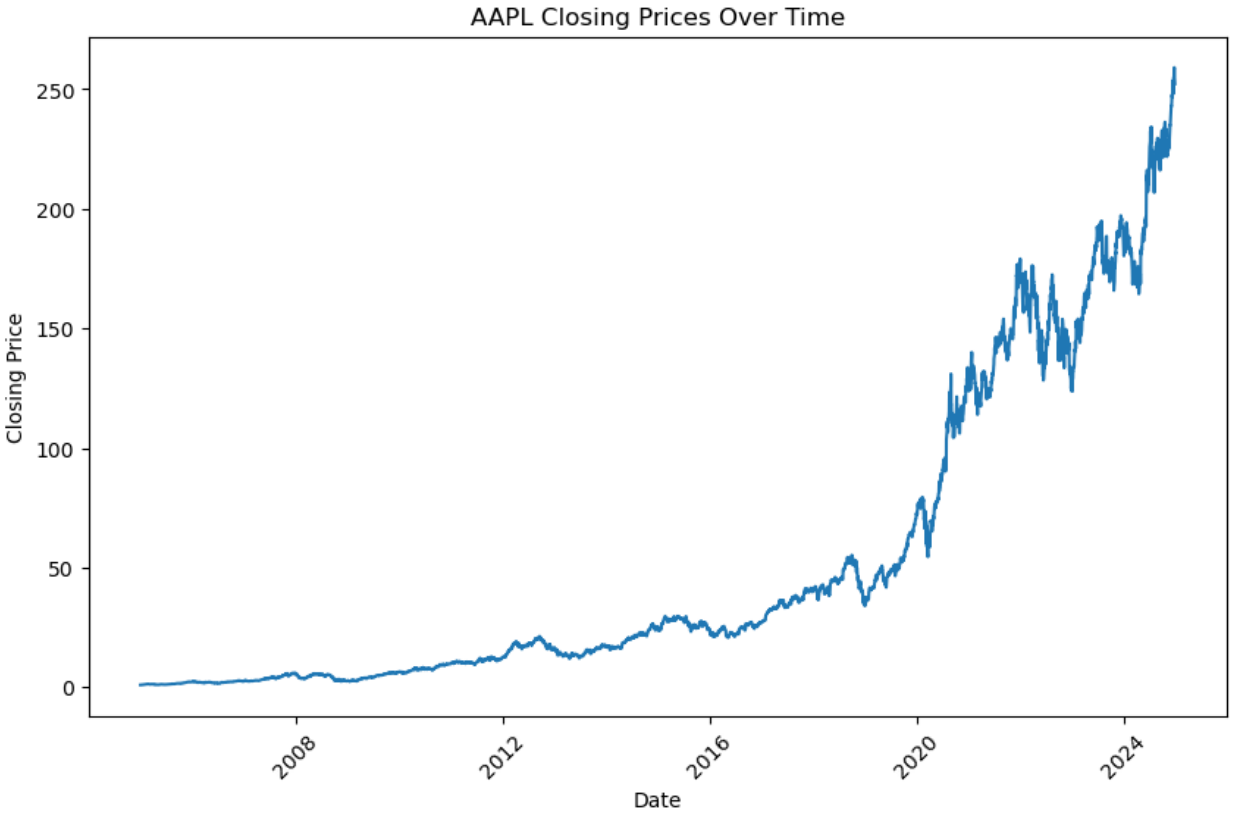

GS Opening Prices Over Time
Opening Price
600
500
400
300
200
100
2008
2012
2016
2020
2024
Date

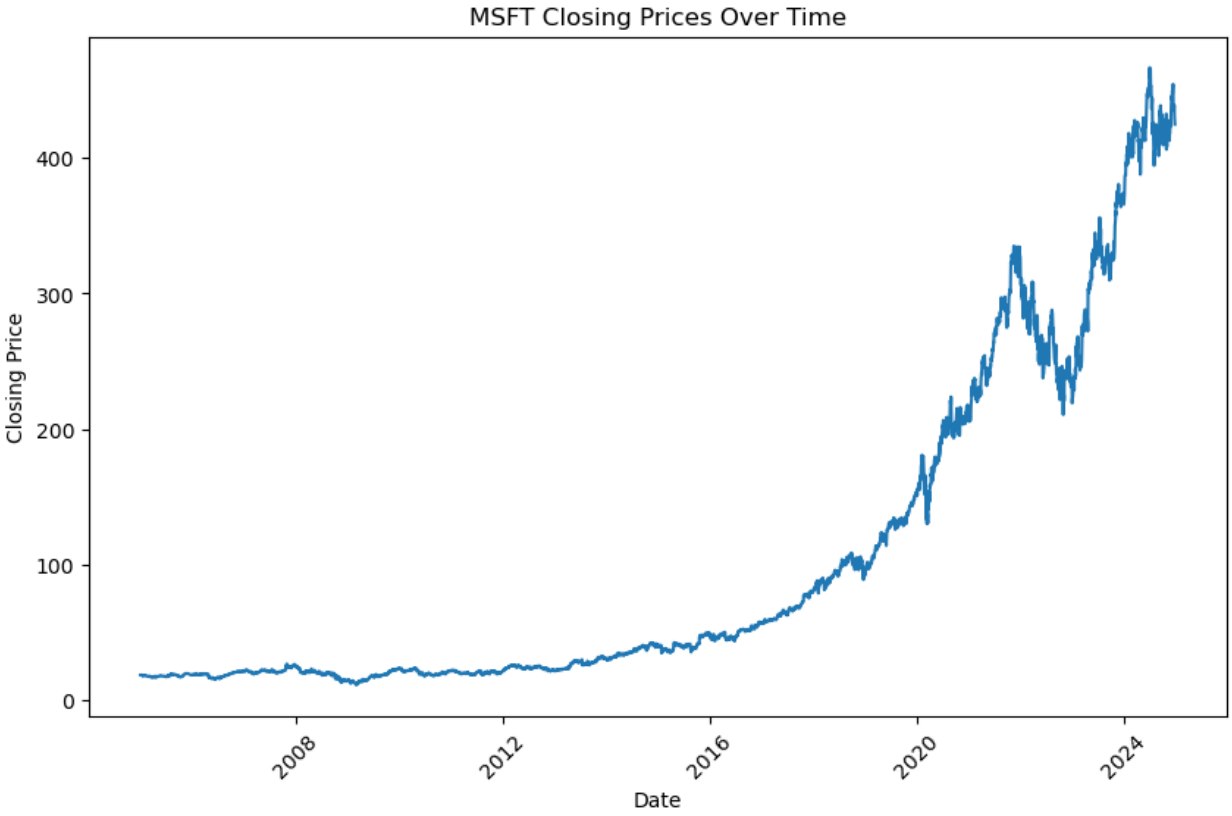

IBM Opening Prices Over Time
Opening Price
225
200
175
150
125
100
75
50
2008
2012
2016
2020
2024
Date

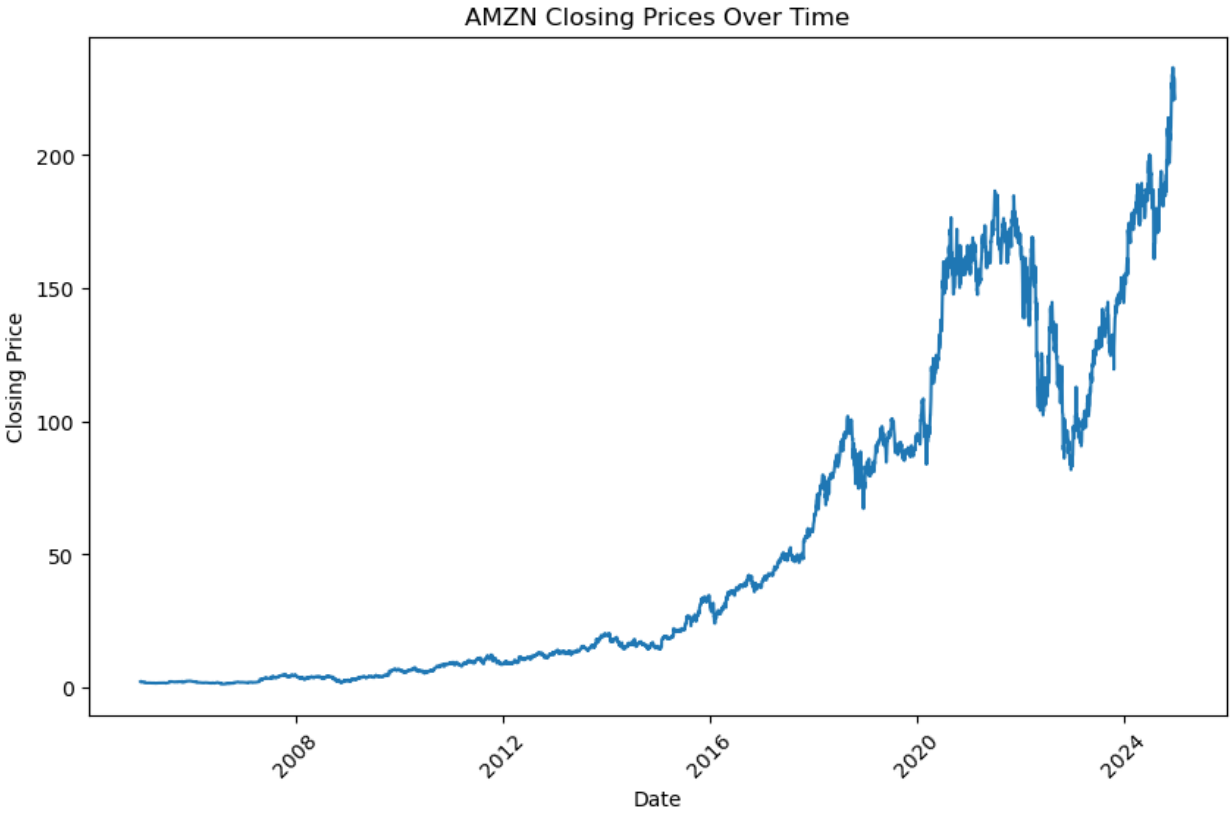

PFE Opening Prices Over Time
Opening Price
50
40
30
20
10
2008
2012
2016
2020
2024
Date

b. Line Plots for Closing Prices:

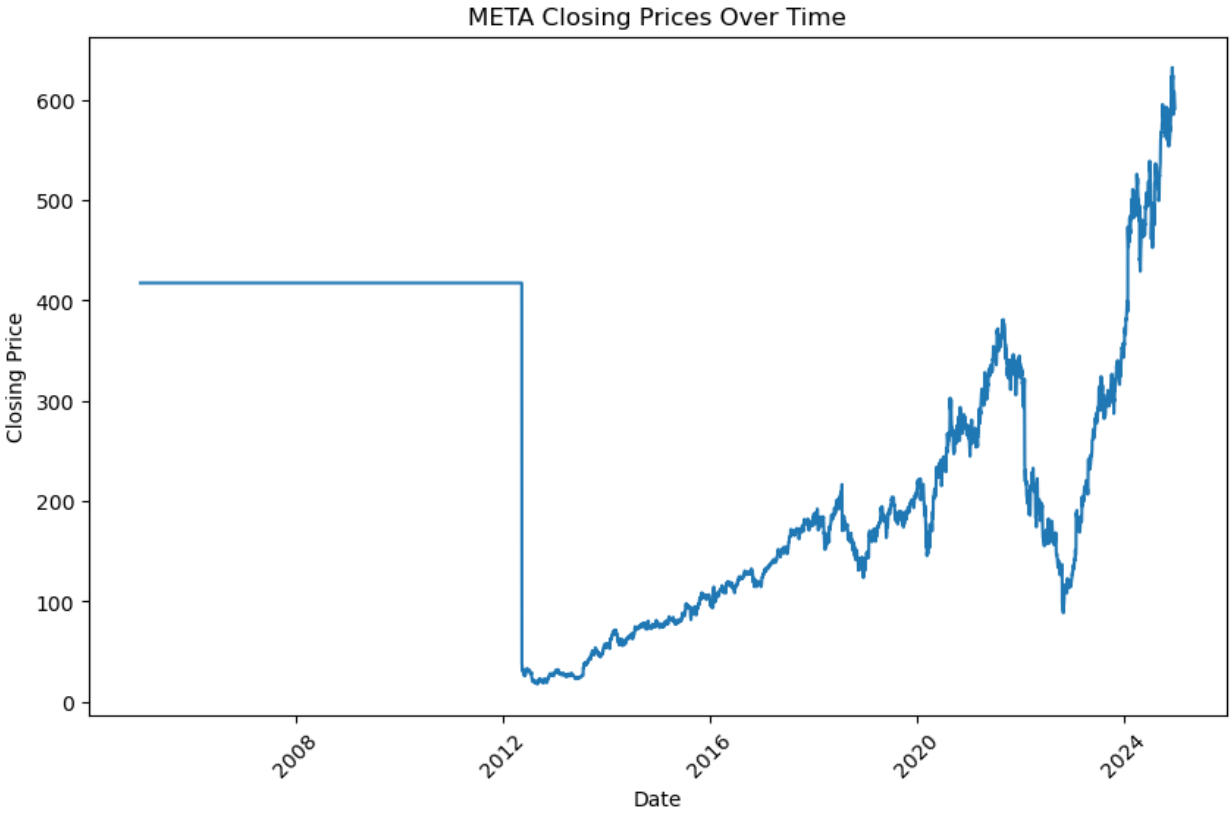
The objective of this line plot is to portray the closing price movements of all firms across a span of time. The data depict the changes in the closing prices over the years for each firm. Trends can tell how the company's stock performed based on years, thus coming with periods of growth, decline, or stability.

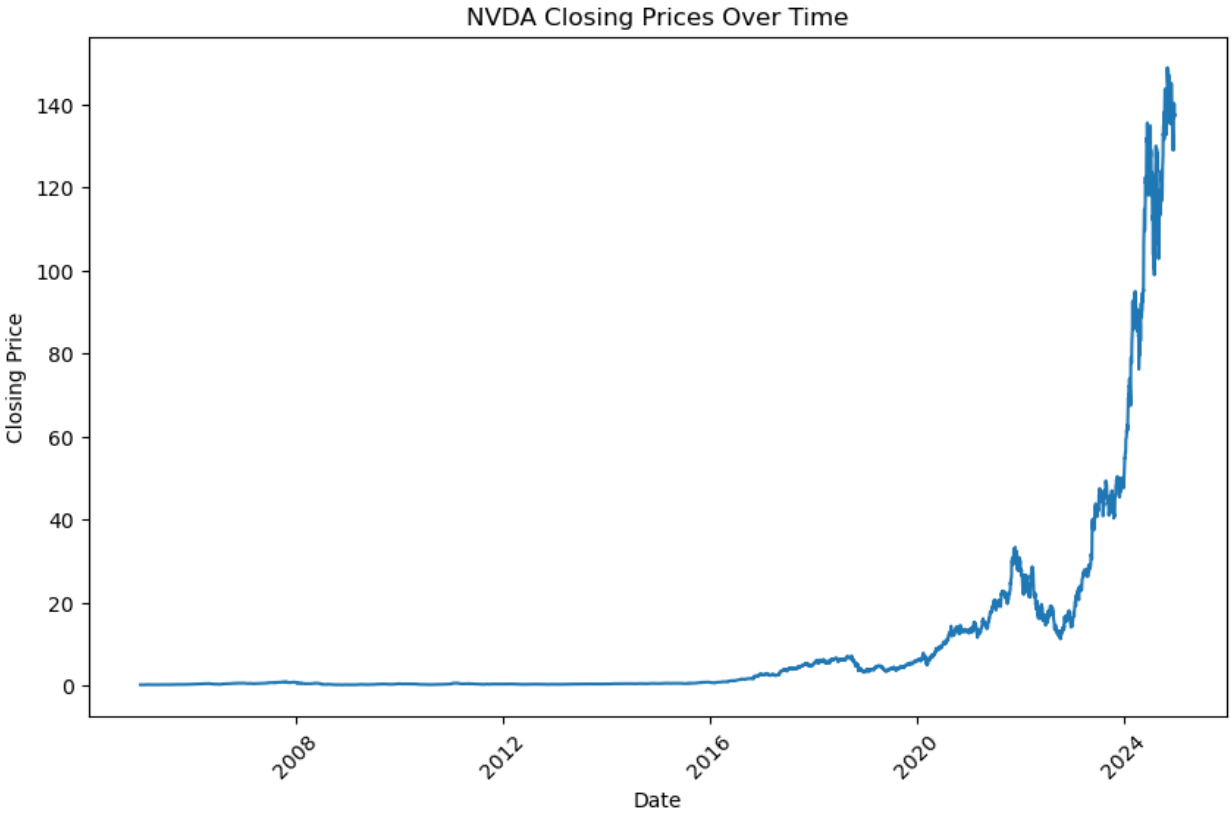



MSFT Closing Prices Over Time
Closing Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

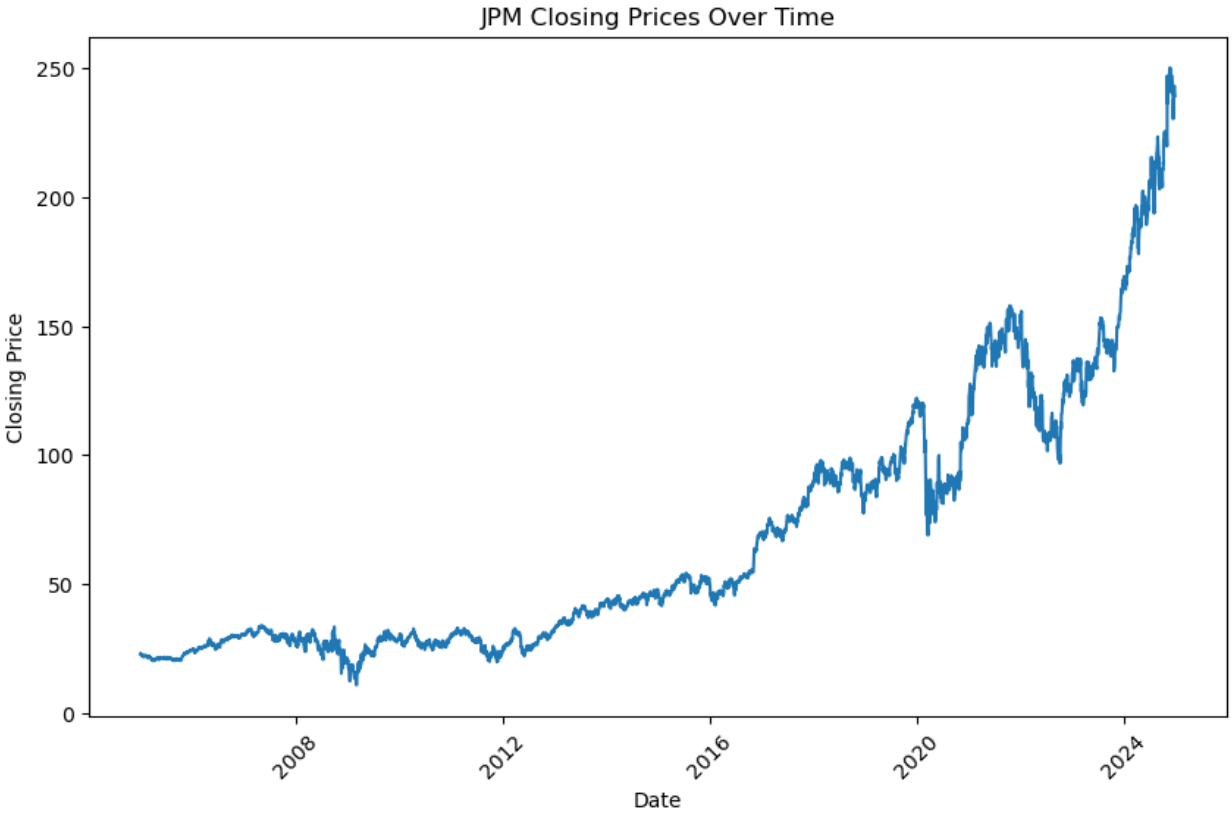

AMZN Closing Prices Over Time
Closing Price
0
50
100
150
200
2008
2012
2016
2020
2024
Date

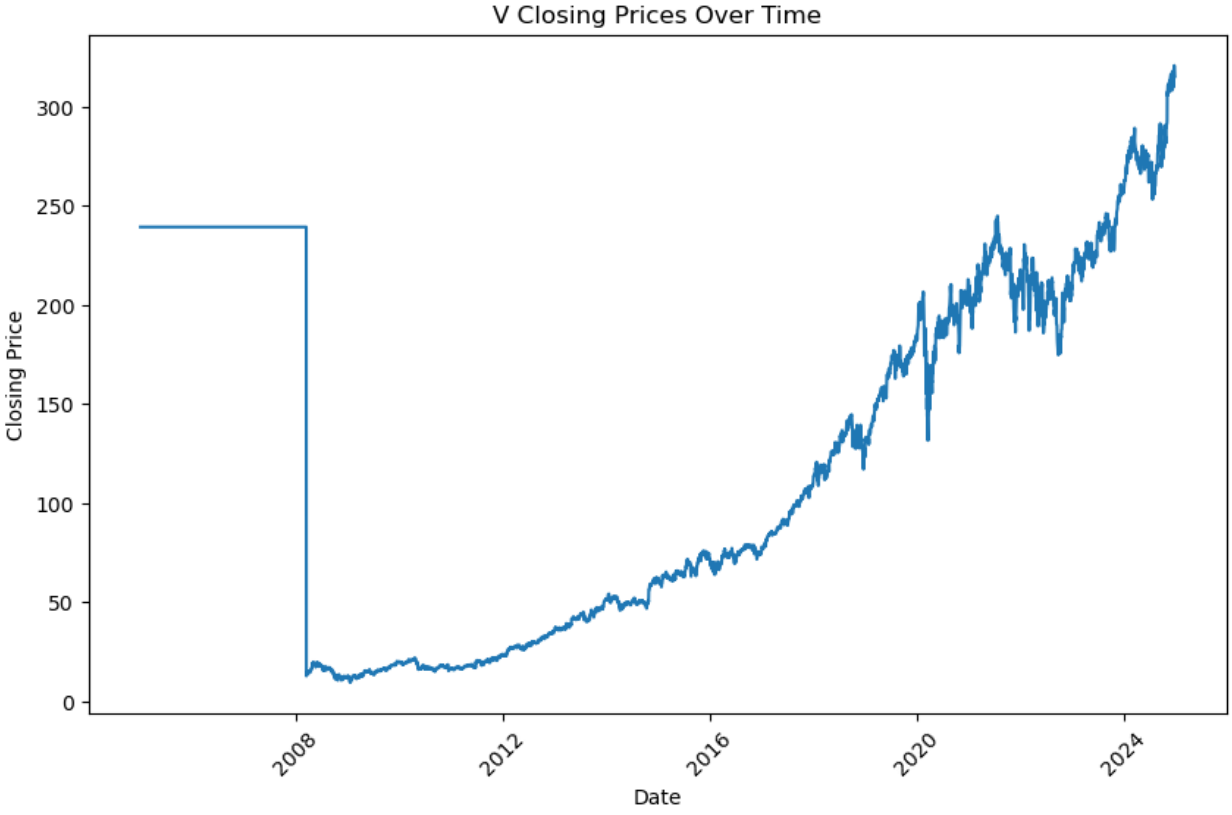

TSLA Closing Prices Over Time
Closing Price
Date
0
100
200
300
400
500
2008
2012
2016
2020
2024

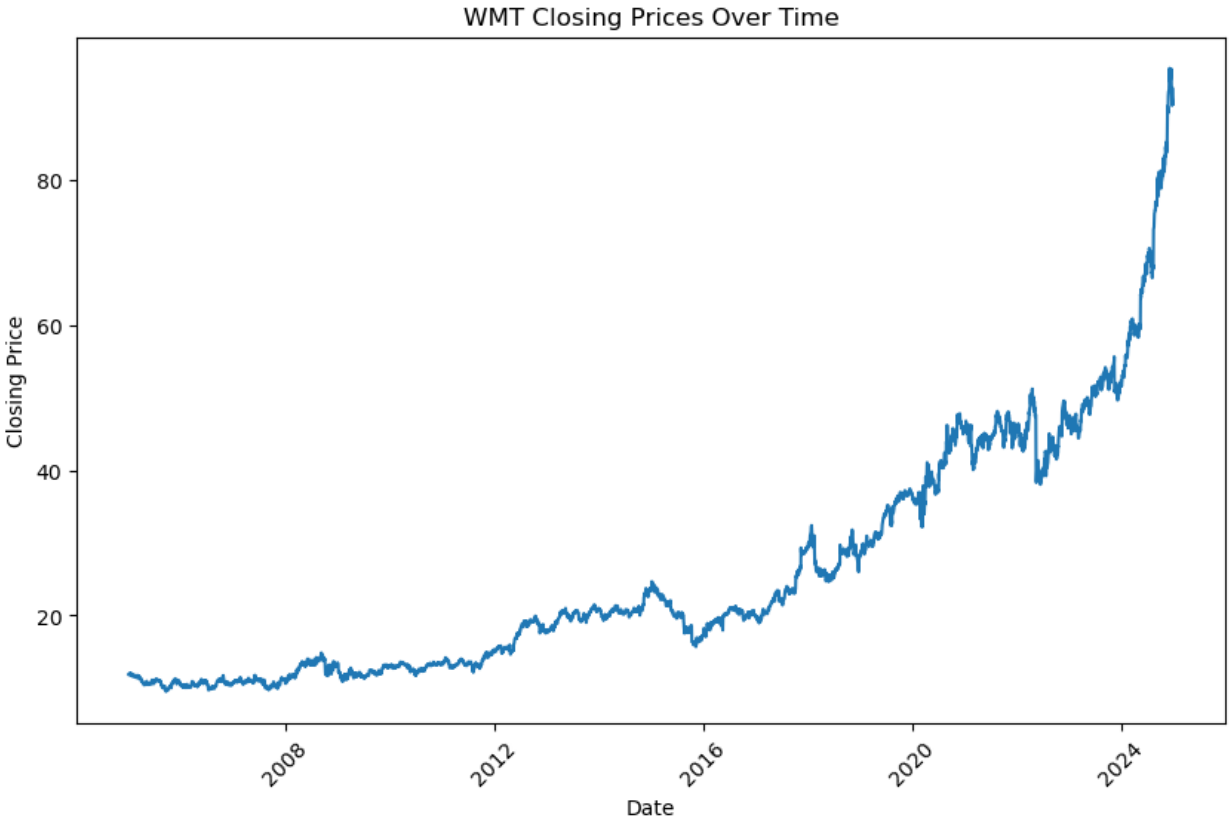

META Closing Prices Over Time
Closing Price
Date
0
100
200
300
400
500
600
2008
2012
2016
2020
2024

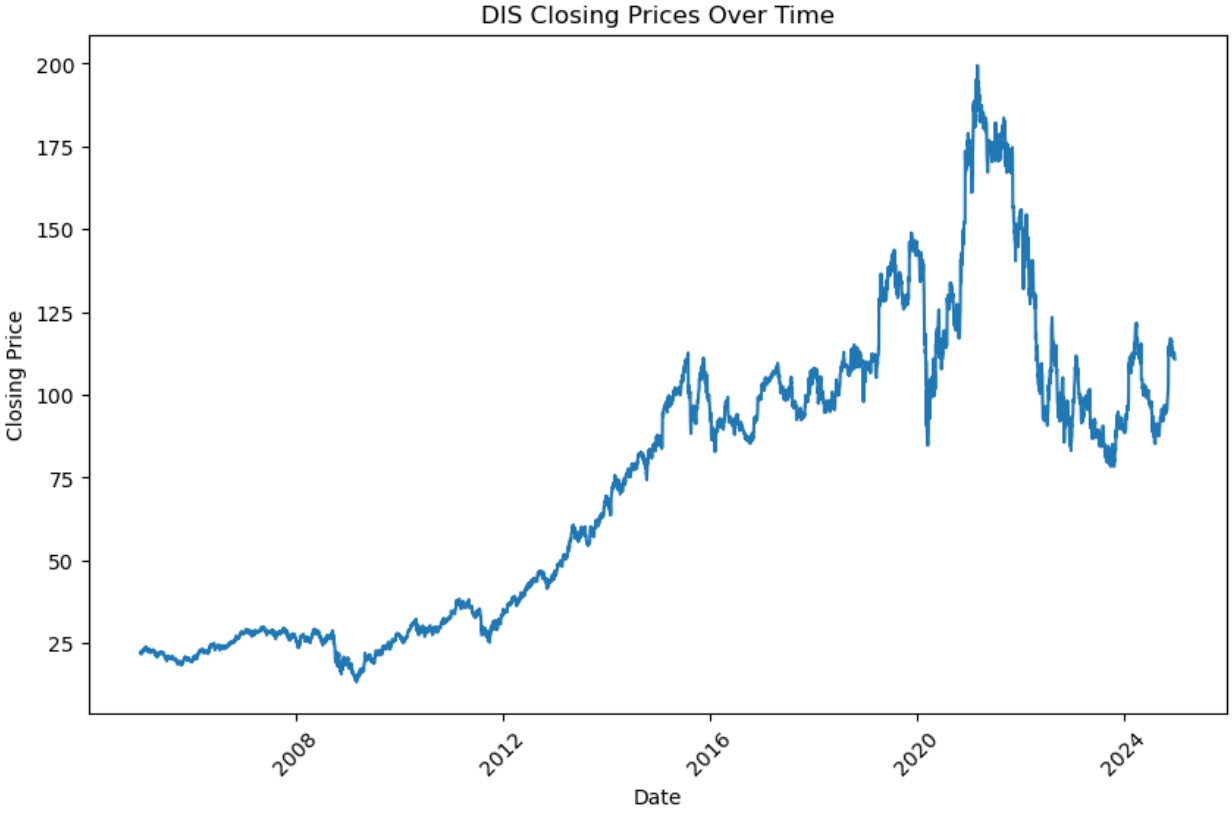

NVDA Closing Prices Over Time
Closing Price
0
20
40
60
80
100
120
140
2008
2012
2016
2020
2024
Date

