PREDICTING STOCK MARKET TRENDS USING MULTI-SOURCE SENTIMENT ANALYSIS AND ADVANCED DEEP LEARNING ALGORITHMS

RAIAN HAFIZ NILOY

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI TEKNOLOGI MALAYSIA
DECLARATION OF Choose an item.

Author's full name	:	RAIAN HAFIZ NILOY		
Student's Matric No.	:	MCS241008	Academic Session	: 2024202501
Date of Birth	:		UTM Email	: raian@graduate.utm.my
Choose an item. Title	:	PREDICTING STOCK MARKET TRENDS USING MULTI-SOURCE SENTIMENT ANALYSIS AND ADVANCED DEEP LEARNING ALGORITHMS		

I declare that this thesis is classified as:

☒ **OPEN ACCESS** — I agree that my report to be published as a hard copy or made available through online open access.

☐ **RESTRICTED** — Contains restricted information as specified by the organization/institution where research was done. *(The library will block access for up to three (3) years)*

☐ **CONFIDENTIAL** — Contains confidential information as specified in the Official Secret Act 1972)

(If none of the options are selected, the first option will be chosen by default)

I acknowledged the intellectual property in the Choose an item. belongs to Universiti Teknologi Malaysia, and I agree to allow this to be placed in the library under the following terms :

1. This is the property of Universiti Teknologi Malaysia
2. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of only.
3. The Library of Universiti Teknologi Malaysia is allowed to make copies of this Choose an item. for academic exchange.

Signature of Student:

Signature :

Full Name : RAIAN HAFIZ NILOY
Date : 17 FEBRUARY 2025

Approved by Supervisor(s)

Signature of Supervisor I:	Signature of Supervisor II
Full Name of Supervisor I PROF. MADYA DR MOHD SHAHIZAN OTHMAN	Full Name of Supervisor II

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

This letter should be written by a supervisor and addressed to Perpustakaan UTM. A copy of this letter should be attached to the thesis.

Date:

Librarian

Jabatan Perpustakaan UTM,

Universiti Teknologi Malaysia,

Johor Bahru, Johor

Sir,

CLASSIFICATION OF THESIS AS RESTRICTED/CONFIDENTIAL

TITLE: PREDICTING STOCK MARKET TRENDS USING MULTI-SOURCE SENTIMENT ANALYSIS AND ADVANCED DEEP LEARNING ALGORITHMS

AUTHOR'S FULL NAME:RAIAN HAFIZ NILOY

Please be informed that the above-mentioned thesis titled _____ should be classified as RESTRICTED/CONFIDENTIAL for a period of three (3) years from the date of this letter. The reasons for this classification are

(i)

(ii)

(iii)

Thank you.

Yours sincerely,

SIGNATURE:

NAME:

ADDRESS OF SUPERVISOR: PROF. MADYA DR. SHAHIZAN OTHMAN

"I hereby declare that I have read this project report and in my opinion this project report is sufficient in term of scope and quality for the award of the degree of Master of Science(Data Science)"

Signature	:	_____
Name of Supervisor I	:	PROF. MADYA DR. MOHD. SHAHIZAN OTHMAN
Date	:	17 FEBRUARY 2025

Signature	:	_____
Name of Supervisor II	:	
Date	:	

Signature	:	_____
Name of Supervisor III	:	
Date	:	

Declaration of Cooperation

This is to confirm that this research has been conducted through a collaboration RAIAN HAFIZ NILOY and University Technologi Malaysia(UTM)

Certified by:

Signature :
Name :
Position :
Official Stamp
Date

* This section is to be filled up for theses with industrial collaboration

Pengesahan Peperiksaan

Tesis ini telah diperiksa dan diakui oleh:

Nama dan Alamat Pemeriksa Luar **:**

Nama dan Alamat Pemeriksa Dalam **:**

Nama Penyelia Lain (jika ada) **:**

Disahkan oleh Timbalan Pendaftar di Fakulti:

Tandatangan :

Nama :

Tarikh :

PREDICTING STOCK MARKET TRENDS USING MULTI-SOURCE SENTIMENT ANALYSIS AND ADVANCED DEEP LEARNING ALGORITHMS

RAIAN HAFIZ NILOY

A project report submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Science(Data Science)

Faculty of Computing
Universiti Teknologi Malaysia

FEBRUARY 2025

DECLARATION

I declare that this project report entitled *"Predicting Stock Market Trends Using Multi-Source Sentiment Analysis And Advanced Deep Learning Algorithms"* is the result of my own research except as cited in the references. The Choose an item. has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature	:	.....
Name	:	RAIAN HAFIZ NILOY
Date	:	17 FEBRUARY 2025

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Mohd Shariff Nabi Baksh, for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisor Professor Dr Awaluddin Mohd Shaharoun and Associate Professor Dr. Hishamuddin Jamaluddin for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Ph.D study. Librarians at UTM, Cardiff University of Wales and the National University of Singapore also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow postgraduate student should also be recognised for their support. My sincere appreciation also extends to all my colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

Stock market prediction is vital in financial study because investors always want data-driven ways to improve prediction efficiency. Most traditional models have relied on stock prices over a historical window but have mostly failed to regard sentiment as an influencing factor. In this investigate, therefore, sentiment analysis is combined along with some machine learning techniques to help stock price prediction through analyzing financial news sentiment using FinBERT and inputting it into predictive models. The area of research encompasses NYSE stock data for a limited time due to lack of long-term sentiment data. Sentiment classification assigns one of positive, neutral, or negative tags to financial news and then associates them with stock price movement. The paradigm shows a Random Forest model as essentially a benchmark in which sentiment-based features indeed show their effectiveness. Still, these models do not compare with deep learning ones such as LSTM and GPT-4, which can capture sequential dependence for modeling stock trends better than traditional methods. Findings indicate that predictive models, in this case, can be improved using sentiment, with negative sentiment registering a very robust correlation with stock market volatility. Feature engineering-fusing technical stock indicators and sentiment scores-indeed play an important role in improving the model performance. The study has been confirmed that sentiment is one of the facts responsible for the stock movement in the financial market, as it shows the worth of FinBERT, as well as LSTM and GPT-4, in financial forecasting. Future work would aim to cover more such sources of sentiment, including social media, as well as developing better models for real-time prediction, resulting in much stronger stock market forecasting.

ABSTRAK

Ramalan pasaran saham adalah penting dalam kajian kewangan kerana pelabur sentiasa mahukan cara dipacu data untuk meningkatkan kecekapan ramalan. Kebanyakan model tradisional telah bergantung pada harga saham melalui tetingkap sejarah tetapi kebanyakannya gagal menganggap sentimen sebagai faktor yang mempengaruhi. Oleh itu, dalam penyiasatan ini, analisis sentimen digabungkan bersama beberapa teknik pembelajaran mesin untuk membantu ramalan harga saham melalui menganalisis sentimen berita kewangan menggunakan FinBERT dan memasukkannya ke dalam model ramalan. Bidang penyelidikan merangkumi data saham NYSE untuk masa yang terhad kerana kekurangan data sentimen jangka panjang. Klasifikasi sentimen memberikan salah satu teg positif, neutral atau negatif kepada berita kewangan dan kemudian mengaitkannya dengan pergerakan harga saham. Paradigma ini menunjukkan model Hutan Rawak pada asasnya sebagai penanda aras di mana ciri berasaskan sentimen sememangnya menunjukkan keberkesanannya. Namun, model ini tidak dibandingkan dengan model pembelajaran mendalam seperti LSTM dan GPT-4, yang boleh menangkap pergantungan berurutan untuk memodelkan trend saham lebih baik daripada kaedah tradisional. Penemuan menunjukkan bahawa model ramalan, dalam kes ini, boleh diperbaiki menggunakan sentimen, dengan sentimen negatif mencatatkan korelasi yang sangat teguh dengan turun naik pasaran saham. Ciri penunjuk saham teknikal yang menggabungkan kejuruteraan dan skor sentimen-sememangnya memainkan peranan penting dalam meningkatkan prestasi model. Kajian itu telah mengesahkan bahawa sentimen adalah salah satu fakta yang bertanggungjawab terhadap pergerakan saham dalam pasaran kewangan, kerana ia menunjukkan nilai FinBERT, serta LSTM dan GPT-4, dalam ramalan kewangan. Kerja masa depan akan bertujuan untuk merangkumi lebih banyak sumber sentimen sedemikian, termasuk media sosial, serta membangunkan model yang lebih baik untuk ramalan masa nyata, menghasilkan ramalan pasaran saham yang lebih kukuh.

TABLE OF CONTENTS

	TITLE	PAGE
	DECLARATION	**iii**
	ACKNOWLEDGEMENT	**v**
	ABSTRACT	**vi**
	ABSTRAK	**vii**
	TABLE OF CONTENTS	**viii**
	LIST OF FIGURES	**x**
	LIST OF ABBREVIATIONS	xi
CHAPTER 1	**INTRODUCTION**	**1**
1.1	Introduction	1
1.2	Problem Background	2
1.3	Problem Statement	3
1.4	Research Goal	4
1.4.1	Research Objectives	4
1.5	Scope	6
CHAPTER 2	**LITERATURE REVIEW**	**96**
2.1	Introduction	**Error! Bookmark not defined.**
2.2	Traditional Stock Market Prediction Approaches	8
2.3	Transition towards Advanced Methods	11
2.4	Deep Learning for Time Series Forecasting	14
2.5	Financial Analytics through Multi-Source Data Integration	17
2.6	Research Gaps and Opportunities	21
2.7	Conclusion	23
CHAPTER 3	**RESEARCH METHODOLOGY**	**24**
3.1	Introduction	24

3.2 Research Framework 24
3.3 Problem Understanding 27
3.4 Data Collection 27
3.5 Data Preprocessing 28
3.6 Exploratory Data Analysis (EDA) 30
3.7 Model Development 31
3.8 Model Evaluation 31
3.9 Deployment and Monitoring 31
3.10 Conclusion 32

CHAPTER 4 INITIAL FINDINGS 33
4.1 Overview **Error! Bookmark not defined.**
4.2 Exploratory Data Analysis **Error! Bookmark not defined.**
4.2.1 Understanding the stock price dataset 35
4.2.2 Feature Engineering 81
4.2.3 Sentiment Analysis 84
4.2.4 Machine Learning 88

CHAPTER 5 DISCUSSION AND RECOMMENDATIONS 90
5.1 Introduction 90
5.2 Summary 90
5.3 Key Findings 91
5.4 Future Work 92
5.5 Conclusion 94
REFERENCES 95

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
Figure 3.1	Overall Research Framework	26
Figure 4.1	Line plots for opening prices of 15 companies	35-42
Figure 4.2	Line plots for closing prices of 15 companies	43-50
Figure 4.3	Line plots for high prices of 15 companies	51-58
Figure 4.4	Line plots for low prices of 15 companies	59-66
Figure 4.5	Trading volume of 15 companies over time	67-74
Figure 4.6	Box plot for distribution of opening price of 15 companies	75
Figure 4.7	Box plot for distribution of closing price of 15 companies	76
Figure 4.8	Box plot for distribution of high price of 15 companies	77
Figure 4.9	Box plot for distribution of low price of 15 companies	78
Figure 4.10	Box plot for distribution of Trading volume Of 15 companies	79
Figure 4.11	Correlation heatmap between various prices	80
Figure 4.12	Visualization of Lagged Features for a sample company (AAPL)	81
Figure 4.13	Closing prices, rolling mean, and rolling standard deviation for AAPL	82
Figure 4.14	The percentage change in closing prices for AAPL	83
Figure 4.15	Sentiment distribution	84
Figure 4.16	Sentiment distribution over time	85
Figure 4.17	Correlation between sentiment score and stock price	86
Figure 4.18	Stock volatility vs Sentiment	87

LIST OF ABBREVIATIONS

ANN	-	Artificial Neural Network
ML	-	Machine Learning
DL	-	Deep Learning
NLP	-	Natural Language Processing
NYSE	-	New York Stock Exchange
CNBC	-	Consumer News and Business Channel
UTM	-	Universiti Teknologi Malaysia
XML	-	Extensible Markup Language
FinBERT	-	Financial Bidirectional Encoder Representations from Transformers
GPT-4	-	Generative Pre-trained Transformer 4
ARIMA	-	AutoRegressive Integrated Moving Average
SVM		Support Vector Machine

CHAPTER 1

INTRODUCTION

1.1 Introduction

In a world where the average annual fixed deposit interest rate is dropping drastically, people with conservative financial management observe literal evaporation of their wealth. So, in order to make profits, more and more people are inclined towards investment in stocks with high returns and high liquidity. To make profit, the accurate prediction of market movement is very crucial. Accurate predictions can significantly heighten decision-making processes, decrease investiture risks and step up profitability. But the complexness and unpredictability of the stock market make it difficult for researchers and investors to predict market movement. Traditionally, statistical techniques such as time series analysis and regression models have been employed to calculate stock prices. However, these methods often fall unforesightful in identifying intricate patterns within stock data, especially when influenced by extraneous variables such as market sentiment and financial news. In today's interconnected world, people's notion toward any product can change at any time. Within the blink of an eye, there will be innumerous opinions & discussions on social media. Financial articles will start publishing articles. All of these combined will inevitably fluctuate the stock price. So, predicting stock market movements without integrating market sentiment is not possible.

In recent years, advancements in machine learning (ML) and Deep Learning (DL) have provided many helpful tools to tackle this problem. Natural Language Processing (NLP), a subfield of artificial intelligence (AI), has been performing sentiment analysis successfully. Basically, sentiment analysis is the process of extracting insights from large volumes of unstructured textual information. In sentiment analysis, the emotional tone or opinion (positive, negative, neutral) available in financial news, social media posts or analyst reports is analyzed. This comes very

handy to predict the probable fluctuation in any stock price. After sentiment analysis, the extracted sentiment features are combined with historical stock prices to make future predictions. But this whole process has to go through different obstacles that hurt accuracy.

This project uses advanced NLP models like FinBERT & GPT-4 to perform sentiment analysis on textual data collected from multiple sources. It then uses an advanced Deep Learning model, LSTM (Long Short Term Memory) network to make predictions from the sentiment scores and numerical data. By taking a hybrid approach, this project promises to enhance prediction accuracy.

1.2 Problem Background

Sentiment analysis lends great help to incorporate public sentiment in predicting stock market movements by converting unstructured textual data into quantifiable sentiment scores. Then these scores are incorporated into predictive models to provide a more holistic take on the factors that influence stock prices. Advanced Natural Language Processing models like FinBERT (Finance Bidirectional Encoder Representations from Transformers) and GPT-4 (Generative Pre-trained Transformer) have improved sentiment analysis. These models can classify sentiment e.g. positive, negative, neutral. By doing so, they can predict how much sentiment may impact stock prices. But according to, classical Machine Learning models like logistic regression, when tuned properly, display more effectiveness. Five key metrics, i.e., Accuracy, Precision, Recall, F1 Score, and ROC AUC, were used to assess the performance.

It is found from recent research that, FinBERT is a financial domain-centric model and it has very high potential in understanding financial terms and concepts with remarkable precision. However, it is resource-intensive in nature. That's why it is likely to hurt computational efficiency. GPT-4 is a versatile language model. It has great abilities to create and realize human-like text. That's why it is perfect choice for processing unstructured news data. But the exploration of its predetermined and heuristic approach may restrict its precision in particular financial situations. On the

other hand, logistic Regression is computationally efficient. It produces dependable results when it is applied precisely and performs better than both FinBERT and GPT-4 across most metrics in spite of their advanced text analysis capabilities. However, logistic regression faces limitation in handling stock data for a number of reasons like non-linear relationships, high-dimensional data, feature interactions, sensitivity to outliers, and multicollinearity.

This project addresses this dilemma & offers a hybrid approach by combining state-of-the-art NLP models like FinBERT & GPT-4 with LSTM network. FinBERT & GPT-4 show prominence in extracting insights from complicated data. LSTM networks are good at handling sequential data, capturing long term dependencies and robust to noise. These characteristics make them an ideal choice for prediction of stock prices.

1.3 Problem Statement

The primary problem this project addresses originates from the dilemma of choosing between traditional Machine Learning model like logistic regression and advanced NLP algorithms like FinBERT & GPT-4. Logistic regression is computationally efficient and simple. With proper adjustment, it outperforms advanced DL algorithms. But when it comes to handling complicated data patterns, it lacks proficiency. Logistic regression cannot handle non-linear data much efficiently but the relationship between stock prices and predictors is often non-linear. It also faces problems in handling high-dimensional stock market data with numerous features as it might struggle with feature selection and regularization. Handling outliers also becomes an issue for Logistic regression. On the other hand, FinBERT & GPT-4 are highly resource intensive. These models are also computationally heavy. But these models are very strong in understanding financial terminology and human generated textual data which eventually make them suitable for sentiment analysis. To solve this dilemma, this project takes a hybrid approach. It combines these advanced NLP models with LSTM networks. LSTM networks can tackle the challenges caused by the limitations of logistic regression. Thus, this project promises to provide a model with more accuracy.

1.4 Research Goal

The goal of this project is to enhance the prediction accuracy of stock prices by combining state-of-the-art NLP models with advanced DL algorithm LSTM. The NLP models will do sentiment analysis on data collected from various sources & generate sentiment scores. LSTM will perform time series analysis on the sentiment scores and historical stock prices to accurately predict stock market movements.

1.4.1 Research Objectives

- To enhance the comprehensiveness of prediction by collecting and preprocessing sentiment data from multiple sources

- To implement FinBERT & GPT-4 for detailed sentiment analysis and develop LSTM networks for improving prediction accuracy

- To rigorously assess the model's performance with a number of evaluation metrics such as Accuracy, Precision, Recall, F1 score etc

1.5 Scope:

This project combines sentiment analysis of financial news and social media posts and performs time series forecasting to make effective stock price predictions. This covers data collection, preprocessing and model development using LSTM networks and the evaluation of predictive performances. However, the research does not include real-time data analysis. It doesn't perform a full-fledged market analysis and doesn't recommend any particular stock.

To enhance the accuracy of stock price predictions, this project comprises the integration between sentiment analysis and time series forecasting. This involves the collection of historical stock prices and trading volumes, financial indicators, as well as sentiment data from financial news, social networks, and -analyst reports-from several trusted sources. It also includes extensive preprocessing steps like cleaning, normalizing, and tokenizing textual data as a measure for data-quality improvement and compatibility. Advanced Natural Language Processing (NLP) models such as FinBERT and GPT-4 will be employed in feature extraction and sentiment analysis. High-level Long Short-Term Memory (LSTM) networks will be developed to synthesize the sentiment features with the traditional financial indicators and will follow extensive training and a thorough hyperparameter tuning as well as evaluation based on Accuracy, Precision, Recall, F1 Score, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The results of the model are further validated using cross-validation and backtesting in a bid to enhance the robustness of the model.

Model creation and analysis of historical data only exclude real-time data processing and live stock price forecasts. This does not aim to be an all-encompassing analysis of market situations to include geopolitical events or macroeconomic conditions. There will not be specific stock recommendations or investment advice covered by the project, nor will it include predictions made for other financial instruments such as bonds, commodities, or derivatives. Also, more sophisticated machine learning architectures like deep reinforcement learning have been excluded from this study. The analysis is limited by the availability and quality of historical data, and the study is constrained in terms of model training and evaluation because of the machine resources available.

1.6 Expected Contribution:

- Introduction of a novel approach combining advanced NLP models and LSTM networks

- Improvement of stock price prediction accuracy

- Providing valuable insights to the investors and mitigate the risk factor

- Developing a robust framework in financial analysis sector

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Research works on predicting stock market trends concerning the implications of such predictions for investors, analysts, and even policymakers. The fact that price movements in financial markets are so volatile and complex renders one unable to predict exactly what trend the particular market will take on a specific day, using price data only. Traditional methods such as time-series analysis and regression models have shown only limited effectiveness in catching the multi-dimensional factors determining trends in the markets (Fama, 1970; Brock, Lakonishok, & LeBaron, 1992). Therefore, researchers have explored other sources of information, such as the public sentiment, to increase the chances of predicating.

Sentiment analysis is a recent concept in natural language processing (NLP) in merger with which its input further develops into a bridge between textual data and market-relevant insights. Extracted sentiment from financial news and social media or corporate announcements will be converted into quantification of probably public opinion and impact on movements in that market. It poses its problems concerning the ambiguity in language, the right selection of predictive models, and the computational strains of deep learning techniques (Loughran & McDonald, 2011; Nassirtoussi et al., 2014).

Starting from the most recent developments in artificial intelligence, transformer-based architectures have transformed the classrooms of most modern NLP tasks. Recently, pre-trained language models such as FinBERT (Araci, 2019) and GPT-4 (Brown et al., 2020) have shown almost unsurpassed qualities in understanding very specific and recognized scenarios of texts, which makes them very suitable for activities such as financial sentiment analyzes. They have very high capabilities in

categorizing and scoring sentiments from huge piles of financial data, surpassing performance in conventional methods with the leaps and bounds of improvements.

With the changes in the analysis of sentiments, the deep learning techniques have changed the way of forecasting time series. Long Short-Term Memory (LSTM) networks, classifies transient behavior between model-based and convergent neural networks and does wonders with the prediction of some sequences (Hochreiter & Schmidhuber, 1997; Fischer & Krauss, 2018). Predictability of LSTM, which considers both historical price data and sentiment features, gives rise to an effective predictor for stock markets.

This literature review examines the relationship between sentiment analysis and time series forecasting and investigates hybrid models that utilize FinBERT, GPT-4, and LSTM. The aim is to contextualize the methodology used, identify gaps in research, and discuss potential uses of this work so as to achieve the coverage of what the field currently holds and how this project will contribute to advancing prediction of stock market trends.

2.2 Traditional Stock Market Prediction Approaches

Historically, stock market forecasting has hinged on statistical models and primitive machine learning methods that emphasized the use of historical pricing data. Examples of these techniques include linear regression, time-series analysis, and other econometric models that try to recognize patterns and trends in market data (Fama, 1970). These techniques are well-suited for predicting short-term fluctuations but are limited because they do not incorporate factors from the outside such as economic indicators, news events, or investor sentiment (Brock, Lakonishok, & LeBaron, 1992).

2.2.1 Statistical Methods

2.2.1.1 Time-Series Analysis

ARIMA (Auto-Regressive Integrated Moving Average) and GARCH (Generalized Autoregressive Conditional Heteroskedasticity) models have been

significant for taking care of stock price modeling and forecasting. The Auto-Regressive Integrated Moving Average is the first model through Box and Jenkins (1970), which suppressed the dependency of a variable on its past values to facilitate short-term predictions. However, Tsay (2005) says that its performance is limited in non-linear and highly volatile markets due to the fact that it is taking advantage of stationary data and linear assumptions. In this case, GARCH models bring volatility clustering by pointing out variance as a function of past errors; however, it still is unable to address non-linear dependencies (Engle, 1982).

2.2.1.2 Regression Models

Linear regression has been one of the basic foundational techniques that have modeled stock prices and predicted stock prices. It has rules that define and relate dependent variables with independent ones; however simple it may seem, it is often found inadequate at times because it usually does not fit into complex interactions and dynamics of the market and gives a rather oversimplified result (Fama & French, 1992). Some extensions like multiple regression and polynomial regression have been made, but they have not gone far, mainly due to the linearity assumption.

2.2.1.3 Econometric models

CAPM and APT are among the heavily used models in financial modeling. CAPM, as theorized by Sharpe (1964), predicts the expected return of an asset on the basis of an asset's risk relative to the market. Even though CAPM has been widely accepted, Roll (1977) critiques it for simplification by relying only on one factor-that of market risk.

2.2.2 Early Machine Learning Models

2.2.2.1 Support Vector Machines (SVM)

SVM has gained some momentum in predicting stock behavior and trend classifications along with their applications for historical market data. Huang et al.

(2005) illustrated the potential of the technique in using a rather limited database as well as its capability to avoid overfitting. However, SVMs are limited by scalability and interpretation when applied to very large and complex datasets (Li et al., 2014).

2.2.2.2 Decision trees and random forests

Decision tree is a non-parametric model that makes splits in a dataset into branches as per feature thresholds; a random forest is an ensemble method aggregating the predictions of multiple trees. This technique has proved better than any linear model in non-linear capturing, as per the findings of Chen et al. (2013). However, it cannot be applied to any unstructured inputs such as textual sentiment since it relies on structured data.

2.2.3 Limitations of Traditional Approaches

2.2.3.1 Inadequate Integration of Data

Models rely mostly on monetary data at the expense of many things like investor sentiment, macroeconomic indicators, or geopolitical events. This factor renders them incapable of grasping completely the context in which moves on market markets operate (Malkiel, 1973).

2.2.3.2 Inability to Recognize Nonlinear Relationships

Financial markets have a complex, non-linear behavior, which could arise from several factors, including human psychology, world events, etc. Most traditional statistical models fail to meet the complexities in their prediction capabilities and hence provide poor results (Lo, 2004).

2.2.3.3 Overfitting and Generalization

Older machine-learning systems could learn sophisticated patterns but were easily overfitted because of inadequate training data and lack of regularization, so they would not generalize to the new state of the markets (Hastie, Tibshirani, & Friedman, 2009).

2.2.3.4 Case Studies and Real-World Applications

Several studies point to the limitations of the conventional models in real-life conditions. Fama (1970)-the efficient market hypothesis had ergonomical effects on investment in that past data alone cannot predict markets. Brock, Lakonishok, and LeBaron (1992) used time series of non-causal type to apply statistical properties, though ultimately broad brush found little predictability from technical trading rules. Engle (1982) states that while the GARCH model is found useful in forecasting volatility, it fails accounting for non-linear dependencies.

2.3 Transition towards Advanced Methods

The limitations imposed for the prediction of stock markets pertaining to traditional techniques paved a way for advanced techniques, especially those based on deep learning and sentiment analysis. Failure of traditional models to harmonize information from different sources, failure in forming non-linear relations and inability to adapt dynamically makes the advanced models, for instance, deep neural networks with transformer-based approaches, solutions by allowing the entire data analysis and predictive modeling robustly.

2.3.1 Sentiment Analysis in Advanced Methods

Sentiment analysis has completely transformed the predictive frameworks in financial analytics. Advanced sentiment analysis is unlike traditional models- it

employs machine learning and deep learning, unlike the traditional ways of extracting insights from unstructured textual data. For example, natural language processing (NLP) tools analyze financial news, social media posts, and earnings call transcripts to quantify market sentiment (Loughran & McDonald, 2011).

2.3.2 Transformer-Based Models

With transformer models such as BERT and GPT-4 presently revolutionizing sentiment analysis from contextual understanding to transfer learning, there comes an end to pandemic-era approaches where valuable semantic implications-and, hence, relevancy-open new methods to economic text generation. When used opportunistically, these processes tend to create indexed textual representations that forge semantic links between words based not just on usage, but rather the context in which those words are found, allowing practitioners to understand the nuances of any language. FinBERT, an adaptation of BERT for finance, works on emotional sourcing in finance, making prediction much more accurate (Araci, 2019). On the strength of these features, therefore, generation capacity has made GPT-4 bring out a possibility of appraisal along with analyzing financial narratives that it summarizes, making it a multi-actioning tool in finance (Brown et al., 2020).

2.3.3 Hybrid Models with Sentiment Analysis

The combination of sentiment analysis and predictive modeling from Long Short-Term Memory (LSTM) networks has shown much potential in treating emergent conditions from the changing nature of the financial markets because LSTMs are suitable for handling sequential data and can yield better results when combined with sentiment scores from advanced NLP models. As an option, a user can use sentiment scores derived from FinBERT or GPT-4 as input features for LSTM-based models to derive better predictions of stock price movement (Fischer & Krauss, 2018).