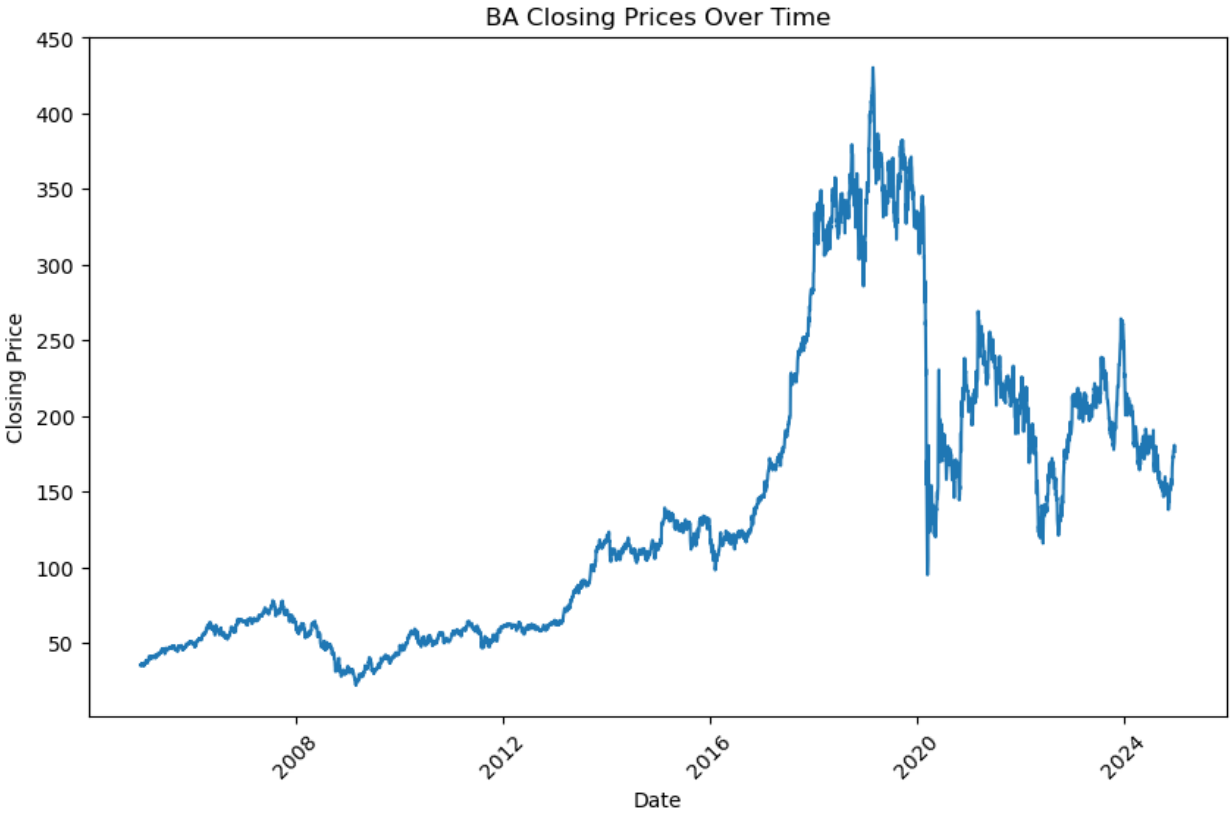

BRK-B Closing Prices Over Time
Closing Price
100
200
300
400
500
2008
2012
2016
2020
2024
Date

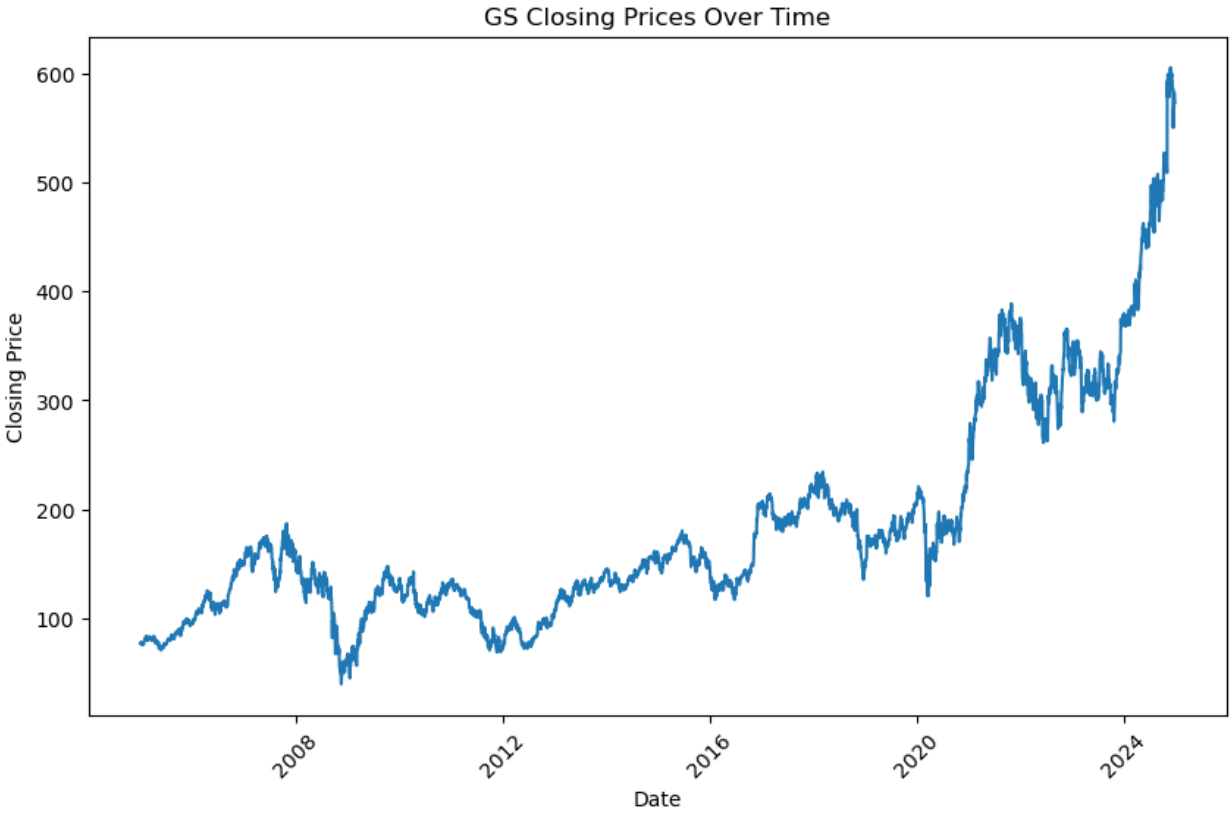

JPM Closing Prices Over Time
Closing Price
0
50
100
150
200
250
2008
2012
2016
2020
2024
Date

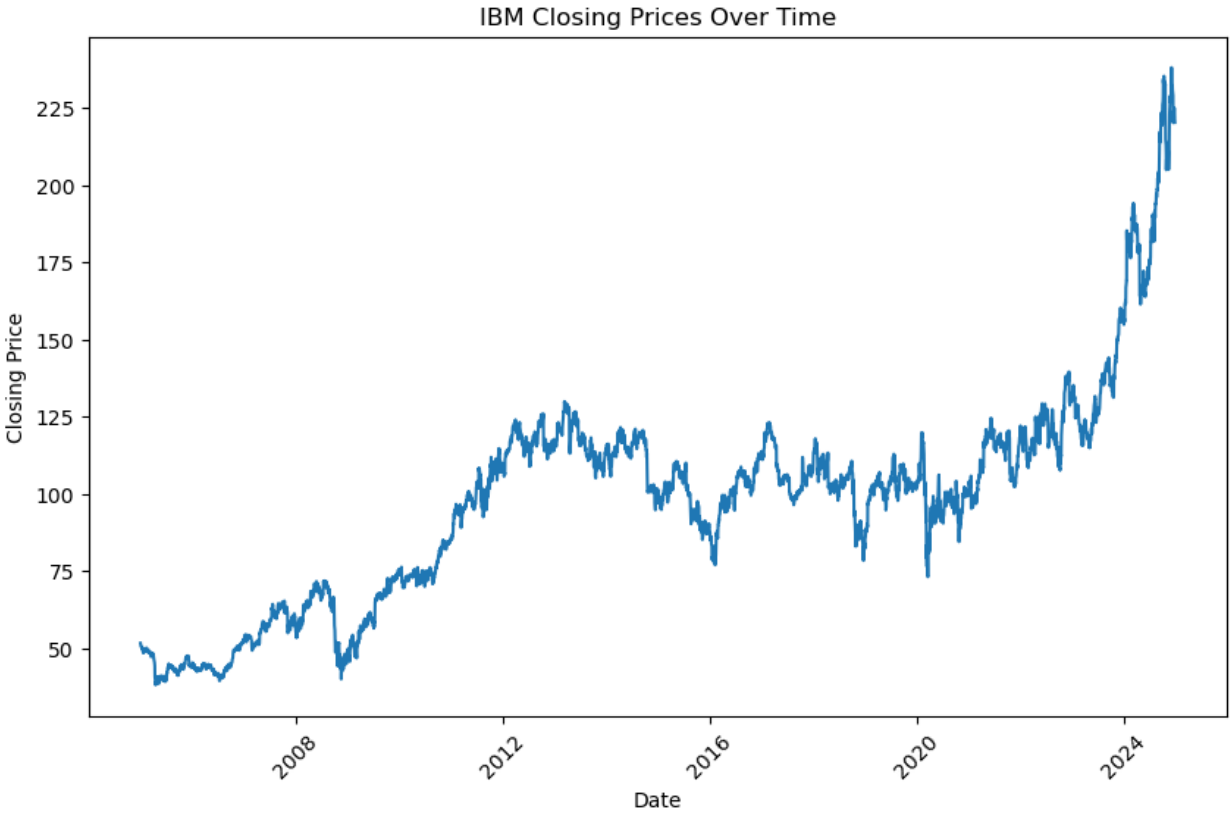

V Closing Prices Over Time
Closing Price
0
50
100
150
200
250
300
2008
2012
2016
2020
2024
Date

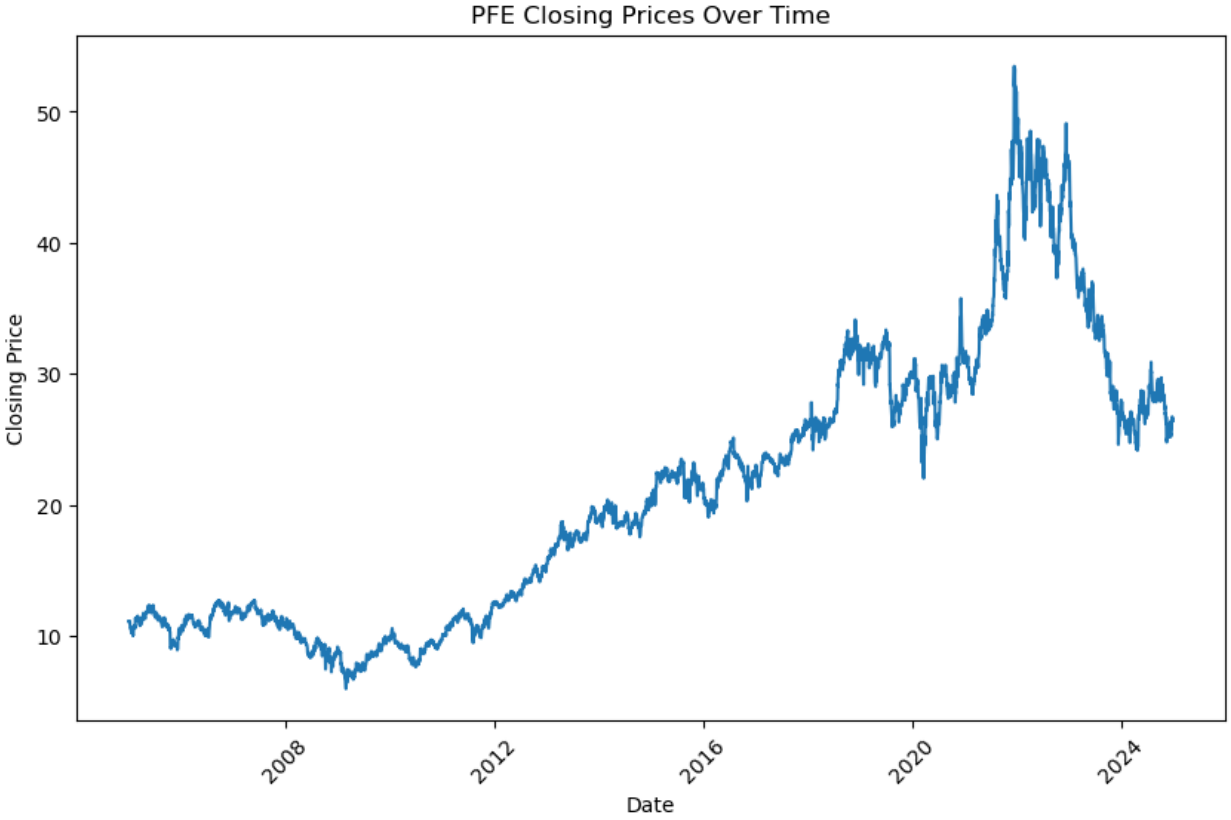

WMT Closing Prices Over Time
Closing Price
80
60
40
20
2008
2012
2016
2020
2024
Date


DIS Closing Prices Over Time
Closing Price
200
175
150
125
100
75
50
25
2008
2012
2016
2020
2024
Date


BA Closing Prices Over Time
Closing Price
Date
450
400
350
300
250
200
150
100
50
2008
2012
2016
2020
2024

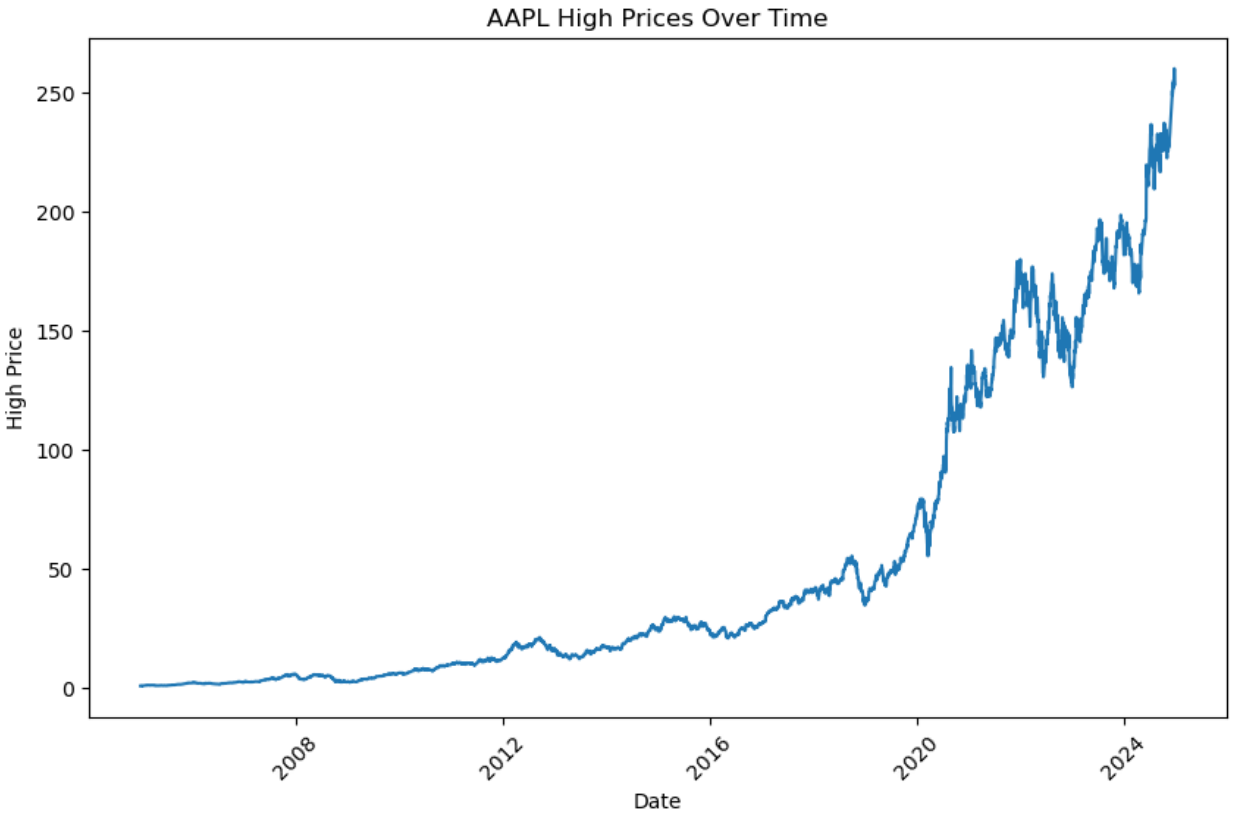

GS Closing Prices Over Time
Closing Price
Date
600
500
400
300
200
100
2008
2012
2016
2020
2024

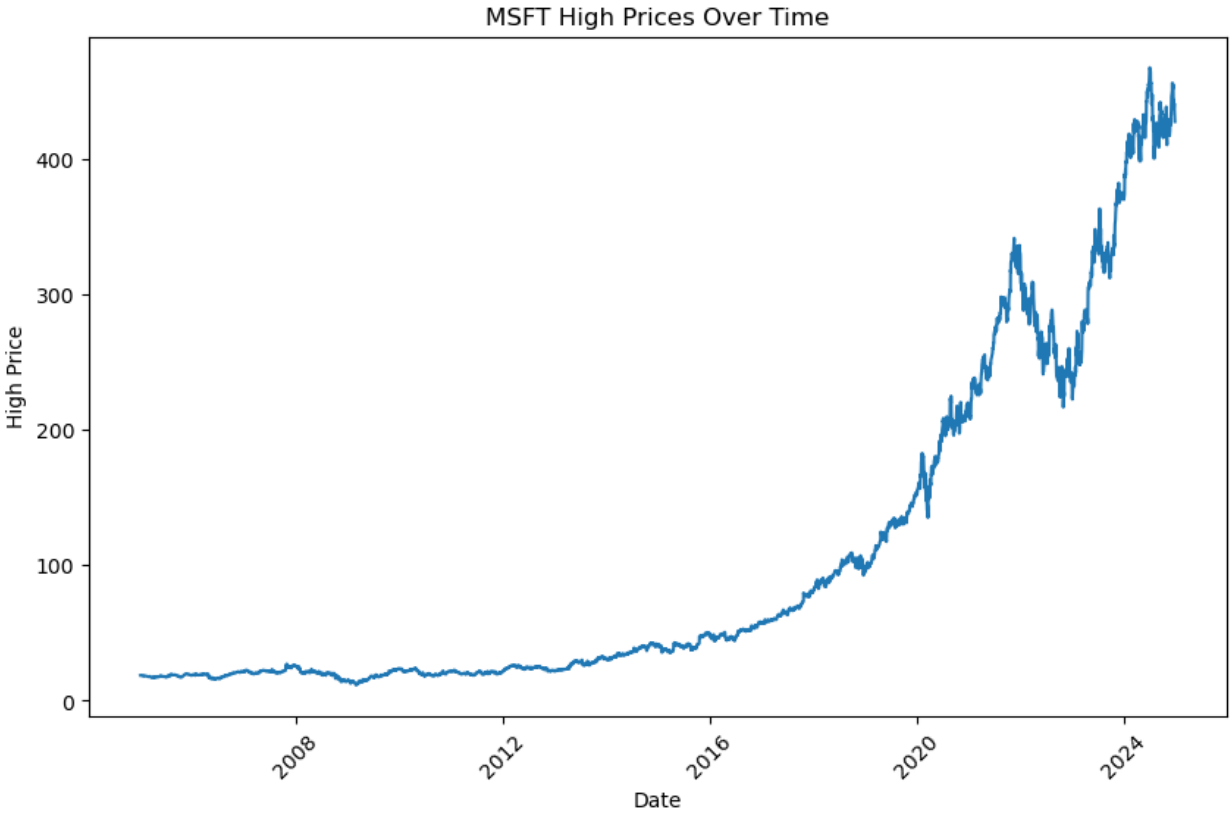

IBM Closing Prices Over Time
Closing Price
225
200
175
150
125
100
75
50
2008
2012
2016
2020
2024
Date

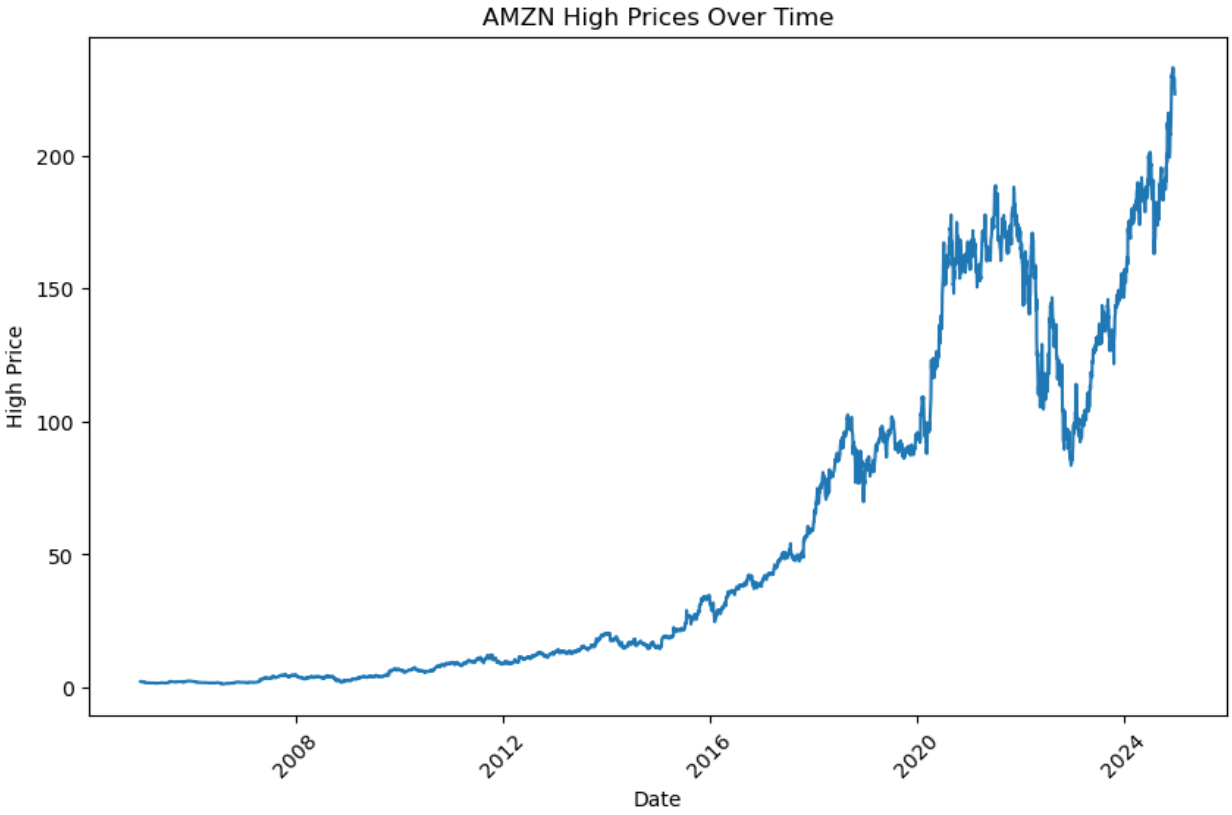

PFE Closing Prices Over Time
Closing Price
50
40
30
20
10
2008
2012
2016
2020
2024
Date

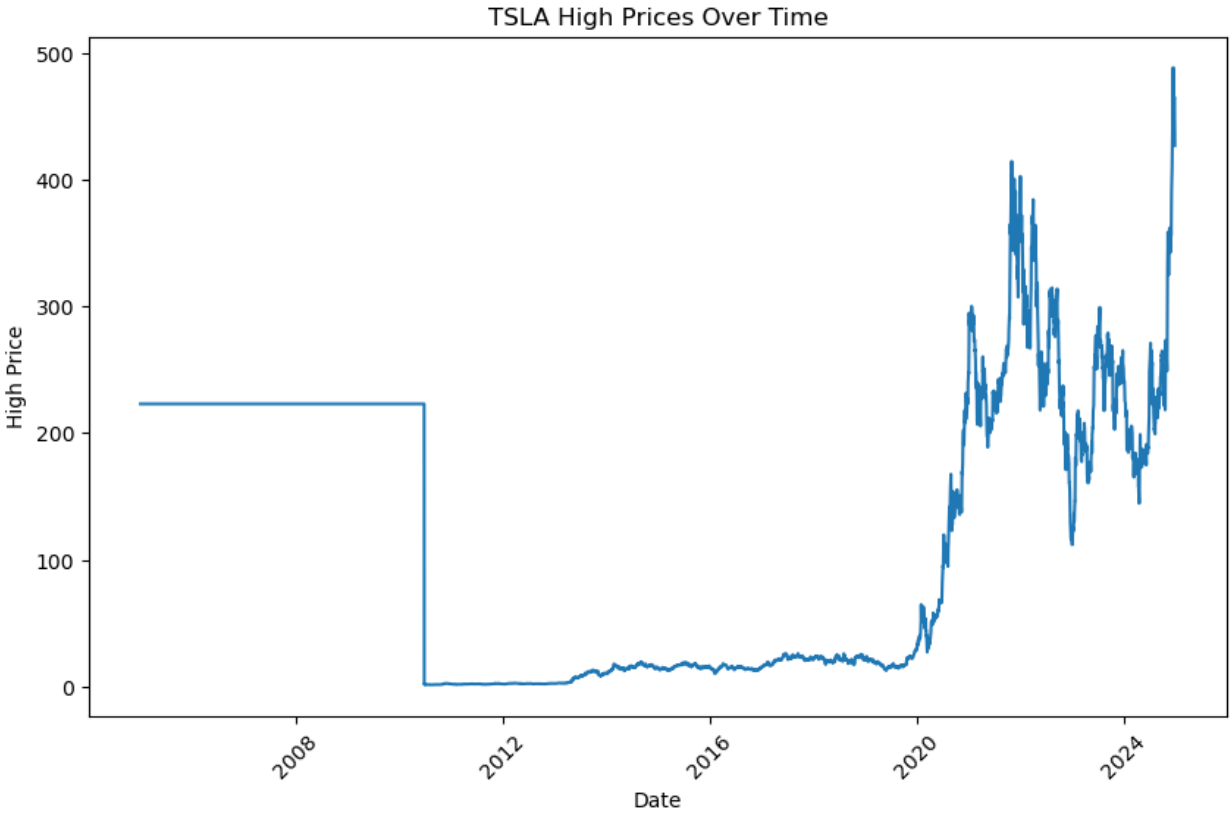
c. Line Plots for High Prices:

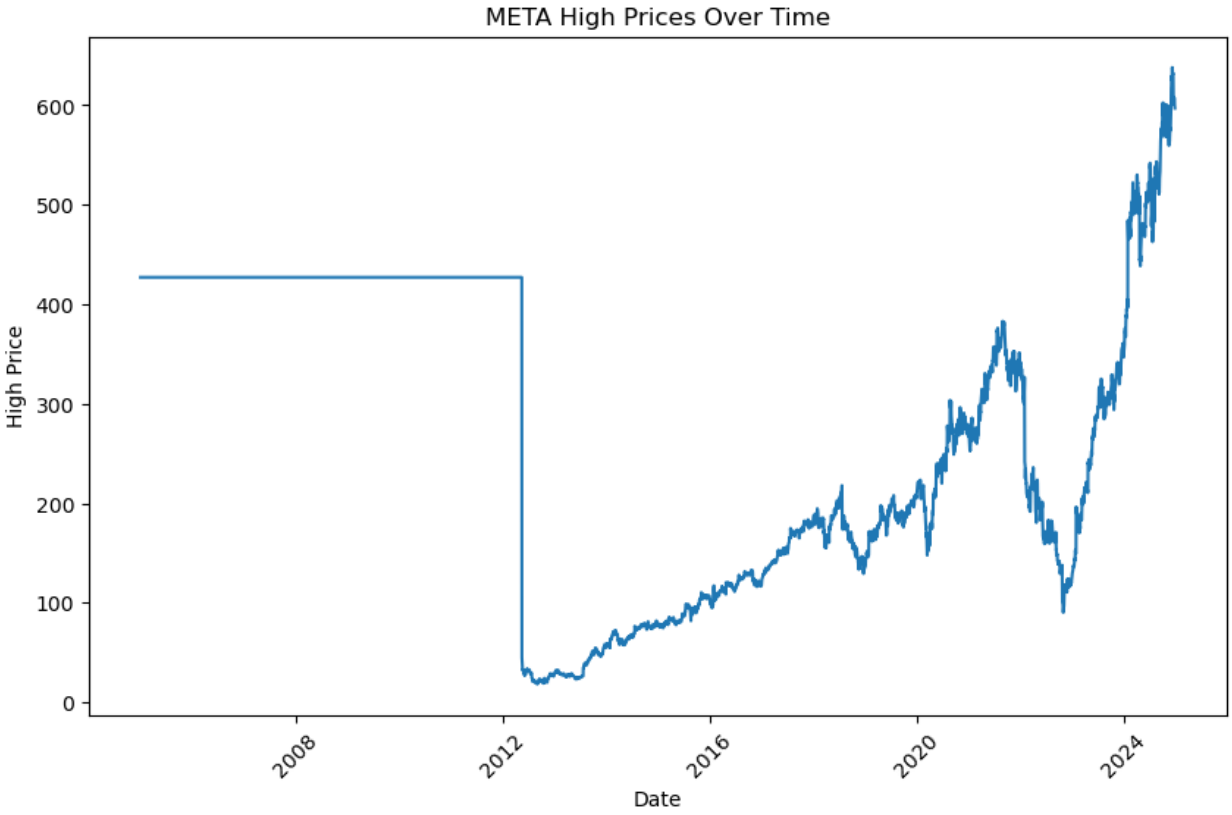
The objective of this line plot is to portray the High price movements of all firms across a span of time. The data depict the changes in the high prices over the years for each firm. Trends can tell how the company's stock performed based on years, thus coming with periods of growth, decline, or stability.

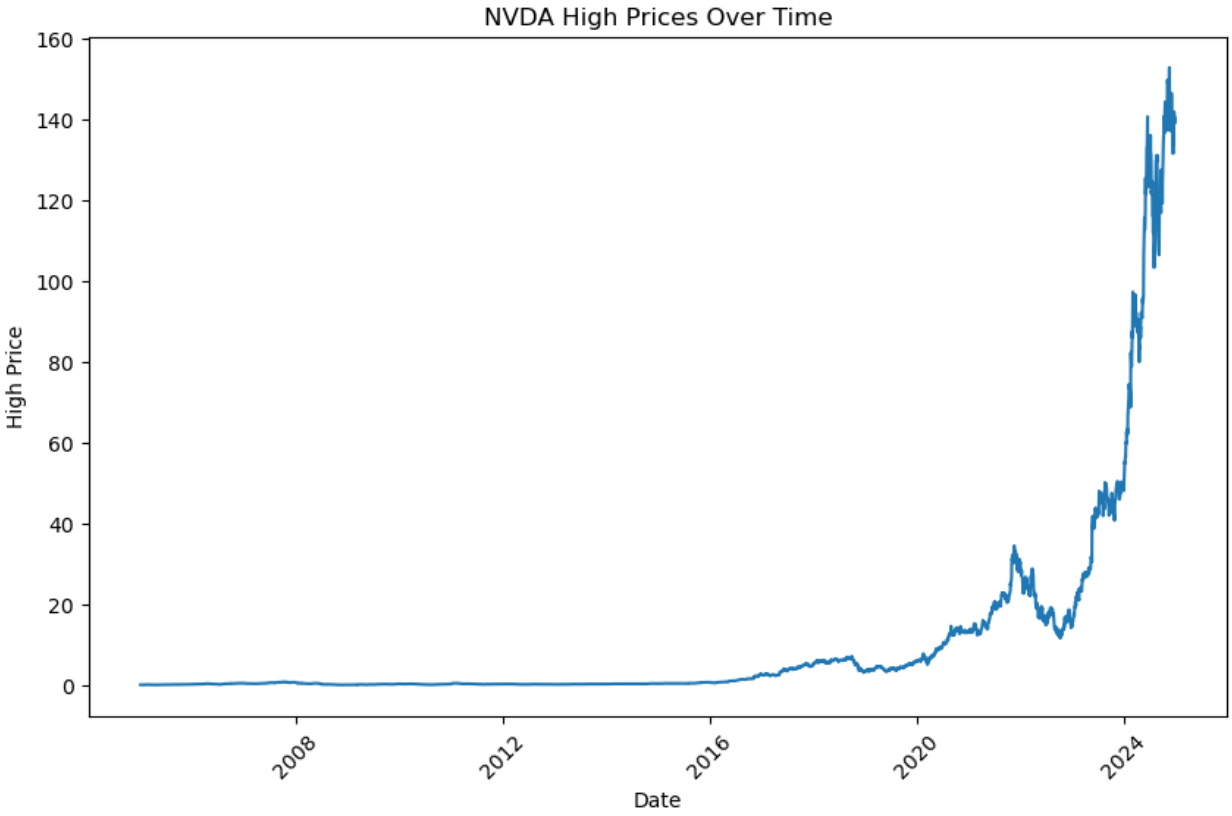


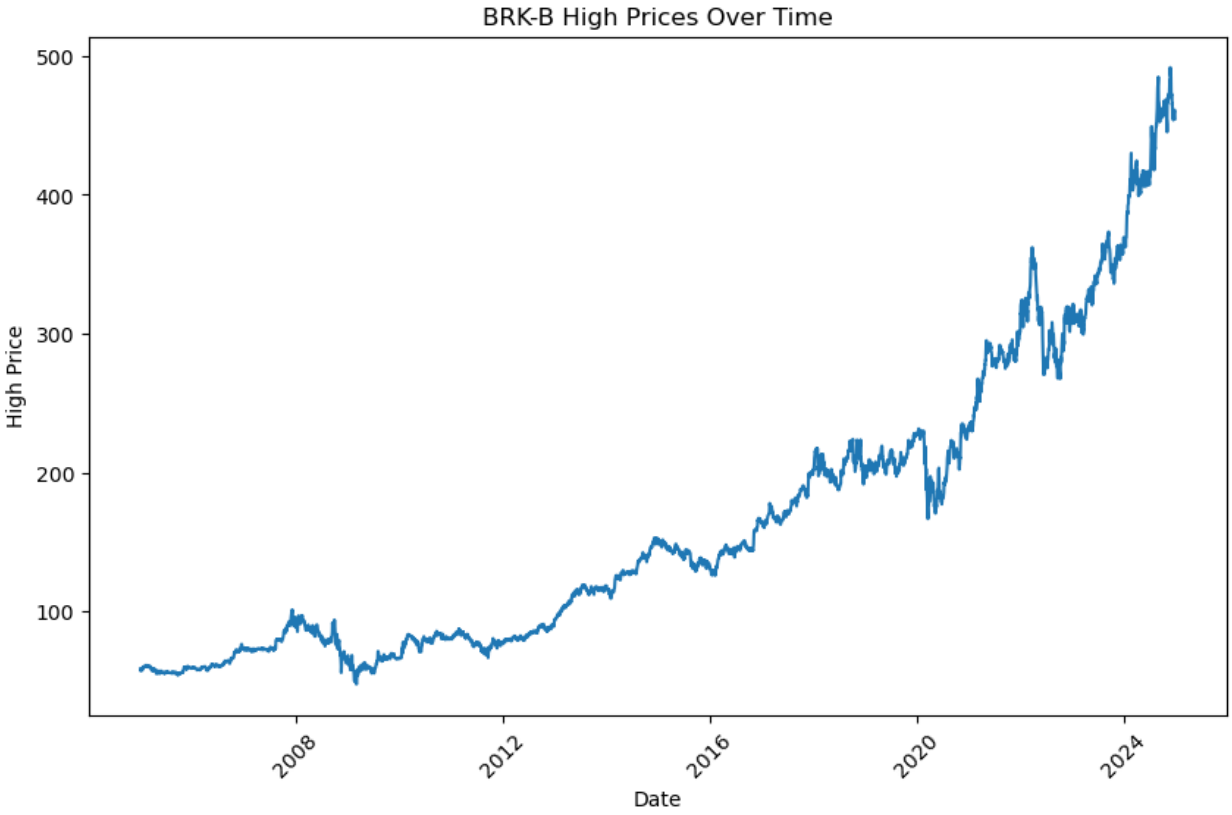

MSFT High Prices Over Time
High Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

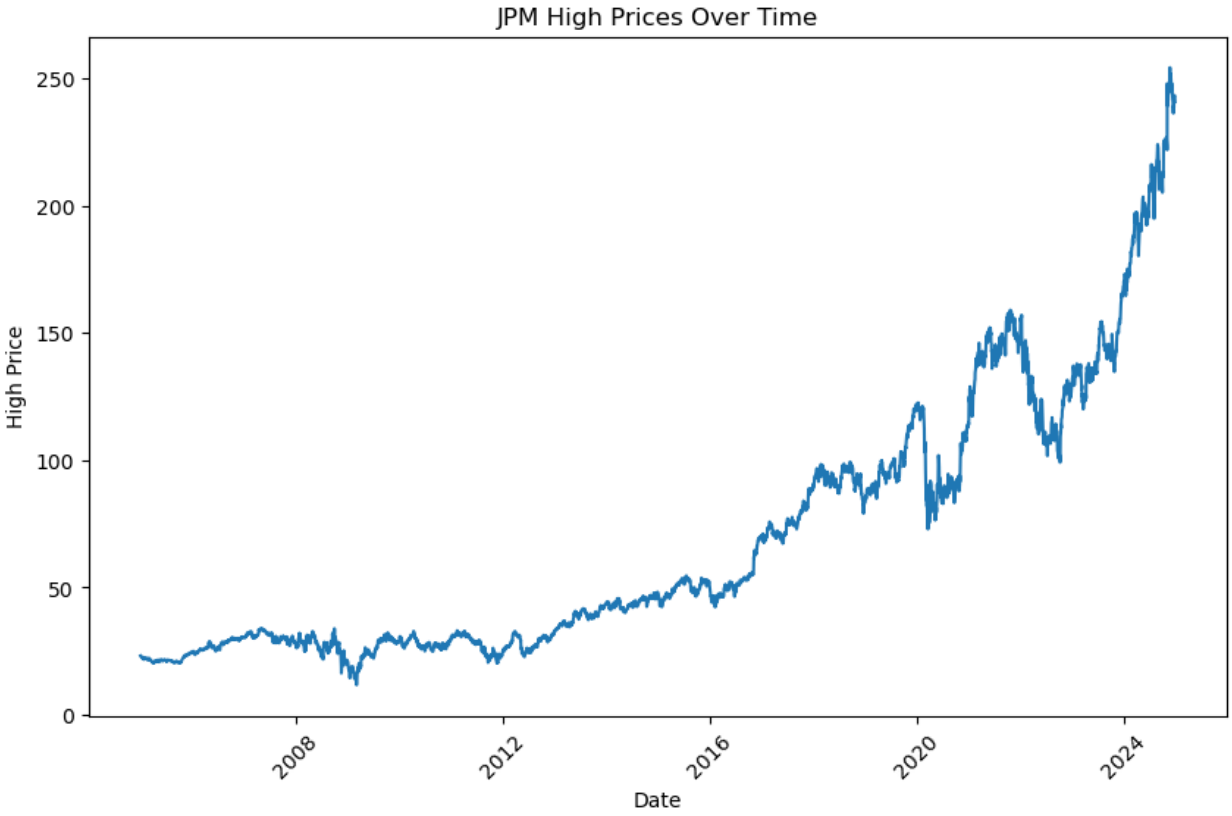

AMZN High Prices Over Time
High Price
0
50
100
150
200
2008
2012
2016
2020
2024
Date

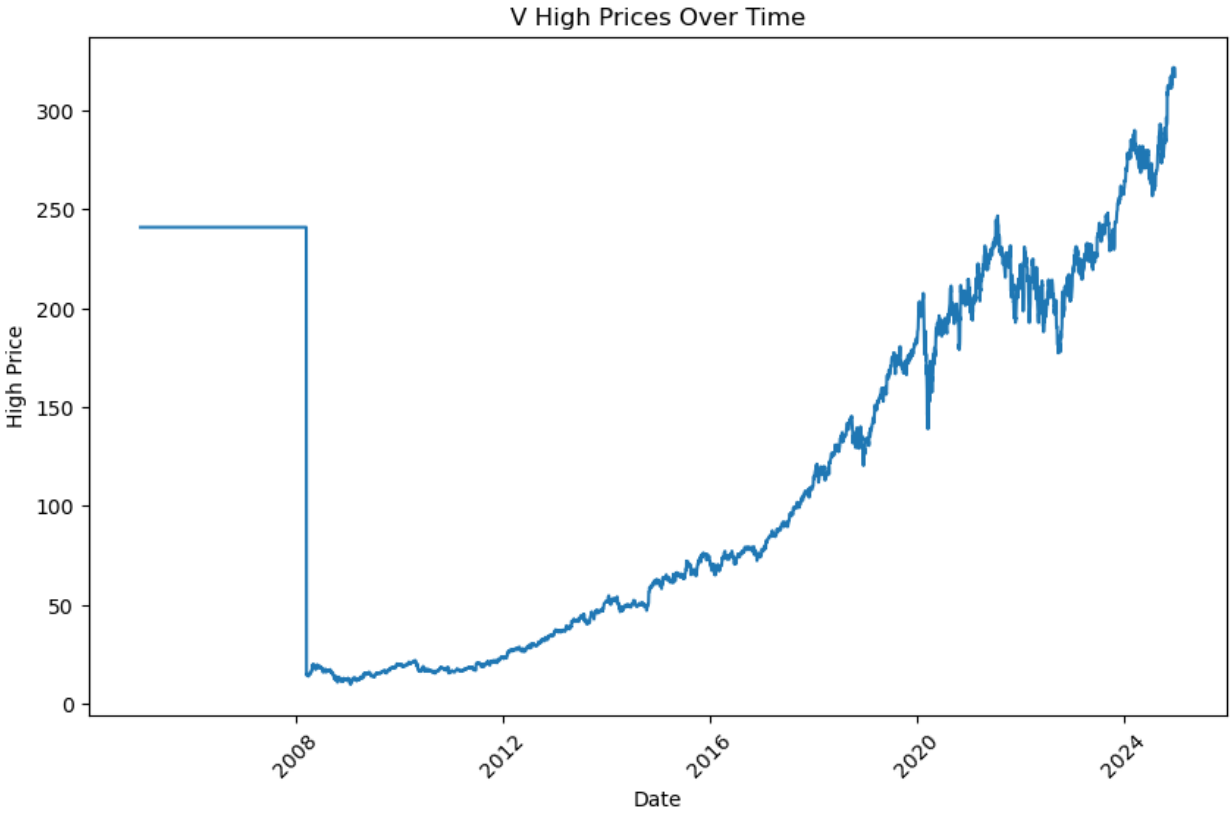

TSLA High Prices Over Time
High Price
Date
0
100
200
300
400
500
2008
2012
2016
2020
2024

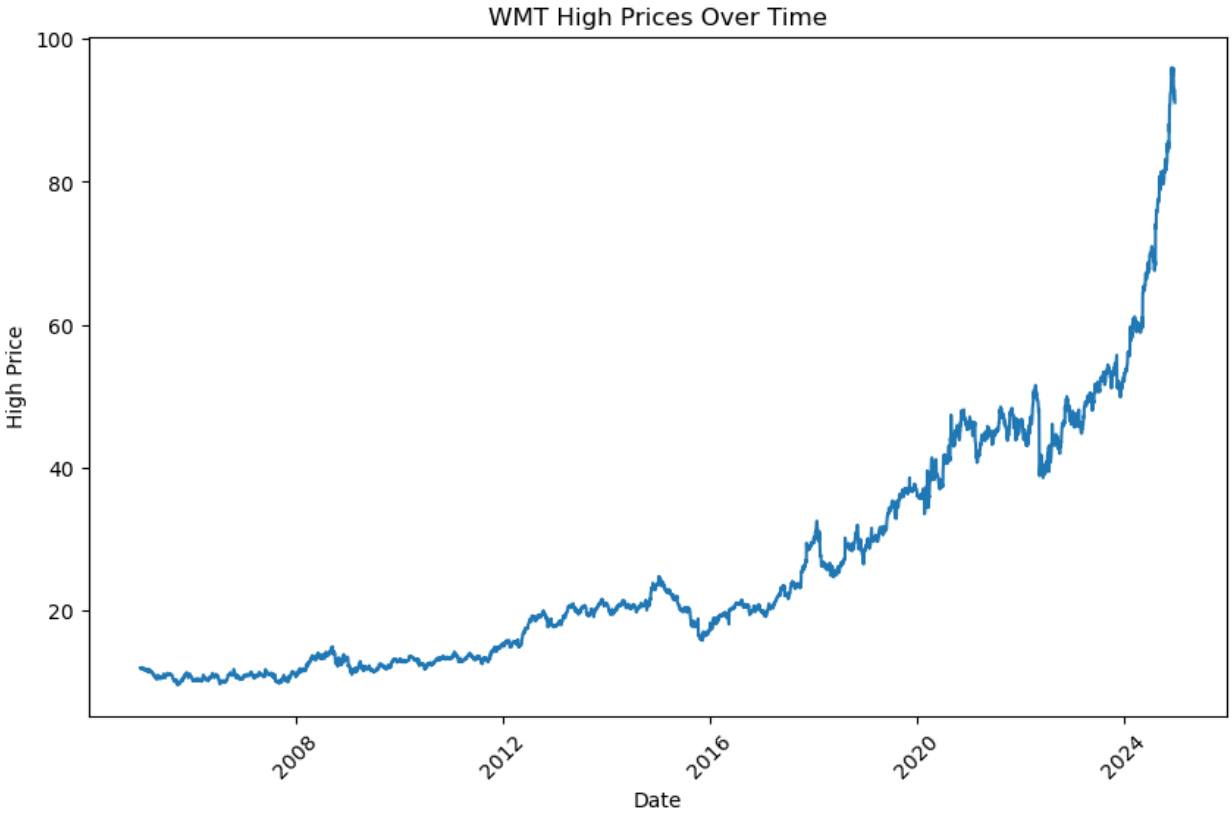

META High Prices Over Time
High Price
Date
0
100
200
300
400
500
600
2008
2012
2016
2020
2024

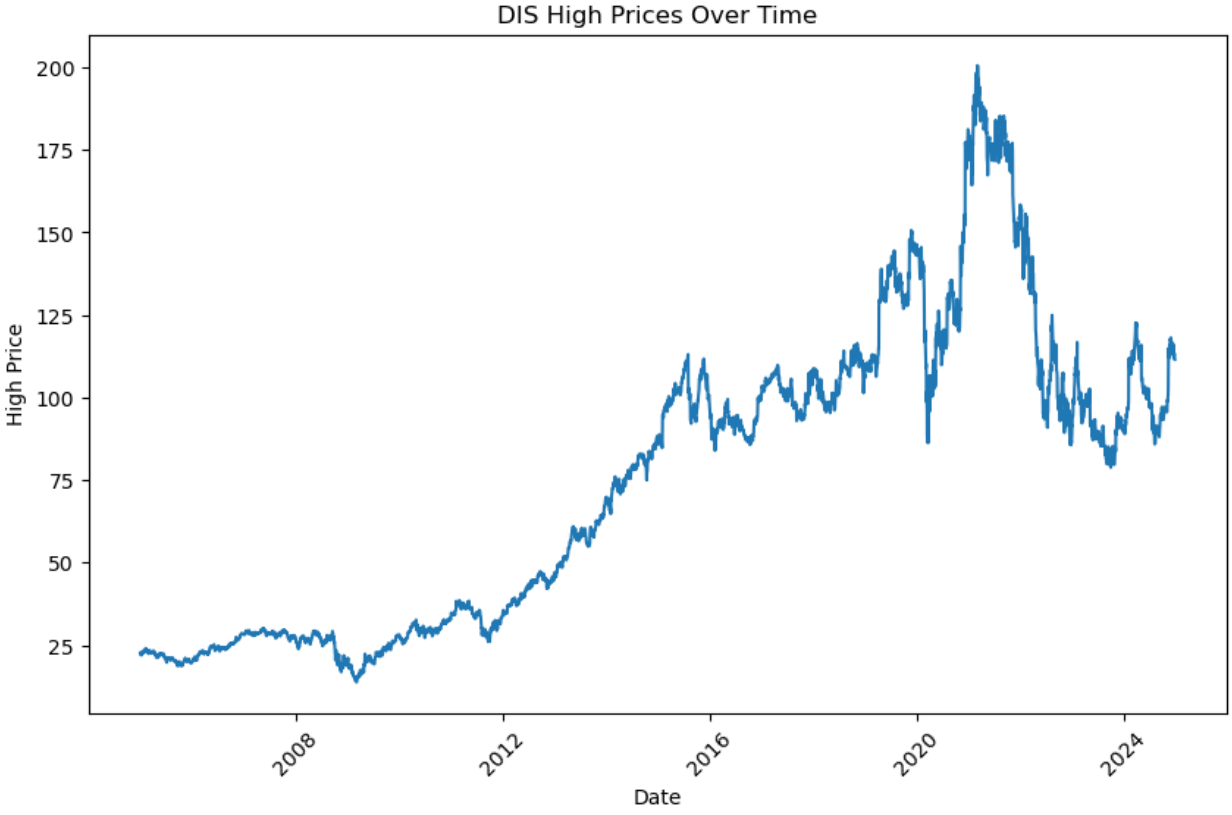

NVDA High Prices Over Time
High Price
Date
0
20
40
60
80
100
120
140
160
2008
2012
2016
2020
2024

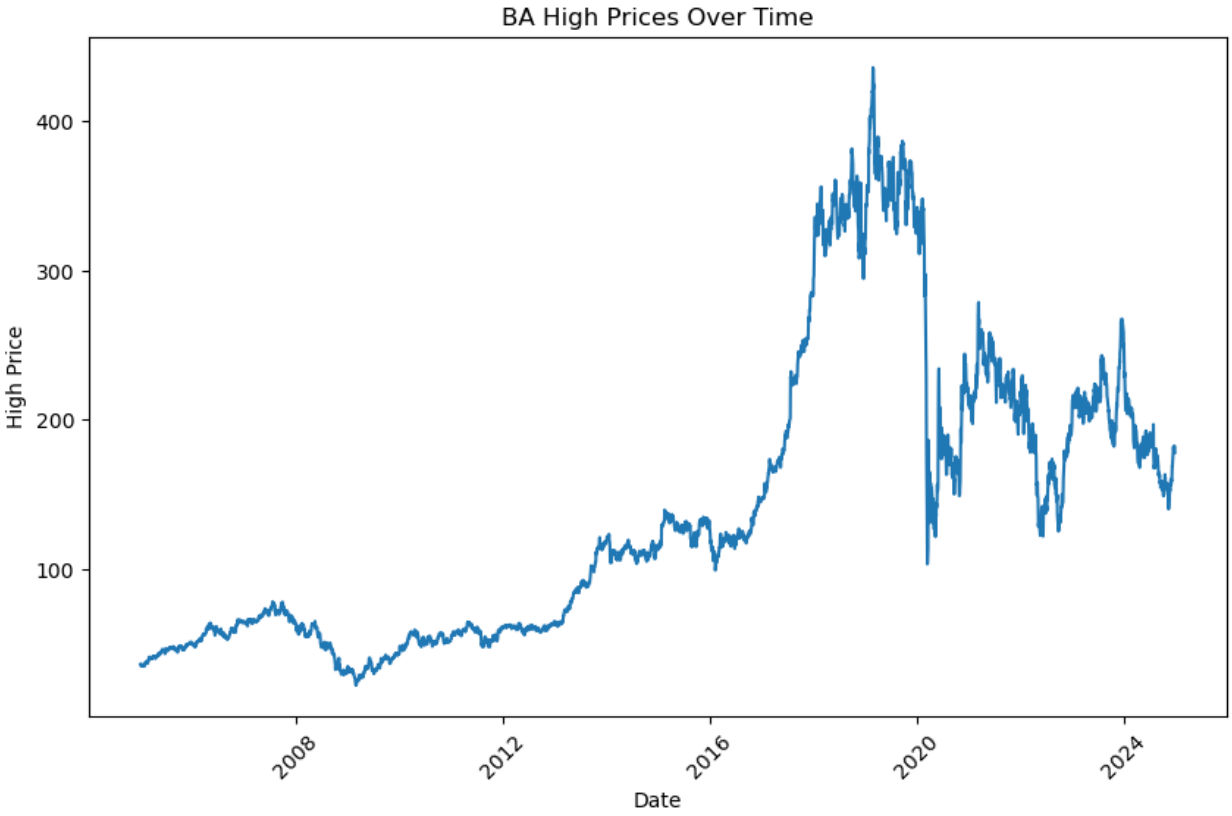

BRK-B High Prices Over Time
High Price
Date
100
200
300
400
500
2008
2012
2016
2020
2024

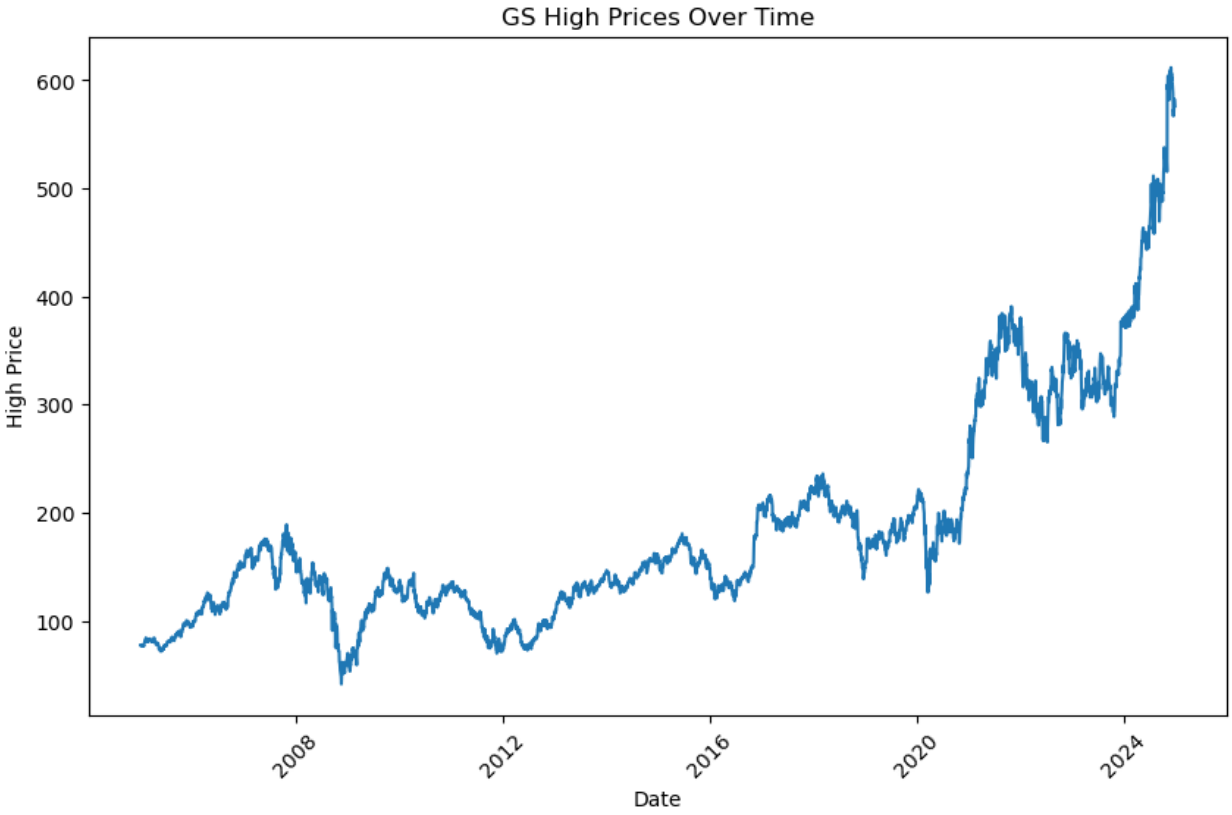

JPM High Prices Over Time
High Price
0
50
100
150
200
250
2008
2012
2016
2020
2024
Date

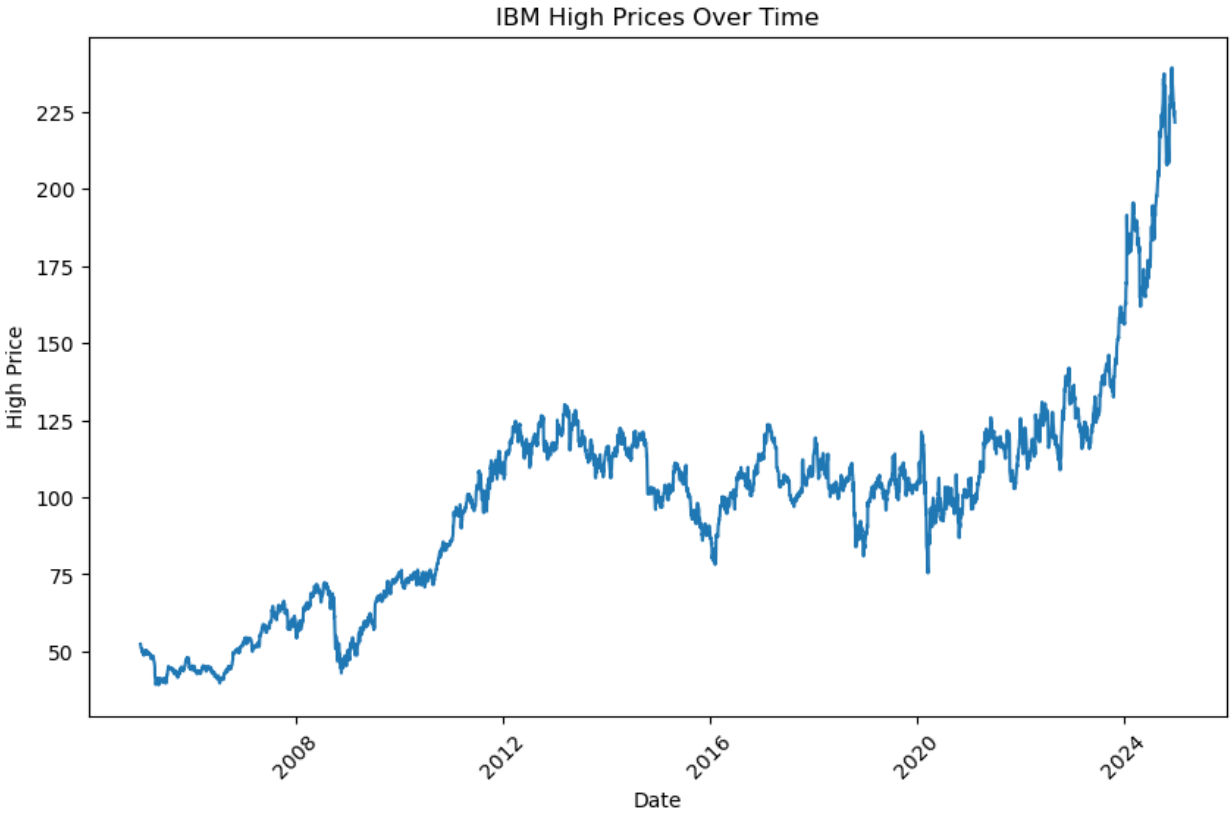

V High Prices Over Time
High Price
0
50
100
150
200
250
300
2008
2012
2016
2020
2024
Date