2.3.4 Deep Learning for Time-Series Analysis

New techniques have emerged in the field of deep learning as very strong tools which can enable mathematical modeling of obviously non-linear linkages in financial data. Where traditional time-series models, for example, ARIMA, are limited by their linear assumptions and do not allow the inclusion of external variables, these conditions are overcome with deep learning methods. For instance, they can **Capture Long-Term Dependencies:** RNNs and their different versions, LSTMs included, can do this quite well for time-series data (Hochreiter & Schmidhuber, 1997). **Exogenous Features:** Specialized complex under-constructed models can even allow the usage of features, such as market sentiment, macroeconomic indicators, or geopolitical events, integrated holistically in consideration of market dynamics (Goodfellow etal., 2016).

2.3.5 Some Case Studies in Hybrid Modeling

1. Nassirtoussi et al. (2014): Provided the multi-source sentiment analytics frame for integrating news and social media data into predictive models for improved forecasting accuracy.

2. Kim and Won (2020): Proved the power of combining LSTM networks with sentiment scores for stock market predictions, surpassing the predictions of conventional machine learning models

These advanced methods have indeed drastically improved stock market predictions; however, challenges still pose threats to the use of deep learning. Foremost among these include the computational cost of deep learning models, the need for large labeled datasets, and real-time data streams. Future exploration will have to include the development of efficient algorithms and data augmentation techniques in conjunction with frameworks for real-time analytics.

The move to advanced methods heralds a new era of financial analytics: more, and better articulated, accurate predictions. Hybrid models, which combine such divergent techniques as sentiment analysis and deep learning, will be a strong foundation for predicting, more accurately, the stock market trends in a more complex financial landscape.

2.4 Deep Learning for Time Series Forecasting

Time-series forecasting is most crucial as far as stock markets are concerned, as it refers to the data recorded in order to analytically forecast trends to come. Methods such as simple ARIMA or Exponential Smoothing became increasingly limited under such a description, as they do not capture the dynamic, non-linear character of financial data due to their linear assumptions. However, in this era of developments, it has become a trendy way of method for advancing capabilities to transform actual dependence to much more complex interdependence and integrate individual data types.

2.4.1 Limitations of Conventional Models

Traditional time-series models such as ARIMA rely on stationary assumptions and consider linear relationships, which is insufficient to model the volatile and multifaceted nature of stock returns (Box & Jenkins, 1970). For example, although they are quite valid for short-term forecasts, Exponential Smoothing techniques do not capture long-term dependence or nonlinearities from financial data (Hyndman & Athanasopoulos, 2018).

2.4.2 Recurrent Neural Networks (RNNs)

To overcome such restrictions, recurrent neural networks (RNNs) introduce feedback loops, advising on the modeling of temporal dependencies. RNNs, however, have vanishing gradients, which prevent them from learning long-distance dependencies (Bengio et al. 1994).

2.4.3 Long Short-Term Memory (LSTM) Network

An RNN specifically designed for learning long sequences, LSTMs use a gating mechanism paired with memory cells. This means that the LSTM can keep new input and clear contents by separating the operations needed to be set at different times for banking applications (Hochreiter & Schmidhuber, 1997).

2.4.3.1 LSTM Networks' Mechanism

In long-term storage memory cell such that it can also reduce the effect of the vanishing gradient. Input, Forget and Output Gates regulate the current flowinside, which allows the important patterns to remain, while discarding the irrelevant information. Sequel Data Modelling defines the dependency through the time steps thus enables accurate trend analysis.

2.4.3.2 LSTM Applications in Stock Market Prediction

-Predictions of Price Movements: LSTMs have been successfully used to predict stock prices from historical data, with performance levels exceeding that of ARIMA and other such models (Fischer & Krauss, 2018).

-Sentiment Accounting: LSTMs can be effectively coupled with sentiment data, which means they also consider external factors that affect market trends, such as news events and social media activity (Kim & Won, 2020).

2.4.4 Hybrid Models

Hybrid models are designed to take advantage of the benefits of LSTMs and other advanced techniques, such as NLP-based sentiment analysis, in achieving enhanced predictive accuracy. Sentiment-oriented scores derived from FinBERT and GPT-4 will be used as input features of LSTM networks in a holistic approach to stock market forecasting.

2.4.4.1 Benefits of Hybrid Models

Multi-dimensional Insight: Combining historical prices with sentiment-derived features for a complete analysis. Increased Accuracy: Both structured and unstructured data contribute to a comprehensive understanding of what constitutes market influences. Nassirtoussi et al. (2014): Showed how aggregating multiple sources of sentiment analysis with predictive models could be effective. Chen et al. (2020): Used LSTM networks with sentiment features for outperforming the models using sentiment features.

2.4.4.2 Tools and Techniques

- Frameworks: TensorFlow, PyTorch, and Keras are robust frameworks used in the implementation of LSTMs.

- APIs for Data Retrieval: Tests are with the following: Tweepy, BeautifulSoup for real-time sentiment data acquisition.

- Visual Tools: Matplotlib and Seaborn complete performance analysis of models with visual metrics.

2.4.4.3 Challenges

1. Computer Resource Needs: It requires a huge amount of computational power and memory to train LSTM networks.

2. Data Quality: The quality of numerical and sentiment data and its preprocessing influence the accuracy of hybrid models.

3. Predictions in Real-Time: High-latency integration and processing of data make real-time analytics almost impossible.

2.4.4.4 Future Opportunities

1. Real-Time Analytics and Forecasting Integration: Makes possible real-time forecasting applications with improved hardware and software.

2. More Accurate Sentiment Analysis: Fine-tuning sentiment tools-such as GPT-4-for the financial context should increase the efficiency of these tools.

3. New Data Sources: Usages of other types of data-such as geopolitical news and investor behavior-would add new sources for prediction.

Deep learning for time-series prediction is a huge breakthrough in stock market prediction. By overcoming the weaknesses of conventional models and building hybrid approaches, these methods indeed have great benefits for improving forecasting accuracy and enabling an informed investment decision.

2.5 Financial Analytics through Multi-Source Data Integration

The confluence of data from various oases now carries the title weight in the advanced financial analytics space. Multi-source data integration opens new vistas for an investor where data from financial news, social media, and price history converge to understand more about market dynamics. This is most relevant for stock market prediction as both sentient and exogenous factors can influence price.

2.5.1 Challenges Associated with Data Collection and Preprocessing

2.5.1.1 Volume and Diversity of Data

Financial data is recorded in an unprecedented humongous scale, and the components include Structured Data: Historical prices, economic indicators, trading volumes. Unstructured Data: News articles, tweets, forums.

The heterogeneity of data types makes it more difficult to integrate and analyze. Processing unstructured data is particularly tricky: cleaning, tokenization, and

entity recognition pose some of the greatest challenges in doing so (Loughran & McDonald, 2011).

2.5.1.2 Data Quality and Noise

Most sentiment data from social media and forums contains irrelevant or noisy information such as: Spam and bot-created content. Non-financial discussions that cover up meaningful signals of sentiment (Bollen et al., 2011).

2.5.1.3 Real-Time Data Processing

Real-time data streams need to be integrated and analyzed with low-latency pipelines that can be expensive in terms of compute and technically rather complex to implement.

2.5.2 Examples of Tools and Techniques

2.5.2.1 Data Collection Tools

- APIs: Facilitate extraction from data in real-time by an API such as Twitter, Alphavantage, Google News, etc.
- Web Scraping: BeautifulSoup, scraping websites from forums or blogs

2.5.2.2 Data Cleaning and Preprocessing

NLP - The techniques under NLP include tokenization, lemmatization, and stopword removal.

Noise filtering: This process uses a machine learning model to identify irrelevant content and exclude that from input data.

2.5.2.3 Data Storage and Management

Databases: NoSQL databases like MongoDB and cloud-based solutions such as these will provide scalable storage solutions to work on huge data sets.

Data lakes: These facilitate bringing both structured and unstructured data into a shared repository, where it can be accessed and processed.

2.5.3 Admixing Variants of Data Sources

2.5.3.1 Congruent Historical and Sentiment Data Analysis

Historical price data combined with sentiment data portrays a two-fold: Historical Trends: Reveal long-term market patterns and dependencies. Sentiment Concerns: Any short, direct market reactions to news and public opinion.

2.5.3.2 Multisource Data Feature Engineering

Feature engineering plays a vital role in multi-source data integration such as: Sentiment Scores. For example, derived from text using models such as FinBERT and GPT-4. Lagged variables. Historical price movements act as features.
Categorical variables Event-wise, such as earnings-related announcements or global news.

2.5.3.3 Examples of Case Studies

Basel has developed a framework that incorporates news and social media sentiment for stock prediction. Mittal & Goel (2012): Predictive power of social media sentiment with the historical data then market prediction.

2.5.4 Obstacles to Integration

Data Mismatches: Offsetting the publication times or datetime stamps of news articles, newspapers, and tweets just complicates the alignment.

Overfitting Risks: Very high dimensionality data may give rise to overfitting because of increased usage of deep learning models with comparatively small samples in training (Goodfellow et al., 2016).

Scalability and Computer Costs: Quite enormous amounts of computational resources in processing and storing this multi-source data feed are likely needed for real-time applications.

2.5.5 Future Directions in Multi-Source Integration

1. Real-Time Analytics

Real time data processing can be achieved only with efficient development of algorithm and hardware that can process data in real time. The speed and precision of the decision- making would be enhanced by developing these.

2. Advanced Techniques for Data Fusion

Attention mechanisms and graph-based learning approaches distinguish data fusion improvement systems from the other methods.

3. Advanced Alternative Source Data

Inclusion of voice transcripts, satellite imagery, and IoT data would deepen dimensionality of market analysis.

Multi-source data integration is a revolution for financial analytics. The future pay-off in term of new insights and predictions is likely to be very high by opening up

data quality and alignment problems, together with computational scalability. New observations should then account for behavior" in markets.

2.6 Research Gaps and Opportunities

While there have been significant improvements in sentiment analysis and deep learning, the accuracy of stock market prediction models has significantly improved. However, challenges still present opportunities for research and new discoveries.

2.6.1 Challenges in Real Time Sentiment Analysis

1. Latency: Real-time sentiment analysis requires low-latency systems for data collection, preprocessing, and analysis. Current models, such as FinBERT and GPT-4, are computationally heavy and even might not reach the real-time performance levels (Mittal & Goel, 2012).

2. Sentiment Ambiguity: Sarcasm and context-dependent sentiment are but few examples of ambiguous factors in the textual content. Although the above models have been developed, the current capacity of the models is limited in achieving a consistently well-resolved ambiguity (Cambria et al., 2013).

3. Integration with Market Dynamics: Synchronizing real-time sentiment information with other dynamic market measures-how order flows-as well as including volatility indicators-suggests the problems in the area of feature selection, which continues being an open issue (Loughran & McDonald, 2011).

2.6.2 Computational Limitations of Hybrid Models

1. Resource Requirements: Essentially, such hybrid models combining sentiment analysis with time series forecasting, that is FinBERT with LSTM, require

huge computing power and memory in their training phase, making accessibility for smaller organizations limited (Goodfellow et al., 2016).

2. Scalability: It becomes quite challenging to scale up systems in proportion to data and make the model increasingly complex over decreasing margins of performance.

3. Data Disbalance: There exist properties of hybrid models as related to unbalanced collections, to be precise when the models under study include analysis of event-related sentiment with respect to swaps. Such imbalances may cause inappropriateness in predictions and may reduce generalization.

2.6.3 Prospects for Future Research

1. Real-Time Processing Frameworks: Developing weightless and efficient algorithms to enable real-time data ingestion and processing would change the landscape of financial forecasting. Techniques like distributed computing and edge processing could be a possible answer.

2. Improved Sentiment Models: Tune transformer constructions like GPT-4 to build financial-specific ambiguity resolution and sentiment extraction from its domain for greater accuracy.

3. Use of Alternative Data Sources: More unconventional data sources could include other forms of satellite imaging, web traffic measures, and even voice analysis from earnings calls captures to provide some more interesting insights into market behavior (Daas et al., 2015).

4. Hybrid Model Explainability: Improve the explainability of such hybrid models as FinBERT-LSTM to attract more investors and regulators to them by making them more actionable insights than predictions of the black-box type.

5. Ethical Artificial Intelligence in Financial Analytics: Addressing ethics regarding data privacy and algorithmic bias would ensure responsible application of AI in predicting the stock market (Jobin et al., 2019).

2.7 Conclusion

This literature review provided insight into sentiment analysis, deep learning, and hybrid methodologies for stock market prediction. Traditional models were very useful in the early days, but compared with recent developments, they fell very short in terms of successful modeling of modern developments in financial markets. Advanced research like FinBERT, GPT-4, LSTM, etc. has shown promising potential in conditionally alleviating the integration of multi-sourced data and non-linear dependencies.

Despite their promise, limitations concerning computational inefficiency, real-time processing, and data integration need to be overcome. Gaps will be filled with future studies to develop lightweight, scalable, and interpretable models that control data diversity and ethical AI.

Through the advancement of the domain of financial analytics, hybrid models have the potential to change the complete outlook with which stock market predictions are constructed today. More accurate predictions and more informed decision-making will soon take place in an increasingly complex and dynamic market environment.

Chapter 3

Research Methodology

3.1 Introduction

This chapter describes the approach used in this research work in detail especially the process of combining sentiment analysis and time series forecasting to predict stock market. This is because to recognize the interrelation between the structured and unstructured data types in financial analytics is challenging given the nature of the data, and by leveraging natural language processing and deep learning approaches in this study, it shall effectively solve this research problem. The practical aspect of the work also meets these objectives since the proceeding from data collection to deployment is best practices, thus offering a blueprint for more investigations.

3.2 Research Framework

To address the research objectives efficiently and with an adequate level of comprehensiveness, the research frameworks follow a clear data science life cycle. It is composed of interdependent subprocesses and each of them results in the production of a predictive model. The design includes problem formulation and data collection and initial assessment, cleaning and exploring, feature engineering and modeling phases. This analysis involves the combination of two techniques proves to be the main working model of the methodology. Assessment and implementation follow the created and improved solutions accompanied by continuous monitoring to make the solutions Roi-oriented.

Prominent in this framework is the ingration of quantitative (Numeric) and Qualitative (Text) data. Text analysis applied to the data corresponds to the attitude of

the public and media, while time series analysis utilizes historical {price behavior}. These insights are then integrated into the hybrid model used in understanding market dynamics. Besides, this framework can not only solve the imminent research concern but also present a potential blueprint for future research on financial analytics.

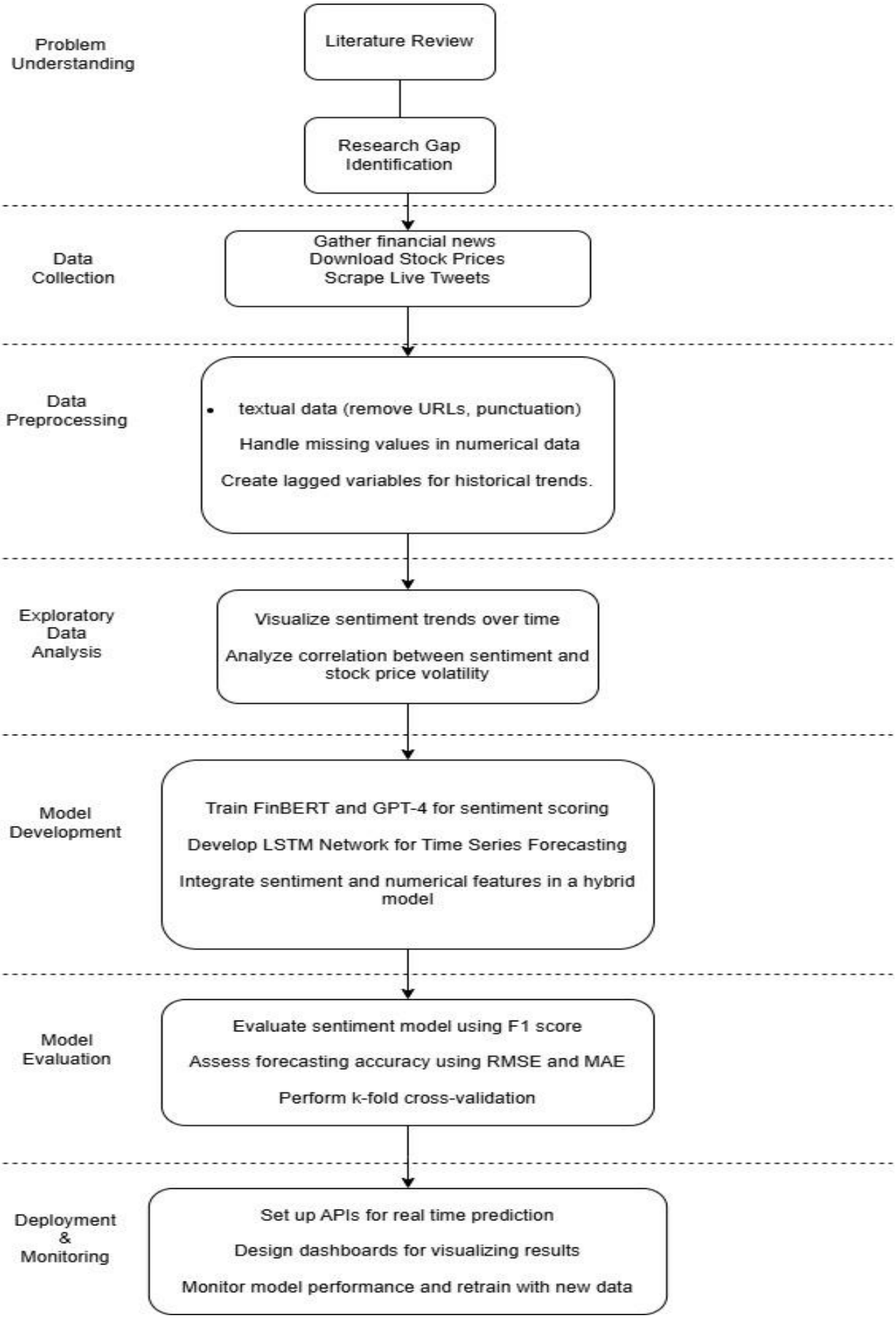

Problem Understanding
Literature Review
Research Gap Identification
Data Collection
Gather financial news
Download Stock Prices
Scrape Live Tweets
Data Preprocessing
textual data (remove URLs, punctuation)
Handle missing values in numerical data
Create lagged variables for historical trends.
Exploratory Data Analysis
Visualize sentiment trends over time
Analyze correlation between sentiment and stock price volatility
Model Development
Train FinBERT and GPT-4 for sentiment scoring
Develop LSTM Network for Time Series Forecasting
Integrate sentiment and numerical features in a hybrid model
Model Evaluation
Evaluate sentiment model using F1 score
Assess forecasting accuracy using RMSE and MAE
Perform k-fold cross-validation
Deployment & Monitoring
Set up APIs for real time prediction
Design dashboards for visualizing results
Monitor model performance and retrain with new data

Figure 3.1: Overall Research Framework

3.3 Problem Understanding

In the context of analyzing the stock market as an object of the investor's focus, implying fluctuations in its dynamics, these factors are an obvious fact. In traditional conjugations primarily linear and static relationships are employed to model these influences, which does not make much sense in particularly non-linear and dynamic fields when textual data such as, financial news or social media posts are included into the prospective analysis. The idea of this study is to fill this gap by integrating raw price numbers with sentiment values estimated from text messages. The integrated strategy allows having a deeper understanding of the behavior in the market and makes forecast more accurate and exact.

Here, the performance evaluation of this project is defined in terms of metrics that consider both sentiment classification and predictive accuracy. The following are the performance indicators for sentiment analysis model: Precision, recall and F1 scores Forcasting performance is measured using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Comparison with other models includes with ARIMA models, as well as stand alone LSTM and other LSTM variants with an aim of proving that the proposed hybrid architecture outperforms the other models.

3.4 Data Collection

Continuing the research framework, the next step is Phase 2: Data Collection. Determining the data source is a crucial step in the sentiment analysis process. In this project, it is essential to ensure that the review data used is original and obtained directly from a trusted source, which in this case is the players themselves, without any modifications or manipulation. Once the data source is determined, the next step is to proceed with data collection. During the data collection phase, the relevance of the data to the project objectives was also considered to ensure that the analysis provided results aligned with the project's goals.

3.4.1 Data Source

The data source for the analysis process was taken from three primary sources. The stock market dataset has 15 NYSE-listed companies stock price for a period of 20 years. The dataset also has features such as Open, High, Low, Close and Volume. For sentiment analysis, the textual information is derived from CNBC news articles. Unfortunately financial news articles dataset spanning over 20 year could not be found. So only the articles from 2018 to 2020 are considered for sentiment analysis. In fact, these are the sources of sentiment analysis in the text of these data, which tells us the perception of market about it.

3.4.2 Data Collection Method

The historical stock prices were collected using Python libraries like “pandas” and “yfinance”. The “yfinance” library was used to fetch the stock data. Then the fetched dataset was stored in a CSV file for further processing. The “pandas” library provided data handling, transformation, and storage. For acquiring financial news data, news articles and headlines were scraped through web scraping techniques elaborately in Python through “BeautifulSoup” and “Requests” libraries. The web scraping process deals with fetching the news from financial online news sources, extracting relevant HTML components, and saving entries for sentiment analysis. Later the textual data were processed through transformers library to apply FinBERT model that scores the sentiment of the news articles as well.

3.5 Data Preprocessing

Data preprocessing is a crucial step aimed at preparing raw data for further analysis. Raw data collected through web scraping often contains incomplete data (e.g., missing values or empty data), inconsistencies, and information that is irrelevant to the study's objectives. These issues can reduce the accuracy of the analysis. The data preprocessing process ensures that the dataset is clean and consistent before moving on to the analysis phase.

3.5.1 Data Cleaning

The raw data undergoes a few preprocessing procedures before the final analysis. The subsequent processes are as follows. For the Stock Price Data, the missing values are dealt using forward-filling techniques in pandas. Duplicate records are identified and removed since they are deemed irrelevant to the entire data integrity. Statistical methods are also adopted to detect anomalies for adjusting extreme price fluctuations. For the News Sentiment Data, the textual data was preprocessed through nltk and re libraries. This involved removal of punctuation marks, special characters, and stop words. Tokenization and lemmatization are performed using "nltk.WordNetLemmatizer" to standardize the words and improve the accuracy of sentiment analysis.

3.5.2 Data Transformation

Once the initial dataset is cleaned, added transformations are done with the proper-python libraries. For Stock Price Data, dates have been standardized using "pandas.to-datetime()" in order to conveniently conduct the time-series analyses since they are consistent in format. The percentage changes were calculated using "pct_change()" with respect to stock prices which are normalized for variations in stock price. For news Sentiment Data, numerical sentiment scores were obtained from the transformation of sentiment labels using the FinBERT model gained through the transformers library. This enables correlation analysis, between sentiment and stock prices.

3.5.3 Feature Engineering

For improving the predictive power, some new features are added to the dataset.

Sentiment Aggregation: This is achieved by aggregating sentiment scores over different time windows with the help of daily, weekly, and monthly rolling of sentiment scores using the pandas rolling function. This helps in bringing the effect of sentiment aggregation across time.

Technical Indicators: Some stock market indicators, including Moving Averages, Relative Strength Index (RSI), Bollinger Bands, among others, are calculated using the ta (technical analysis) library to help predictive modeling.

Lag Features: Past values of stock price and sentiment are included as lag features using shift() in pandas in an attempt to introduce historical dependence into the model.

3.6 Exploratory Data Analysis (EDA)

Exploratory Data Analysis helps users to get a better understanding of many aspects of the given dataset. Where descriptive analyses involve computation of summary statistics such as mean, median, standard deviation and correlation coefficients, then statistical techniques are employed. Heat maps represent the coefficient between the sentiment scores and stock price swing, while the time series plots show series and trends of data over time.

For instance, the inspection of the heatmap of the correlation coefficient uncovers the nature of the correlation between general or sector-specific sentiment coming from the globe's financial media and daily stock price changes that facilitates feature selection. Wherein the stock price trends per security are plotted against the overall sentiment as a function of time through a time-series plot.

3.7 Model Development

In the proposed hybrid model structure, the obtained sentiment analysis informs the time-series forecasting. The sentiment analysis is first done using FinBERT, a transformer model fine-tuned for the financial text. The model therefore generates sentiment scores that will in turn act as inputs for the time series model.

The forecasting component is based on Random Forest that allows considering sequential dependencies. These aspects are in this hybrid architecture, while input layers work with numerical and textual data, hidden layers for working on sequence data, functional layers dealing with sentiment-derived data. Parameter optimization is conducted with grid search with regards to such features as learning rate and batch size and the number of units to optimize the performance of model.

3.8 Model Evaluation

Selecting a performance measure for the hybrid model and for the individual and composite components is given careful consideration. It is classified based on accuracy on how well it categorizes sentiments then using parameters like precision, recall and F1 score. Moreover, for evaluating the time- series forecasting part of the model Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which provide the measure of prediction error.

Validation methods include k – fold cross validation to check the model's stability and compared with conventional model such as SARIMA and only LSTM model. On the basis of this comparison, the enhancement resulting from the use of sentiment analysis and enhanced forecasting are vehicled.

3.9 Deployment and Monitoring

The last of them is to apply the extracted hybrid model for the practical use with some organization or project. APIs imply data exchange in the real-time

environment, thus, the model can be updated continuously. An explorative dashboard is created for visual display of predictions, sentiments, and other KPI's, in a manner that allows for easy tracking via.

Testing and observation allow the model to reach its best consistently. Assessments of the prediction accuracy are done periodically and the algorithm is updated with new data after some time due to performance decline. Triggers of notification and alerts are included here to draw user's attention when there supposed to be fundamental changes in the market behavior.

3.10 Conclusion

This chapter outlined the methodology employed to conduct the project. Beginning with Phase 1, the research identified gaps in existing studies through a literature review, as discussed in Chapter 2. Next, the methodology moved to data collection and data preprocessing phases to ensure clean, structured and meaningful datasets. In addition, feature engineering was carried out to enhance the analytical potential of the datasets. The later steps are mentioned to give an overview of the holistic approach that is taken to materialize the project.

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.