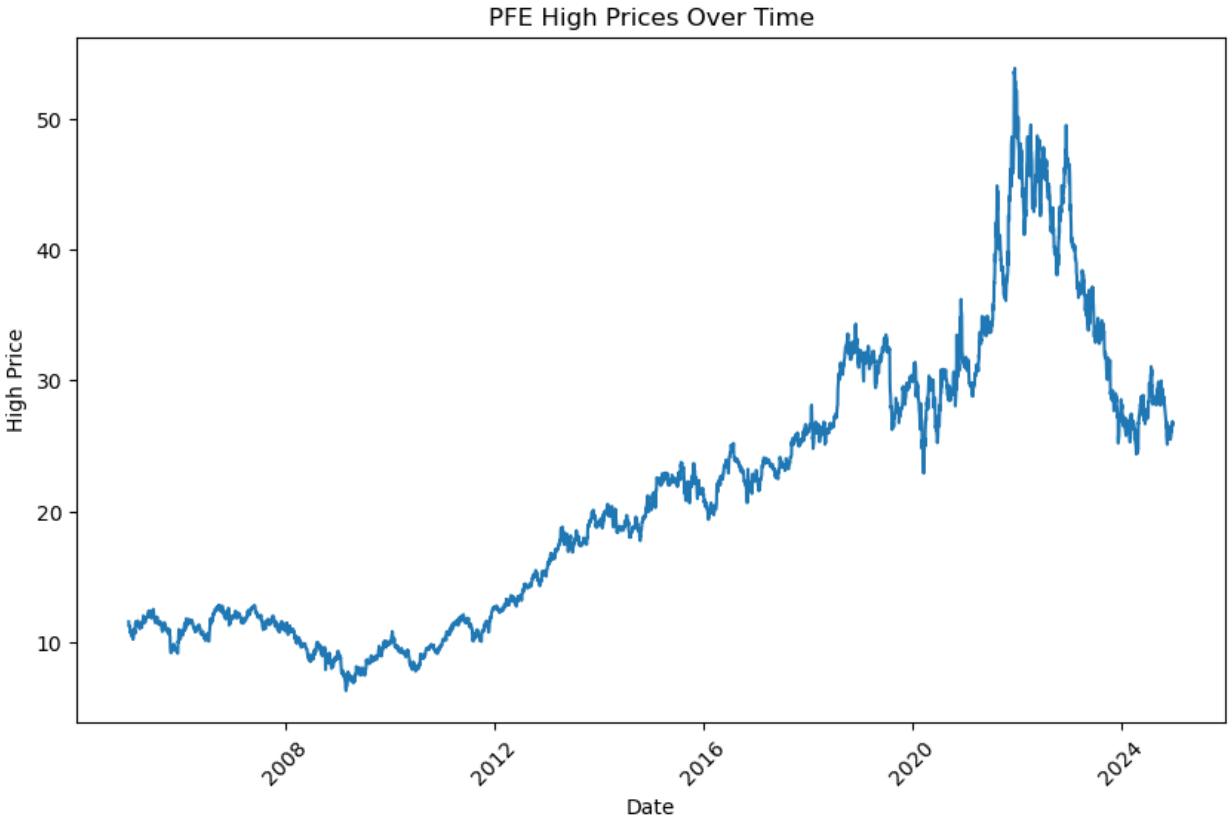

WMT High Prices Over Time
High Price
100
80
60
40
20
2008
2012
2016
2020
2024
Date


DIS High Prices Over Time
High Price
200
175
150
125
100
75
50
25
2008
2012
2016
2020
2024
Date


BA High Prices Over Time
High Price
100
200
300
400
2008
2012
2016
2020
2024
Date

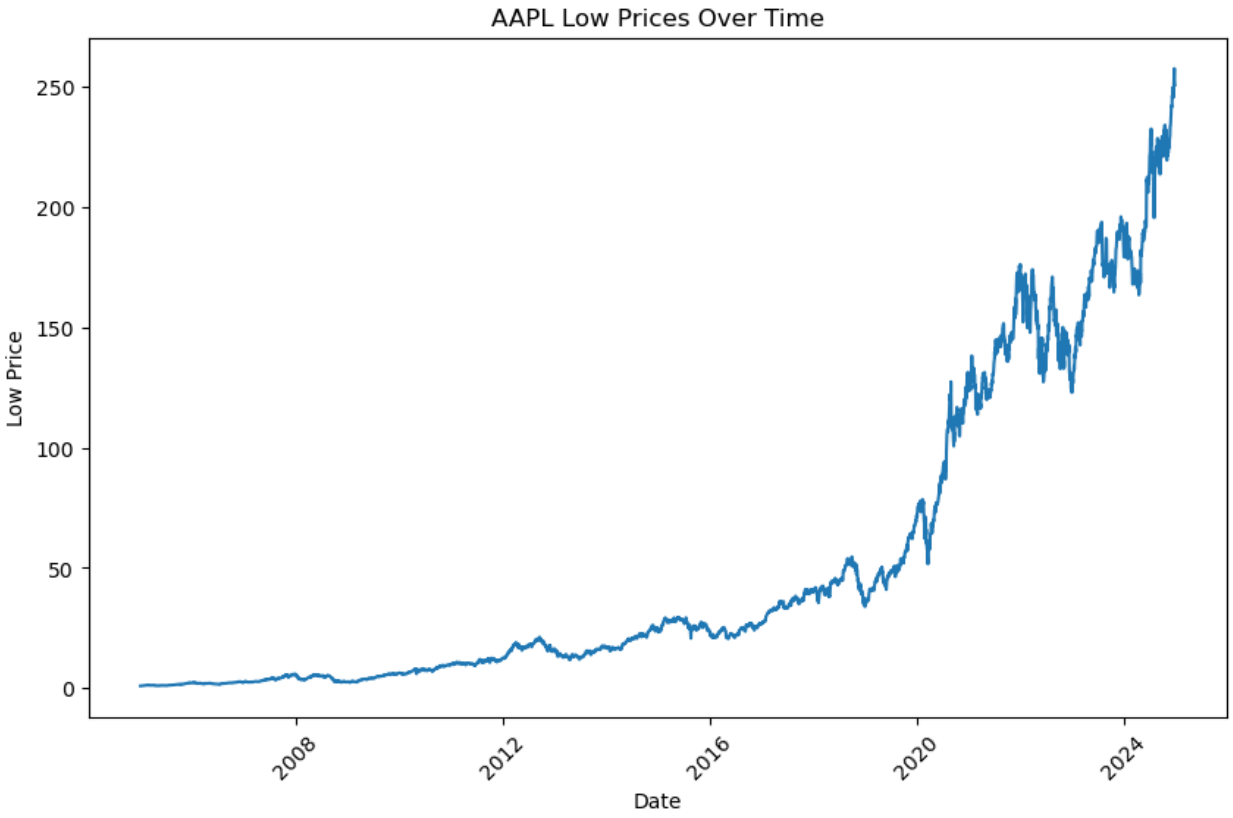

GS High Prices Over Time
High Price
100
200
300
400
500
600
2008
2012
2016
2020
2024
Date

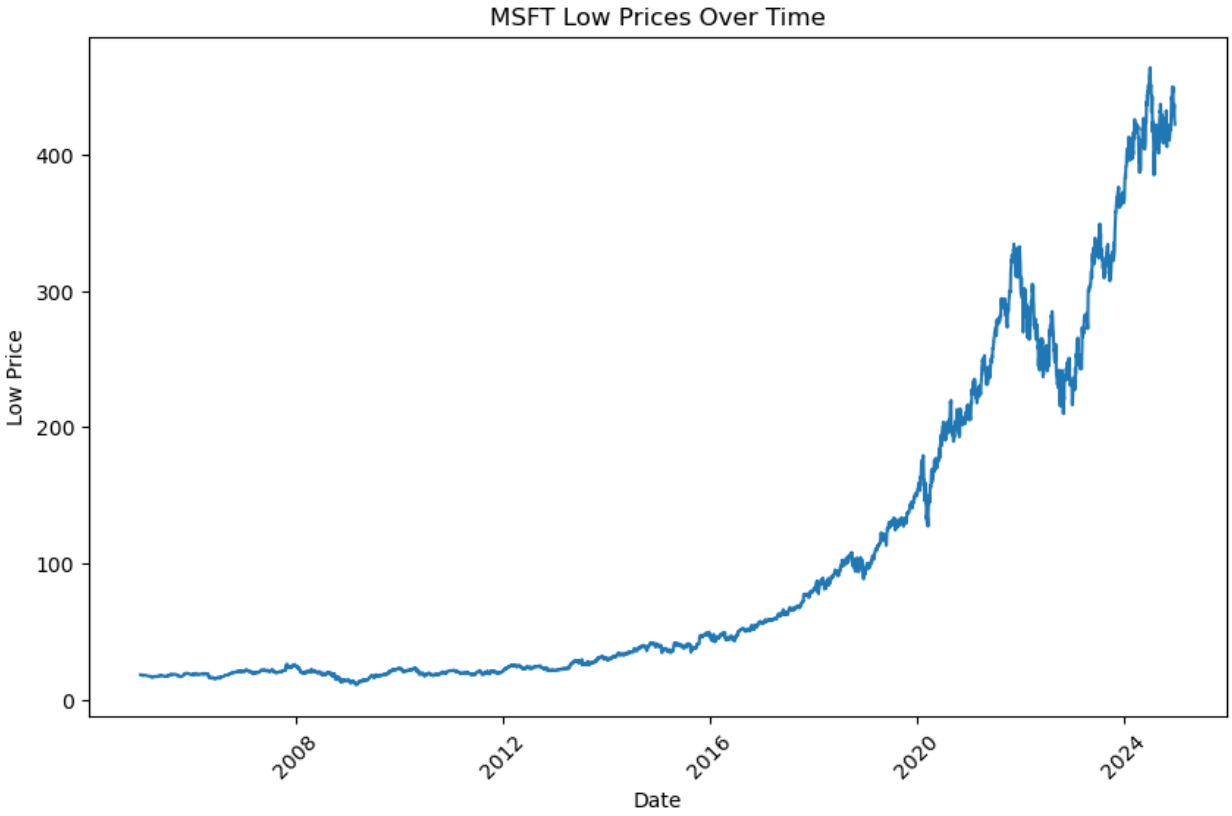

IBM High Prices Over Time
High Price
225
200
175
150
125
100
75
50
2008
2012
2016
2020
2024
Date

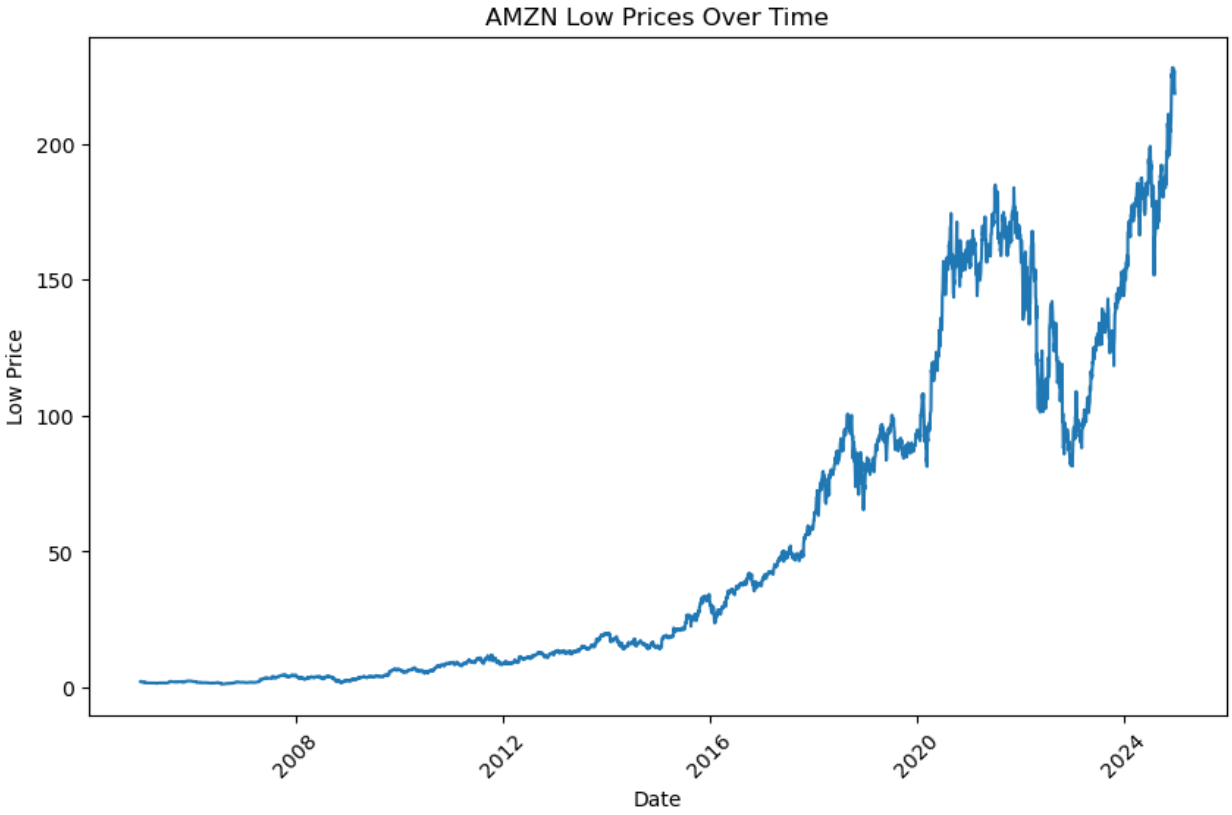

PFE High Prices Over Time
High Price
50
40
30
20
10
2008
2012
2016
2020
2024
Date

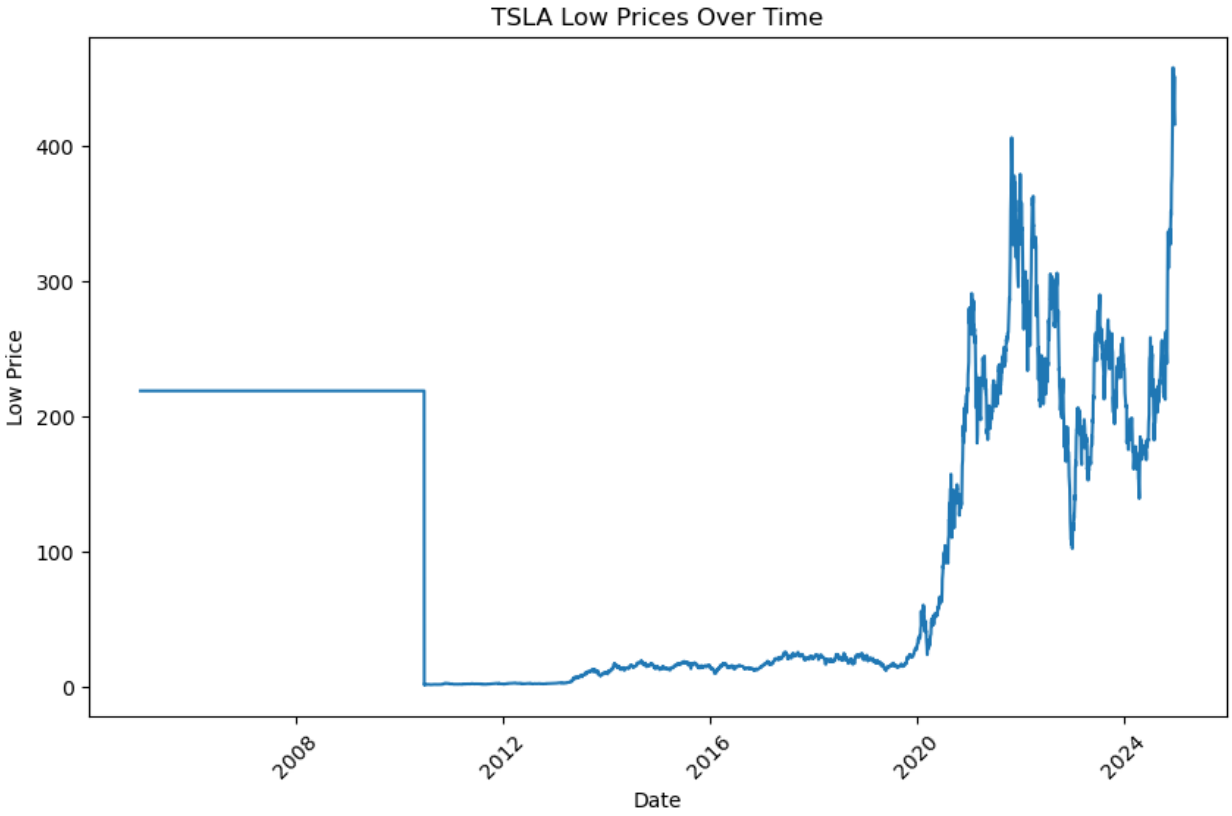
d. Line Plots for Low Prices:

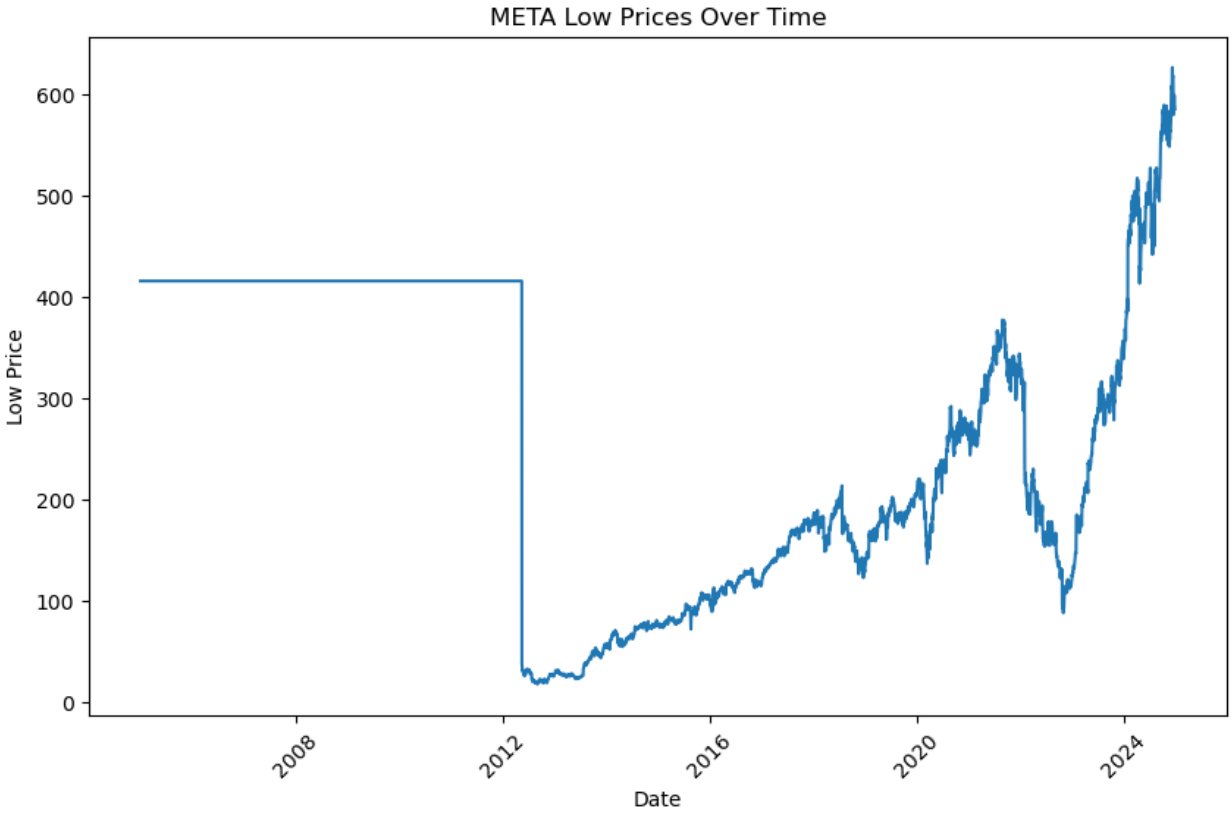
The objective of this line plot is to portray the low price movements of all firms across a span of time. The data depict the changes in the low prices over the years for each firm. Trends can tell how the company's stock performed based on years, thus coming with periods of growth, decline, or stability.

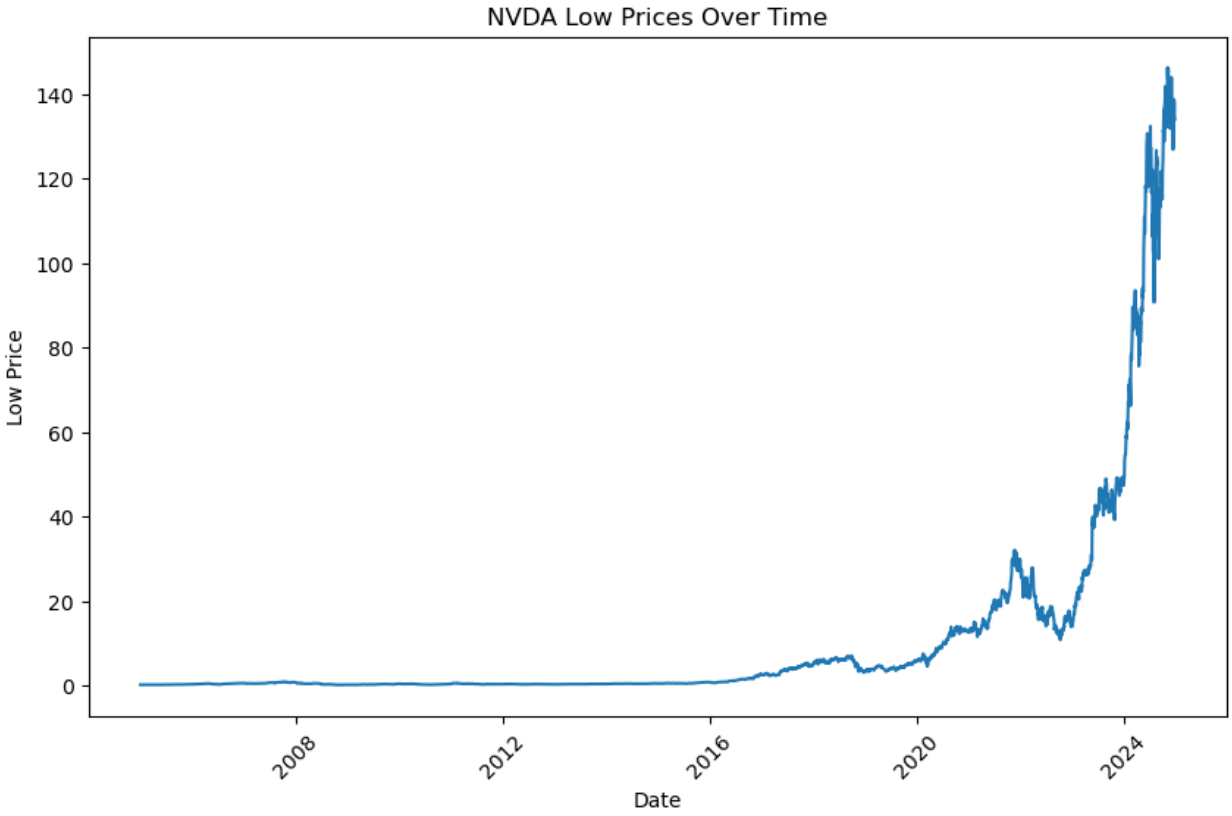


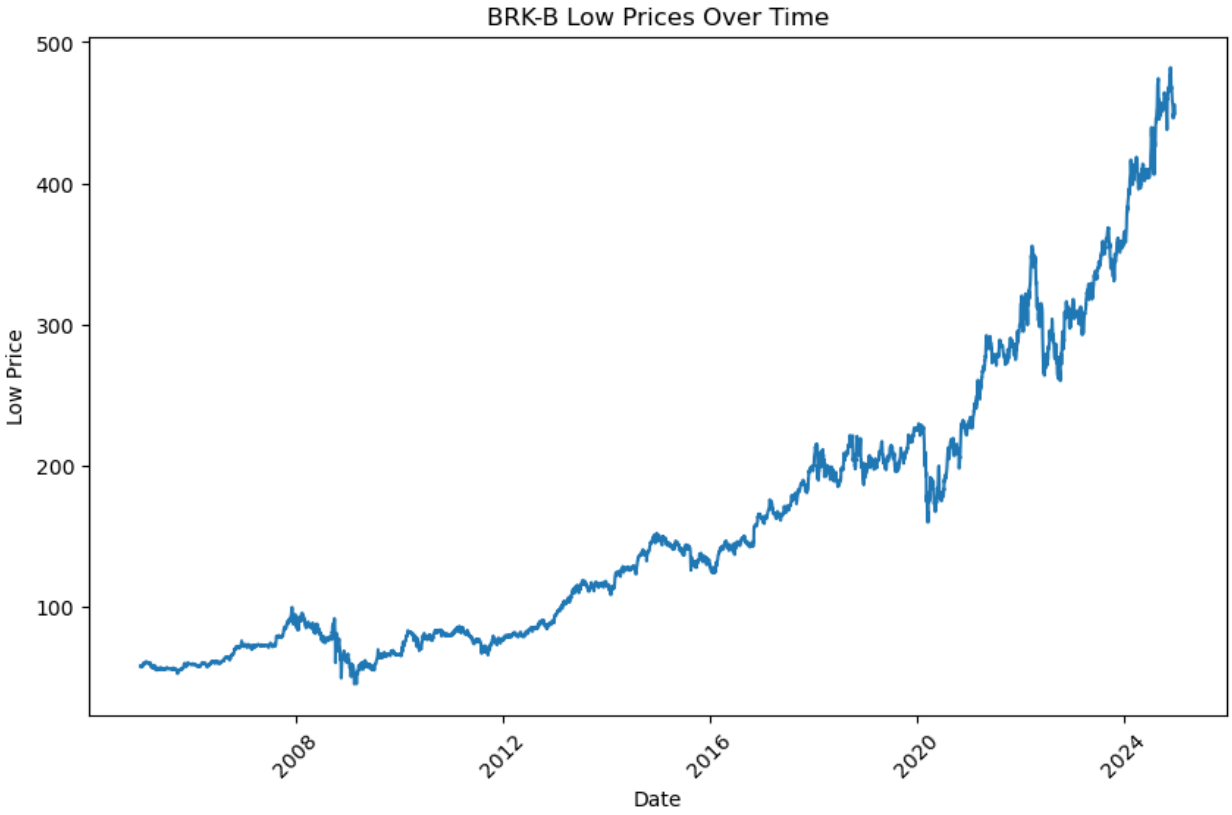

MSFT Low Prices Over Time
Low Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

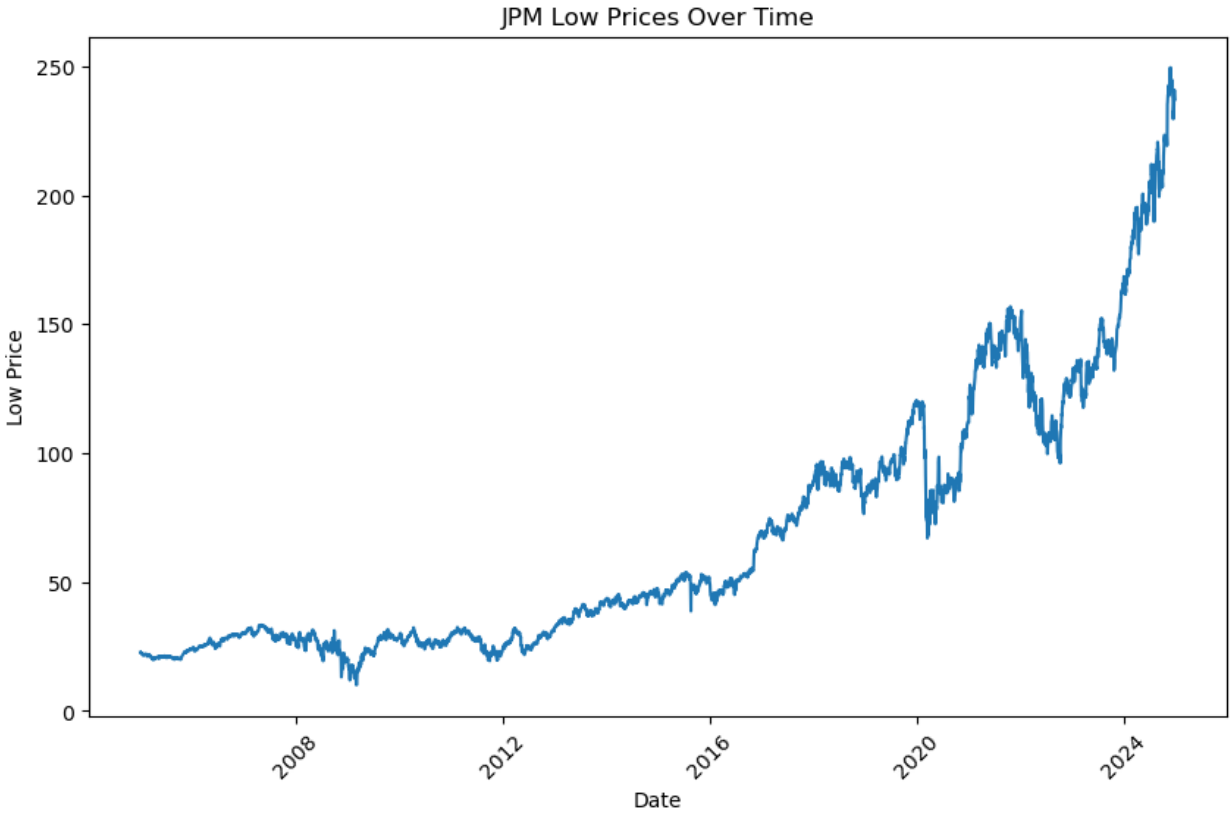

AMZN Low Prices Over Time
Low Price
0
50
100
150
200
2008
2012
2016
2020
2024
Date

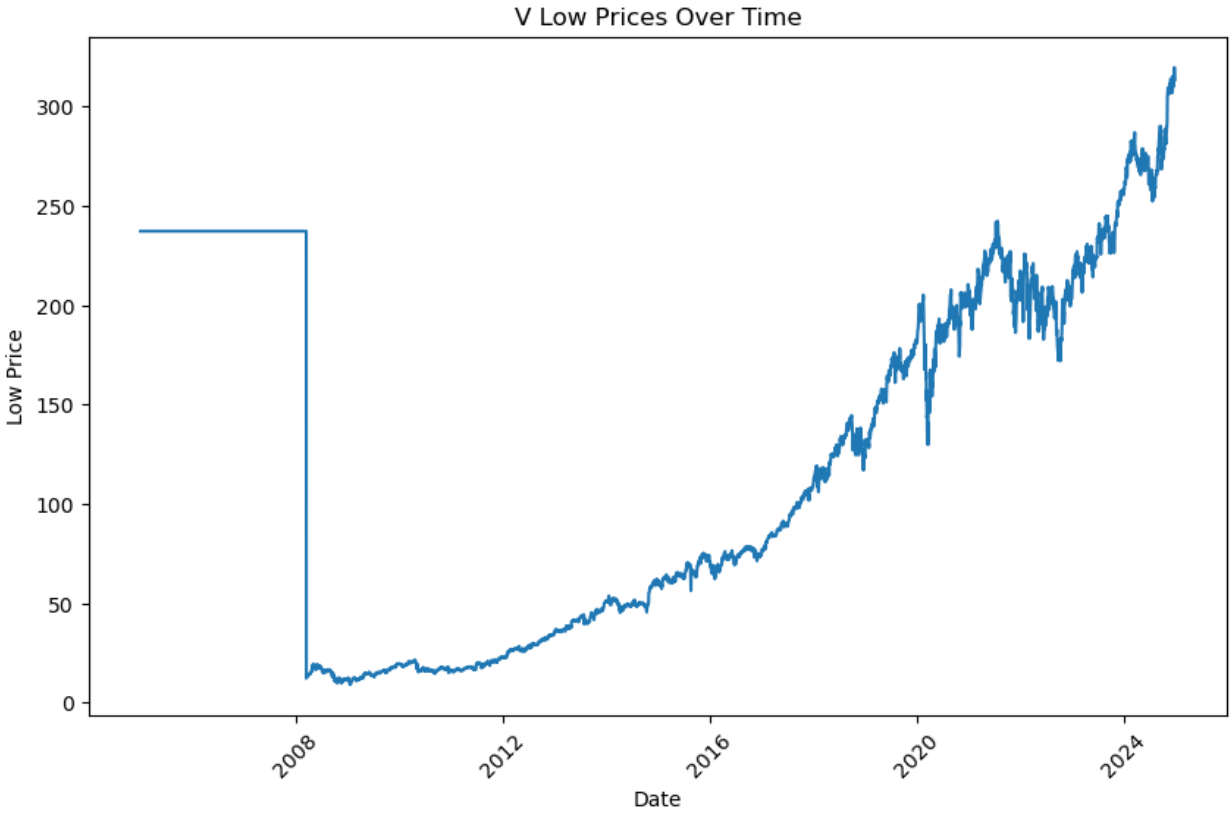

TSLA Low Prices Over Time
Low Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

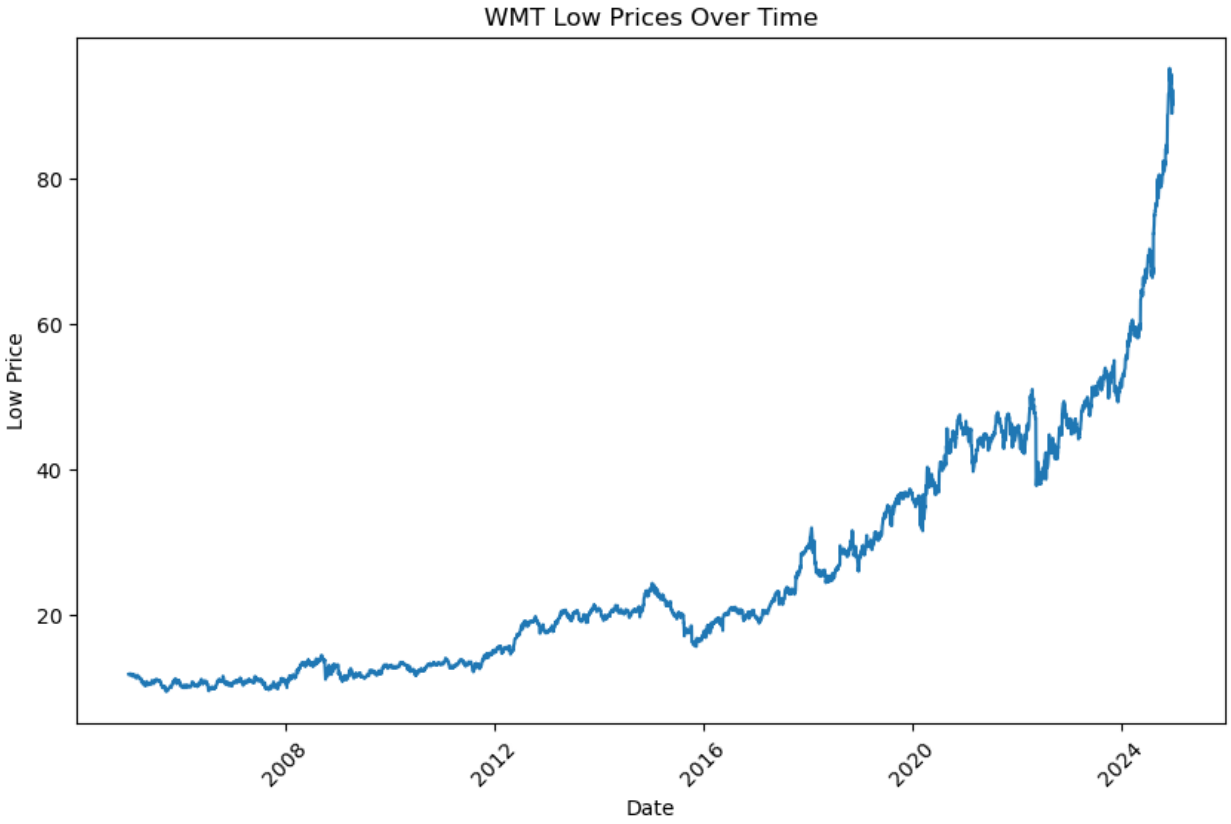

META Low Prices Over Time
Low Price
0
100
200
300
400
500
600
2008
2012
2016
2020
2024
Date

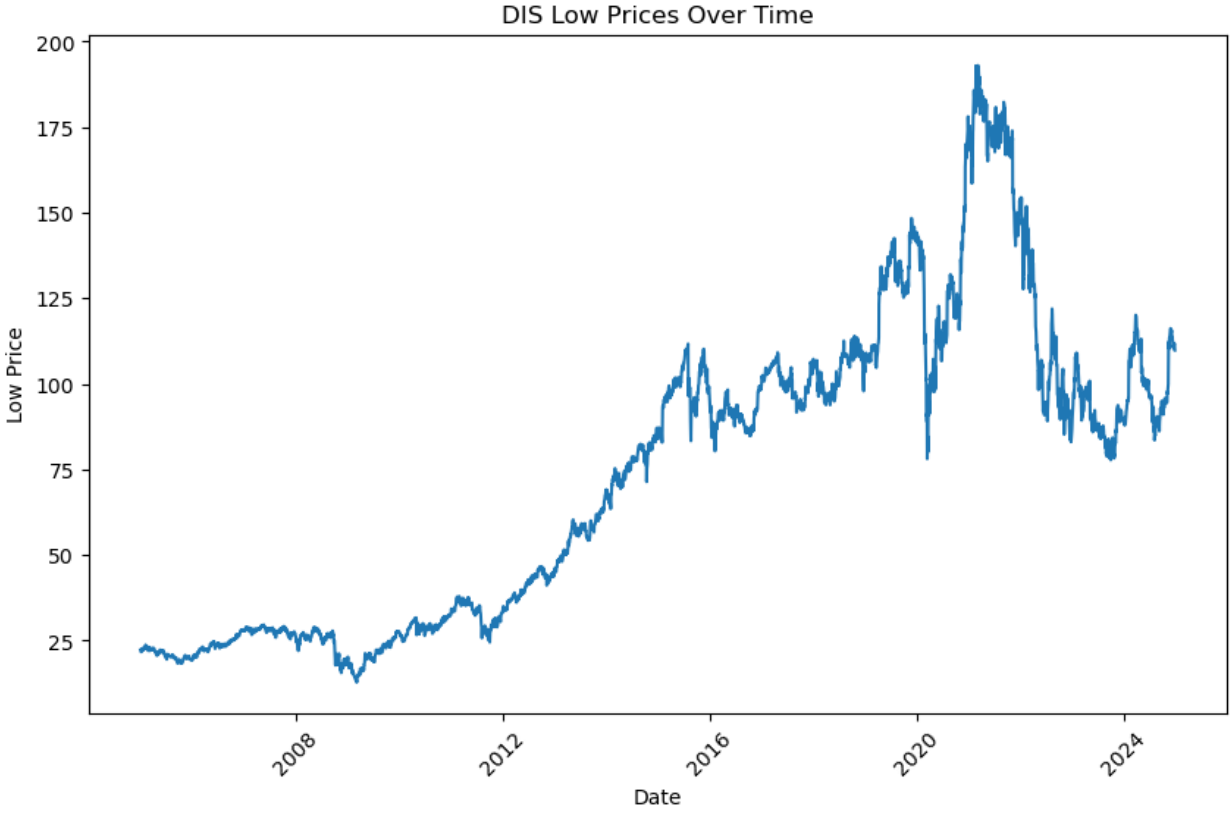

NVDA Low Prices Over Time
Low Price
0
20
40
60
80
100
120
140
2008
2012
2016
2020
2024
Date

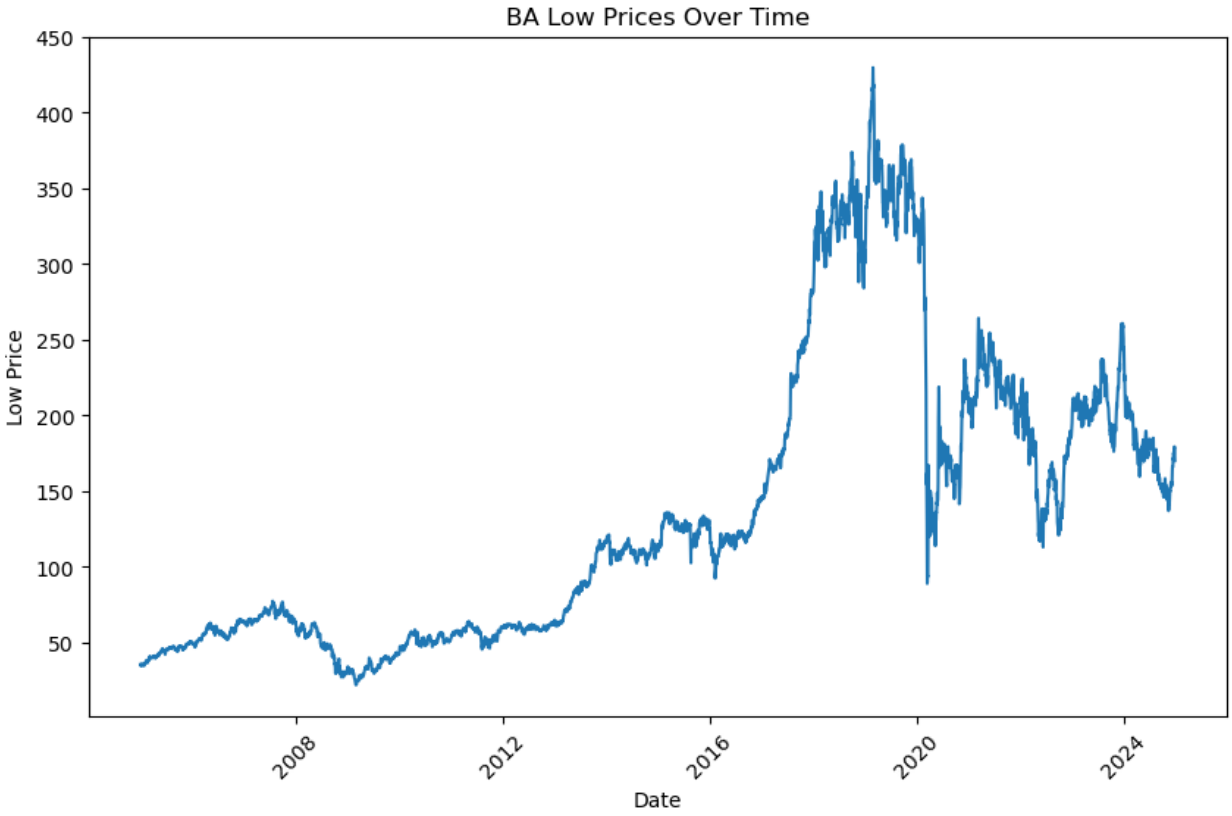

BRK-B Low Prices Over Time
Low Price
100
200
300
400
500
2008
2012
2016
2020
2024
Date

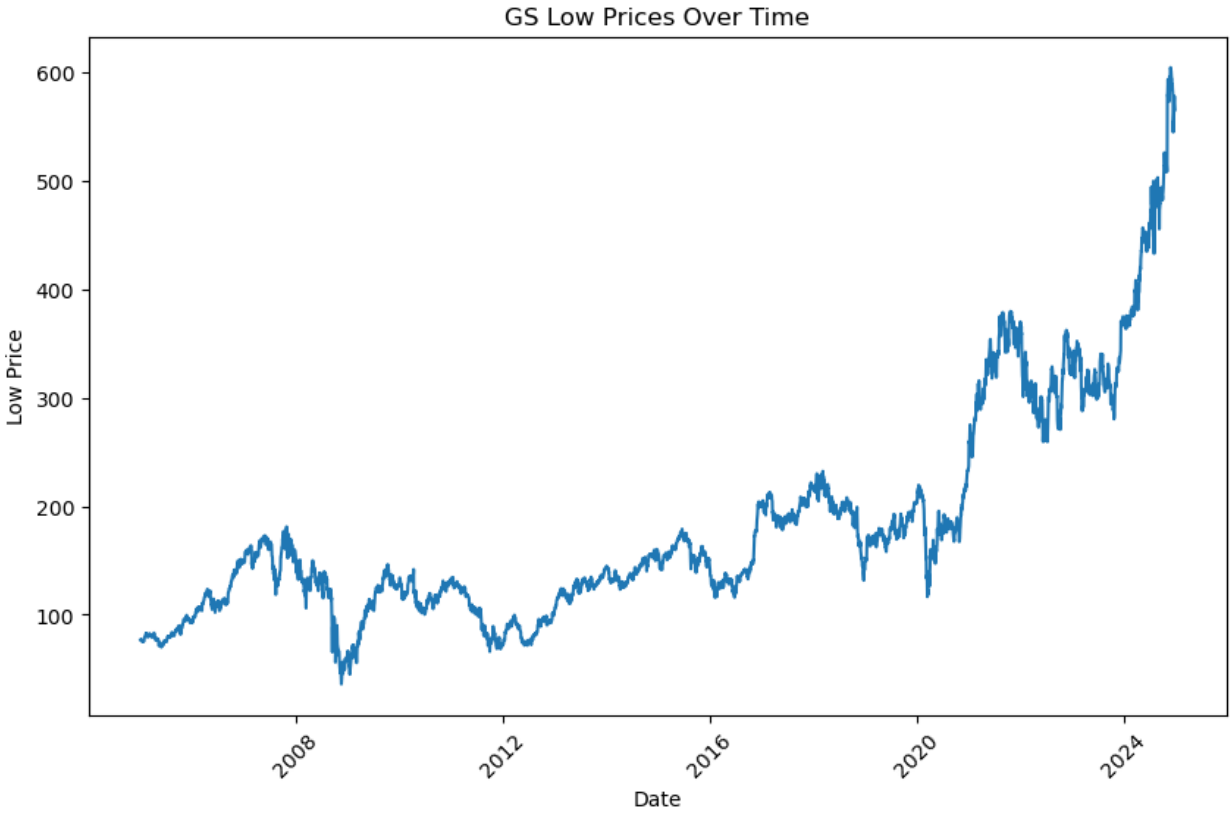

JPM Low Prices Over Time
Low Price
Date
0
50
100
150
200
250
2008
2012
2016
2020
2024

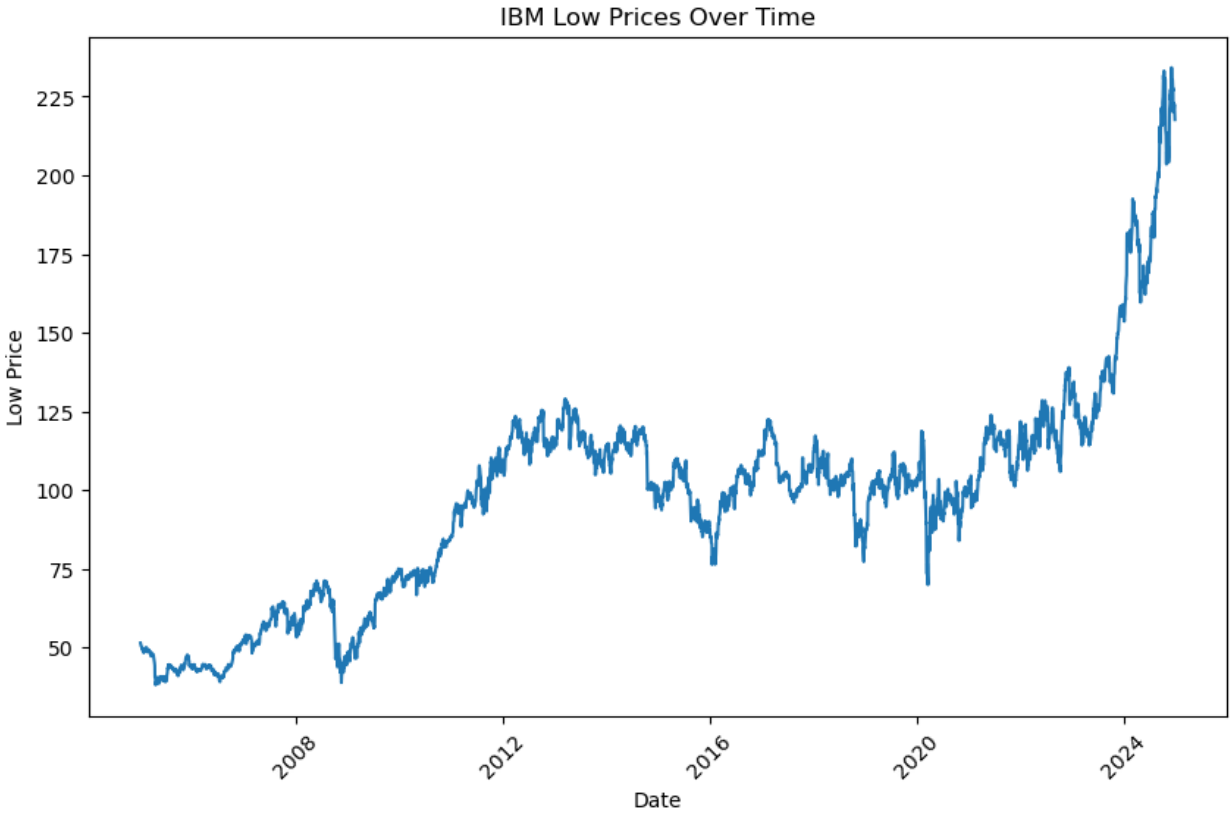

V Low Prices Over Time
Low Price
Date
0
50
100
150
200
250
300
2008
2012
2016
2020
2024

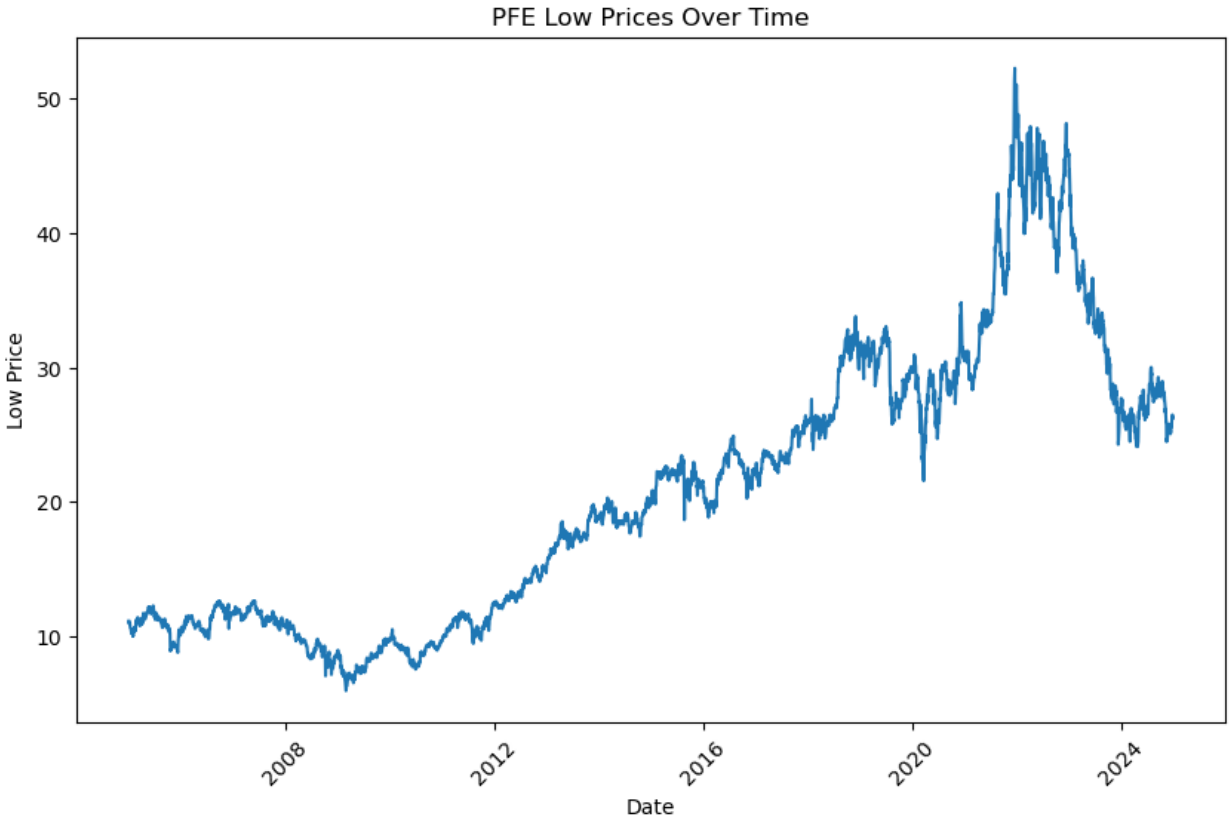

WMT Low Prices Over Time
Low Price
80
60
40
20
2008
2012
2016
2020
2024
Date


DIS Low Prices Over Time
Low Price
200
175
150
125
100
75
50
25
2008
2012
2016
2020
2024
Date


BA Low Prices Over Time
Low Price
Date
450
400
350
300
250
200
150
100
50
2008
2012
2016
2020
2024

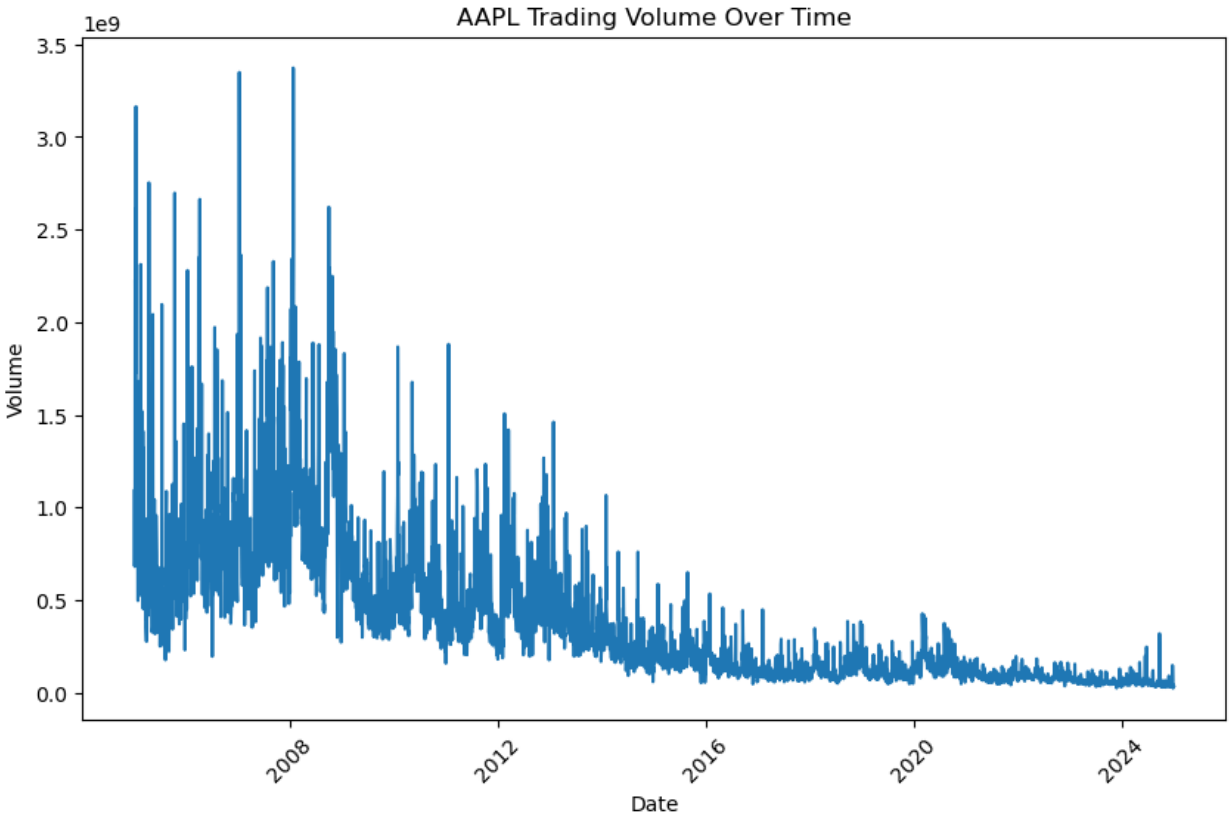

GS Low Prices Over Time
Low Price
Date
600
500
400
300
200
100
2008
2012
2016
2020
2024

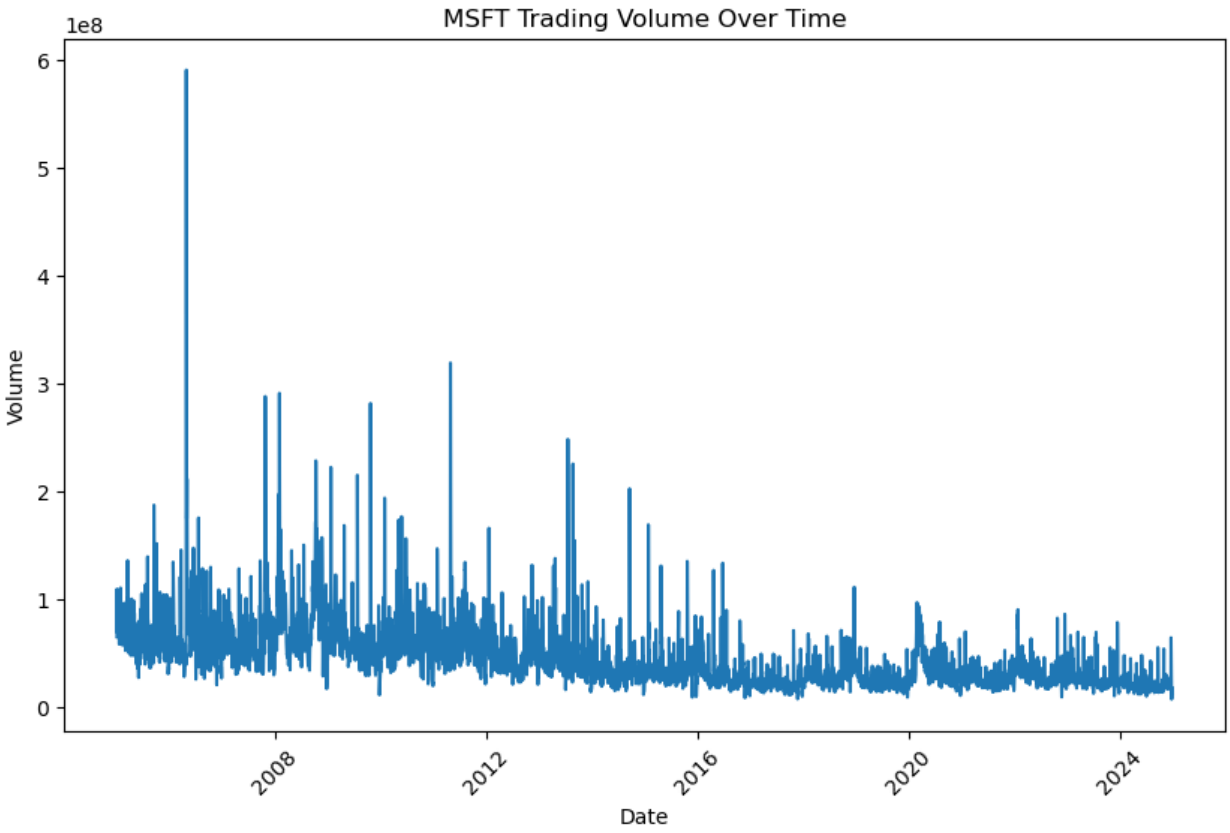

IBM Low Prices Over Time
Low Price
225
200
175
150
125
100
75
50
2008
2012
2016
2020
2024
Date