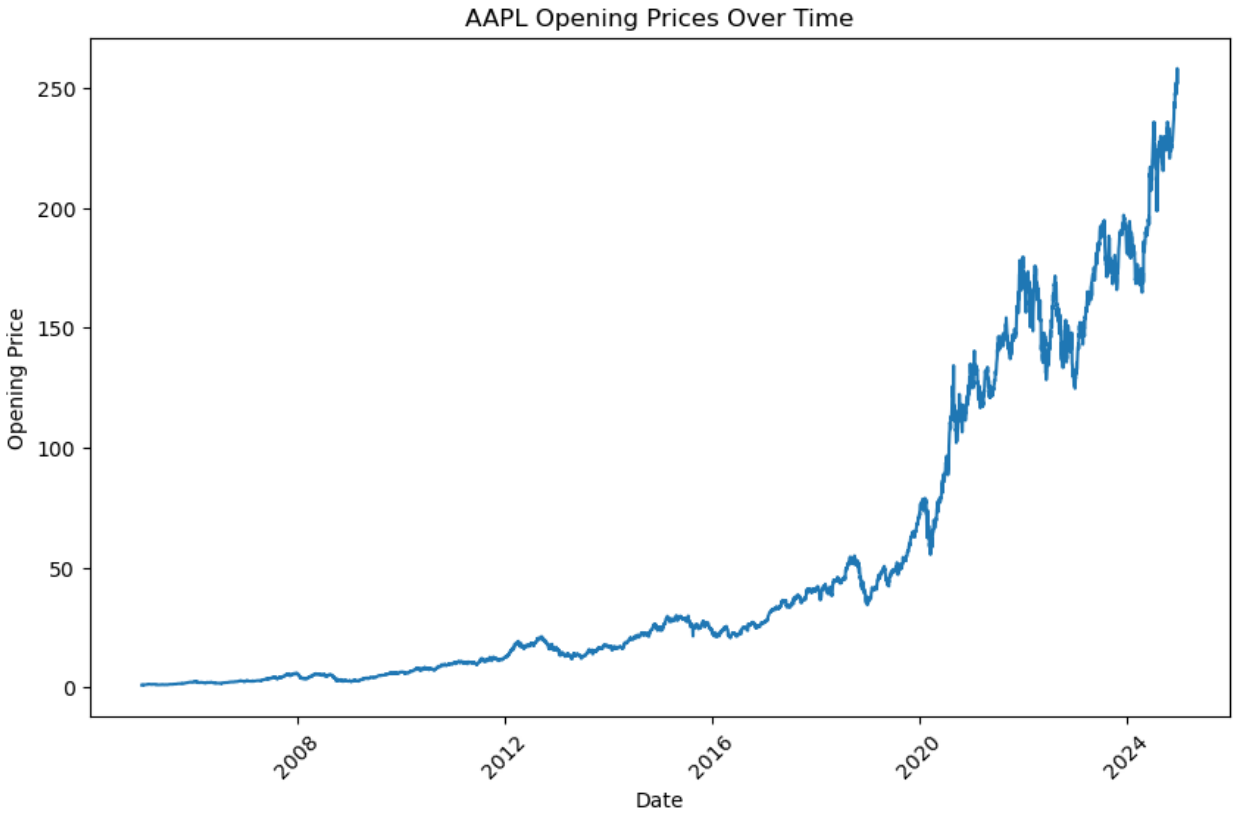
Sentiment Analysis: How sentiments are changing movements regarding stock price throughput-time was analyzed, as these could best reflect the market events that could migrate the sentiments and eventually send the stock prices tumbling or soaring like the earnings reports and political turmoil.

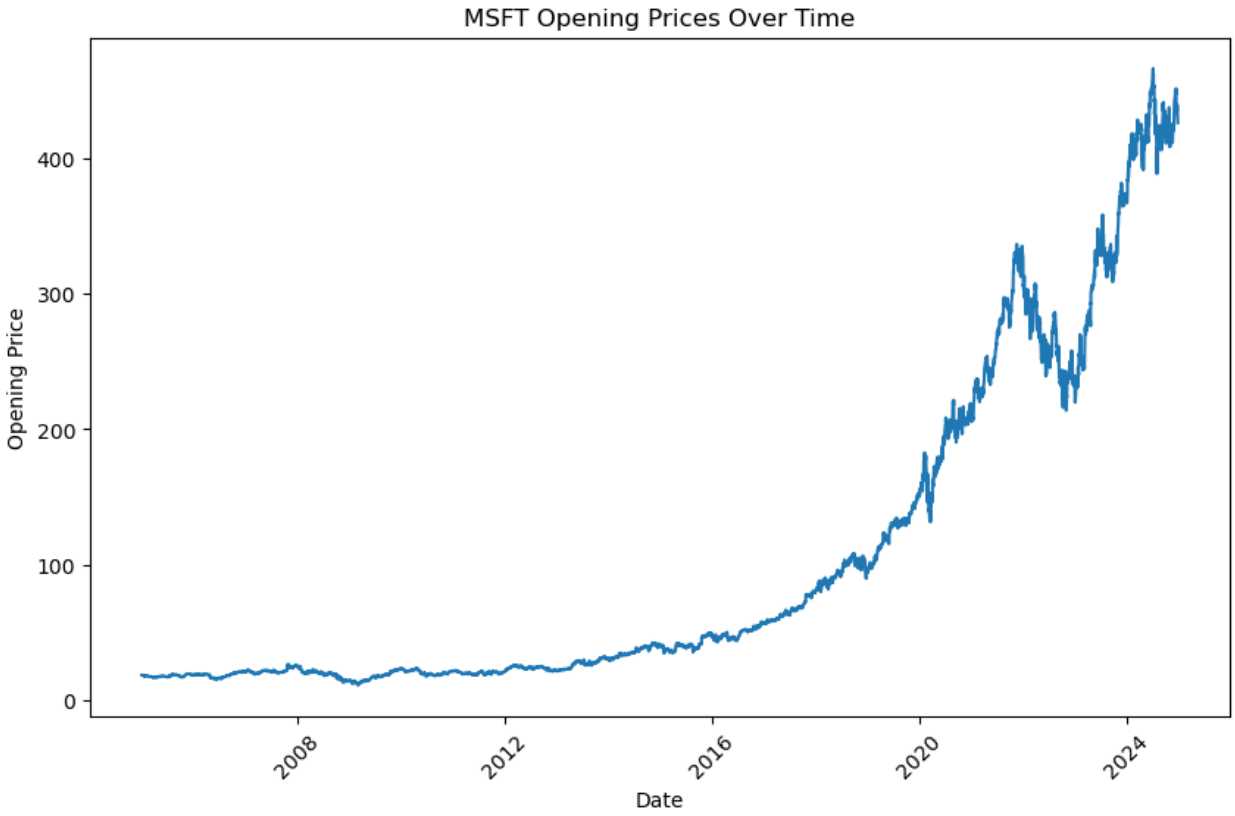
Correlation Analysis: We computed the correlations between sentiment data with stock price movements, particularly focusing on how sentiment influences stock volatility and price returns. It made it possible to determine if such data could act as a predictor for stock price trends.

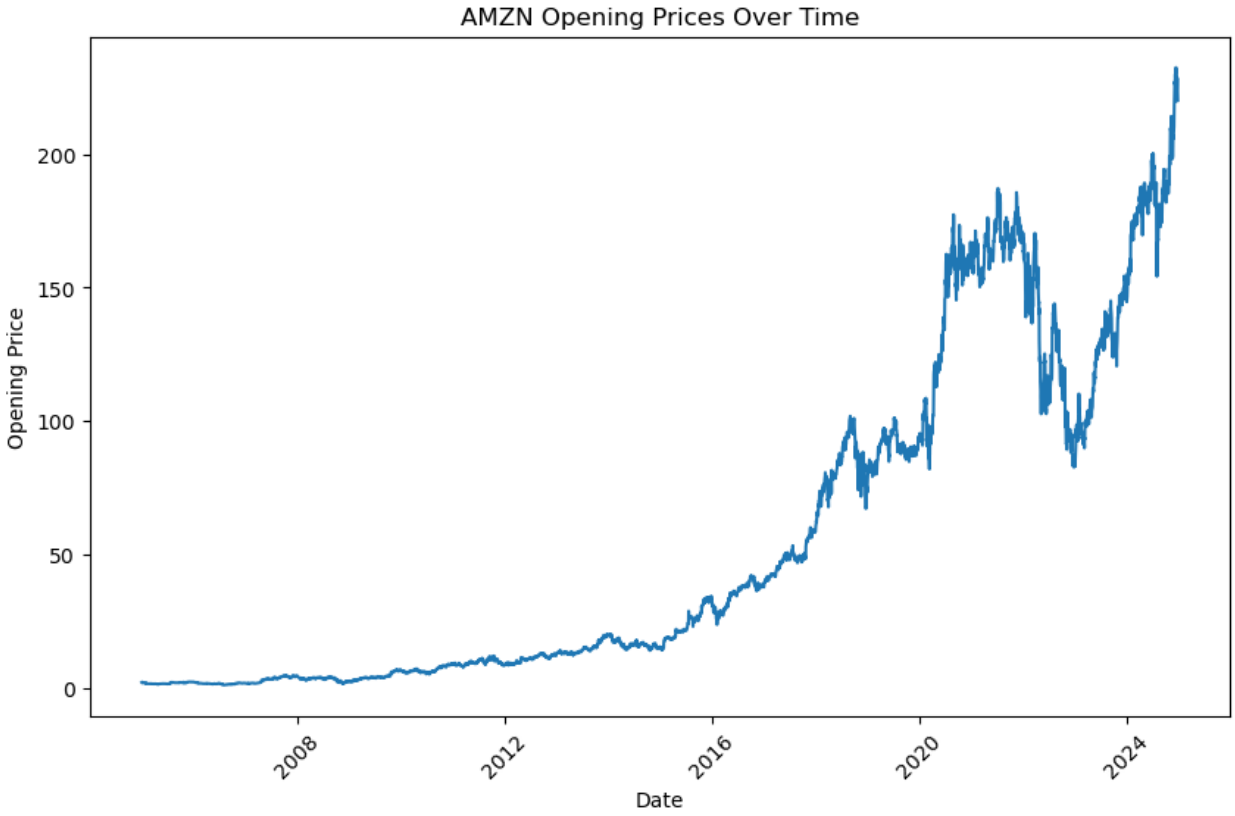
4.2.1 Understanding the stock price dataset

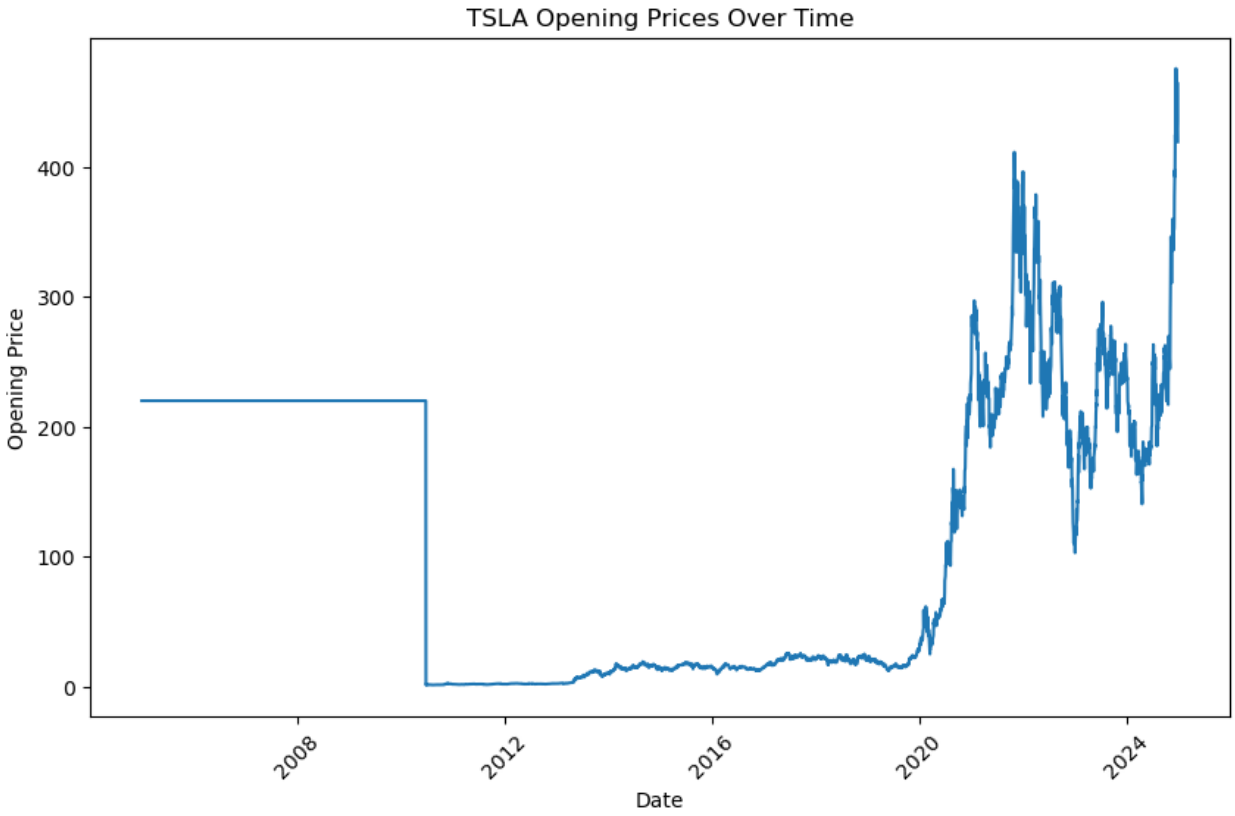
a. Line Plots for Opening Prices:

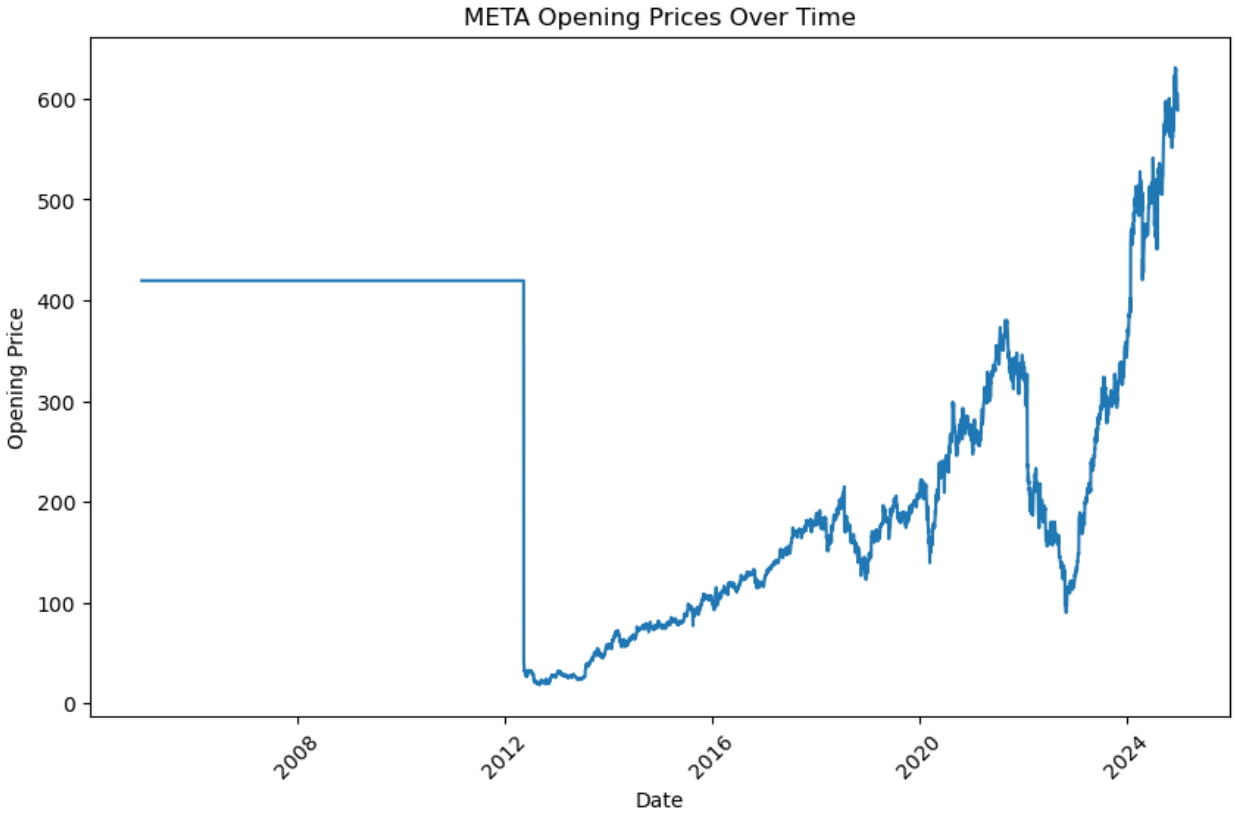
The objective of this line plot is to portray the opening price movements of all firms across a span of time. The data depict the changes in the opening prices over the years for each firm. Trends can tell how the company's stock performed based on years, thus coming with periods of growth, decline,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

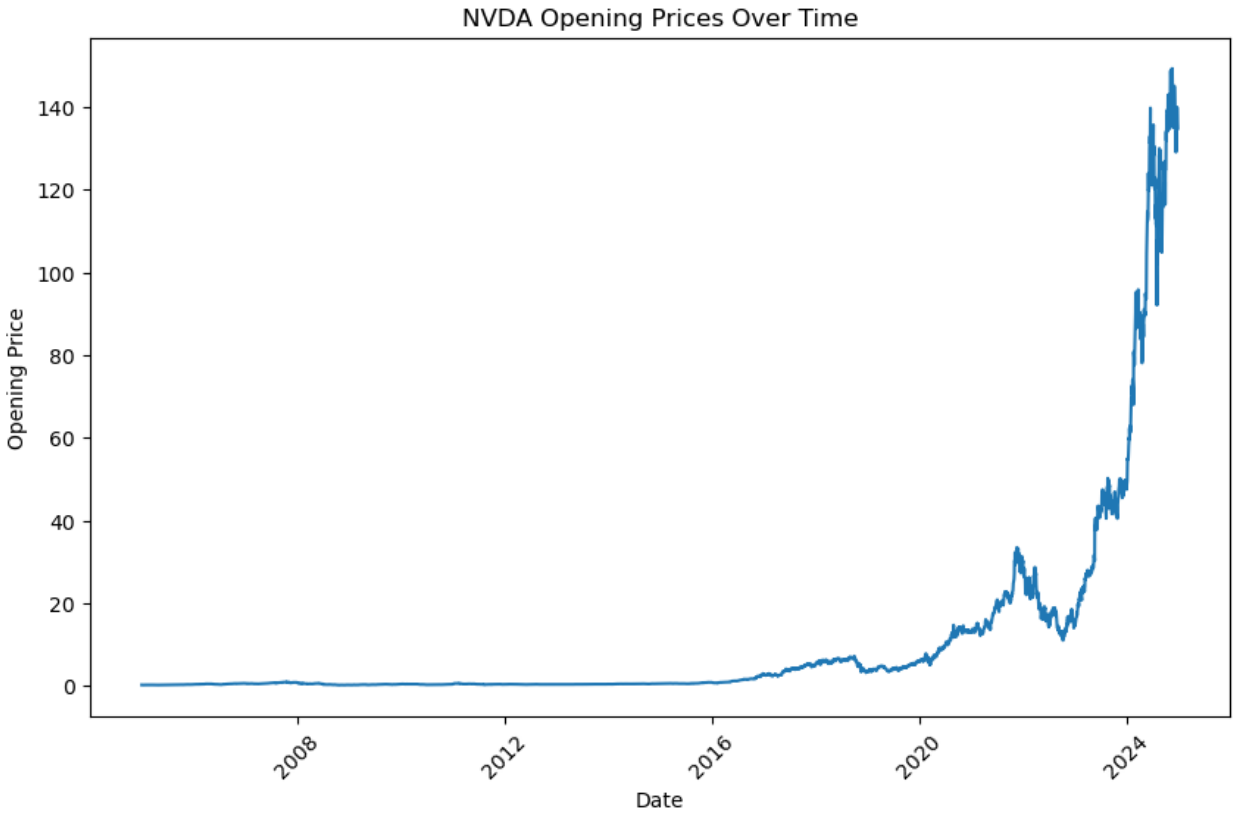


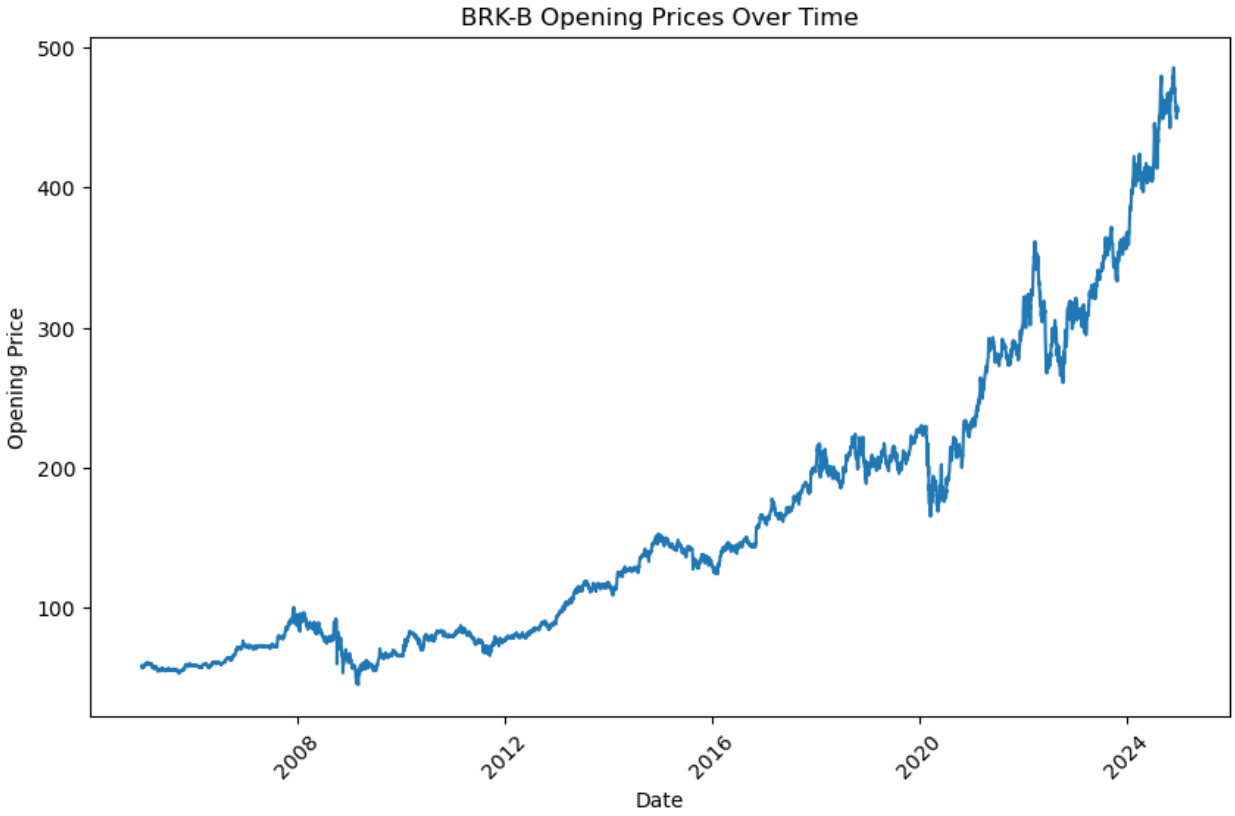

MSFT Opening Prices Over Time
Opening Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

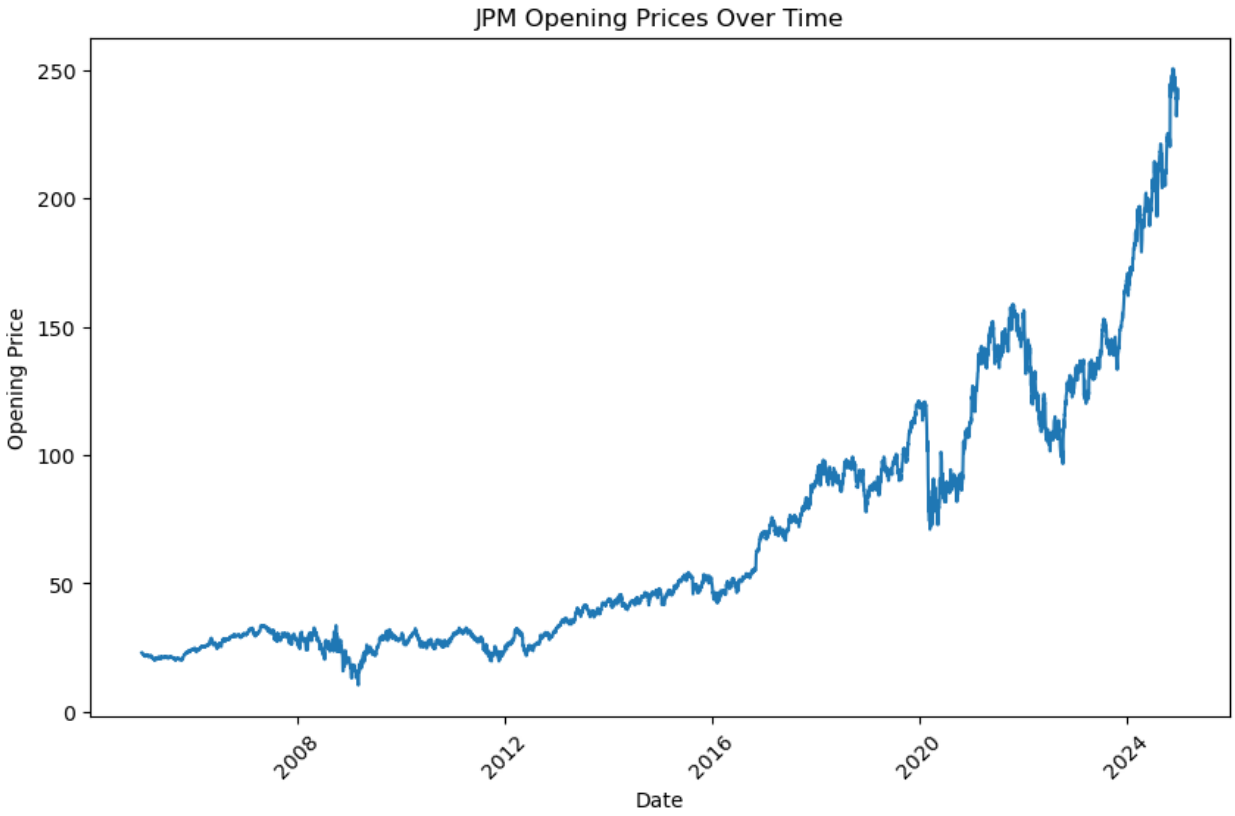

AMZN Opening Prices Over Time
Opening Price
0
50
100
150
200
2008
2012
2016
2020
2024
Date

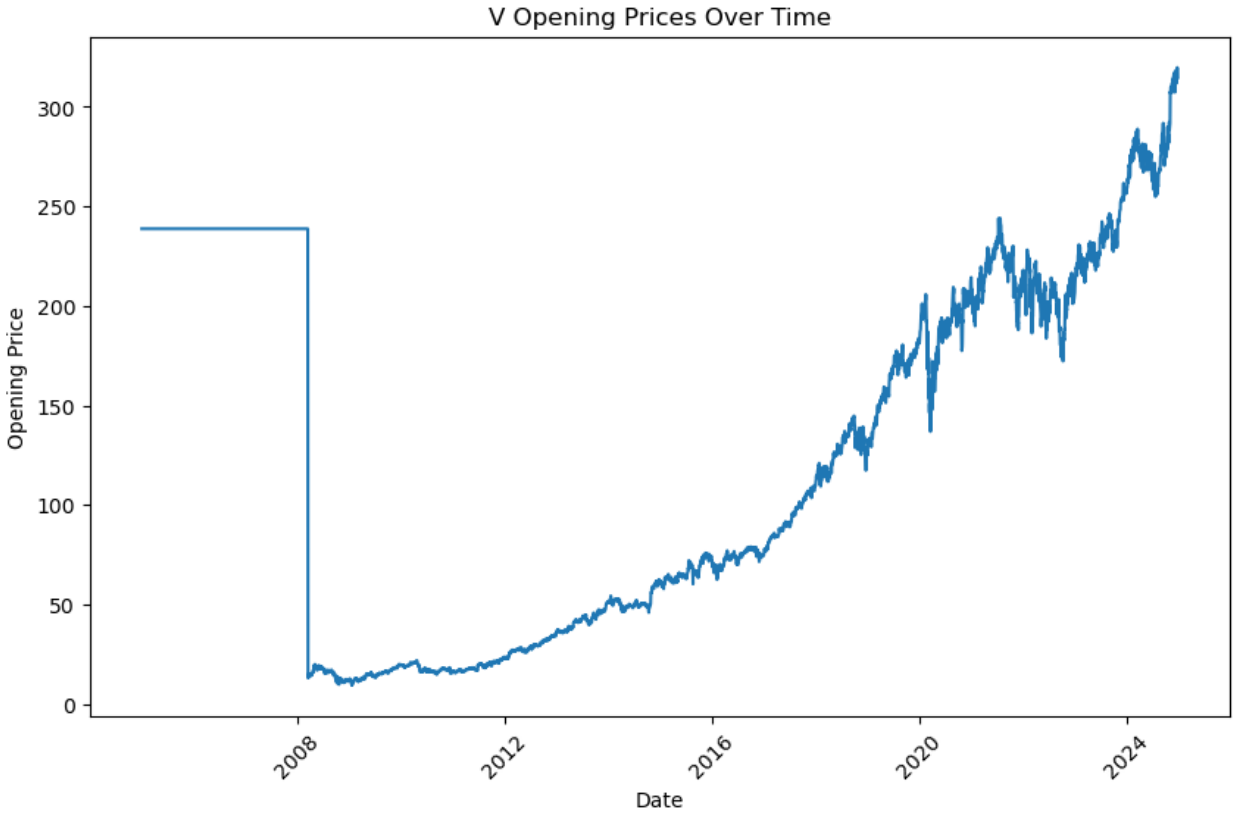

TSLA Opening Prices Over Time
Opening Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