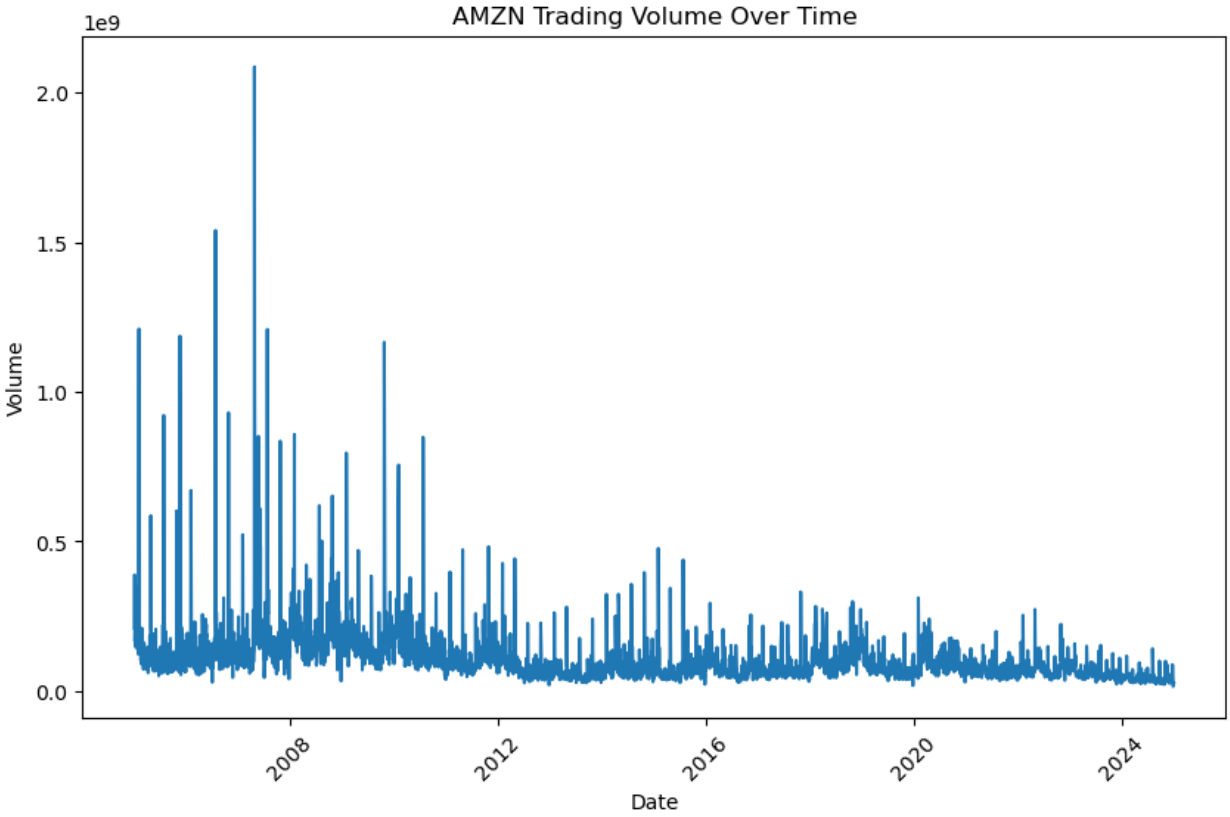

PFE Low Prices Over Time
Low Price
50
40
30
20
10
2008
2012
2016
2020
2024
Date

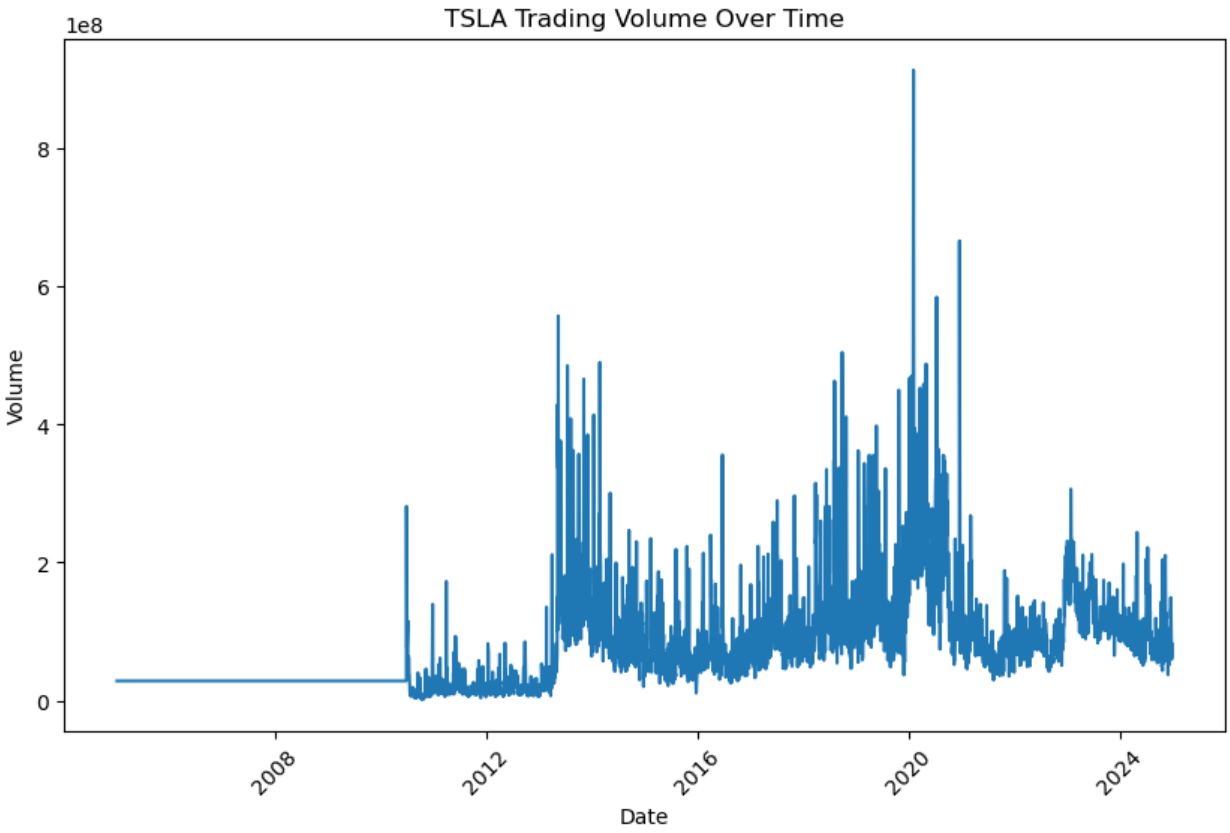
e. Time Plots for Trading Volume:

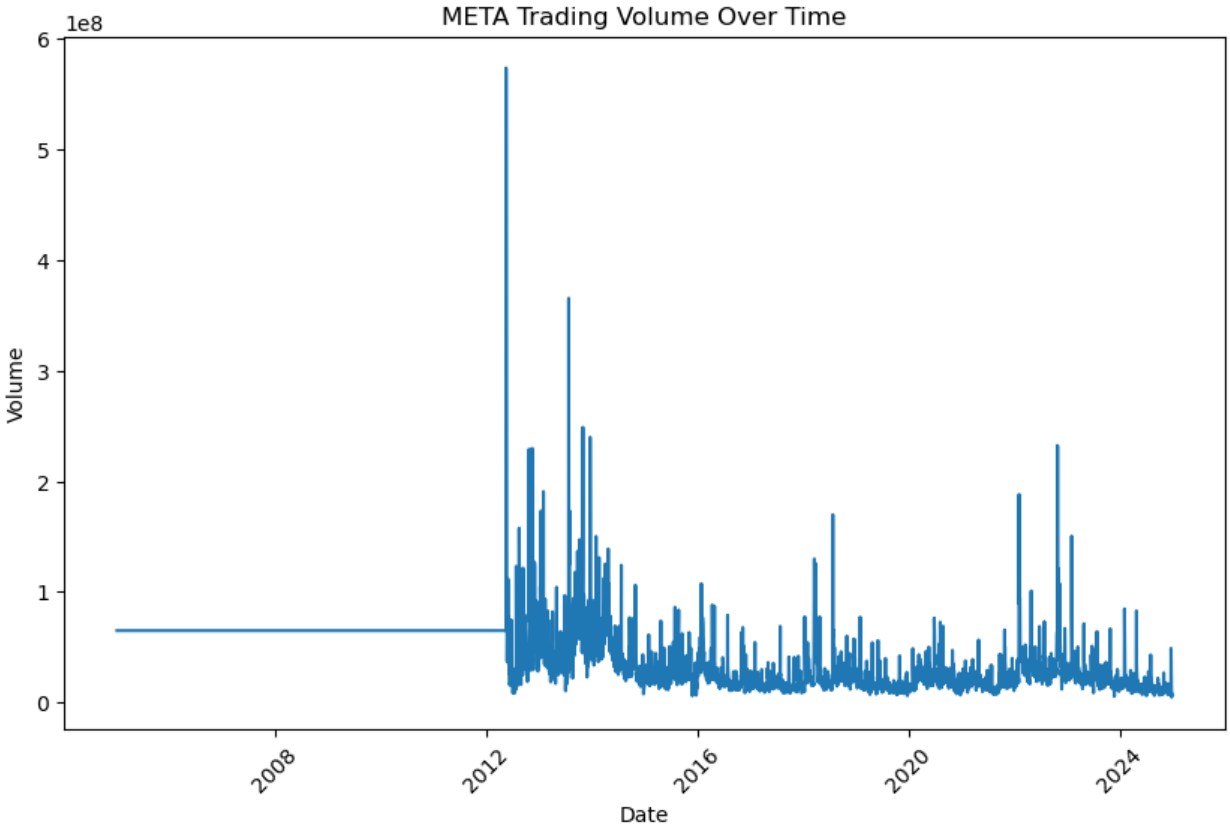
This graph is designed to show the time series for each company for trading volume. The number of shares traded on a particular day can be thought of as a broadcasting unit of market attraction and liquidity, with heftier periods of volume correlating with meaningful market events or company-specific news.

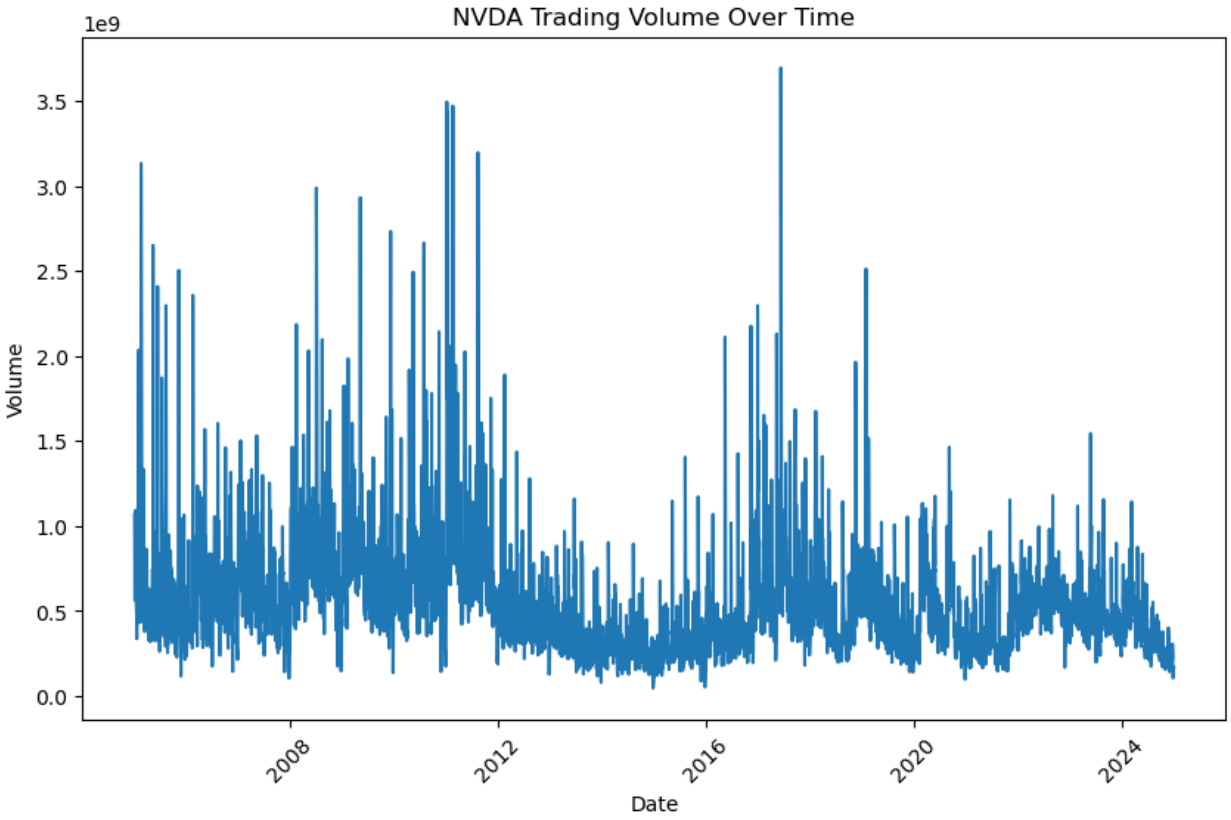


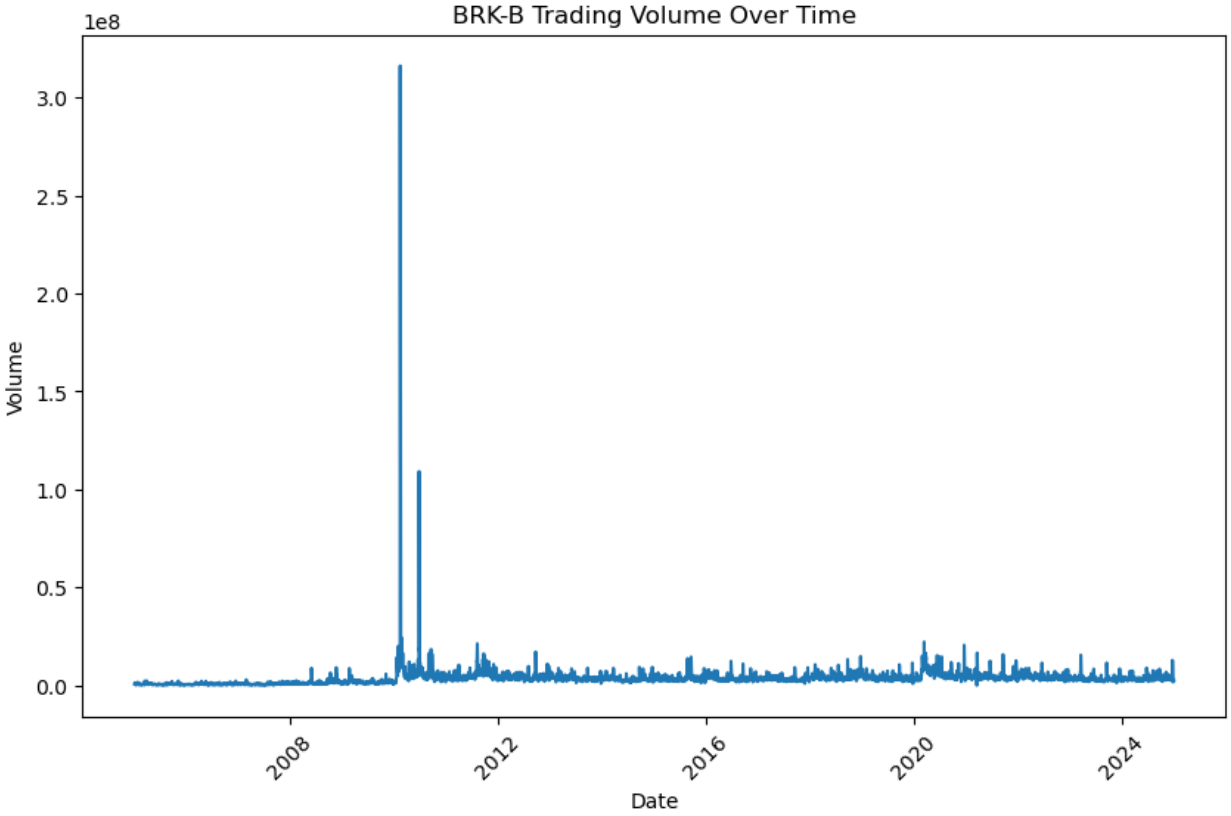

MSFT Trading Volume Over Time
1e8
6
5
4
3
2
1
0
Volume
2008
2012
2016
2020
2024
Date

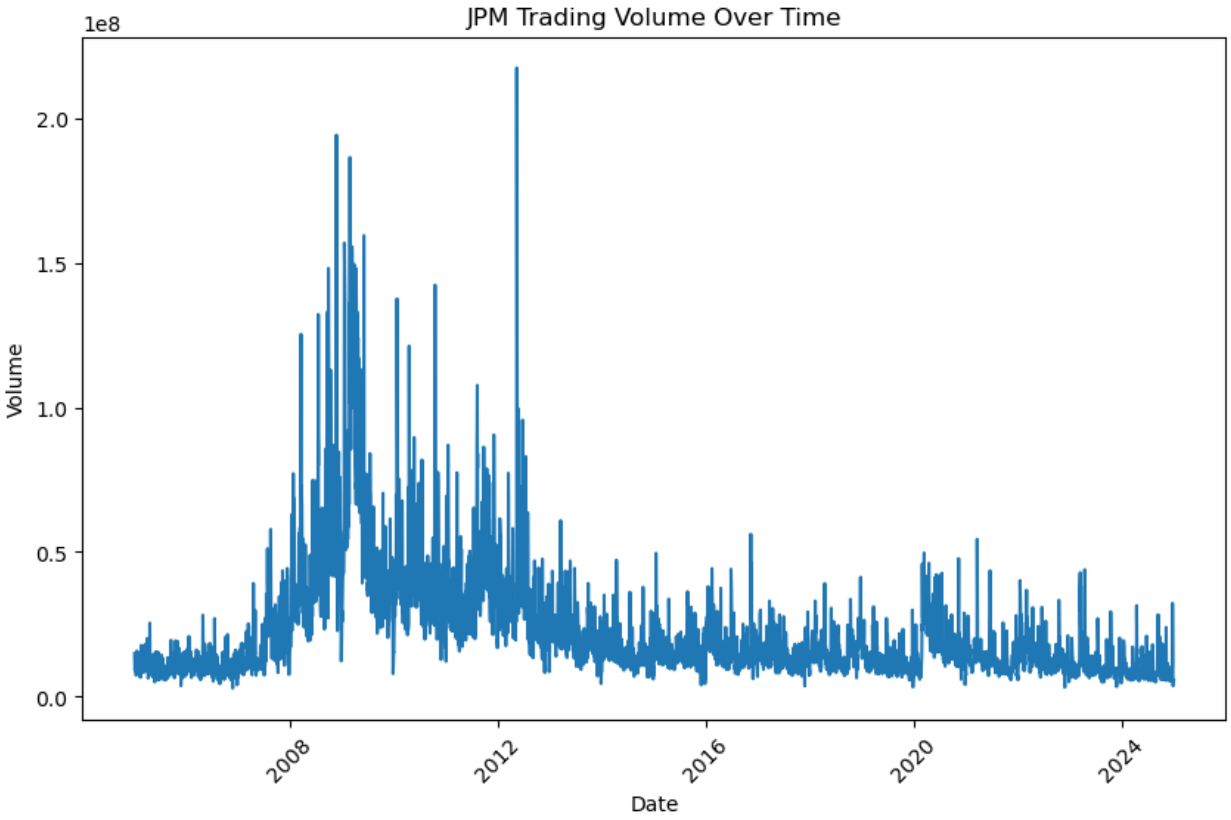

AMZN Trading Volume Over Time
1e9
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date

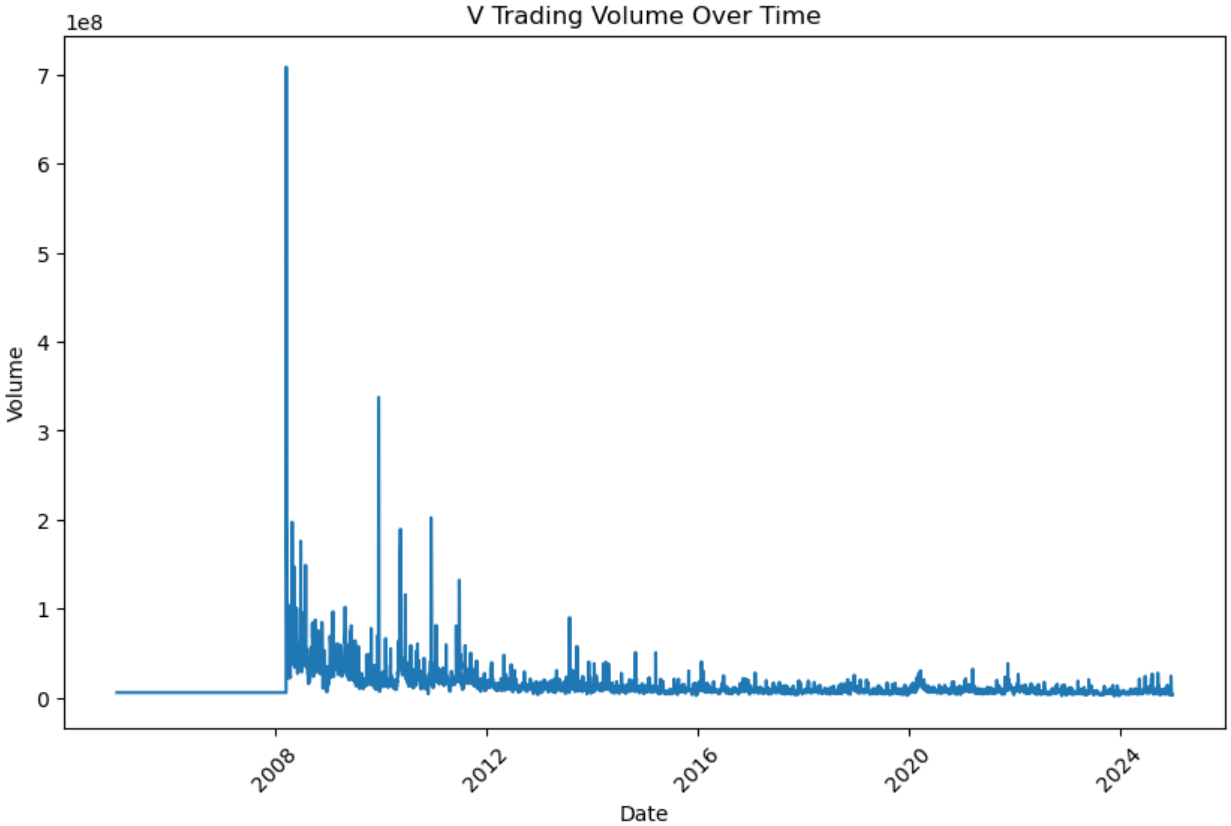

TSLA Trading Volume Over Time
1e8
8
6
4
2
0
Volume
2008
2012
2016
2020
2024
Date

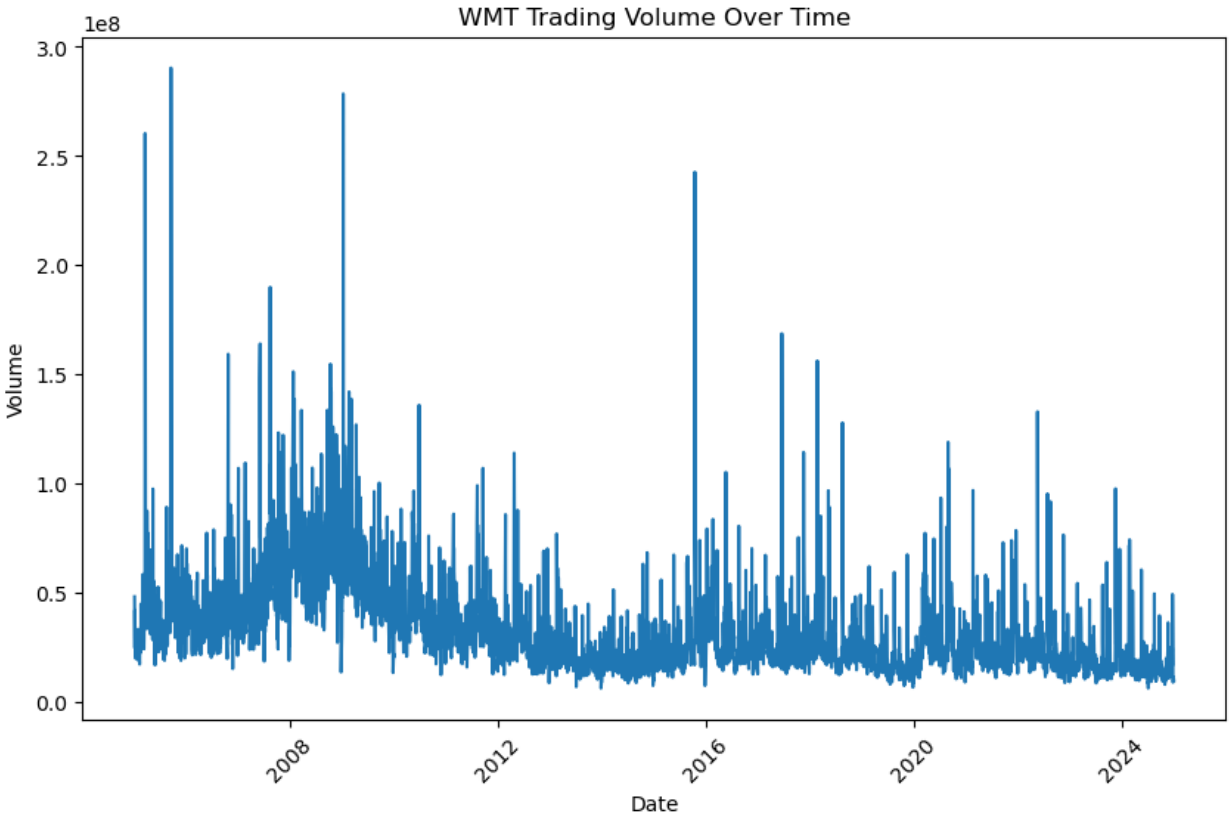

META Trading Volume Over Time
1e8
6
5
4
3
2
1
0
Volume
2008
2012
2016
2020
2024
Date

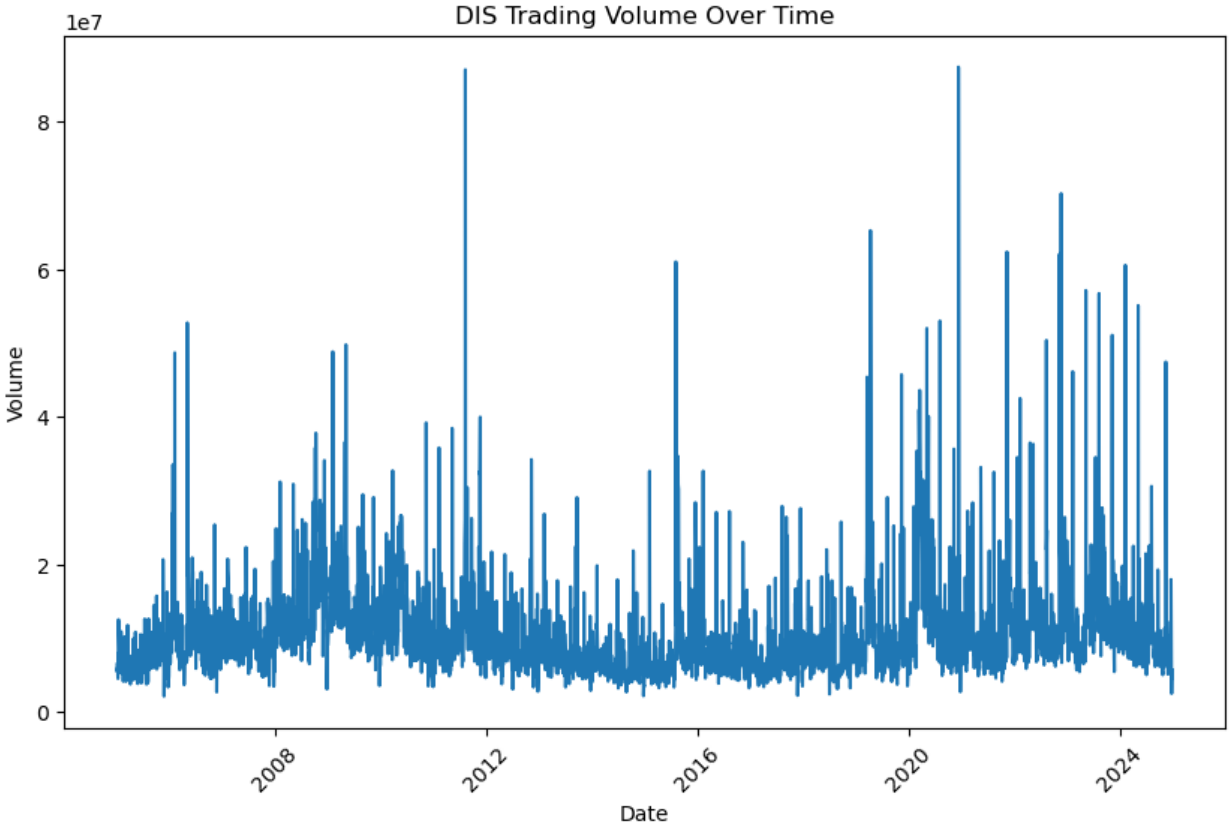

NVDA Trading Volume Over Time
1e9
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date