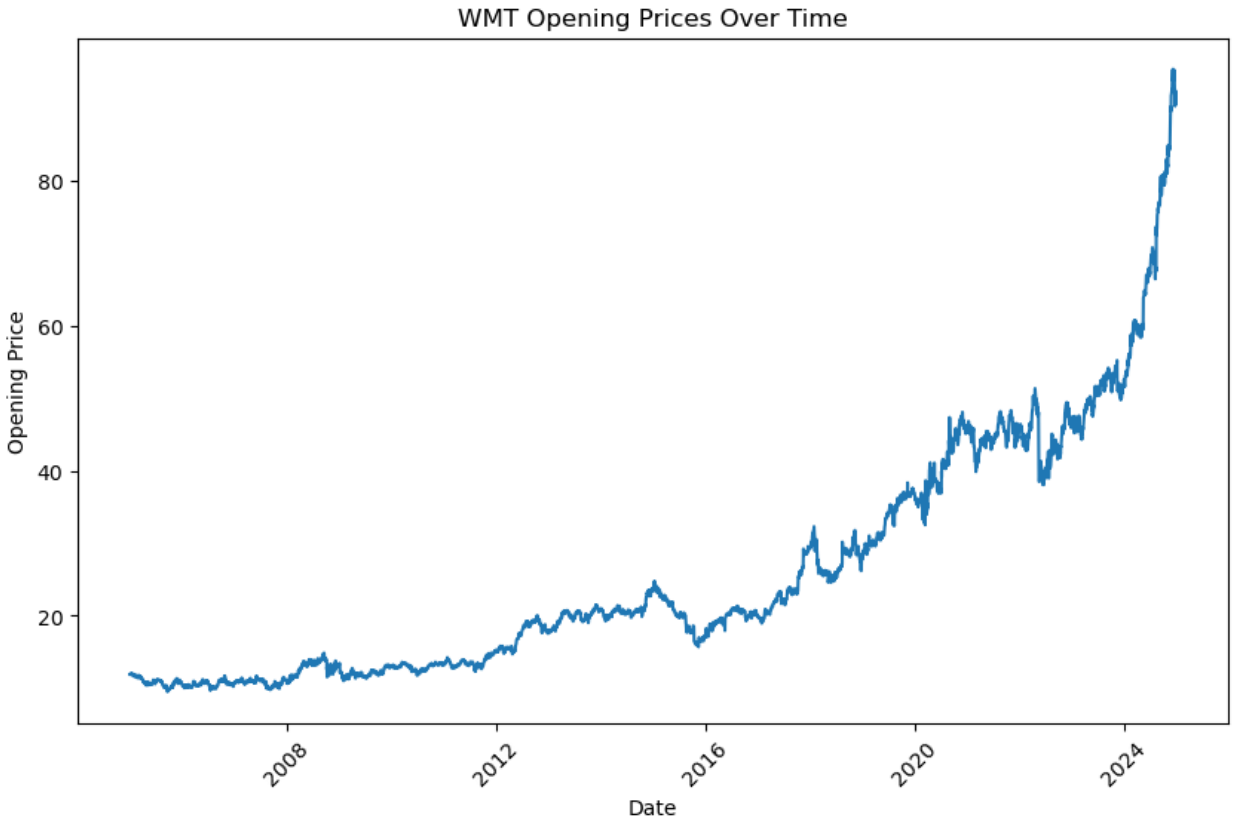

META Opening Prices Over Time
Opening Price
0
100
200
300
400
500
600
2008
2012
2016
2020
2024
Date

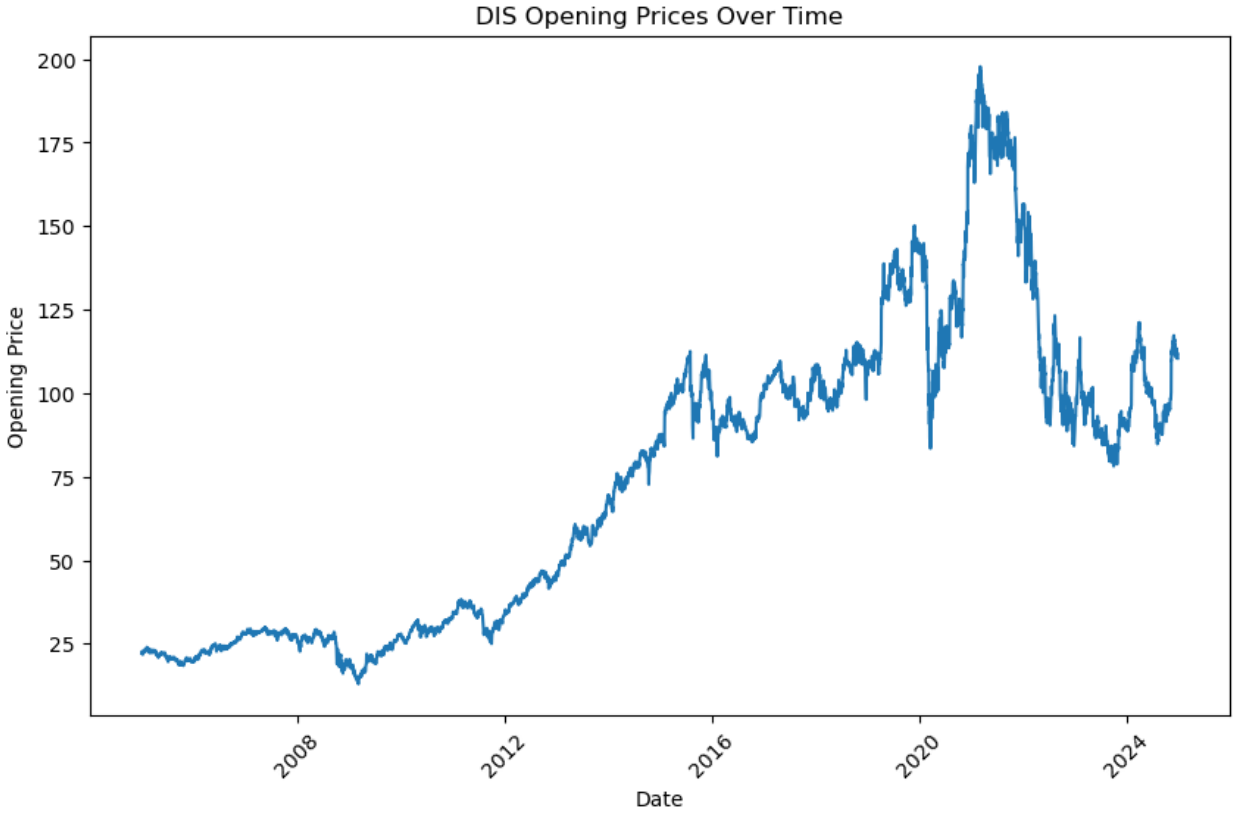

NVDA Opening Prices Over Time
Opening Price
0
20
40
60
80
100
120
140
2008
2012
2016
2020
2024
Date

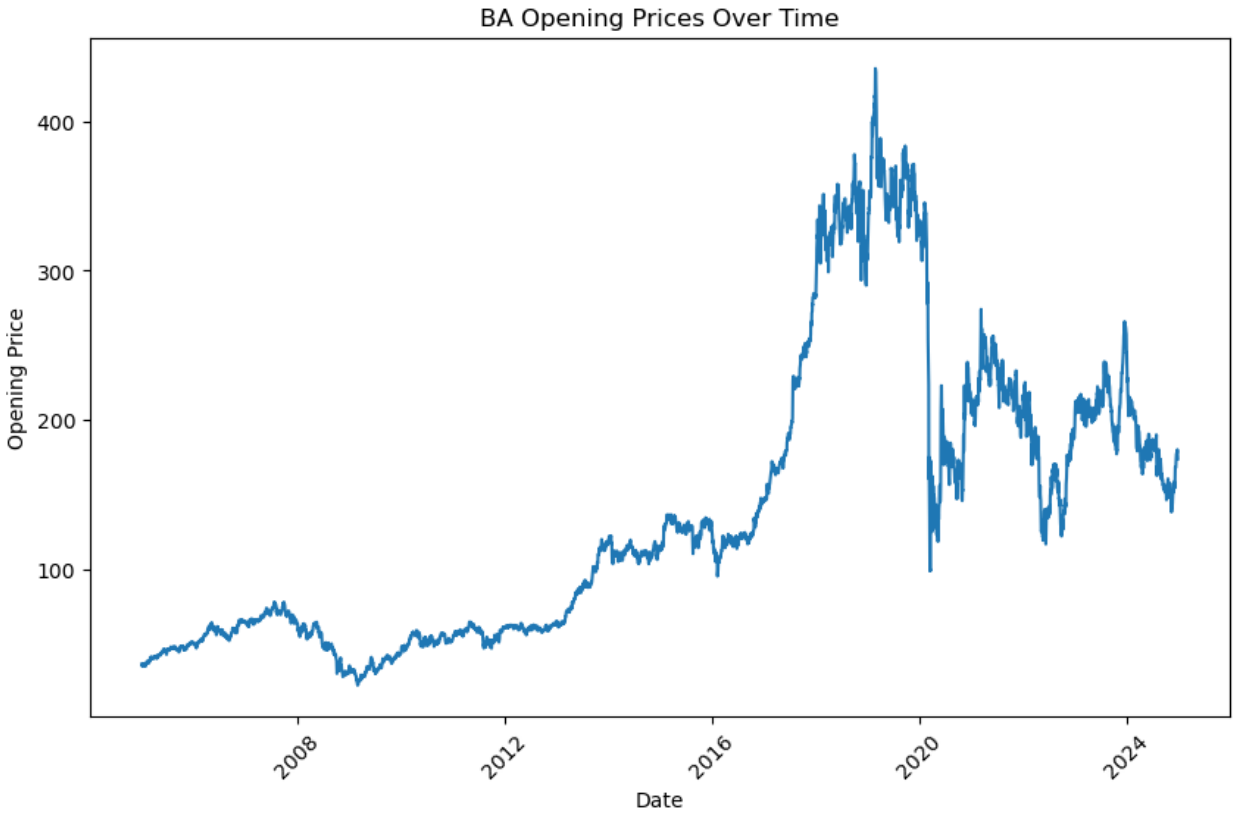

BRK-B Opening Prices Over Time
Opening Price
100
200
300
400
500
2008
2012
2016
2020
2024
Date

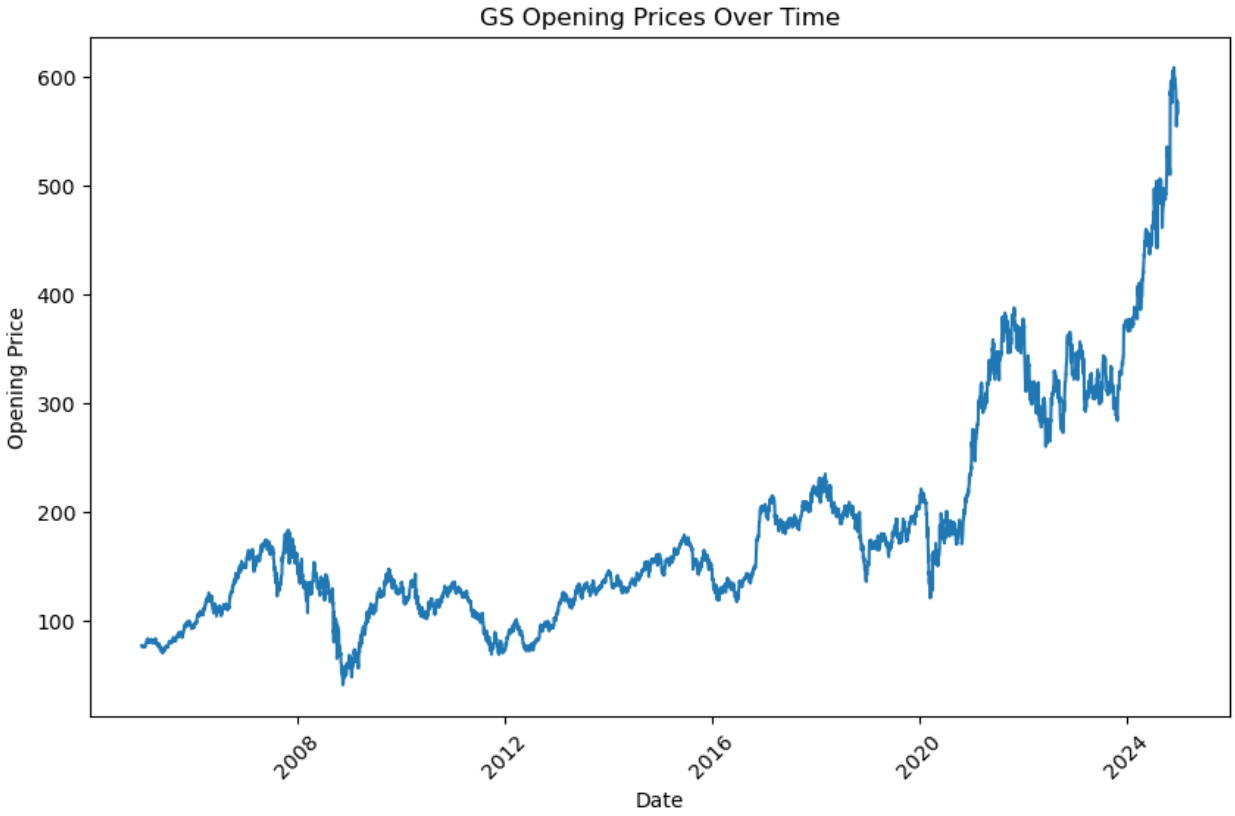

JPM Opening Prices Over Time
Opening Price
0
50
100
150
200
250
2008
2012
2016
2020
2024
Date


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

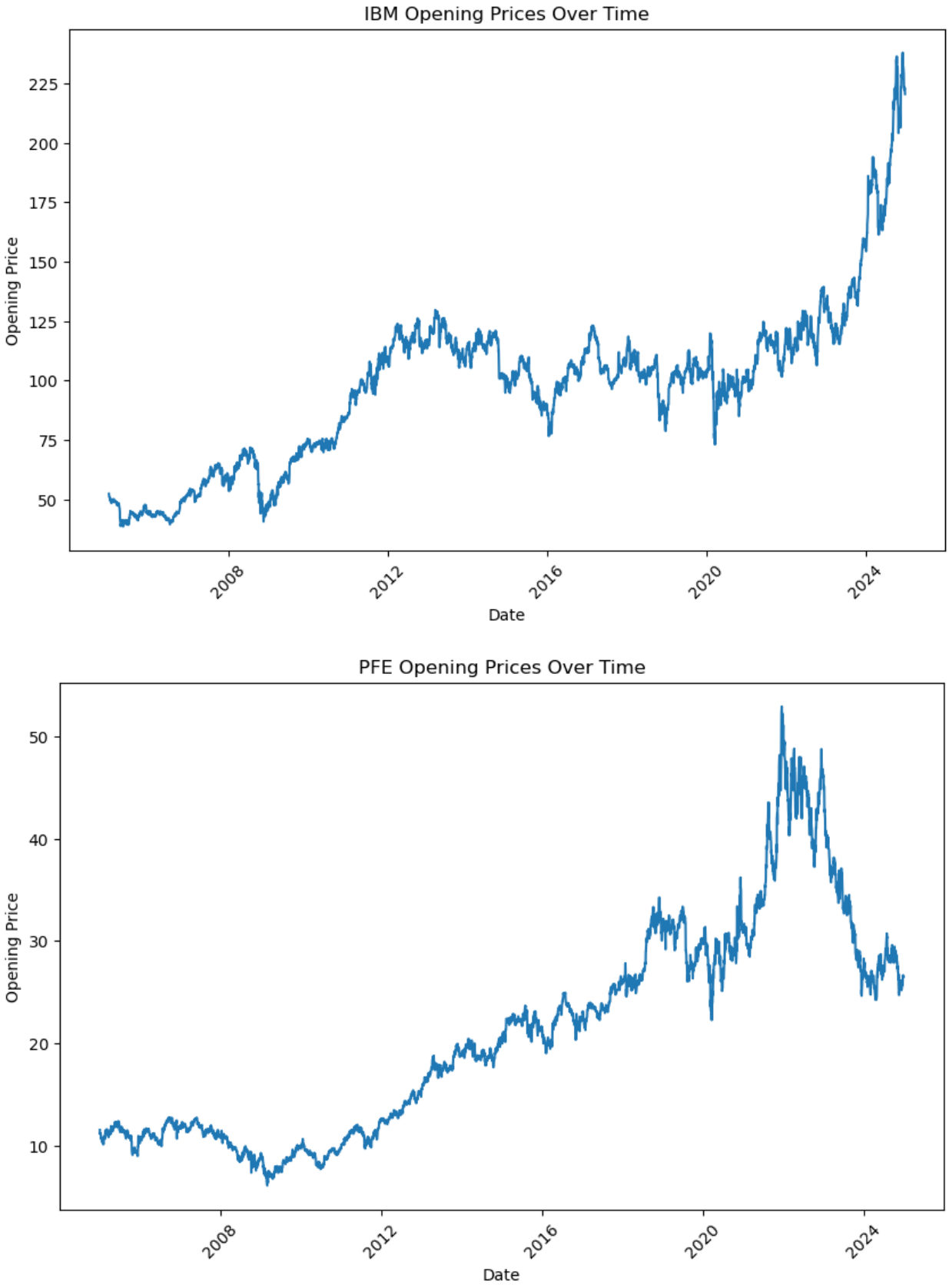
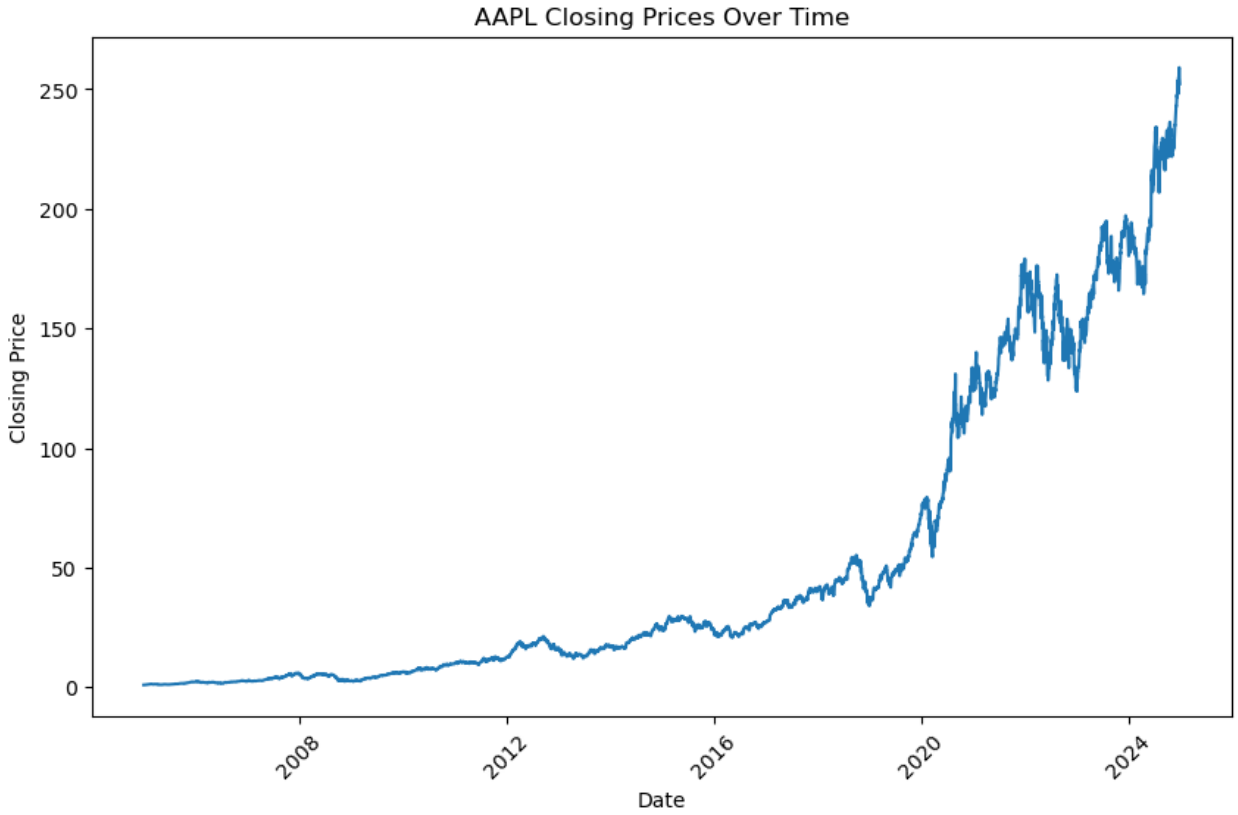
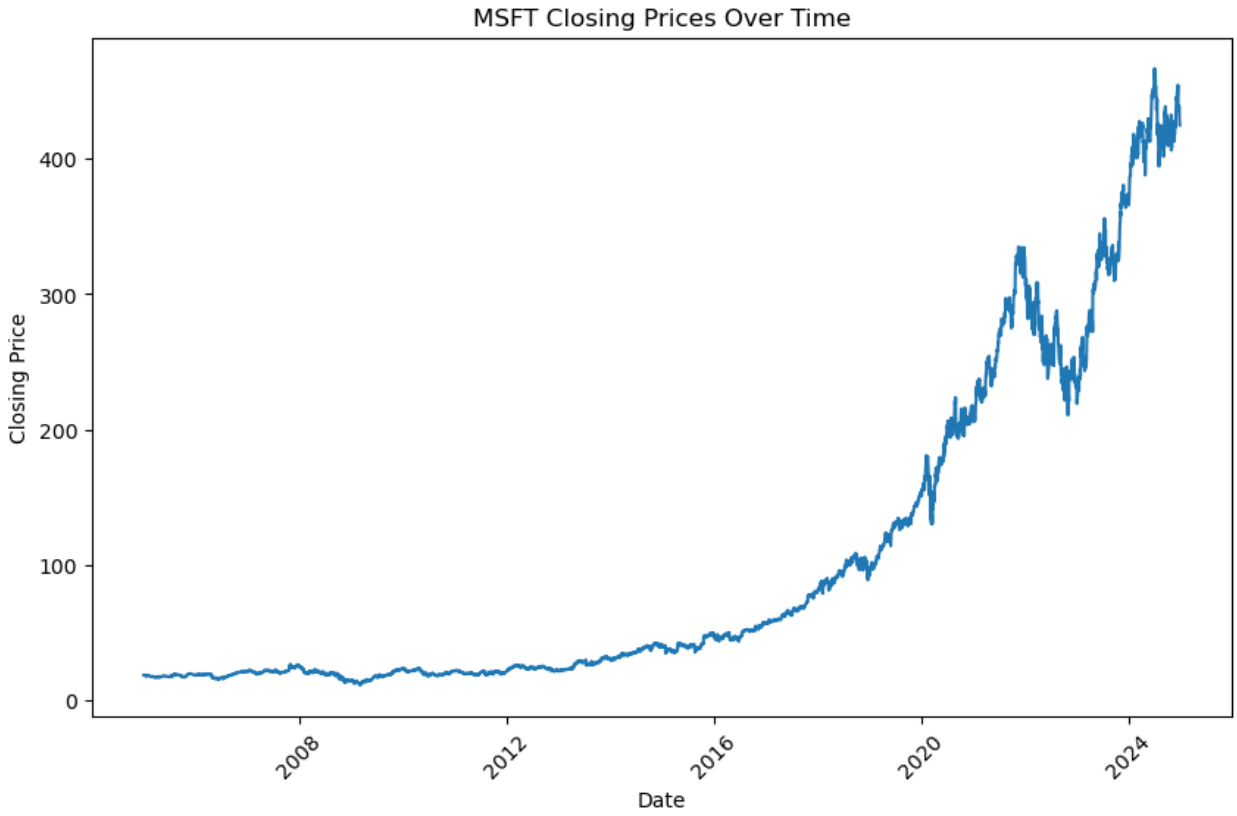


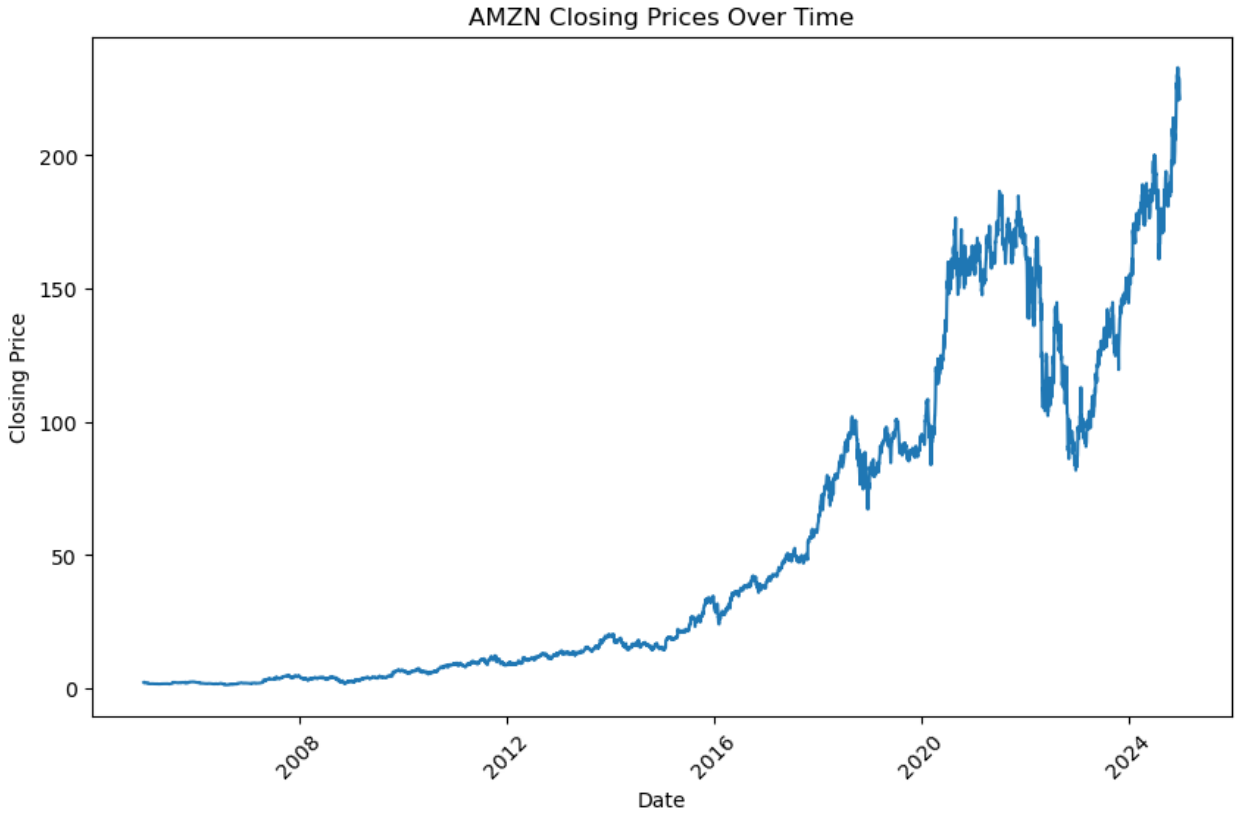
Figure 4.1: Line plots for opening prices of 15 companies

b. Line Plots for Closing Prices:

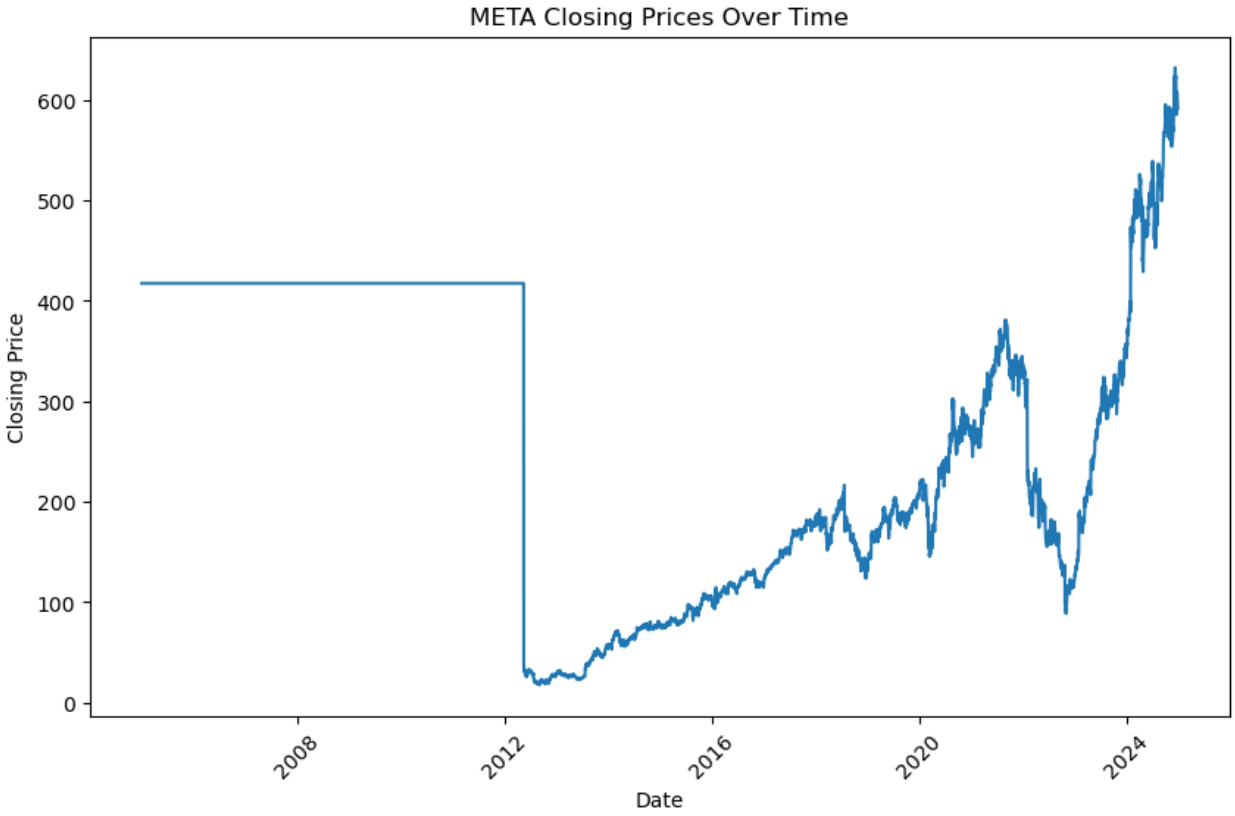
The objective of this line plot is to portray the closing price movements of all firms across a span of time. The data depict the changes in the closing prices over the years for each firm. Trends can tell how the company's stock performed based on years, thus coming with periods of growth, decline,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

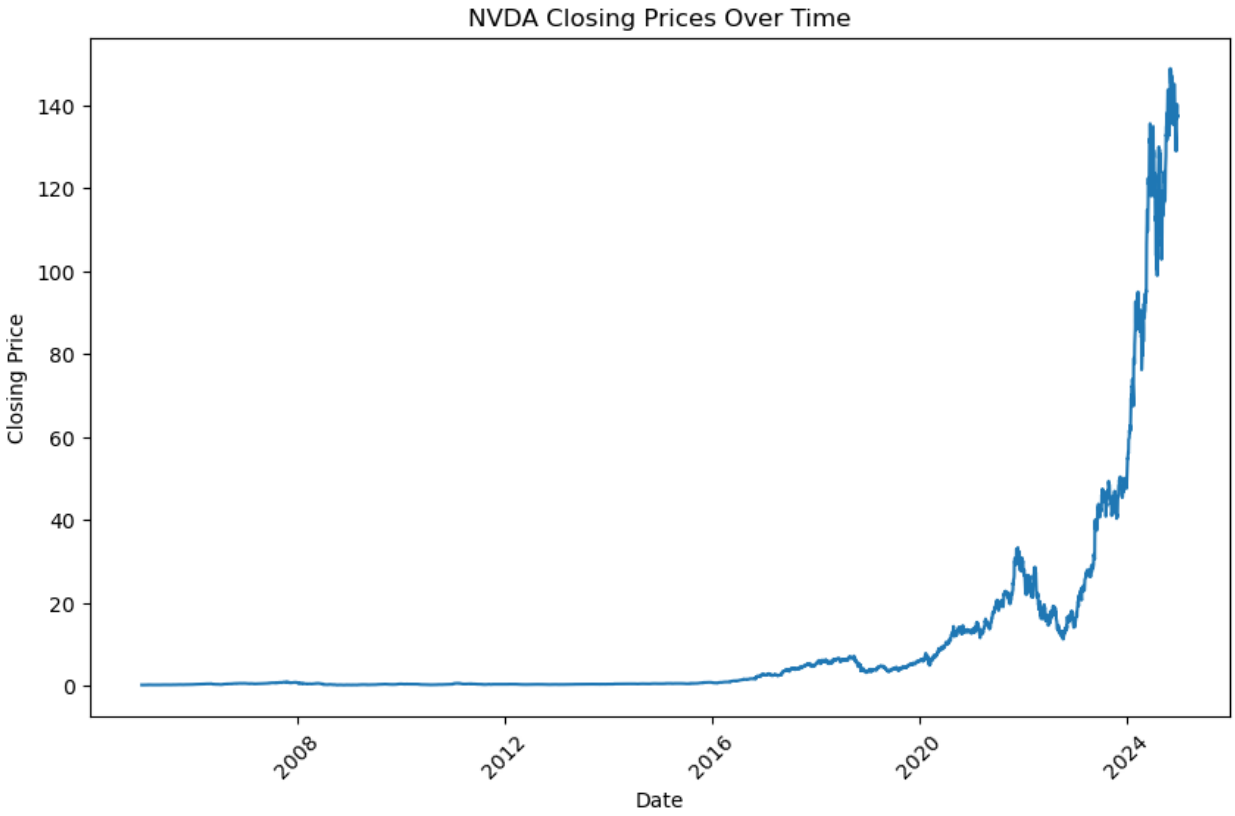


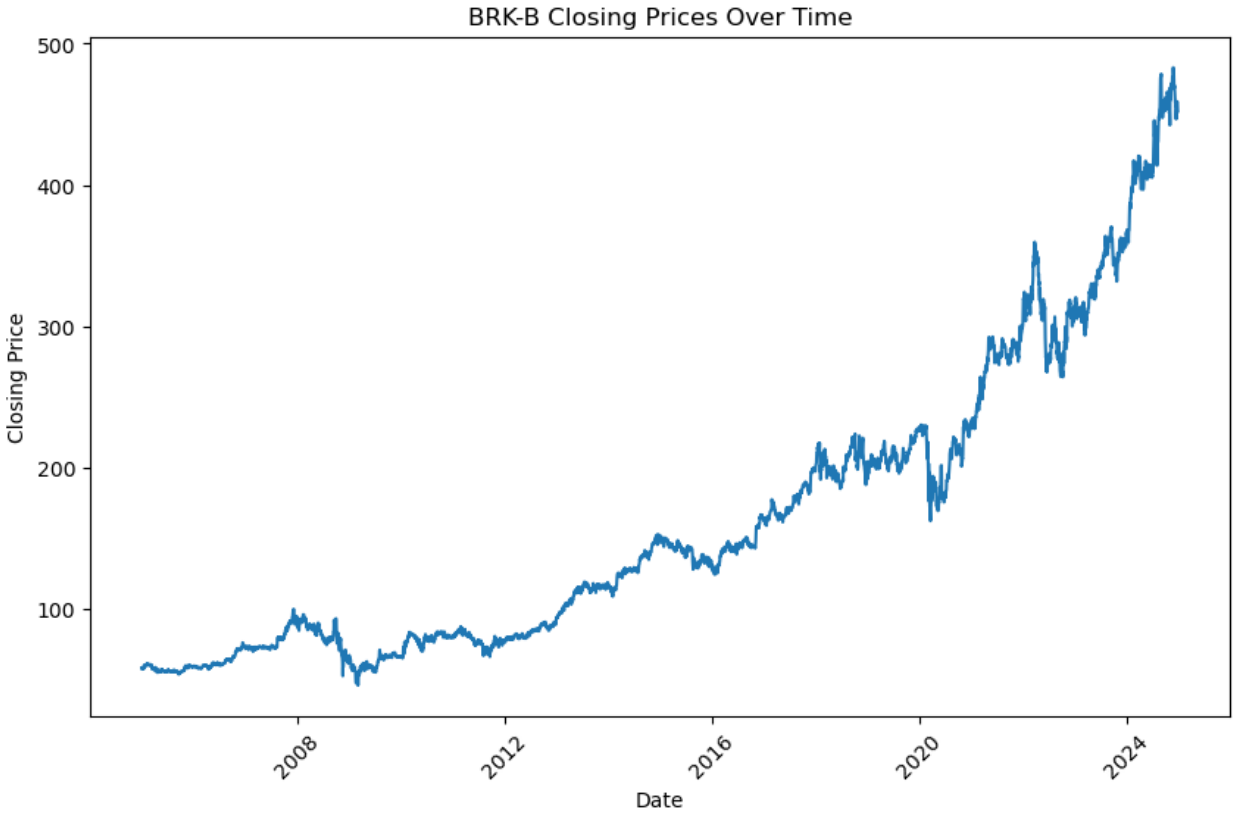

MSFT Closing Prices Over Time
Closing Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

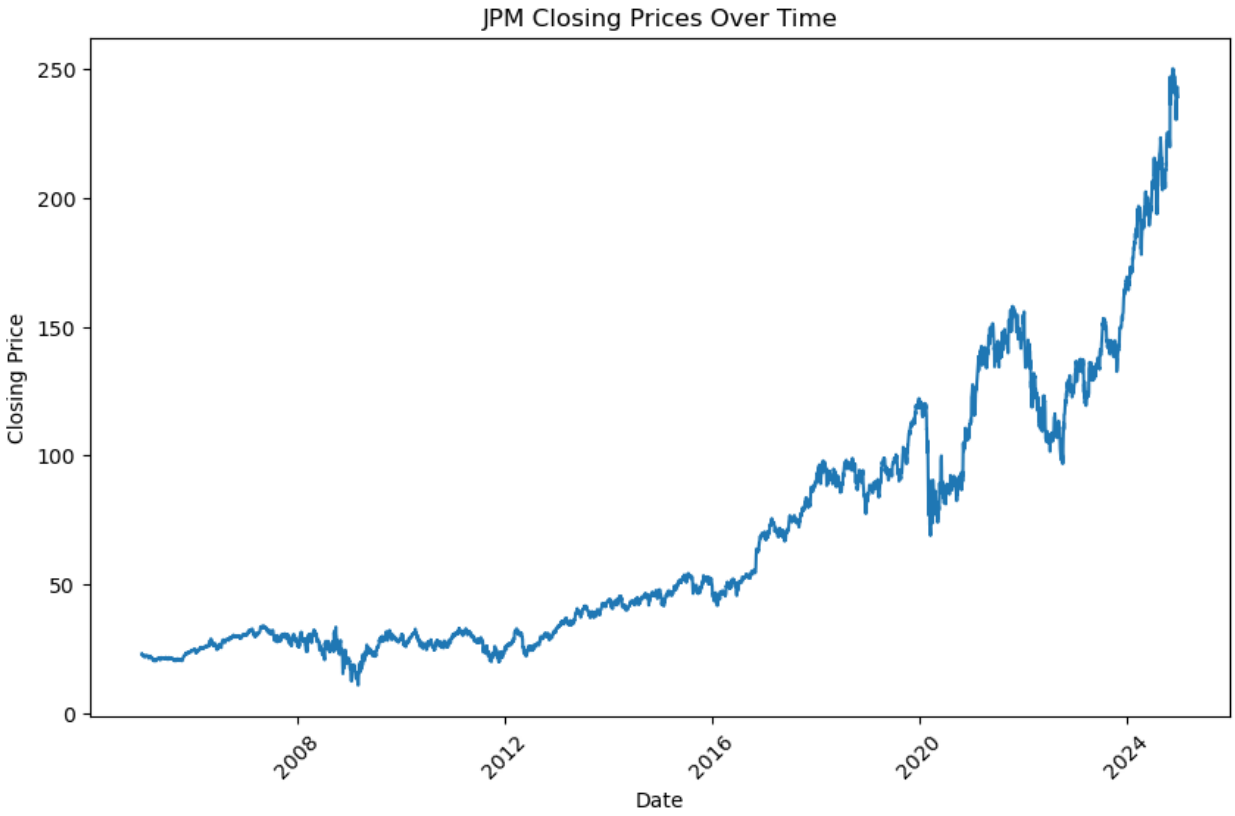

AMZN Closing Prices Over Time
Closing Price
0
50
100
150
200
2008
2012
2016
2020
2024
Date

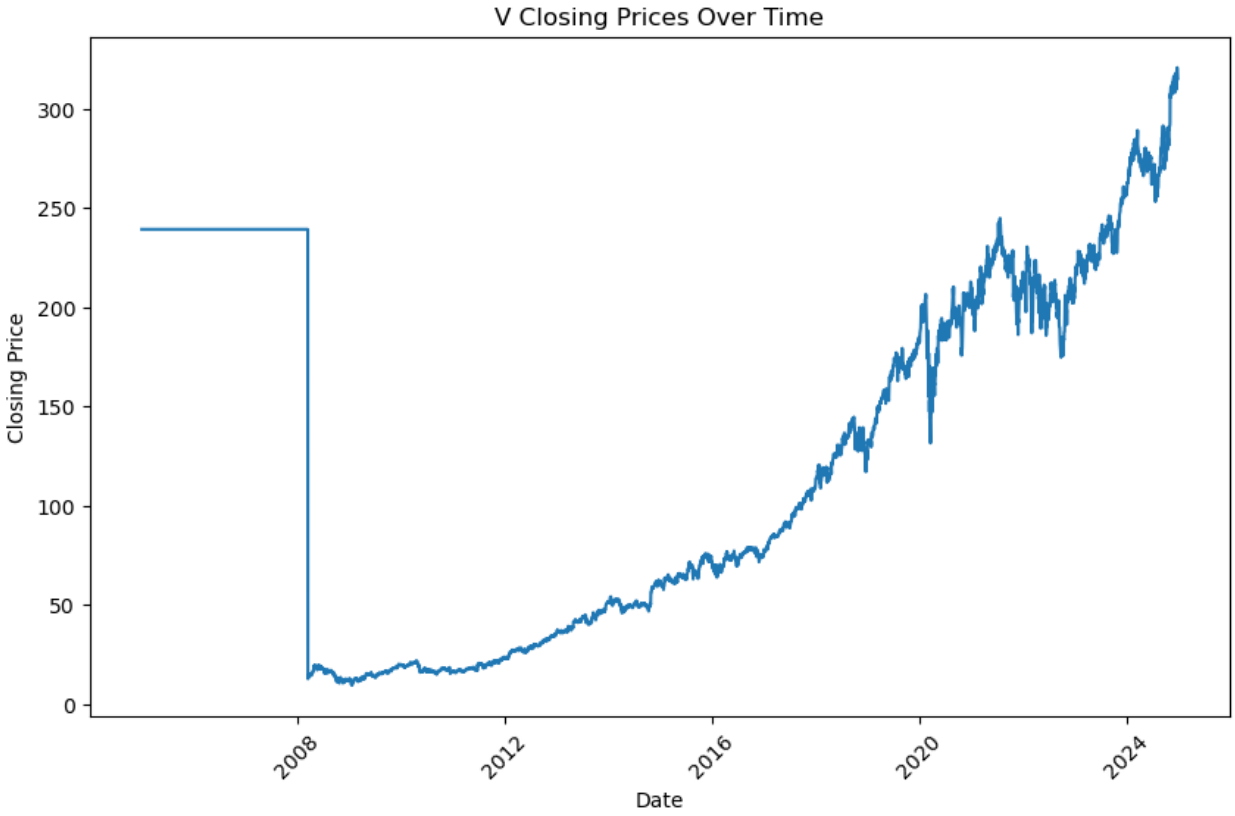

TSLA Closing Prices Over Time
Closing Price
500
400
300
200
100
0
2008
2012
2016
2020
2024
Date

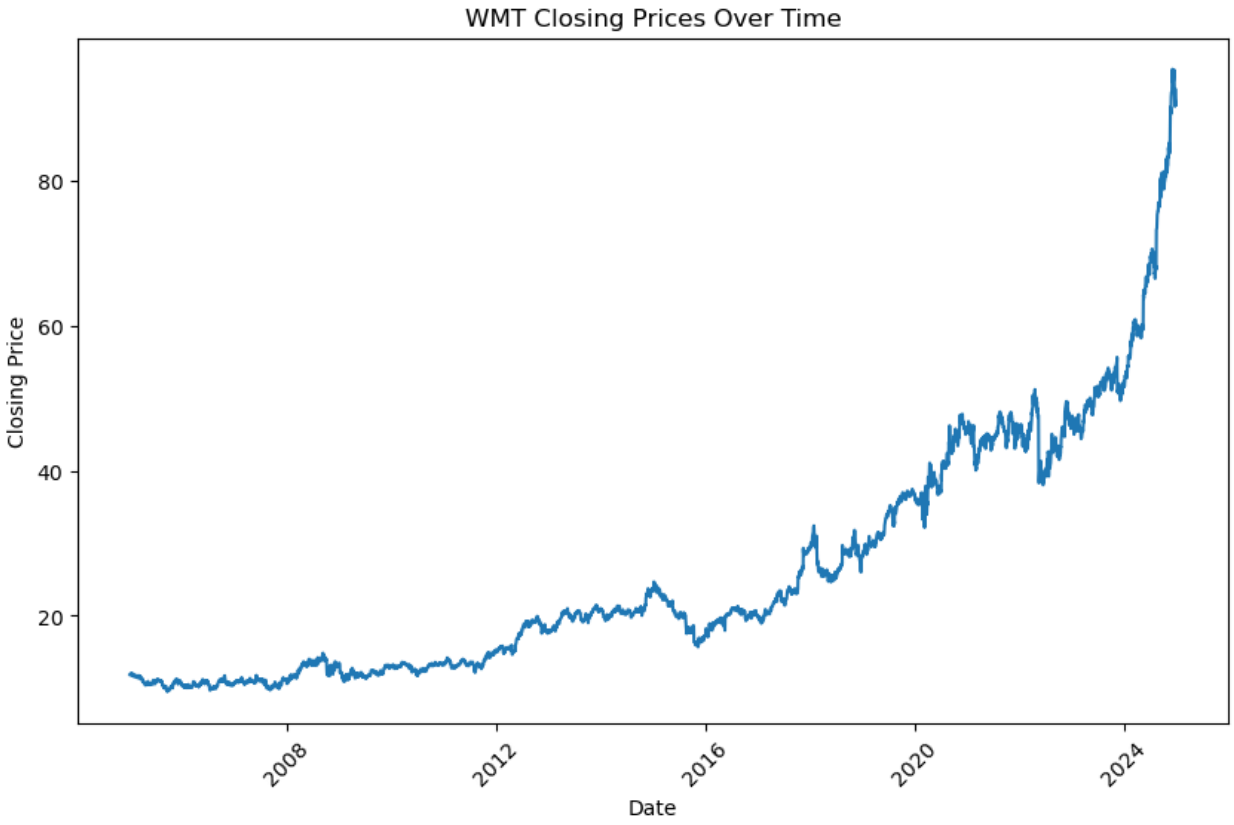

META Closing Prices Over Time
Closing Price
600
500
400
300
200
100
0
2008
2012
2016
2020
2024
Date

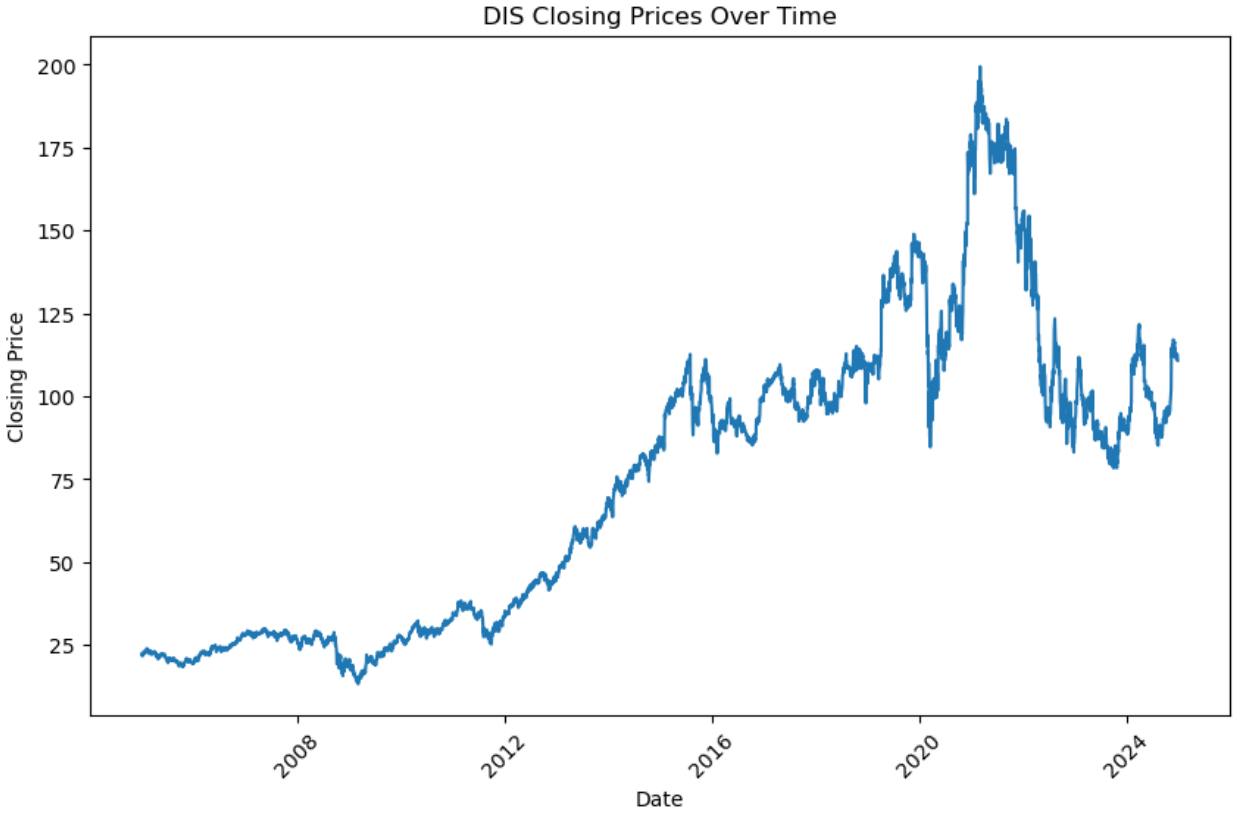

NVDA Closing Prices Over Time
Closing Price
0
20
40
60
80
100
120
140
2008
2012
2016
2020
2024
Date

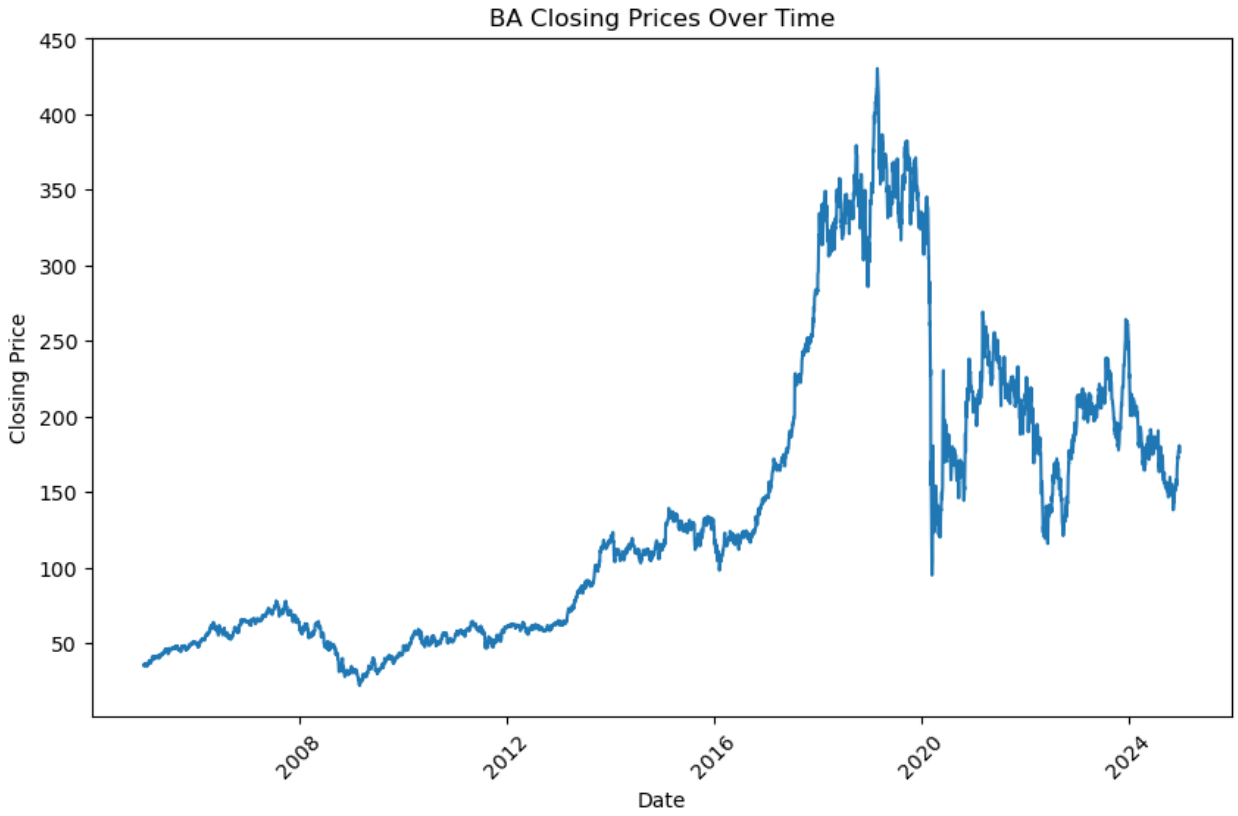

BRK-B Closing Prices Over Time
Closing Price
100
200
300
400
500
2008
2012
2016
2020
2024
Date