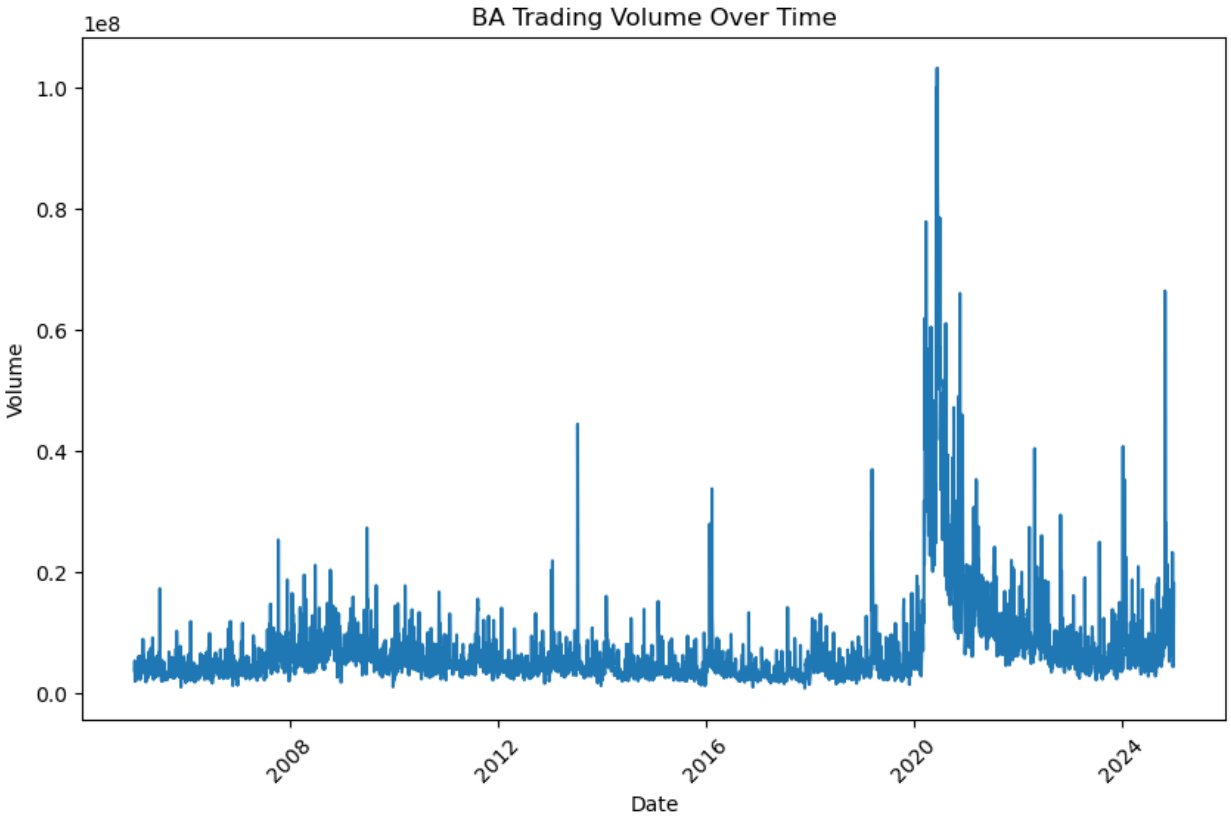

BRK-B Trading Volume Over Time
1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date

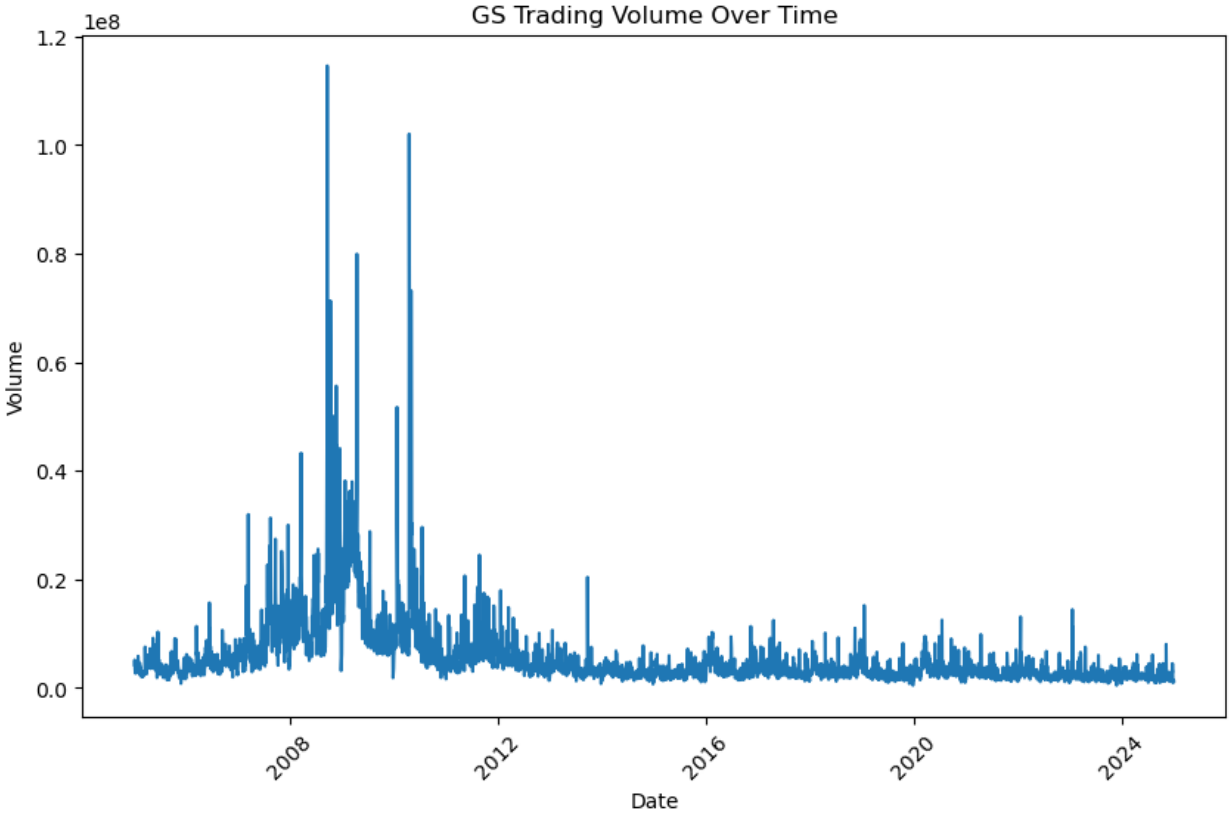

JPM Trading Volume Over Time
1e8
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date

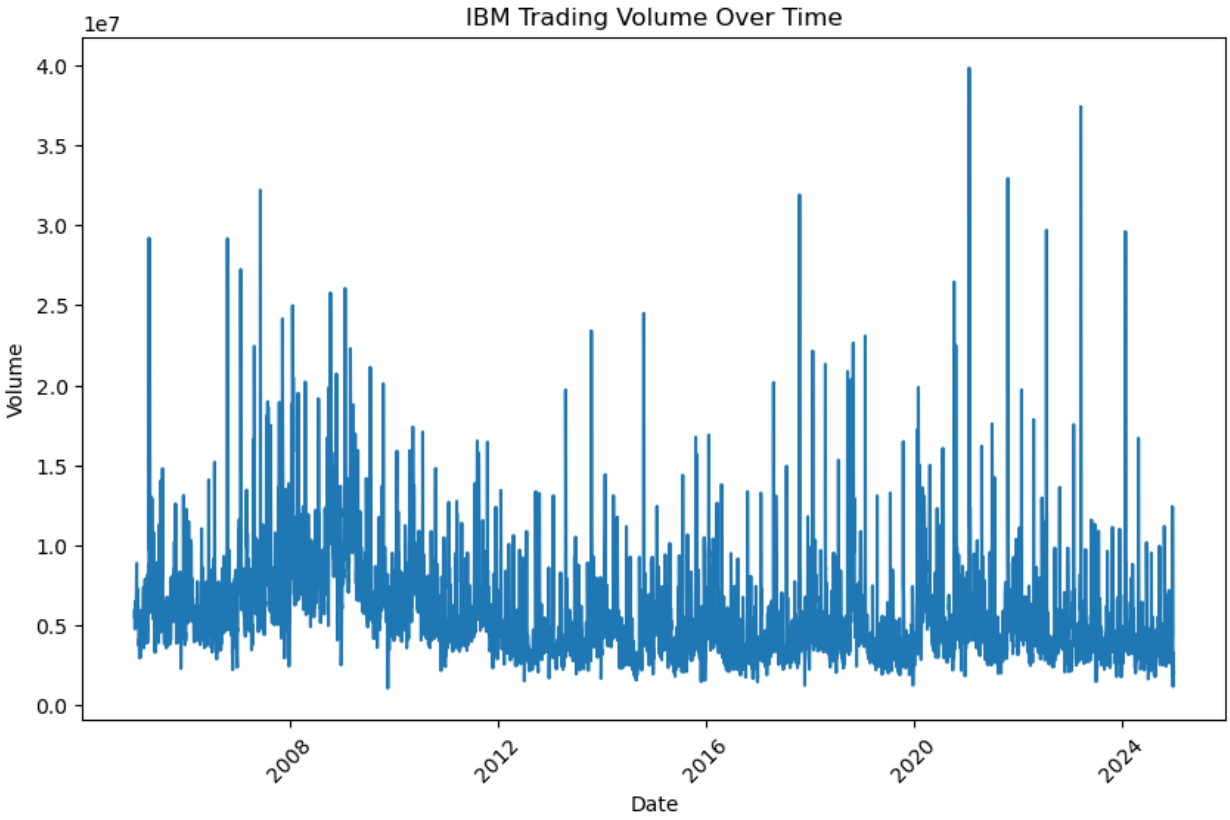

V Trading Volume Over Time
1e8
7
6
5
4
3
2
1
0
Volume
2008
2012
2016
2020
2024
Date

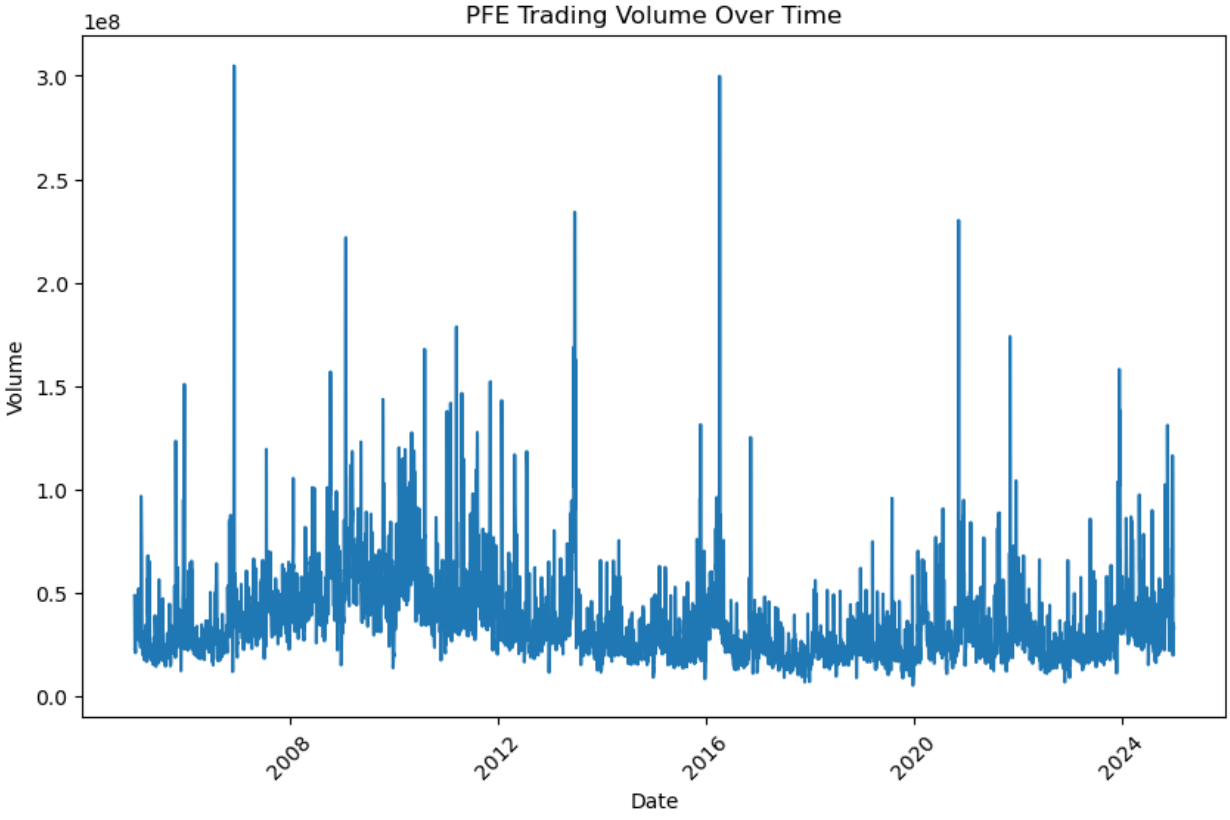

WMT Trading Volume Over Time
1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date


DIS Trading Volume Over Time
1e7
8
6
4
2
0
Volume
2008
2012
2016
2020
2024
Date


BA Trading Volume Over Time
1e8
1.0
0.8
0.6
0.4
0.2
0.0
Volume
2008
2012
2016
2020
2024
Date

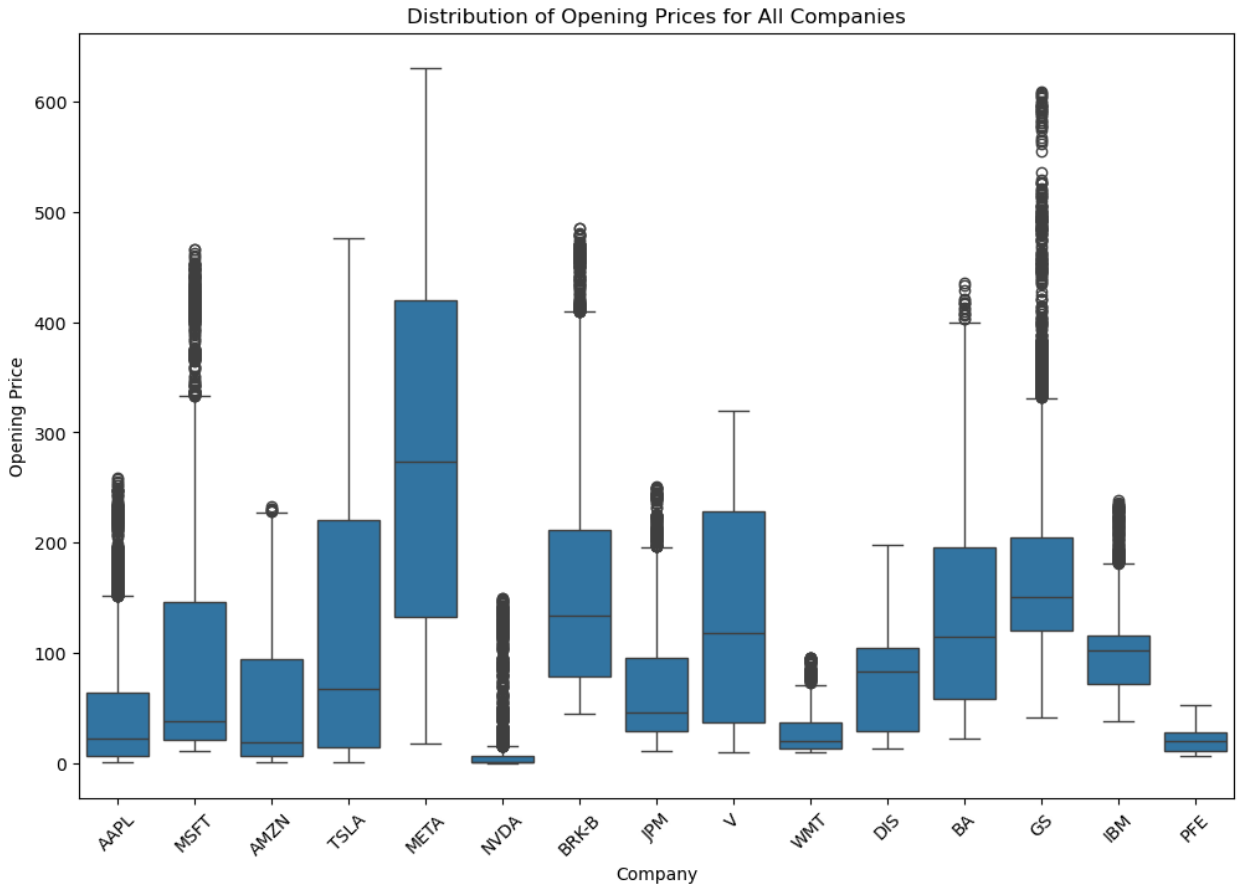

GS Trading Volume Over Time
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Volume
2008
2012
2016
2020
2024
Date


IBM Trading Volume Over Time
1e7
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date


PFE Trading Volume Over Time
1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date

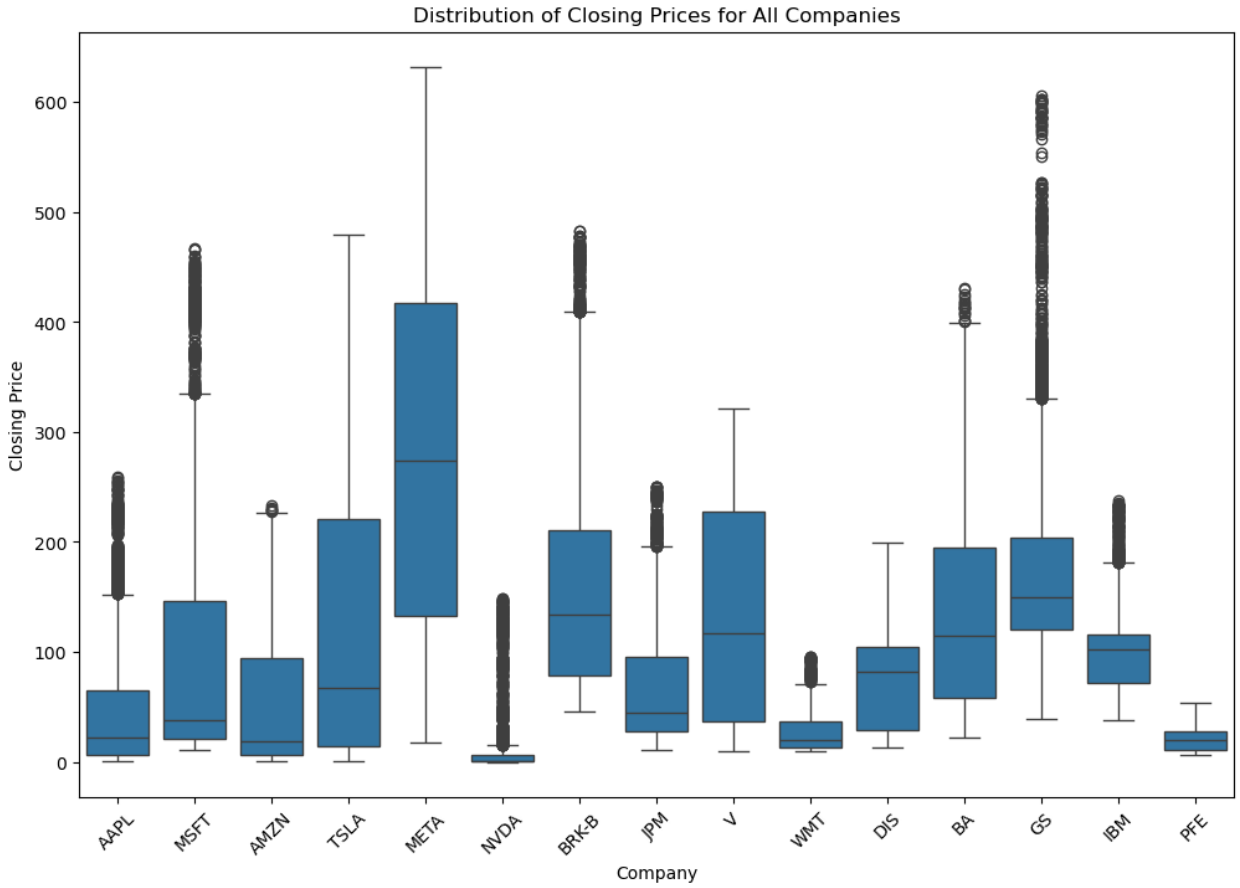
f. Box plot for distribution of opening price:

The goal is to show the opening prices for all companies. The plots show the entire range of opening prices indicating median, quartiles, and outliers. It serves the purpose of comparing price ranges across different companies.



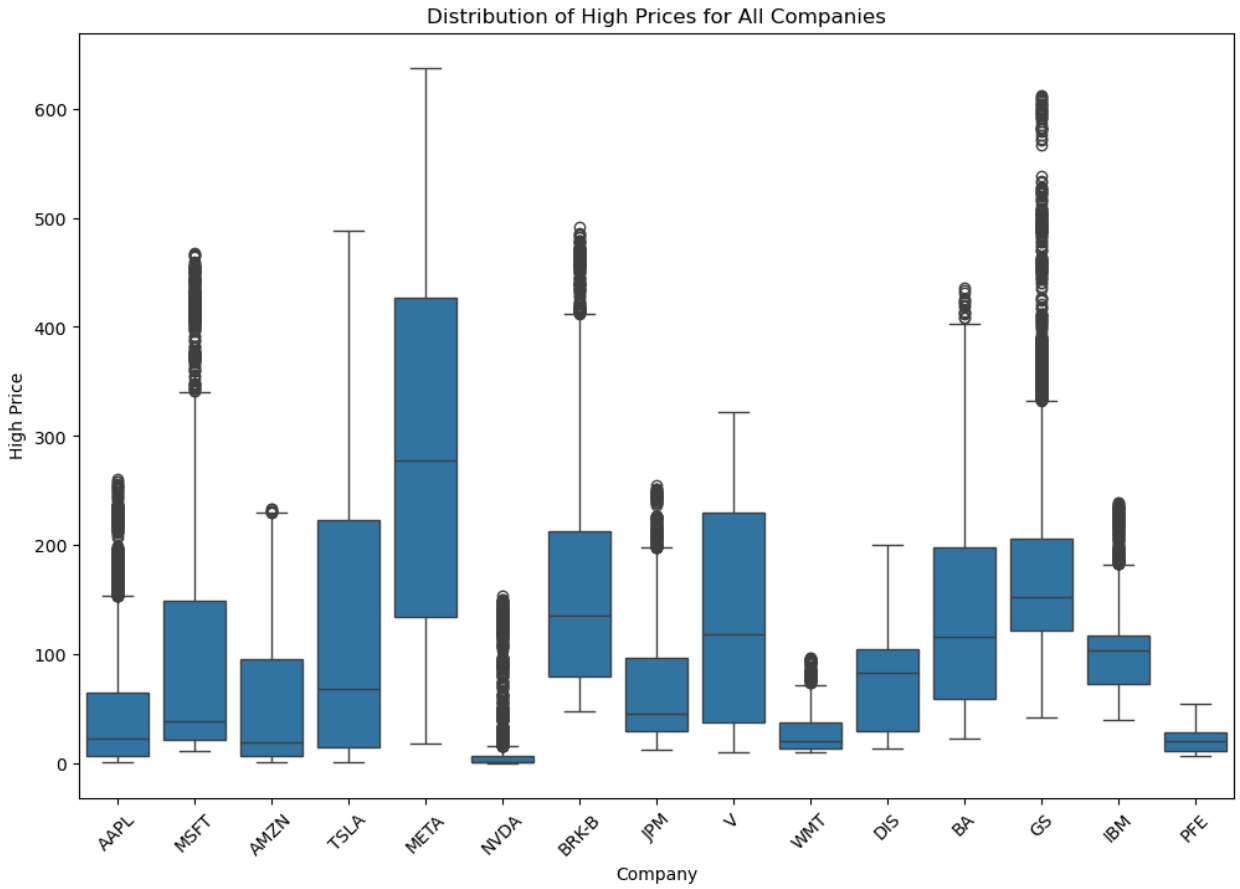
g. Box plot for distribution of closing price:

The goal is to show the closing prices for all companies. The plots show the entire range of closing prices indicating median, quartiles, and outliers. It serves the purpose of comparing price ranges across different companies.



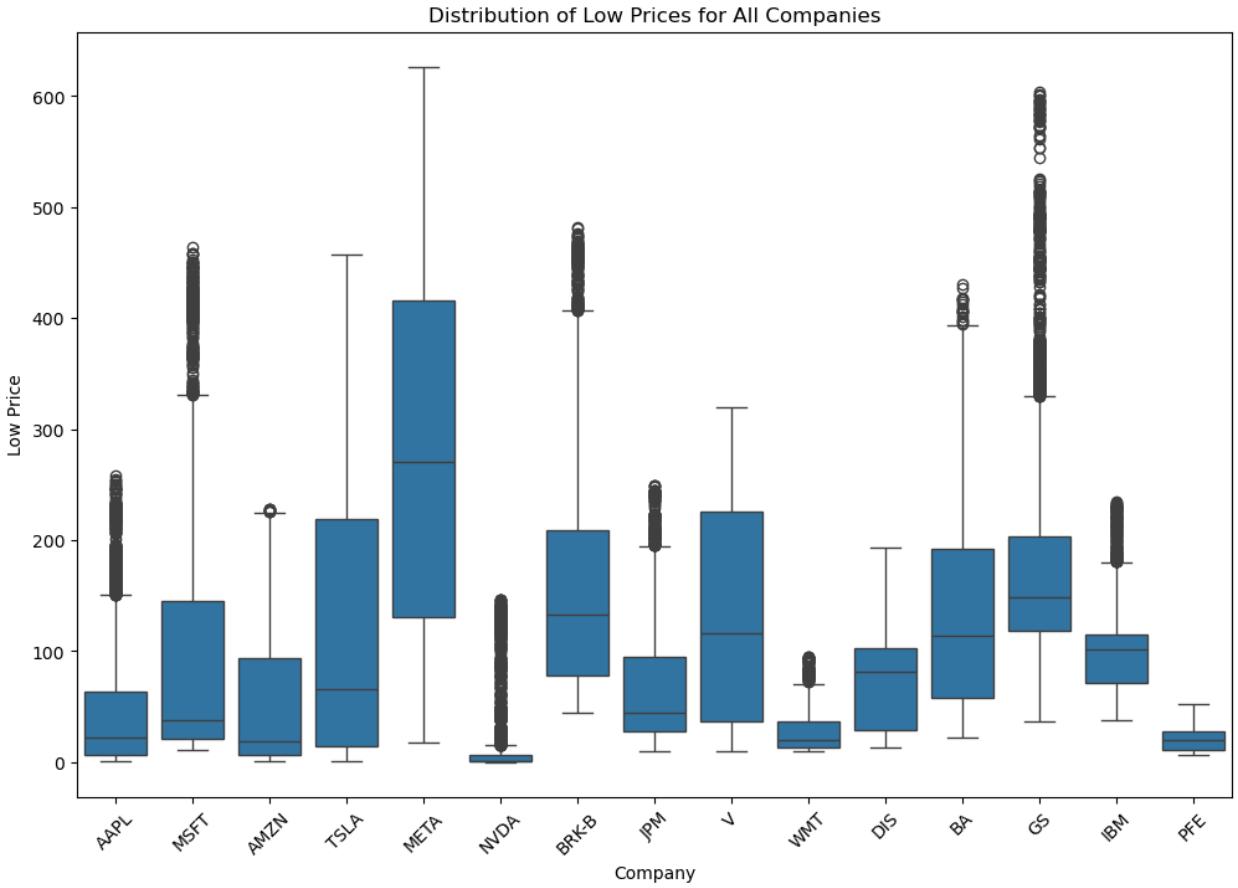
h. Box plot for distribution of high price:

The goal is to show the high prices for all companies. The plots show the entire range of high prices indicating median, quartiles, and outliers. It serves the purpose of comparing price ranges across different companies.



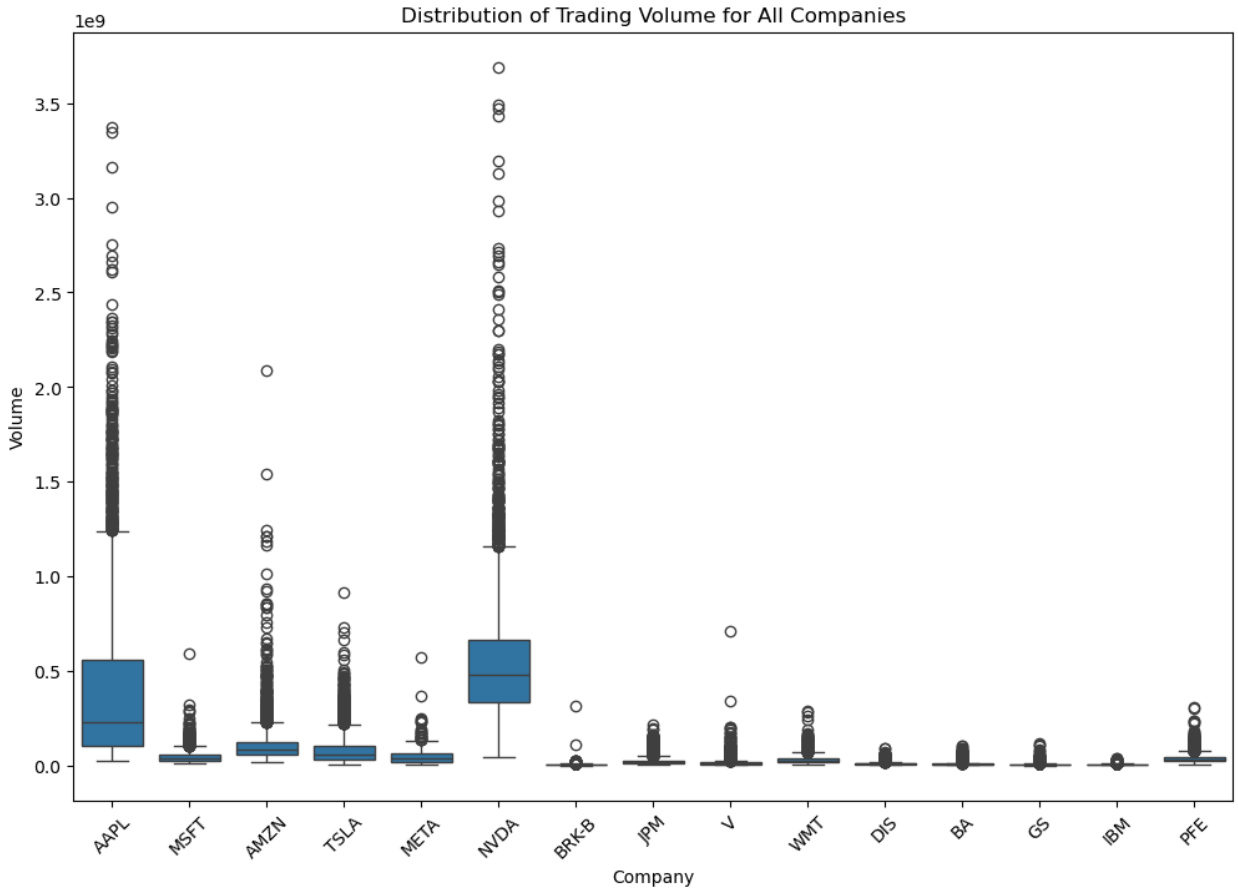
i. Box plot for distribution of low price:

The goal is to show the low prices for all companies. The plots show the entire range of low prices indicating median, quartiles, and outliers. It serves the purpose of comparing price ranges across different companies.



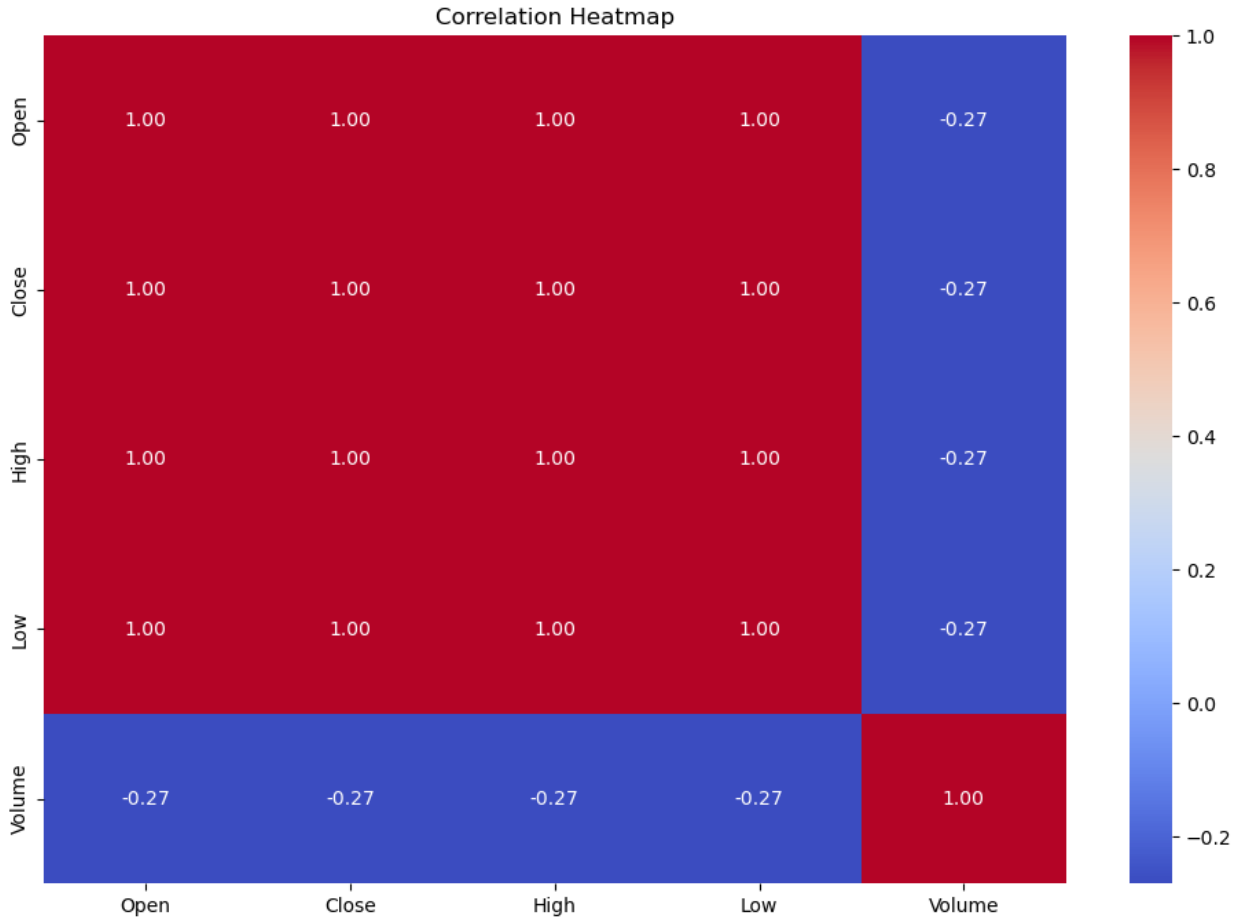
j. Box Plots for Distribution of Trading Volume:

The intention here is to depict the distribution of trading volumes of all companies. Such plots present the variation in trading activities, thereby showing which shares are more often traded, and giving way to peaks in activity.



k. Correlation Heatmap:

The purpose of the heatmap is to visualize the correlation between `Open`, `Close`, `High`, `Low`, and `Volume`. The heatmap shows the relationships between different numerical columns, indicating how closely related they are. High correlation values suggest a strong relationship, which can be useful for feature selection in predictive modeling.



[

From these graphs, several insights are obtained:

- **Trends:** Observed upward and downward trends in stock prices over time for each company.

- Volatility: Noted periods of high volatility in stock prices and trading volumes.

- Outliers: Identified any anomalies or outliers in the data that might need further investigation.

- Correlations: Examined the relationships between different numerical columns to inform feature engineering.

4.2.2 Feature Engineering

Feature engineering was done with the aim of extending the dataset and eliciting more detail in patterns of stock prices by calculating among other things, lagged features, rolling statistics, and percentage change features. These features would hence facilitate the capturing of temporal dependencies directed trends and volatility, hence, important for predictive modeling.

a. Lagged Features

Lagged features are the values of a prior time-step, so that one can capture the momentum and trends in the stock price. Yesterday, we have created lagged features for the `Close` price, covering, in days, the last 5 to 1 days prior [, to that date]. This code creates new columns named `Close_Lag_1`, `Close_Lag_2`, ..., `Close_Lag_5` for each company's closing price

from the previous 1 to 5 days. These lagged features can help identify momentum and trends in stock prices.

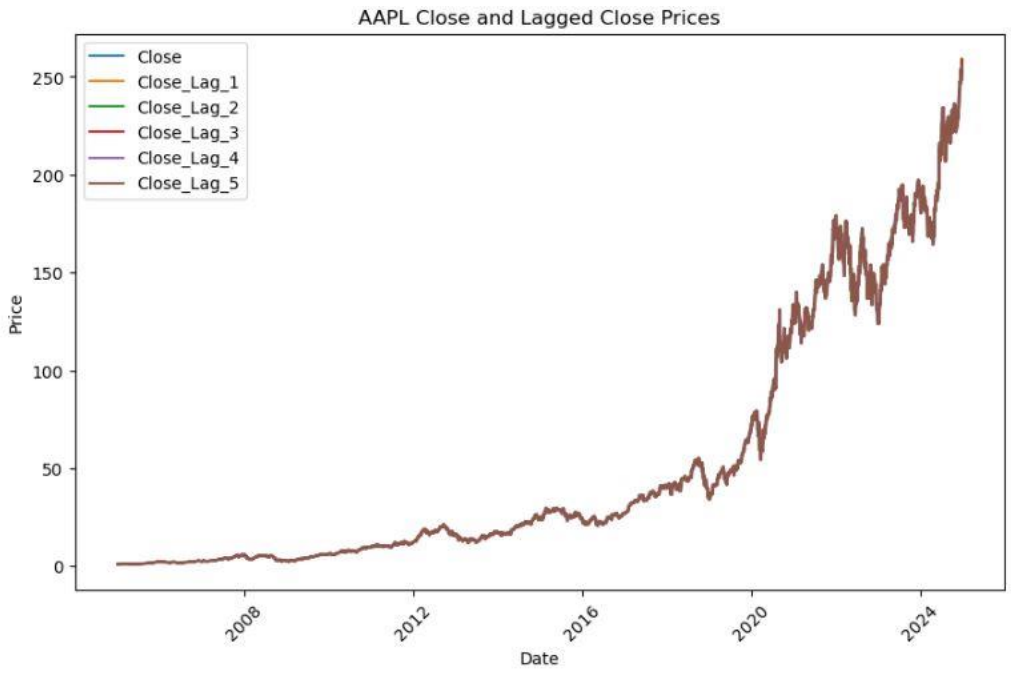


Figure: Visualization of Lagged Features for a sample company (AAPL)

b. Rolling Mean and Standard Deviation

Rolling statistics help capture recent trends and volatility by calculating statistics over a moving window. This code calculates the 7-day rolling mean (`Rolling_Mean_7`) and rolling standard deviation (`Rolling_Std_7`) of the closing prices for each company. These features help capture the recent trends and volatility in stock prices.

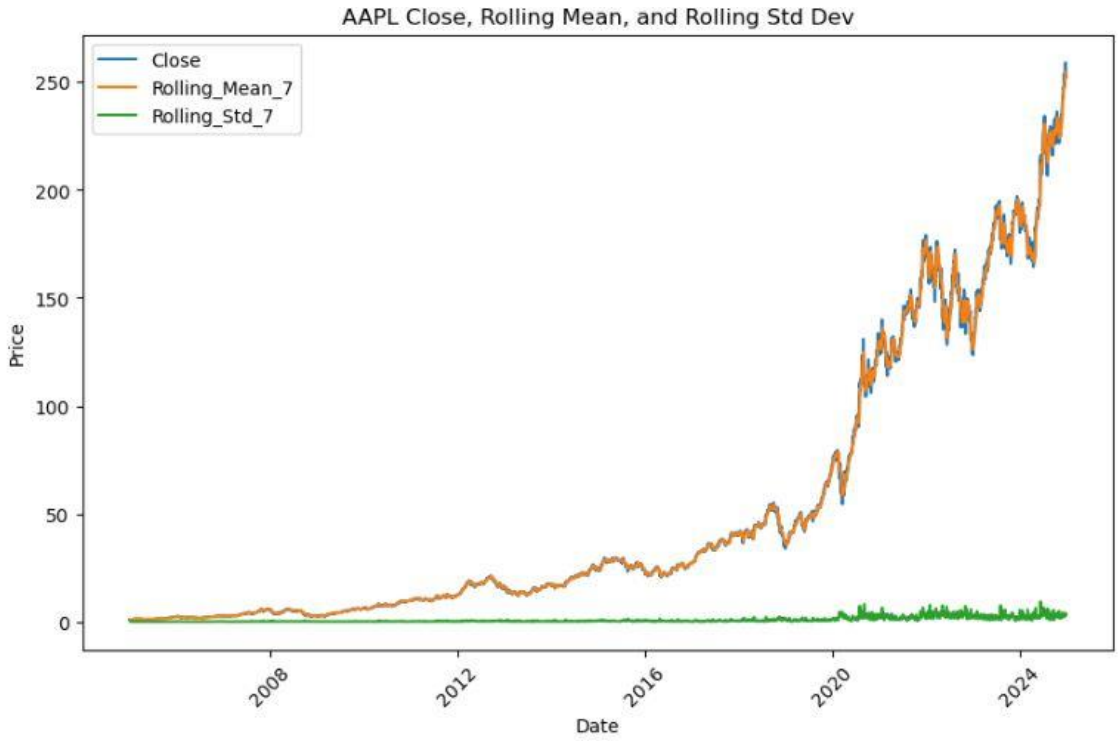


Figure: closing prices, rolling mean, and rolling standard deviation for AAPL

c. Percentage Change

Percentage change features capture the daily momentum by showing how much the stock price has changed from the previous day. The percentage change (`Pct_Change`) in the closing price from the previous day is calculated. This feature helps capture the daily momentum of stock prices.

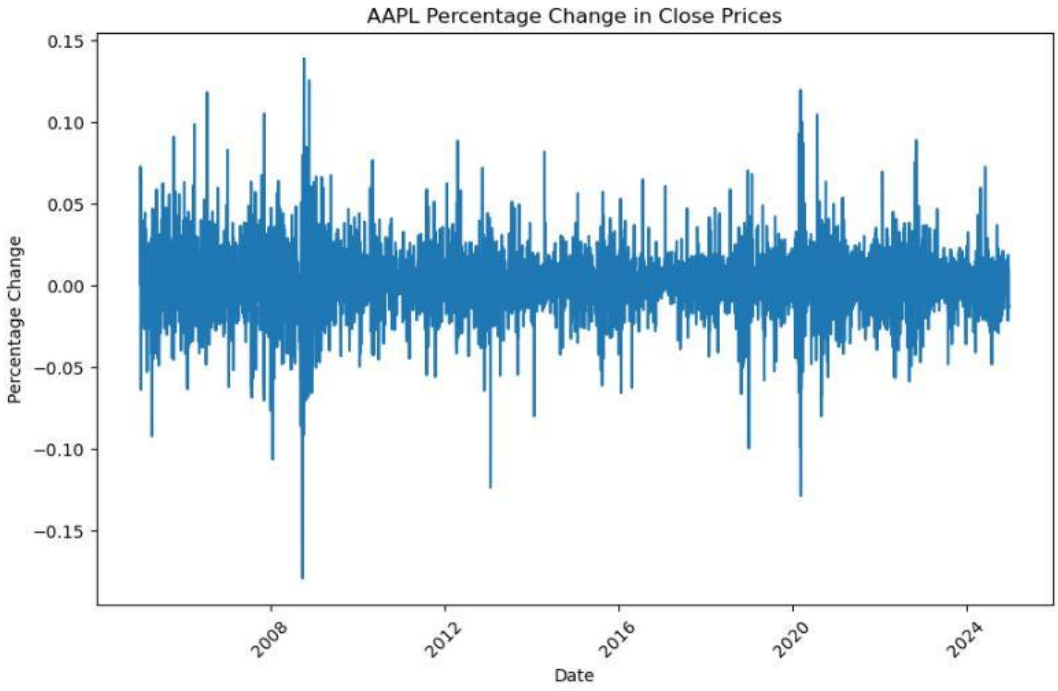


Figure: The percentage change in closing prices for AAPL

4.2.3 Sentiment Analysis

With our sentiment data aligned and prepared, we now embark on our next agenda of analyzing sentiment. By using appropriate tools and resources, we can classify news articles or market reports according to three sentiment labels: one that shows a positive attitude, another that shows a neutral view, and finally, one that shows a negative tendency. These are some of the main procedures involved in the outcome:

a. Sentiment Distribution:

Exploration of sentiment labels (positive, neutral, and negative) in the dataset was the first step in the sentiment analysis process. This is quite important for finding out the distribution of sentiments across the whole period of the data. When we analyze how often each sentiment label occurs, we should be able to tell whether it is dominated by one particular sentiment or is spectrum very balanced. Suitable distribution of sentiments is important for the training of models based on a sentiment as it ensures that there is no overfitting of sentiment in one type.

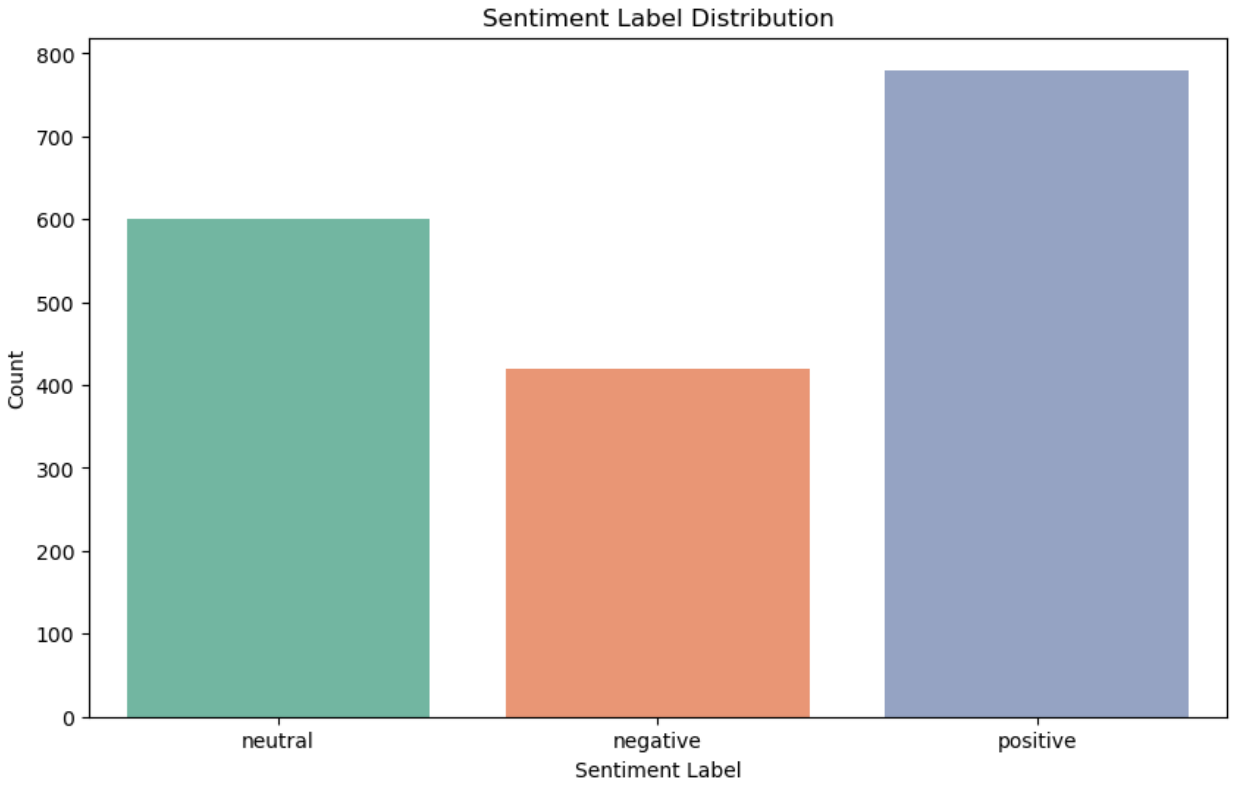


Figure: Sentiment distribution

b. Sentiment Distribution Over Time

We then went on to understand how sentiment labels have been laid over time. This last task would help detect any trends or shifts in sentiment from time to time. For instance, we could observe that positive sentiments were more observed when there was an upturn in the market, whereas bearish sentiments thrived more during downturns in the market or periods of high volatility. Also through time, we looked at the distribution of sentiment at the same time as we correlated it with market events such as major market crashes, earnings announcements, and geopolitical events. This step is useful in judging if the shifts in sentiments could precede stock prices or any events in the market.

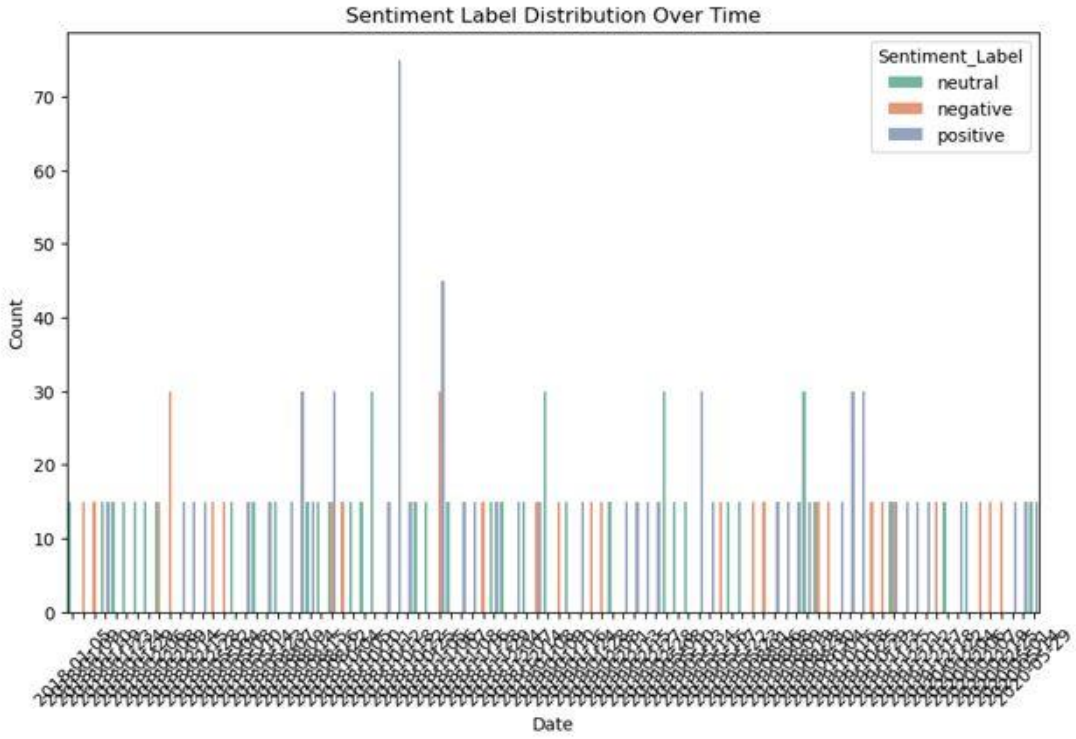


Figure: Sentiment distribution over time

c. Relationship Connecting Sentiment with Stock Prices

Subsequently, sentiments were numerically coded for analysis concerning stock price, thereby leading to positive sentiments scoring 1, neutral 0, and negative as -1. Then, the correlation was made between both sentiment and stock price such as daily returns and price changing features. These steps were taken following understanding whether such relationship, measurable and detectable, exists as it were between sentiment and the movement of stock prices. The further process of correlation analysis resulted in discovering whether positive sentiment would mean a rise in stock prices while negative one would show a fall. Sentiment correlation is not just all that we looked at; we also focused on the effect of sentiment on stock volatility by rolling standard deviation of returns daily. Generally, volatility increases during times of negative sentiments-an indicative occurrence of market uncertainty or fear. On the

other hand, positive sentiment could stand to be associated with periods of market stability or growth.

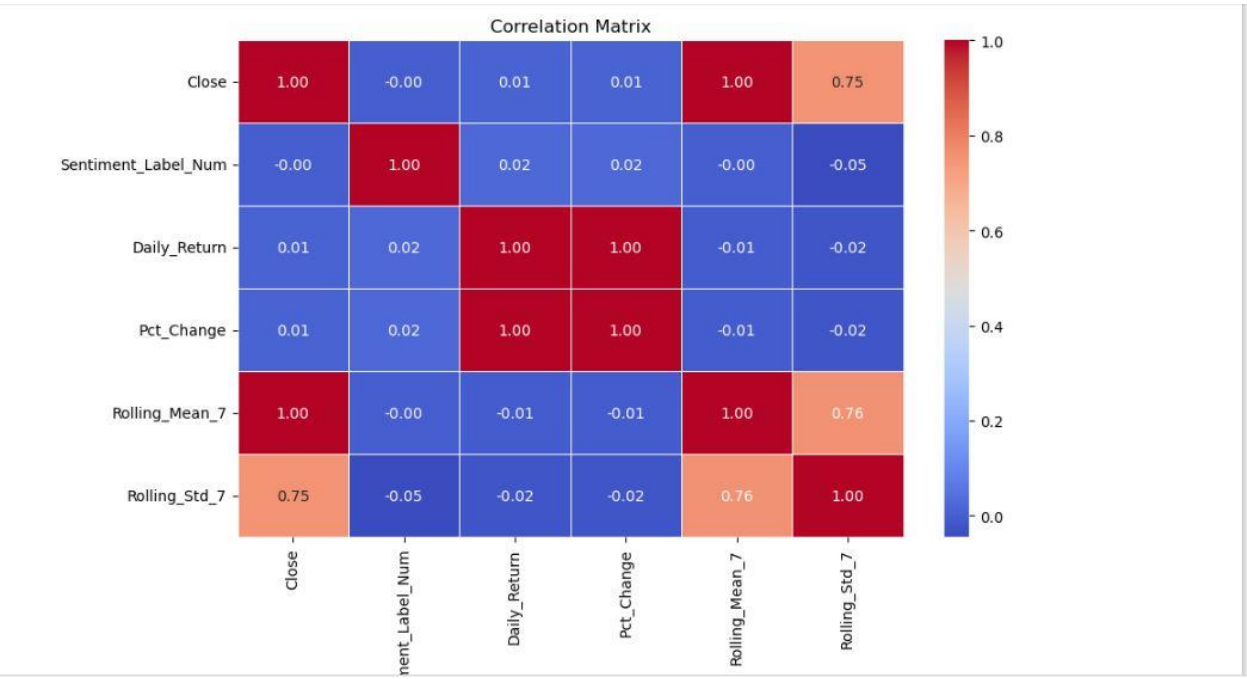


Figure: Correlation between sentiment score and stock price

d. Influence of Sentiments towards Volatility

How the above can be turned out to impact stock price volatility is explored. Visualization showed that there existed differences in volatility during the periods of different sentiment labels. For example, during bad periods, we noted higher volatility as will generally be the case; bad news screams inconsistency or markets' pessimism. This analysis is definitely a deep part of the wider understanding of how sentiment and market behavior relate, since volatility is the most crucial factor with which the investor deals. Most of the time when the market hits high volatility, it is a reflection of uncertainty, and negative feelings may exacerbate the situation.

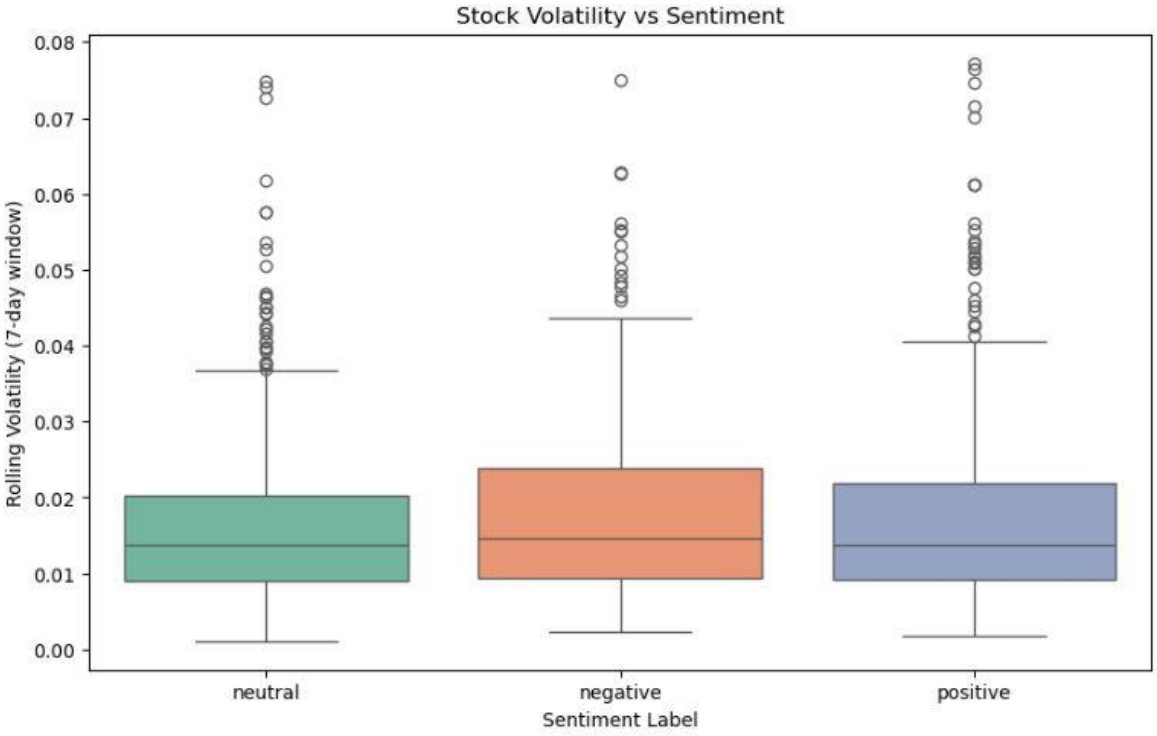


Figure: Stock volatility vs Sentiment

4.2.4 Machine Learning

a. Data Preprocessing

A number of preprocessing steps are run before machine learning model is applied. Some of these steps include extraction of the features desired, filling in missing entries, and categorical encoding- particularly of the sentiment labels. The other features include sentiment scores, daily returns, percentage changes, and volatility measures, which have been considered for training the model. The dataset is divided into a training set (80%) and a test set (20%). The model is trained using the training set while the test set is left for evaluating the model's performance in unseen data.

```
from sklearn.model_selection import train_test_split

# Define features and target variable
features = merged_data[['Sentiment_Label_Num', 'Rolling_Sentiment', 'Daily_Return', 'Pct_Change', 'Rolling_Volatility']]
target = merged_data['Price_Direction']

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.2, random_state=42)
```

b. Model Training

Random Forest is the selected machine learning model for this work. The strength of Random Forest to accommodate complex relationships makes it effective when capturing the relationship between sentiments and stock price movements. It works simply by constructing large amounts of decision trees and then averaging their individual responses to come up with a final prediction. It reduces the overfitting phenomenon so often seen with financial data when training a model. The Random Forest model is trained to predict the stock price movement direction (e.g., Up, Down, or Flat) using the features obtained from sentiment and stock indicator data. After training, the model's predictions are tested on new dataset, and the results are evaluated using many metrics.

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

# Initialize and train the Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

# Predict on test set
y_pred = rf_model.predict(X_test)

# Evaluate the model
print(f"Accuracy: {accuracy_score(y_test, y_pred)}")
print(f"Classification Report:\n{classification_report(y_test, y_pred)}")
```

c. Model Evaluation

The performance of the trained model is evaluated using the standard classification metrics named accuracy, precision, recall, and F1-score.

Basically, accuracy reflects the model's entire proportion of correct predictions in absolute values.

Precision tells us how many of those positive predictions are real.

For recall, it tells us how many of the actual positive instances were collected by the model.

F1 considers the harmonic mean between precision and recall, so it gives a balanced measure between the two.

These indices further interpreted the strength and weaknesses of the model in predicting the stock price direction and indicated how well the model generalized to new, previously unsampled data.