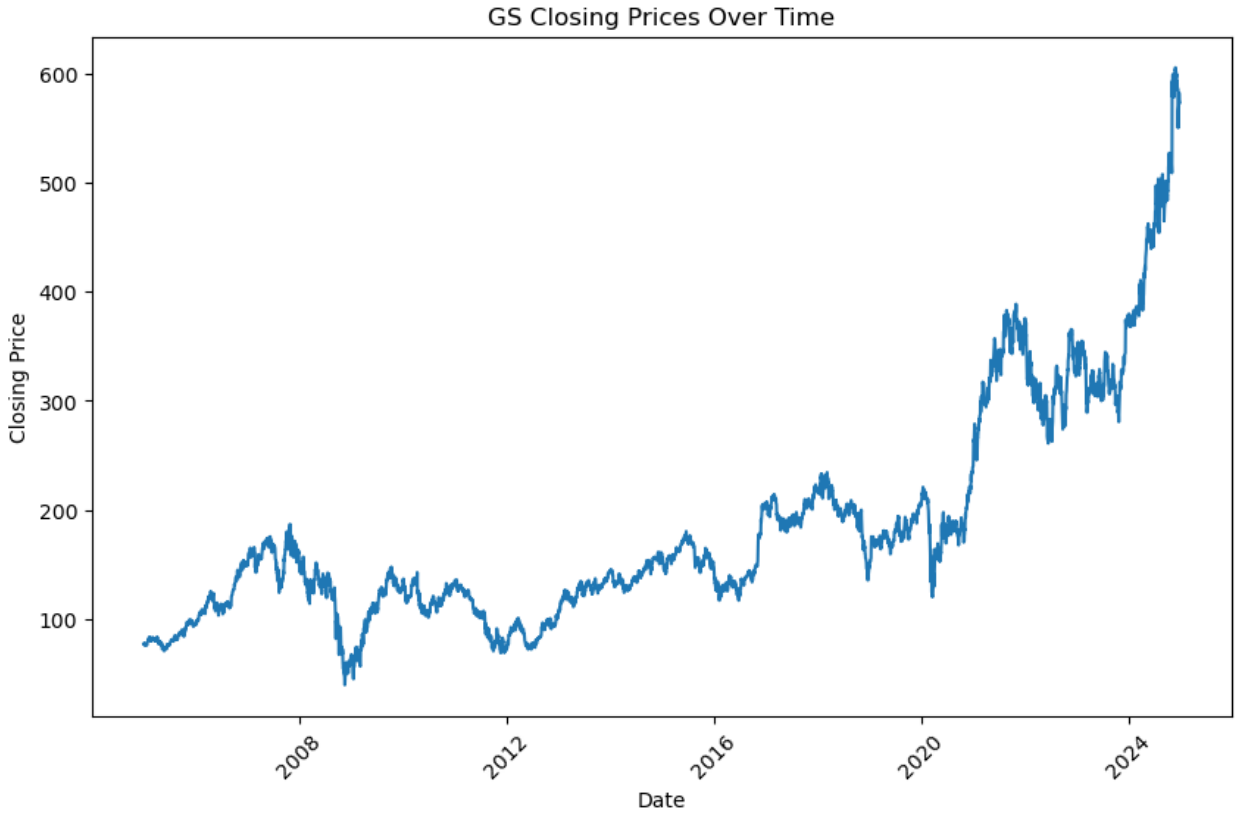

JPM Closing Prices Over Time
Closing Price
0
50
100
150
200
250
2008
2012
2016
2020
2024
Date


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
450
400
350
300
250
200
150
100
50
2008
2012
2016
2020
2024
Date


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

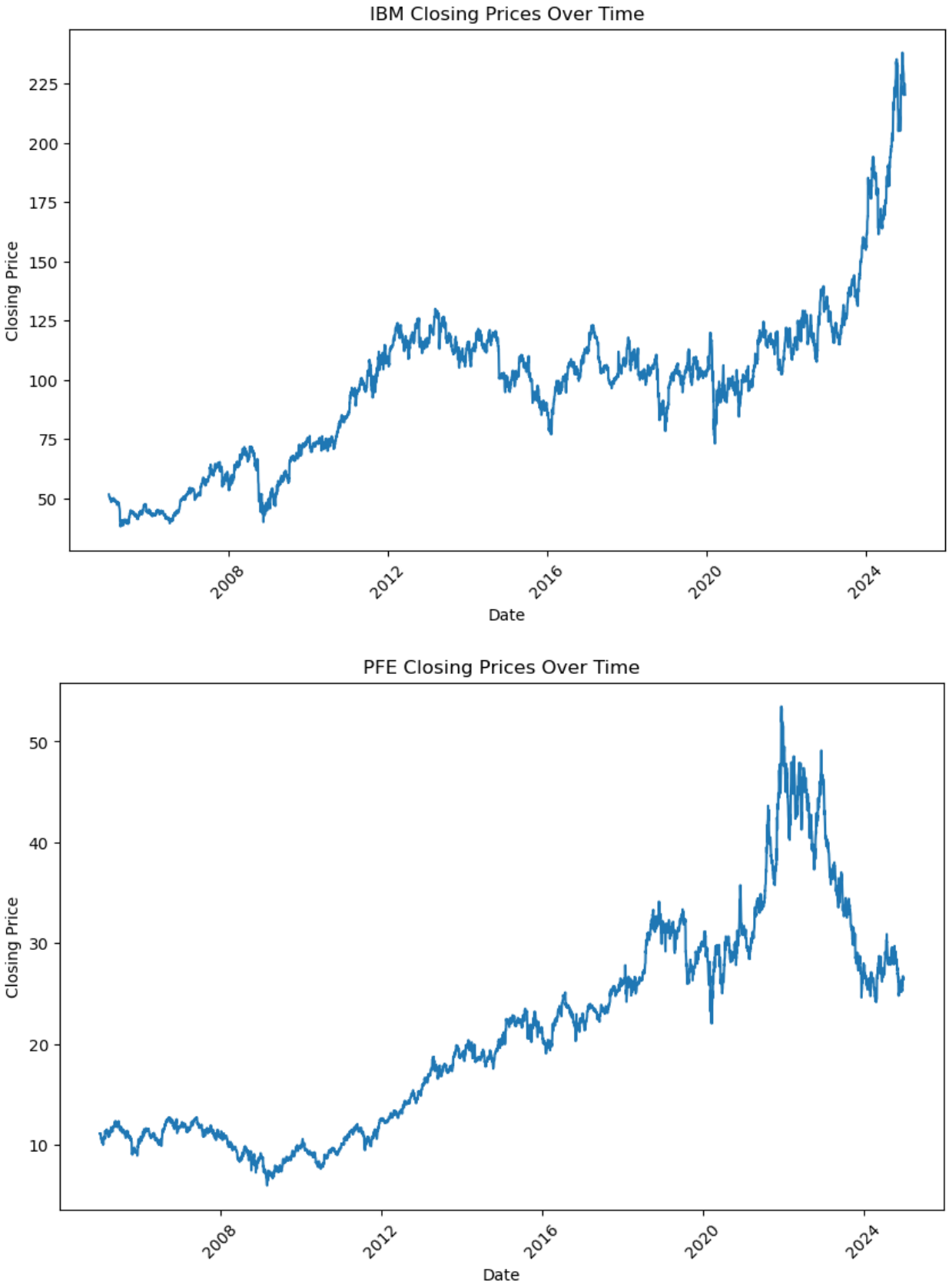
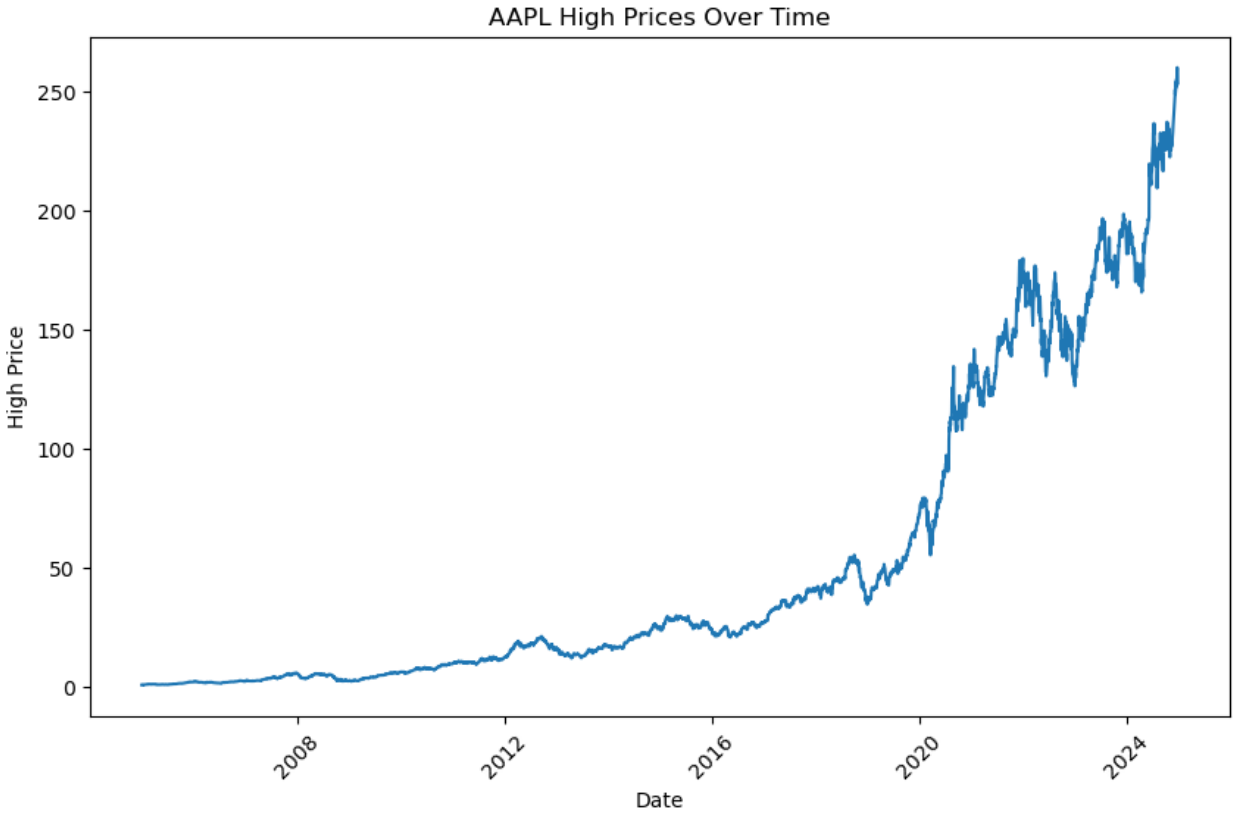
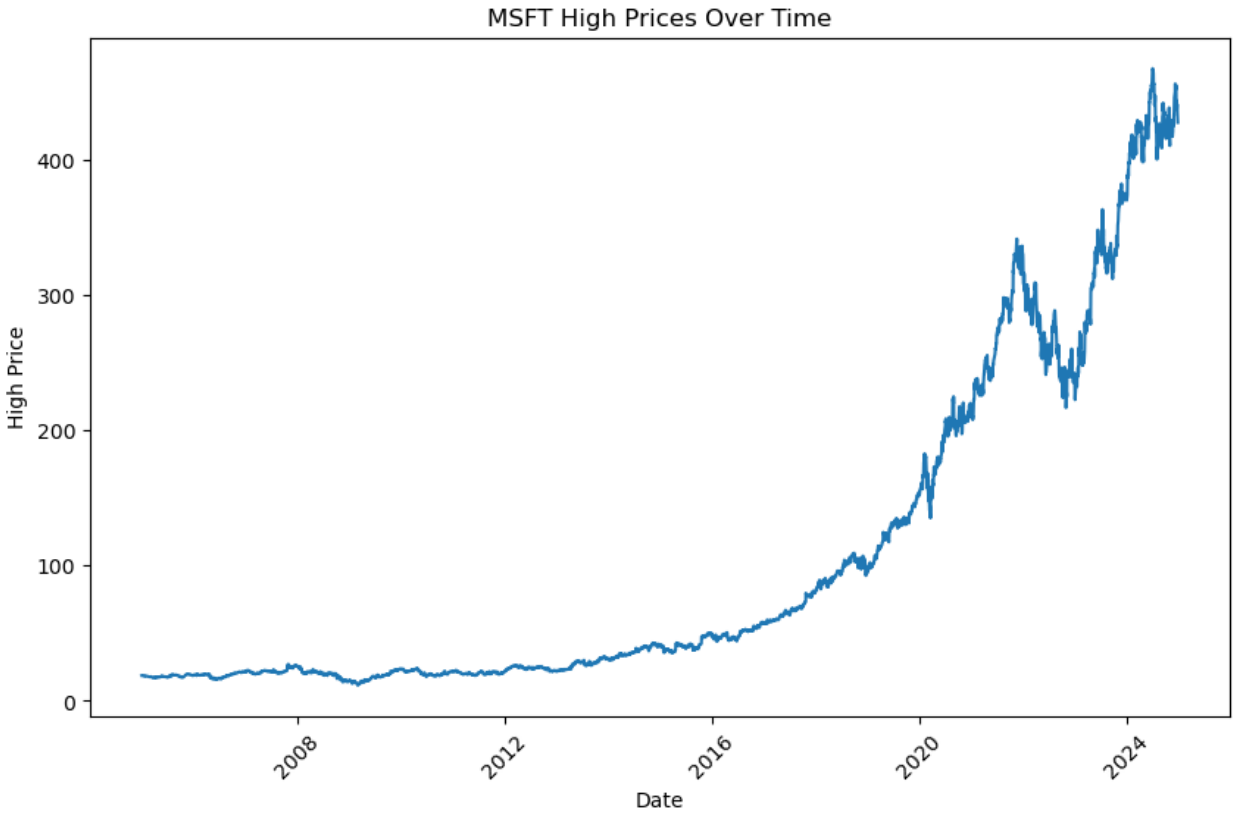


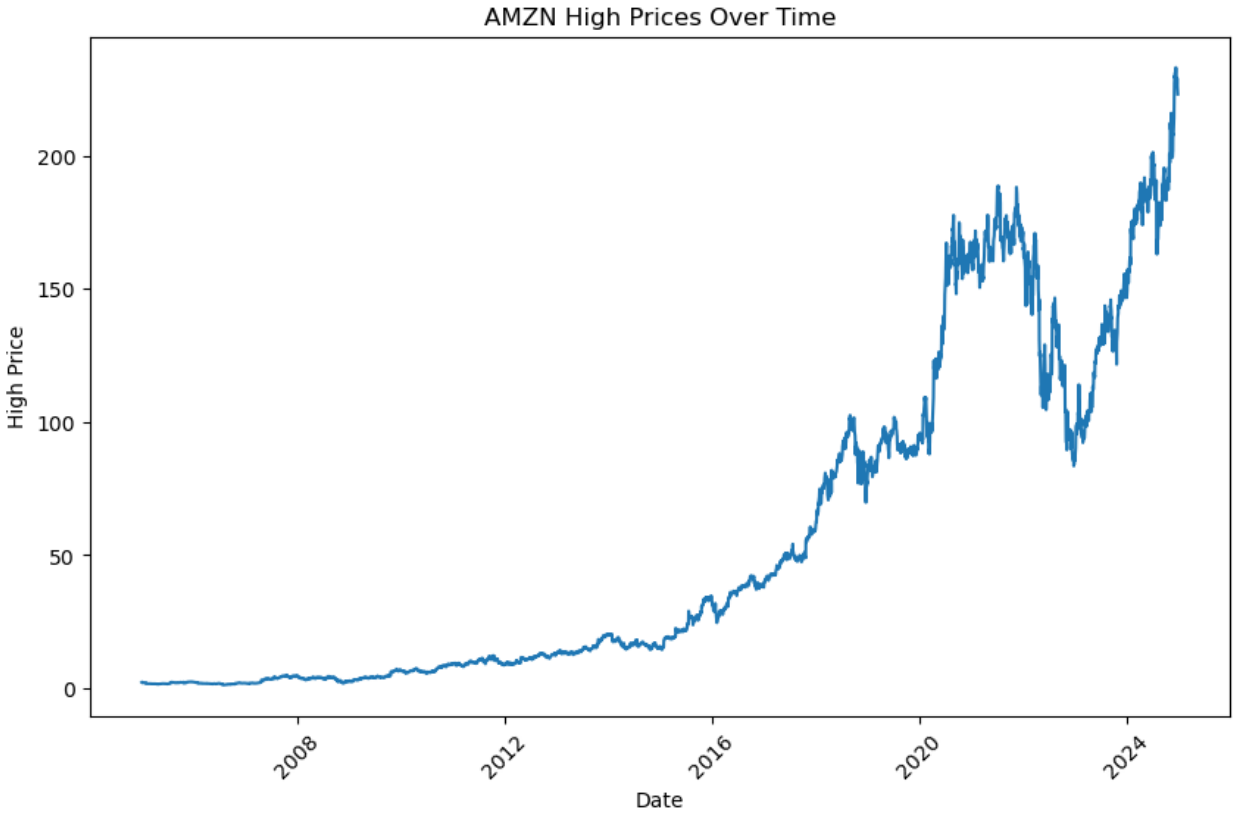
Figure 4.2: Line plots for closing prices of 15 companies

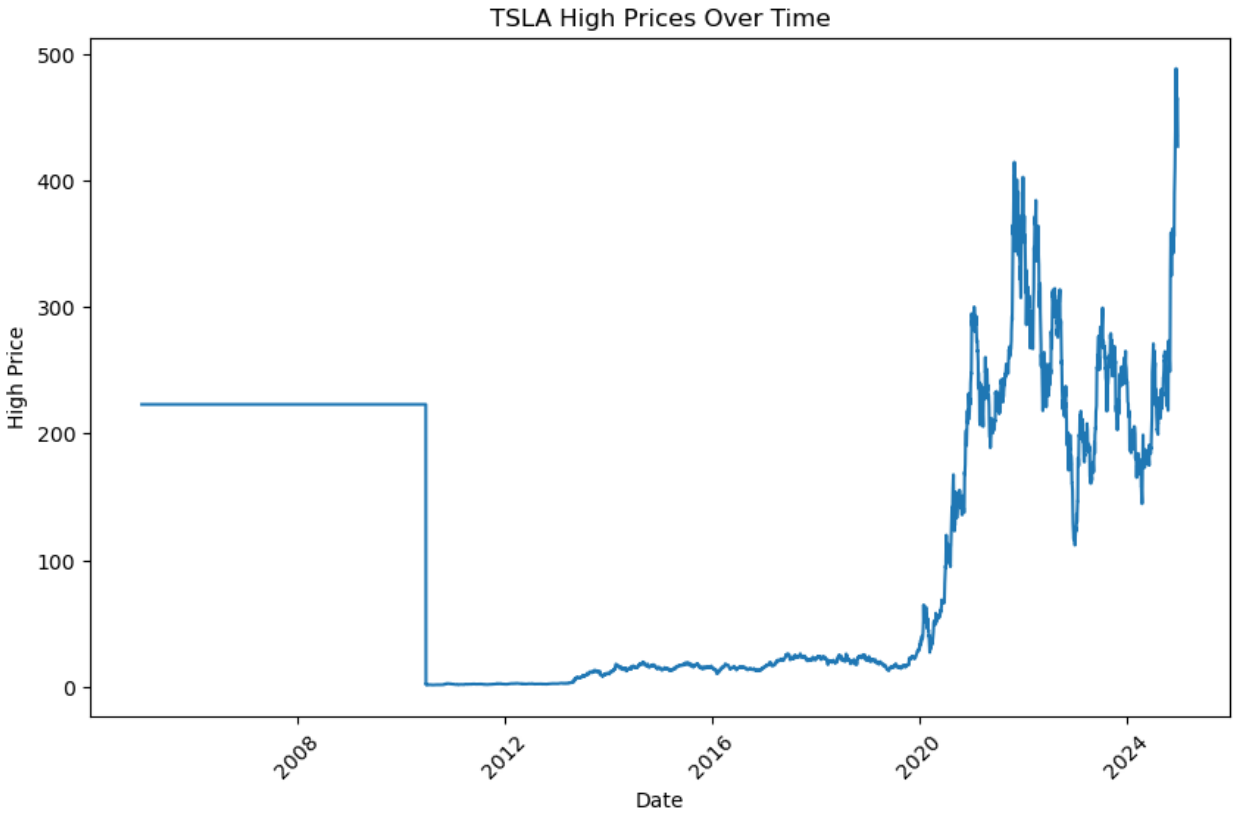
c. Line Plots for High Prices:

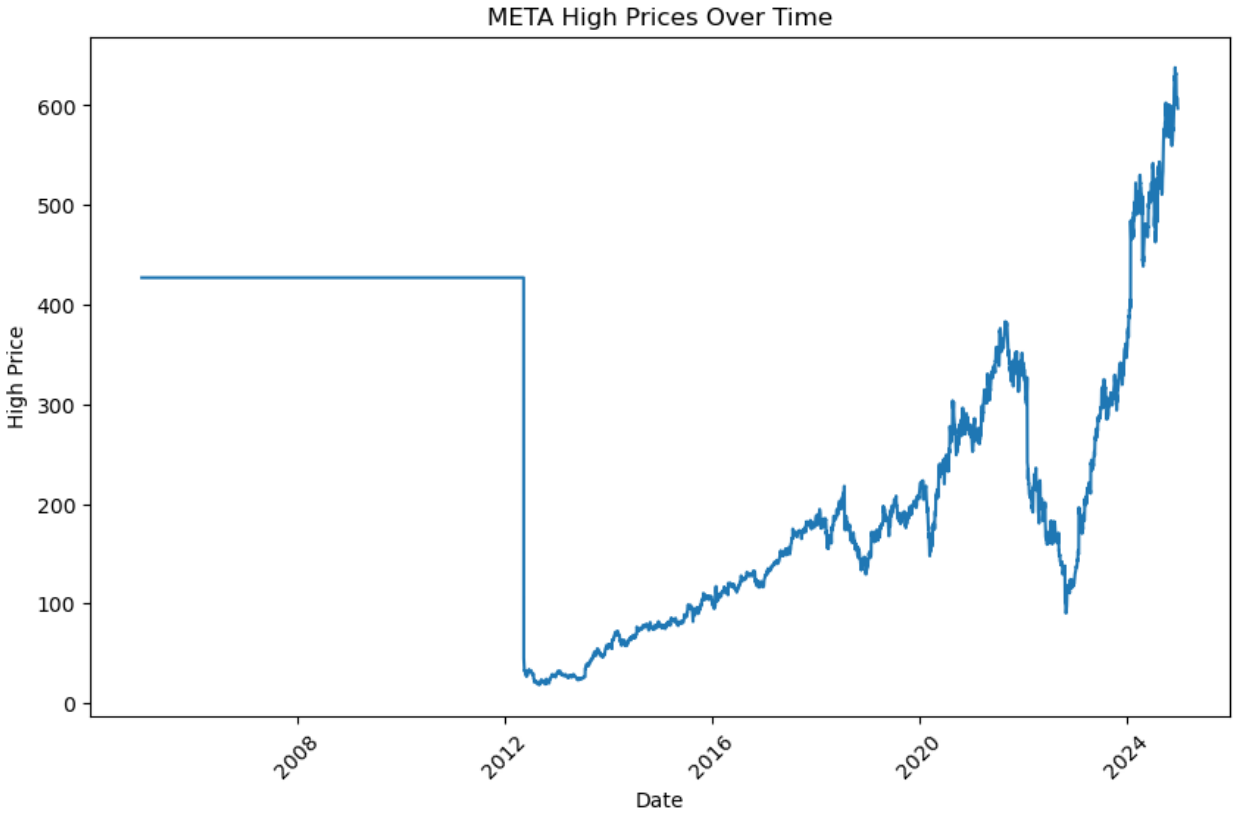
The objective of this line plot is to portray the High price movements of all firms across a span of time. The data depict the changes in the high prices over the years for each firm. Trends can tell how the company's stock performed based on years, thus coming with periods of growth, decline, or stability.

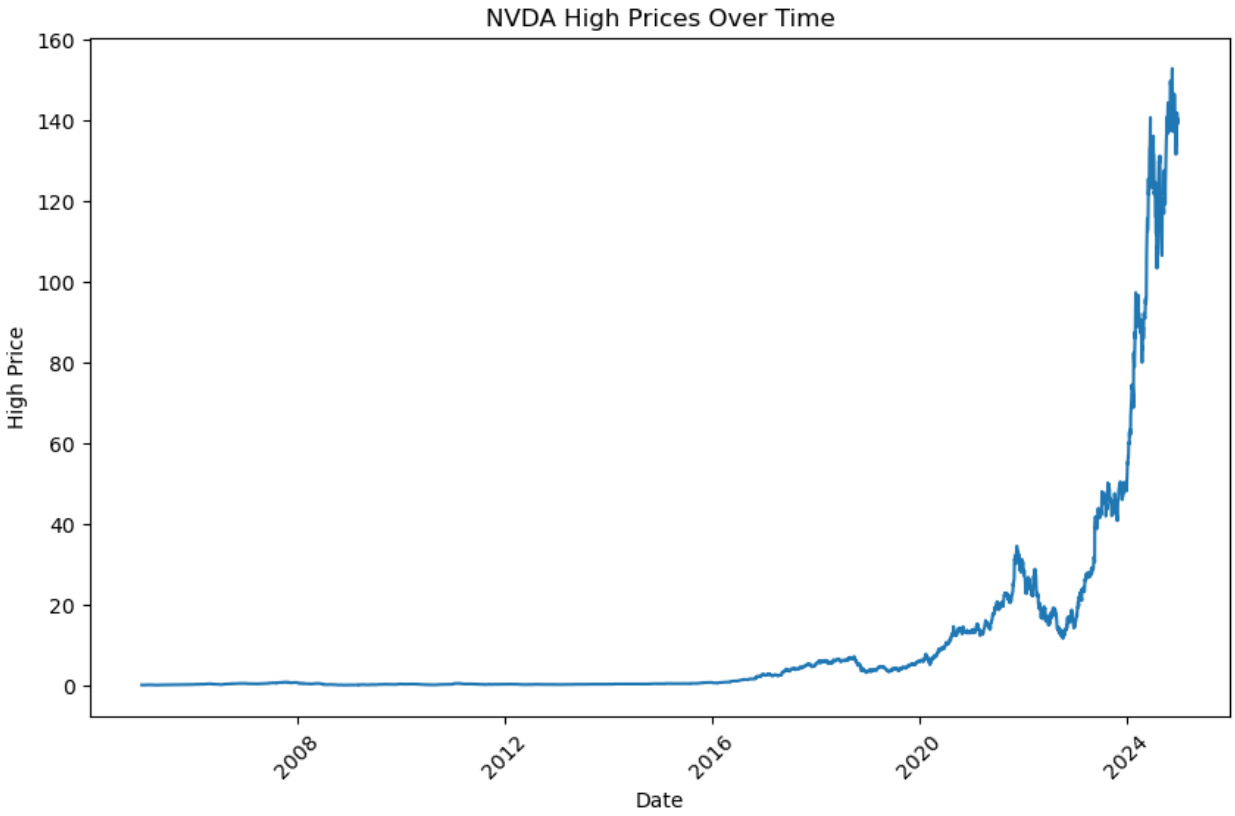


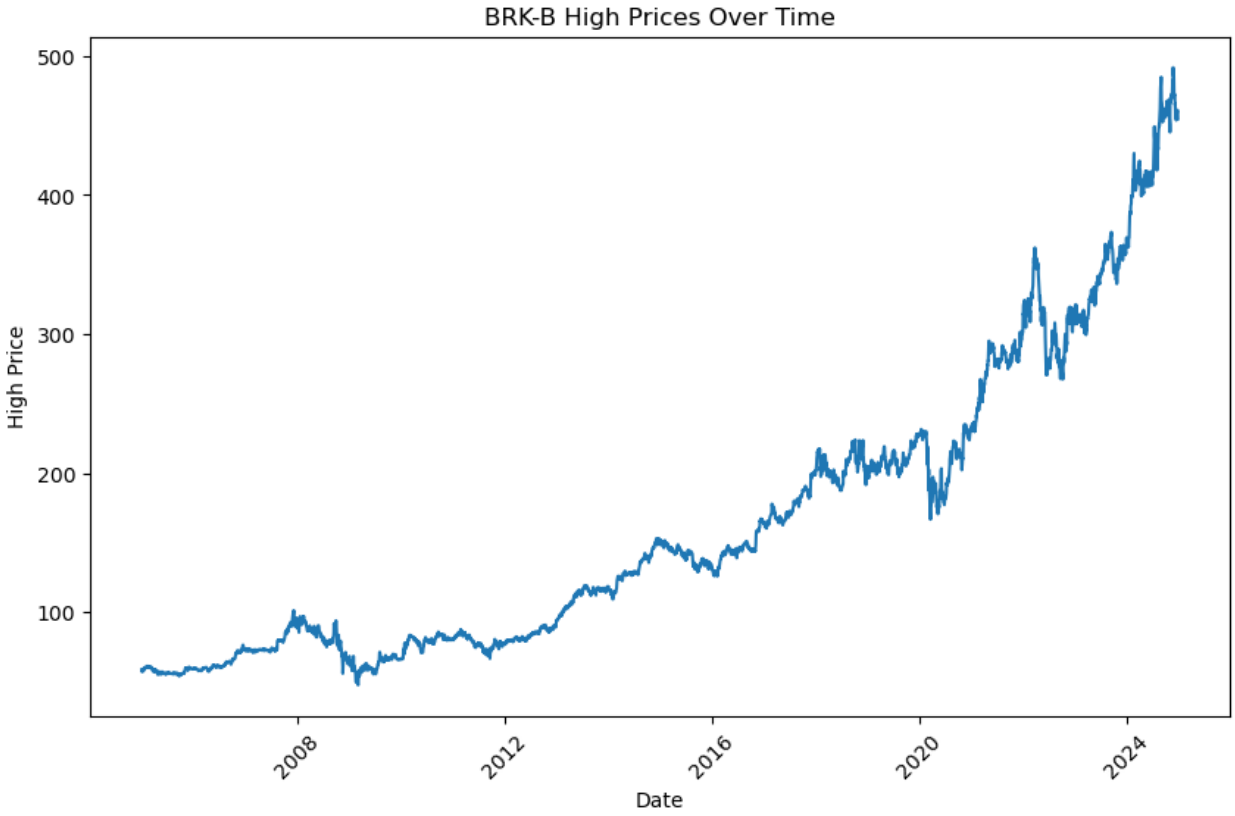

MSFT High Prices Over Time
High Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

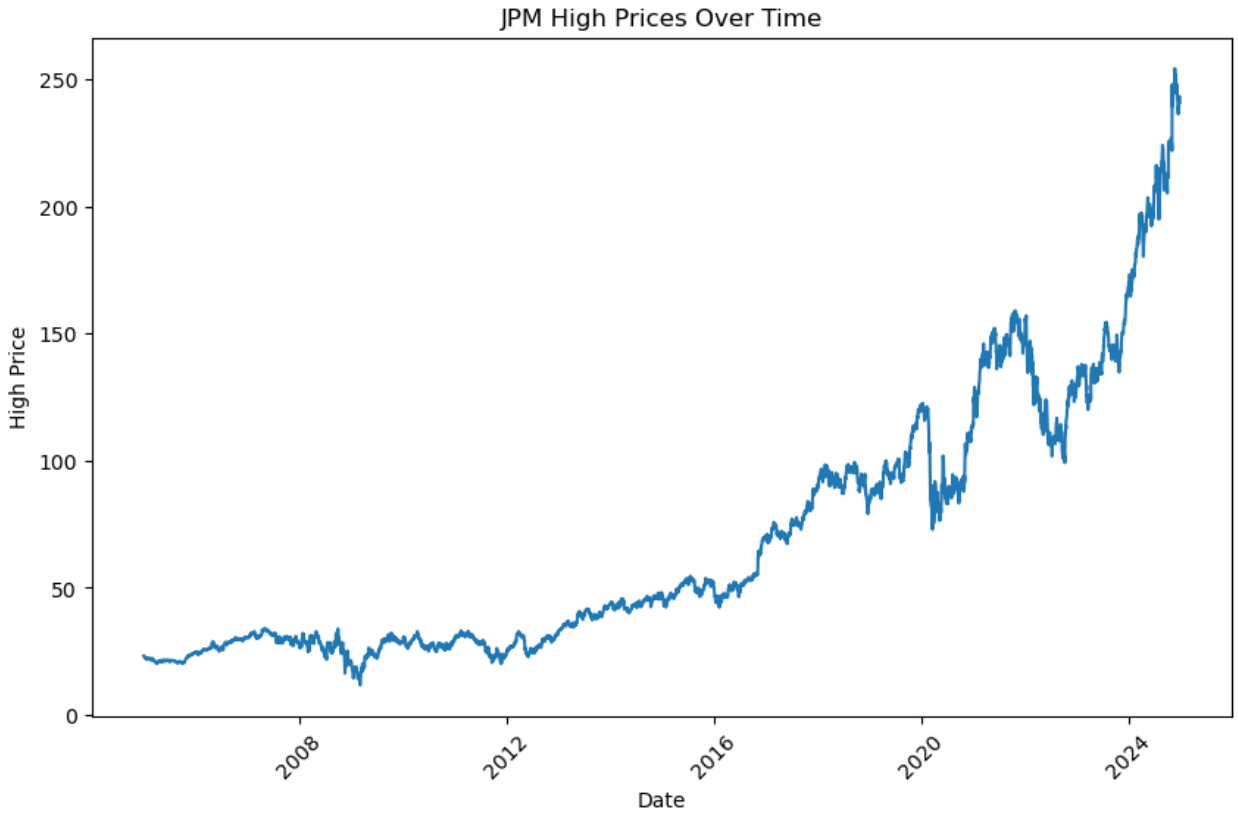

AMZN High Prices Over Time
High Price
0
50
100
150
200
2008
2012
2016
2020
2024
Date

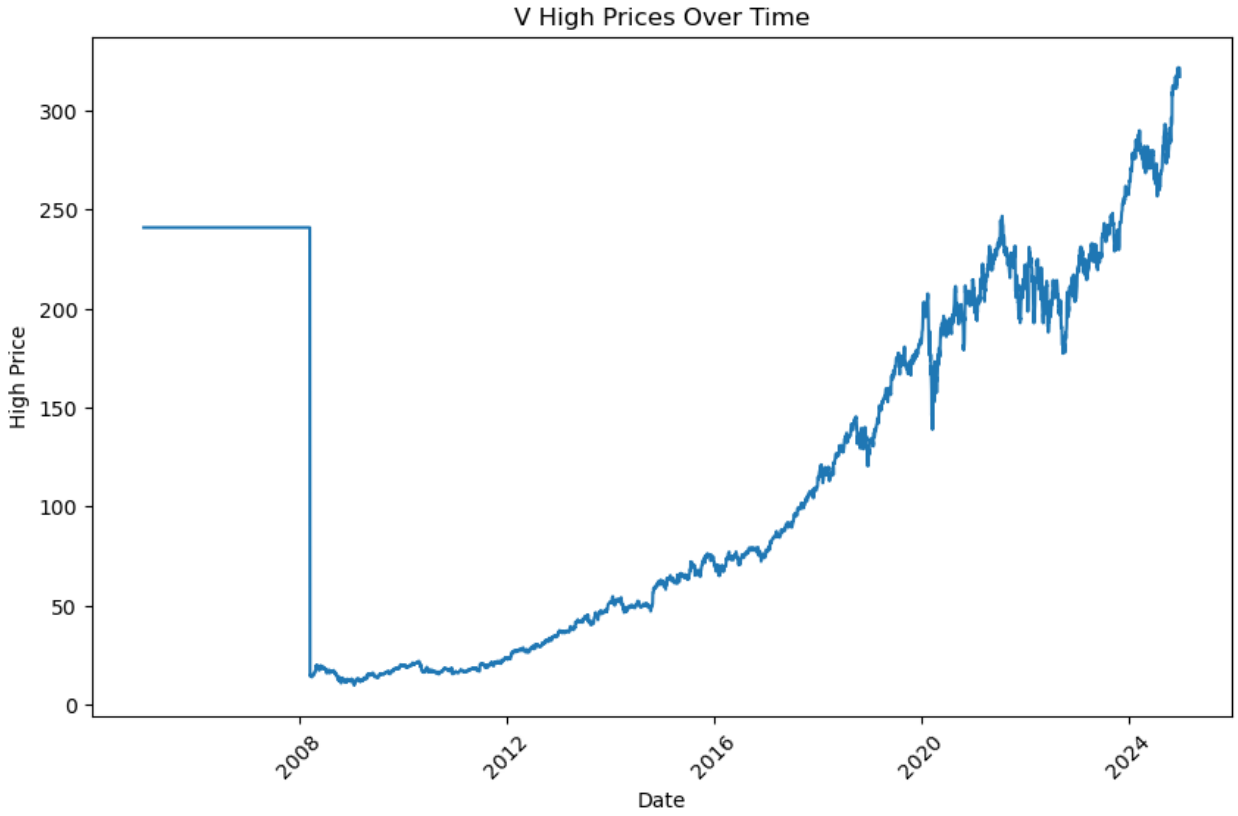

TSLA High Prices Over Time
High Price
500
400
300
200
100
0
2008
2012
2016
2020
2024
Date

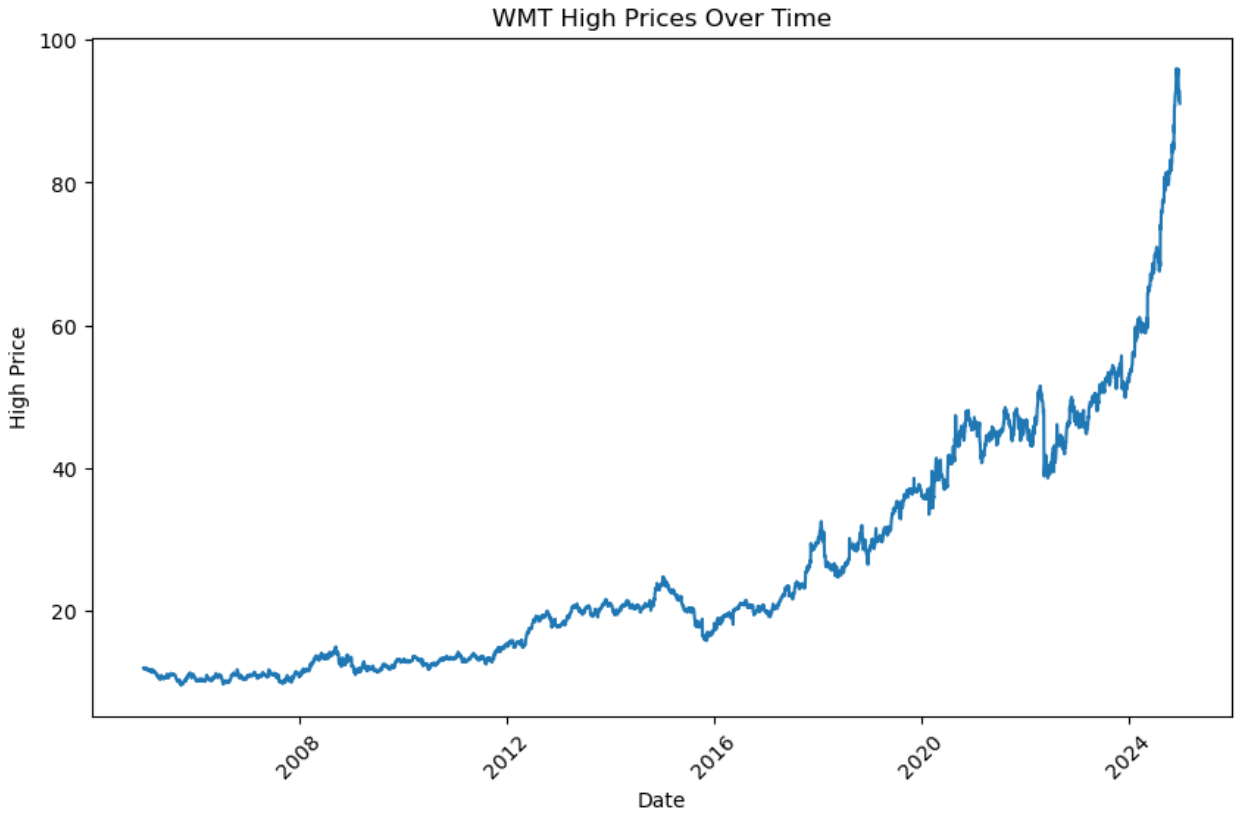

META High Prices Over Time
High Price
600
500
400
300
200
100
0
2008
2012
2016
2020
2024
Date

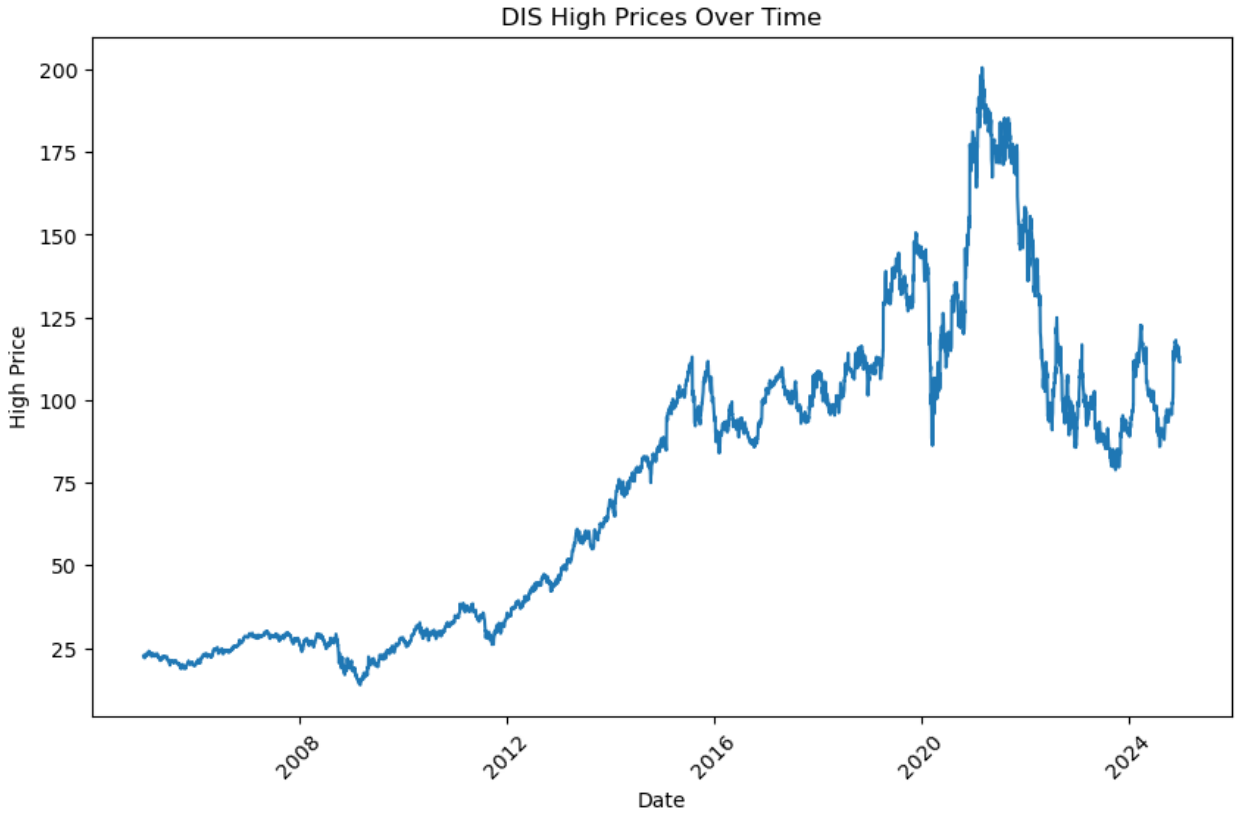

NVDA High Prices Over Time
High Price
Date
0
20
40
60
80
100
120
140
160
2008
2012
2016
2020
2024

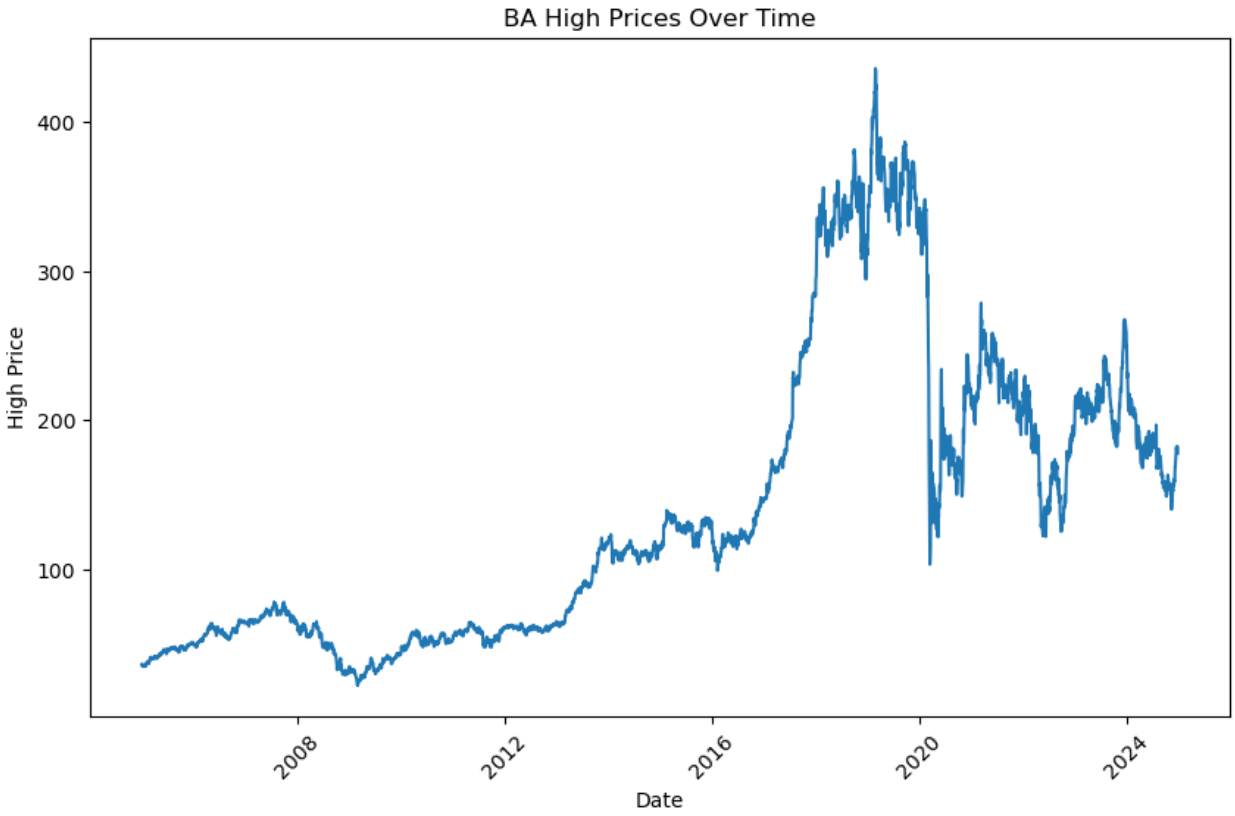

BRK-B High Prices Over Time
High Price
Date
100
200
300
400
500
2008
2012
2016
2020
2024

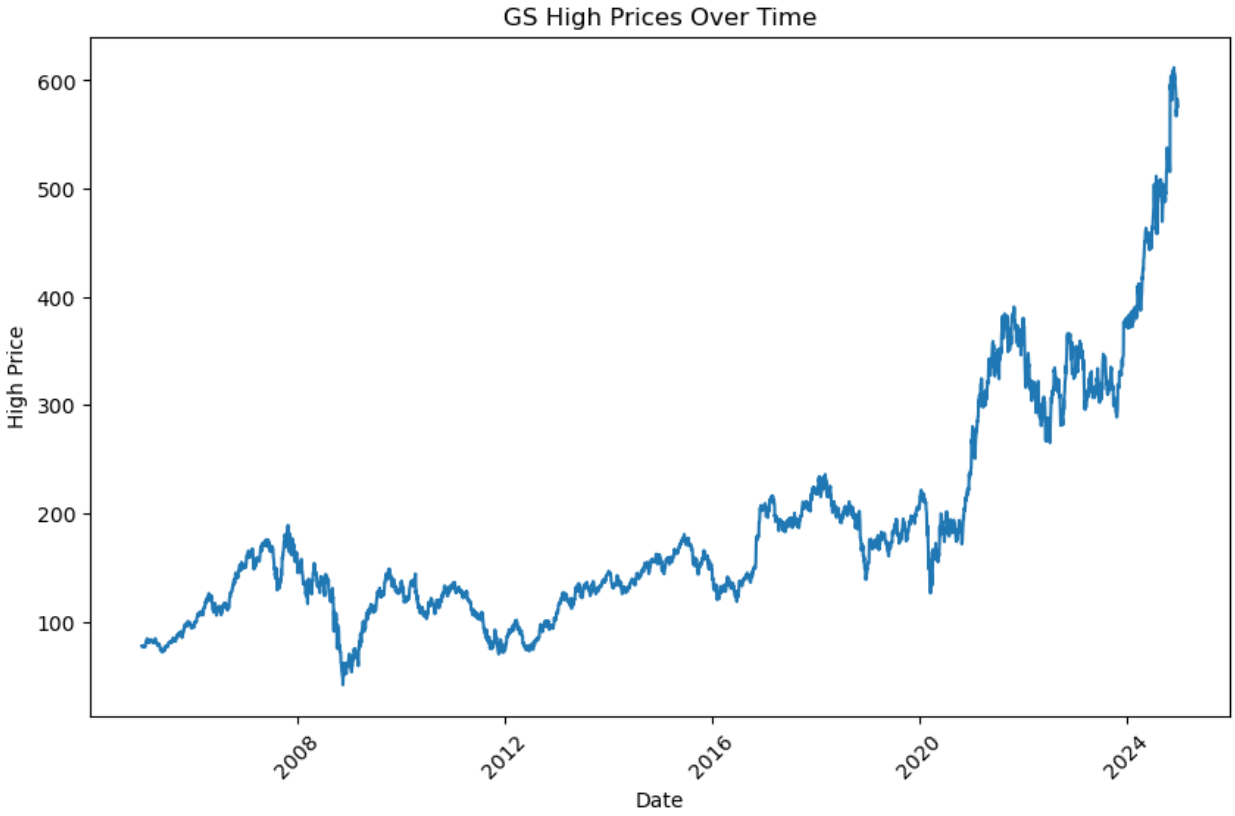

JPM High Prices Over Time
High Price
0
50
100
150
200
250
2008
2012
2016
2020
2024
Date


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
400
300
200
100
2008
2012
2016
2020
2024
Date


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

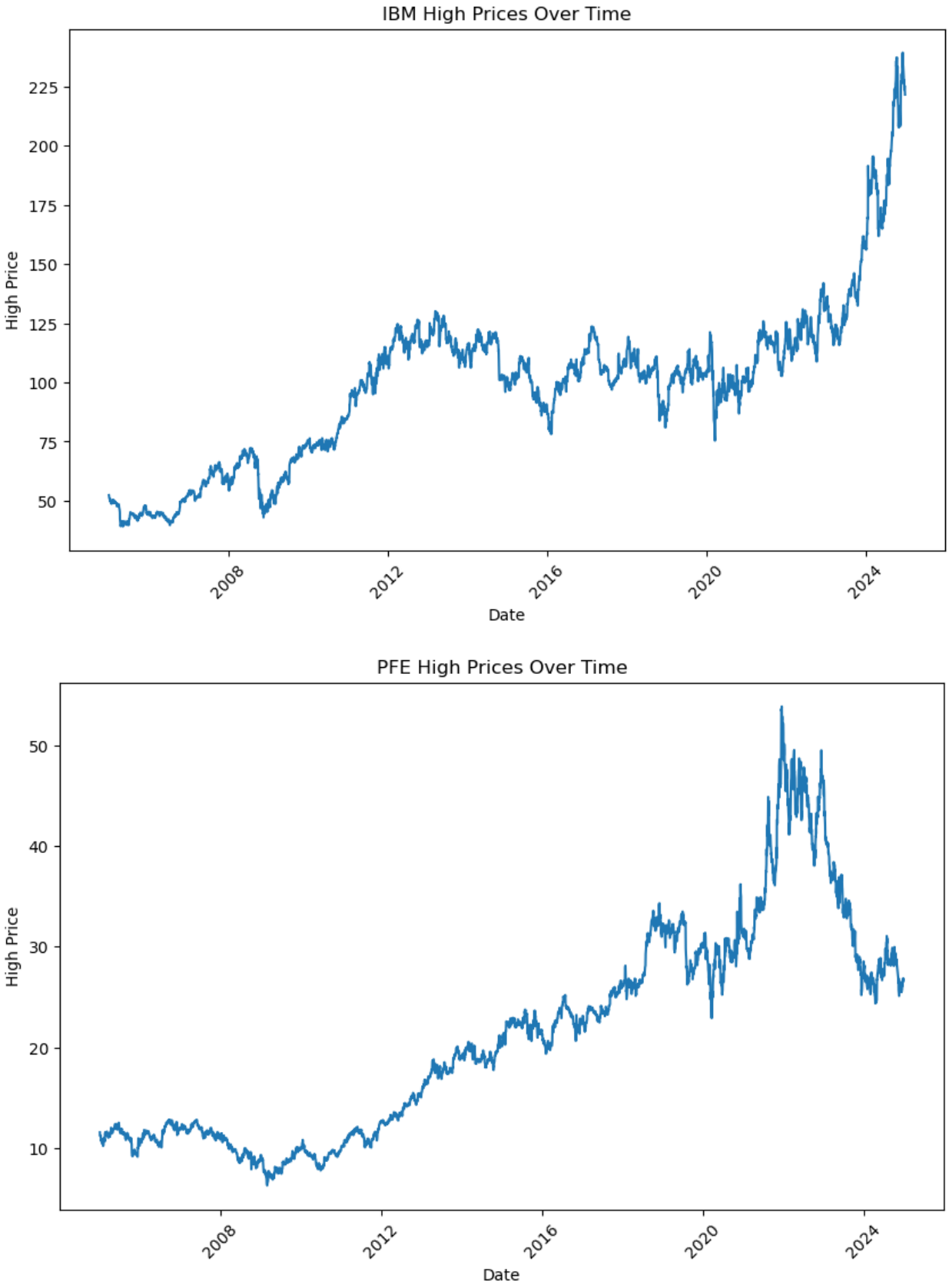
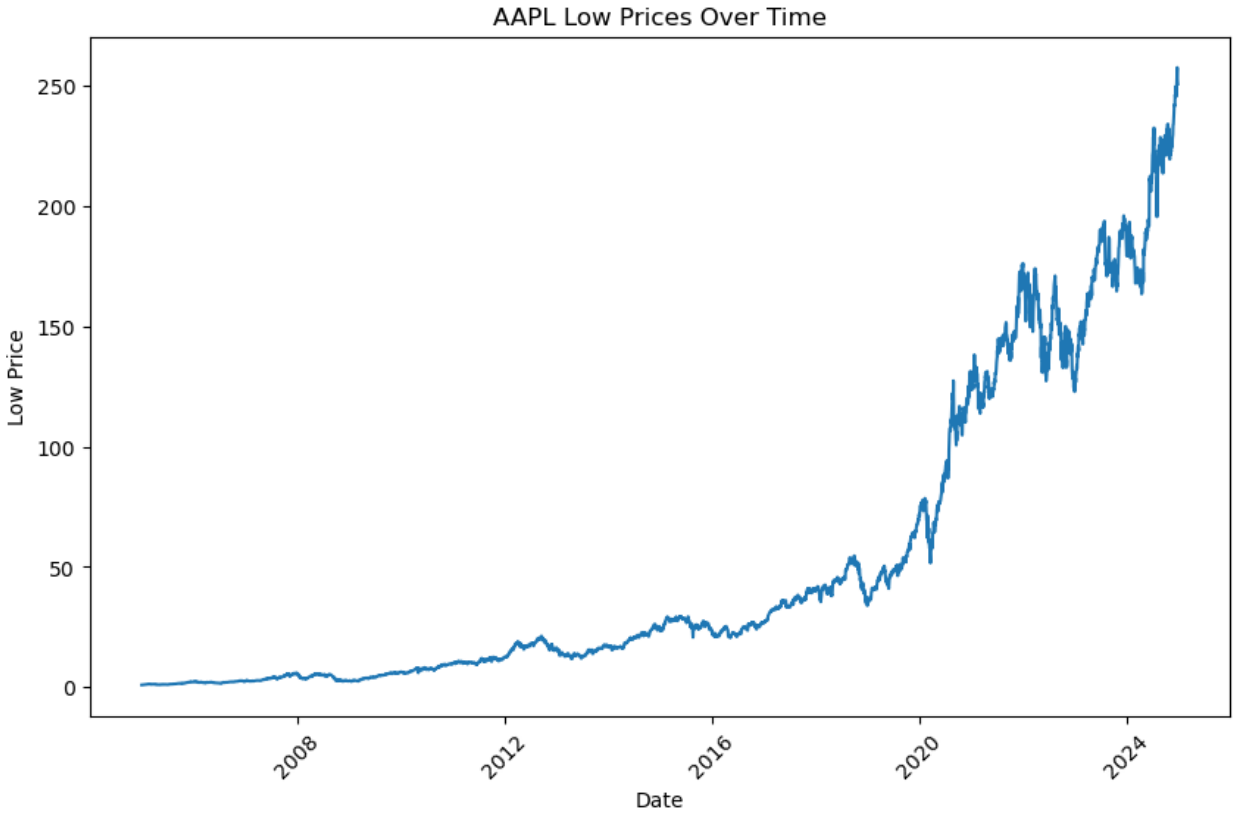
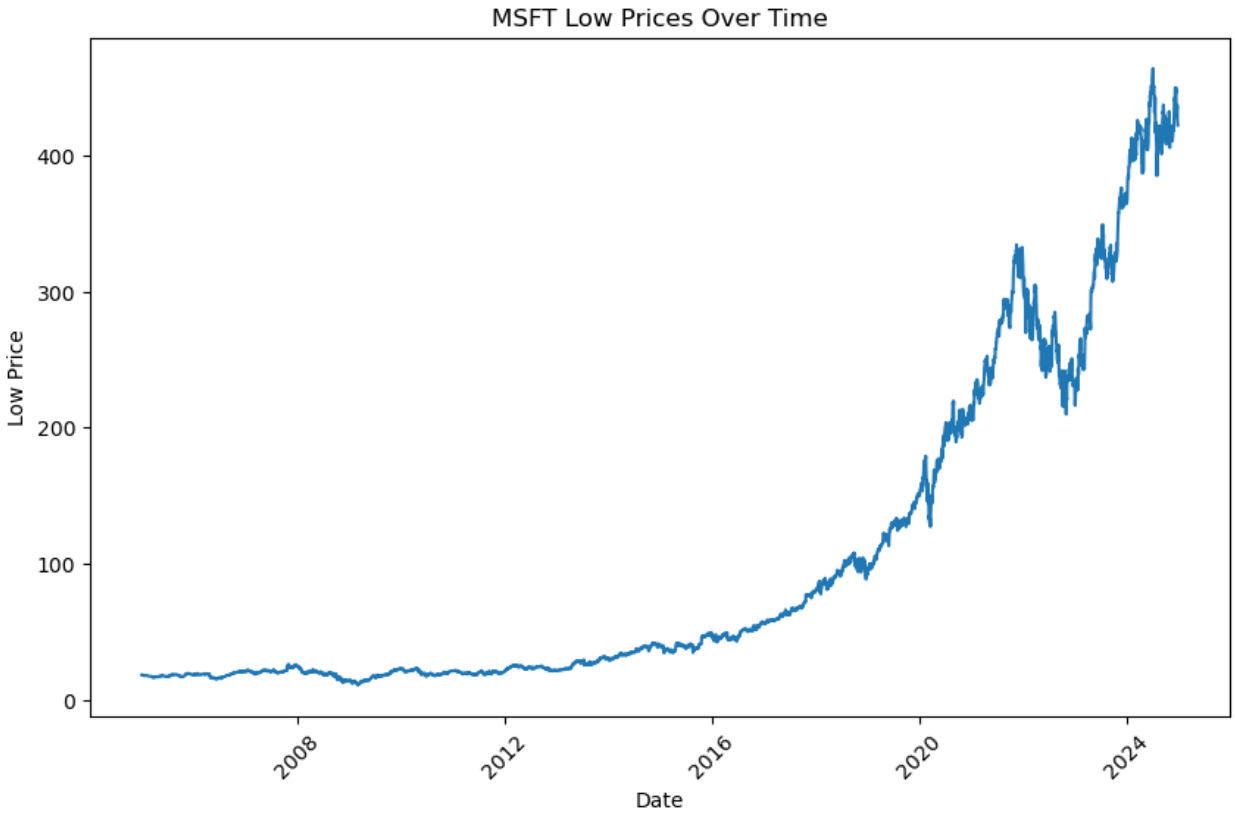


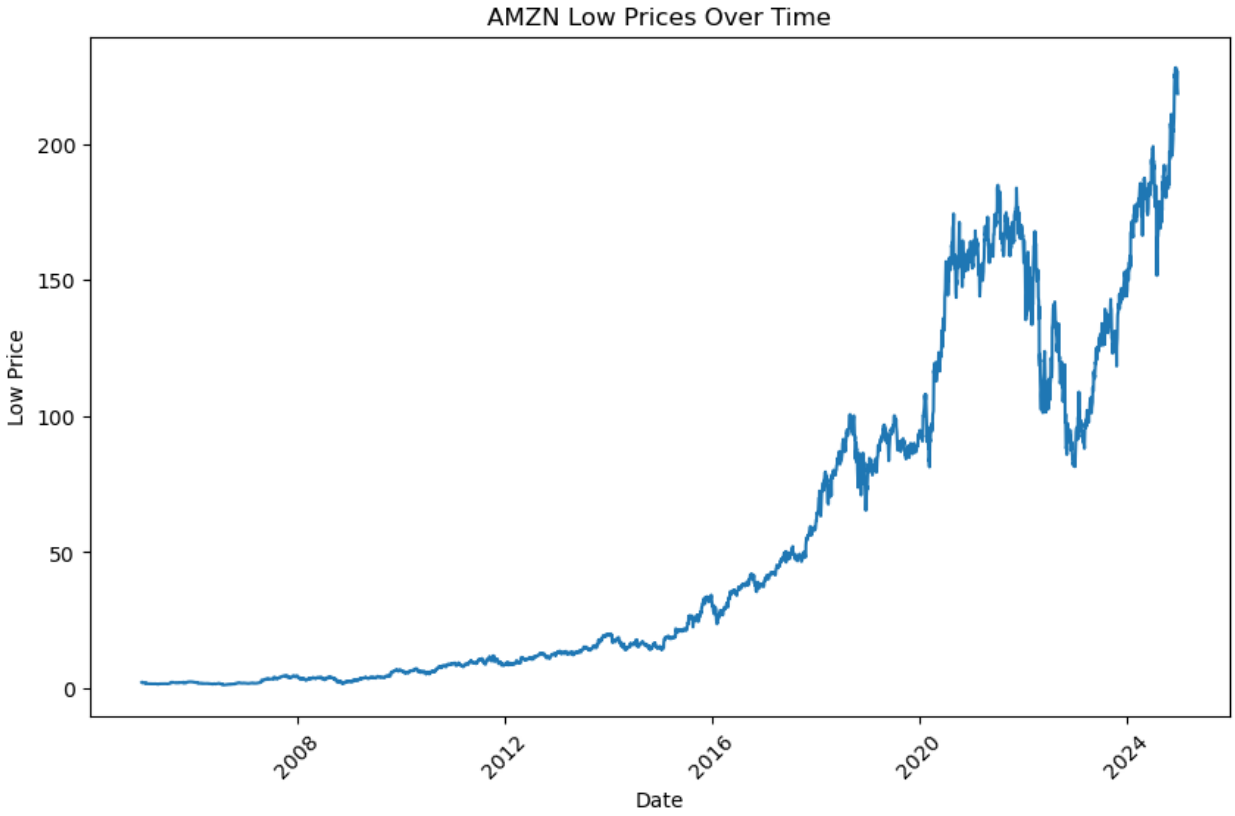
Figure 4.3: Line plots for high prices of 15 companies

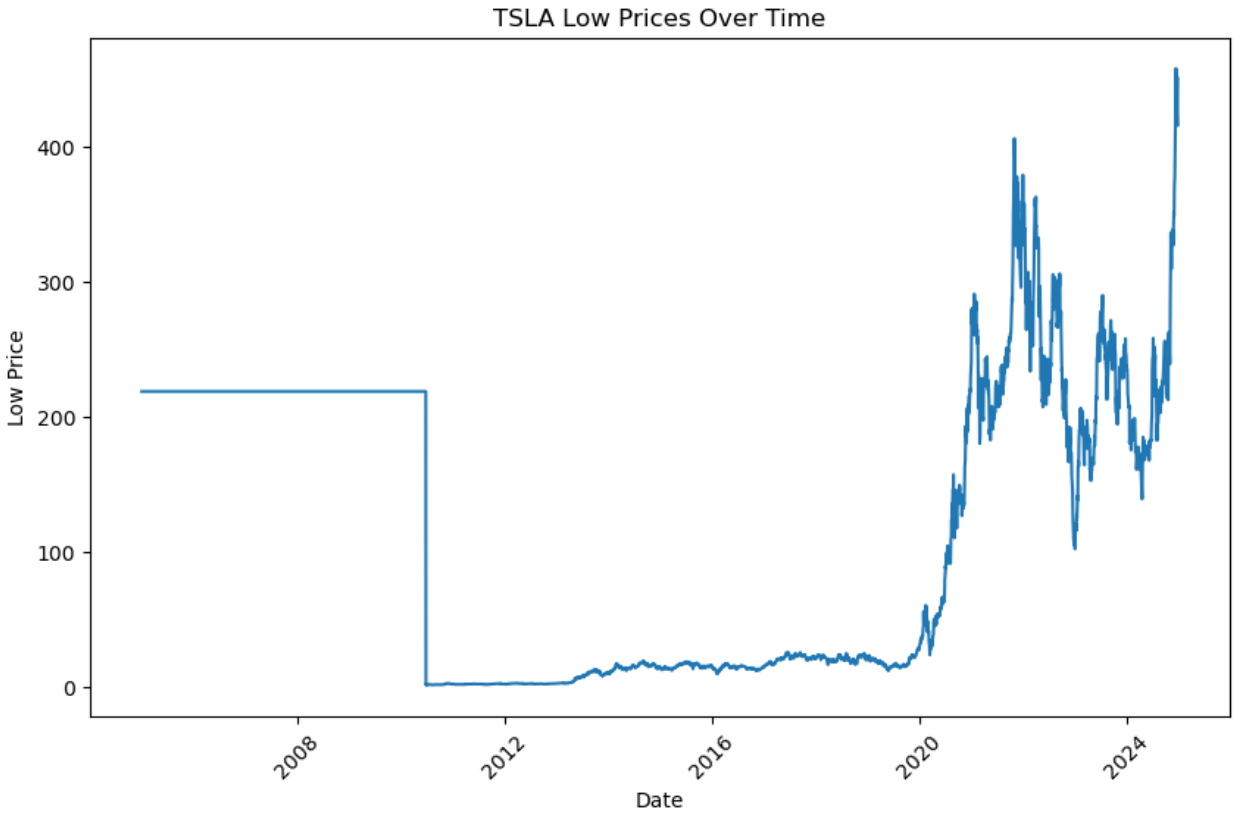
d. Line Plots for Low Prices:

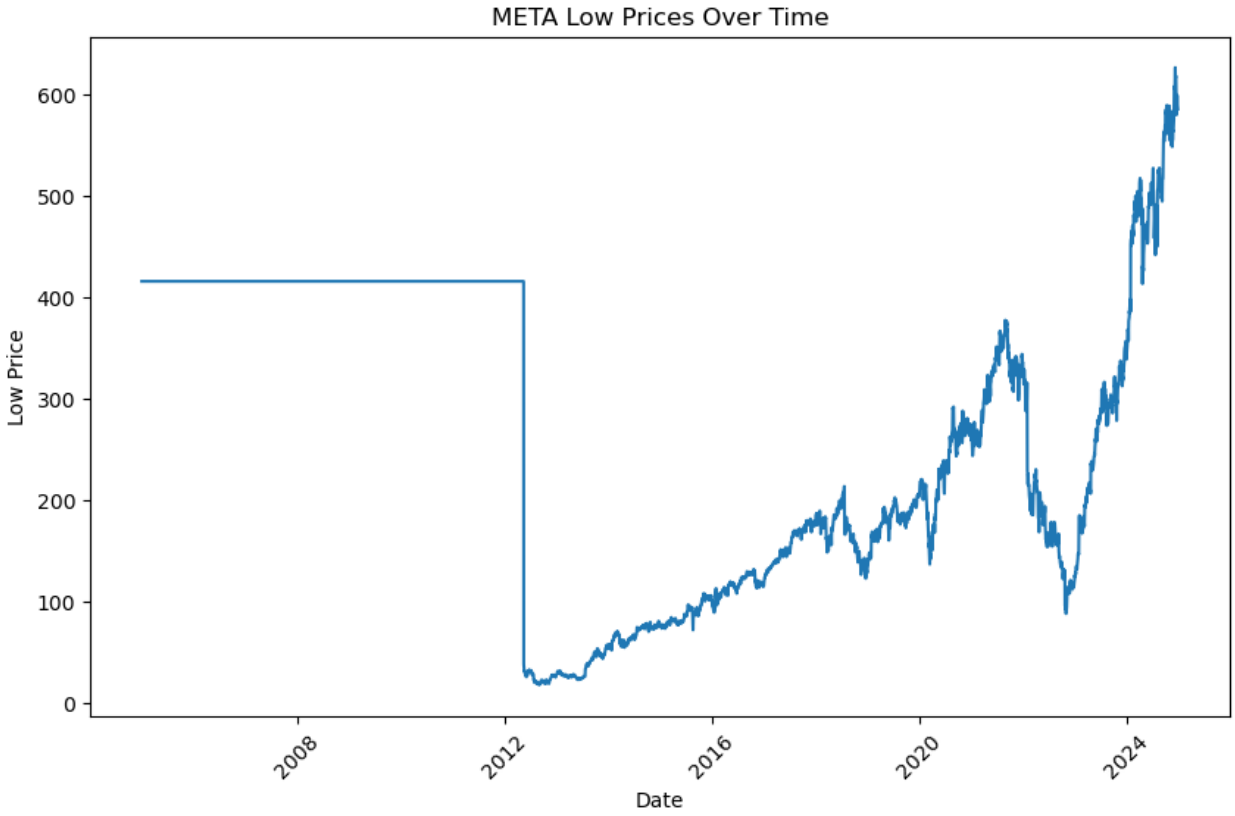
The objective of this line plot is to portray the low price movements of all firms across a span of time. The data depict the changes in the low prices over the years for each firm. Trends can tell how the company's stock performed based on years, thus coming with periods of growth, decline, or stability.

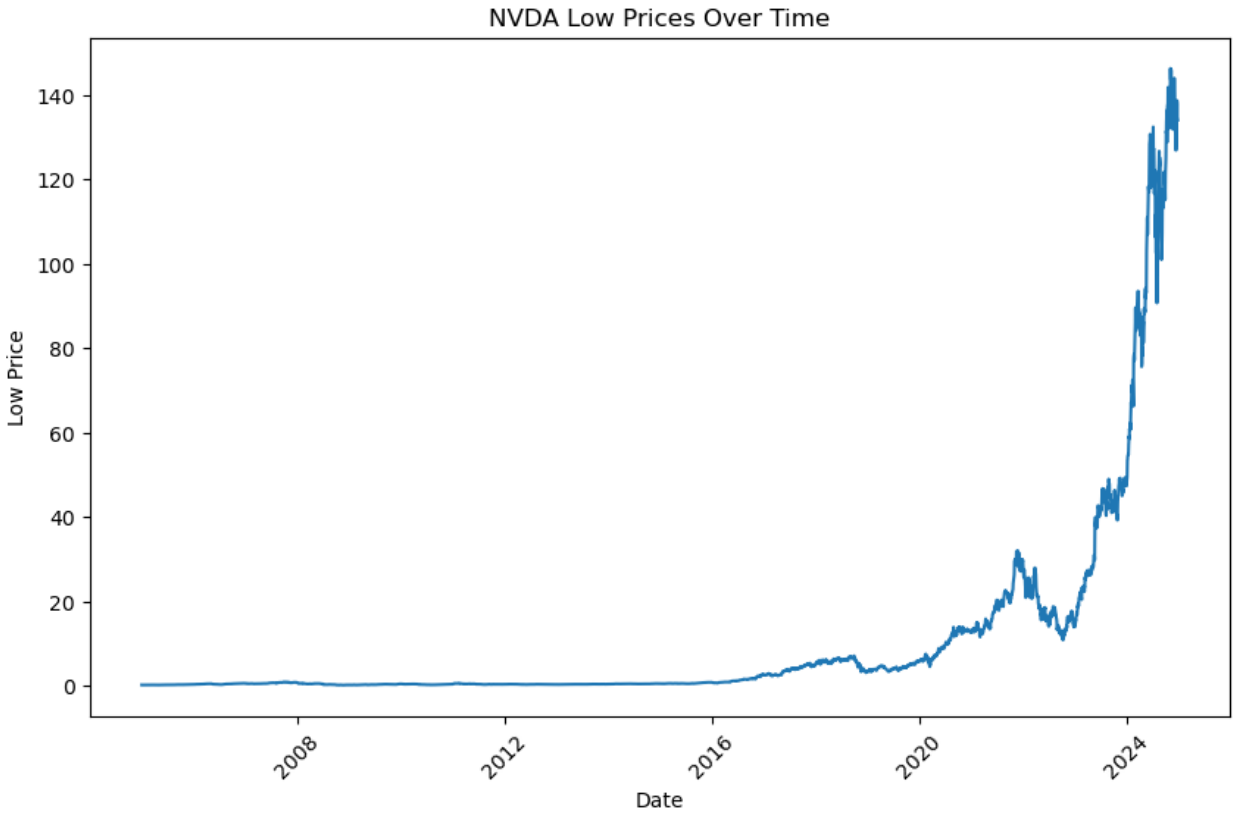


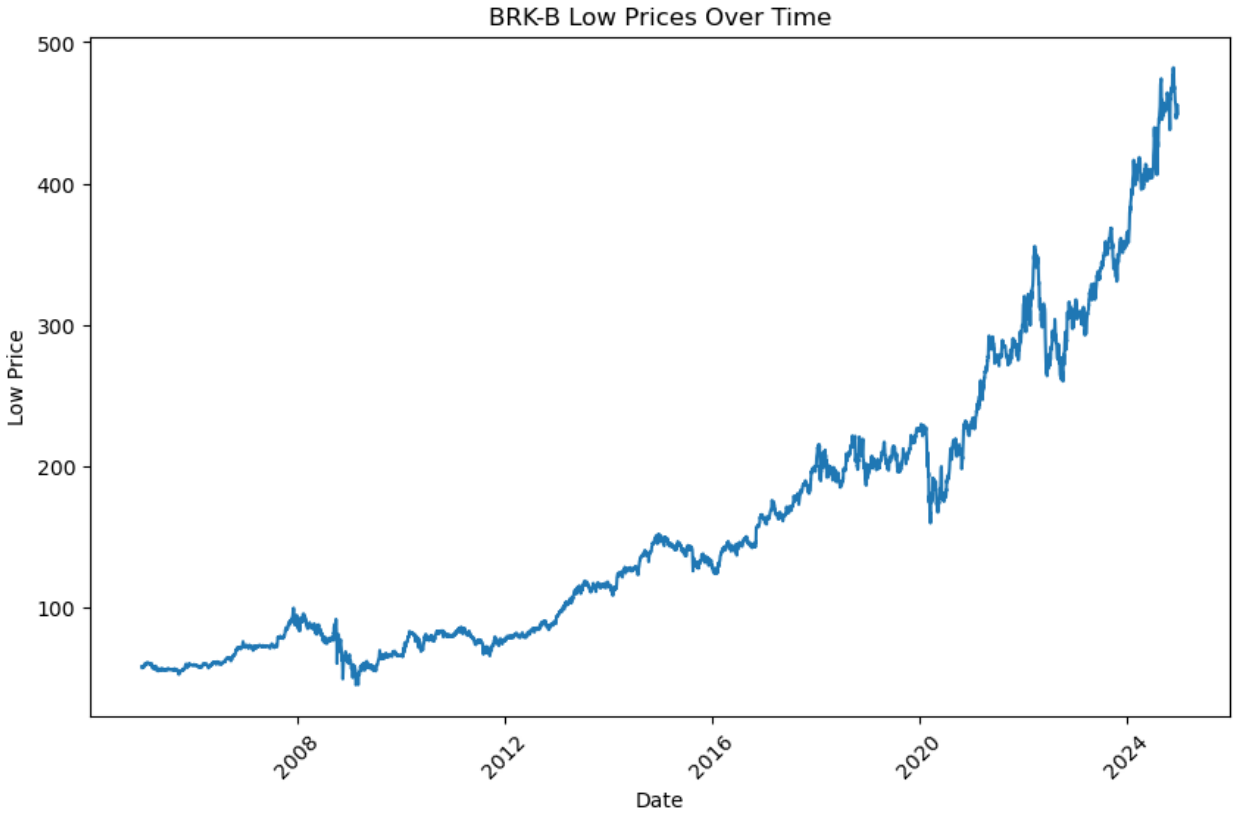

MSFT Low Prices Over Time
Low Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

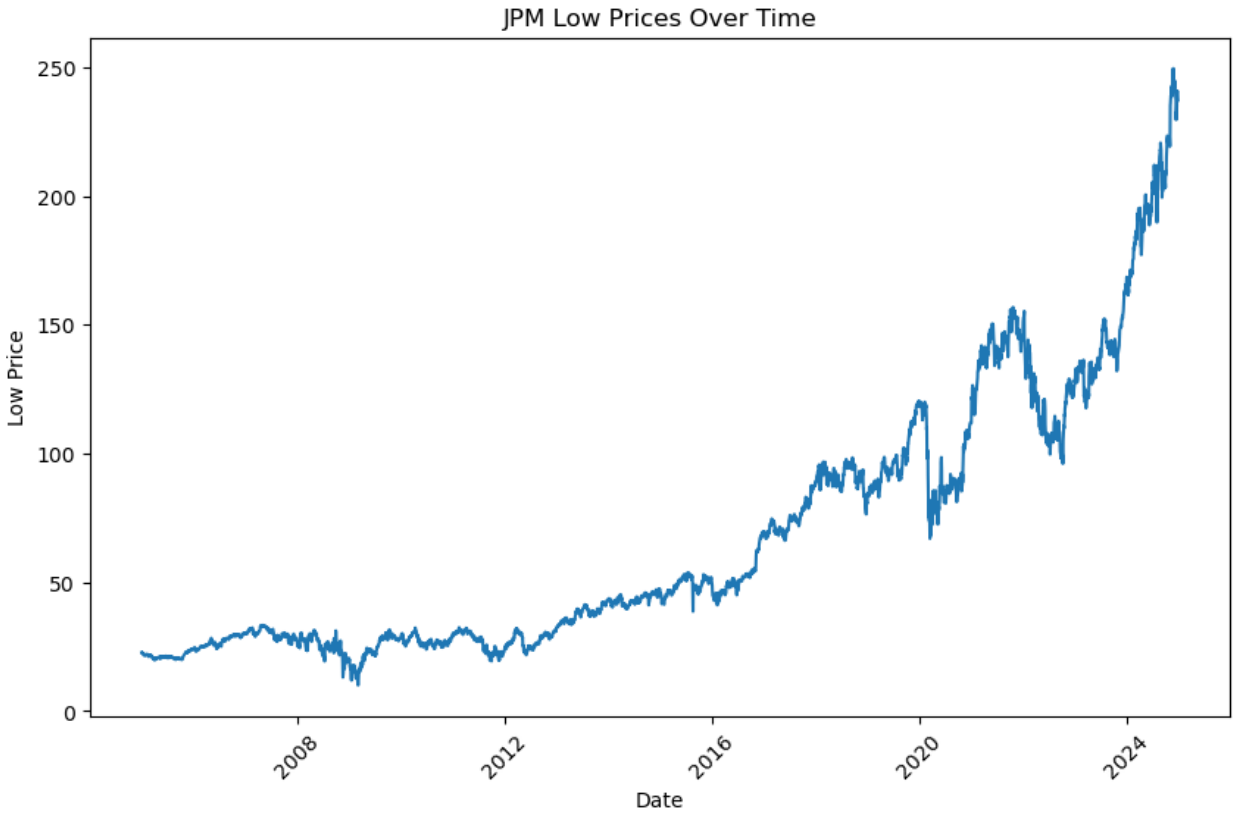

AMZN Low Prices Over Time
Low Price
0
50
100
150
200
2008
2012
2016
2020
2024
Date

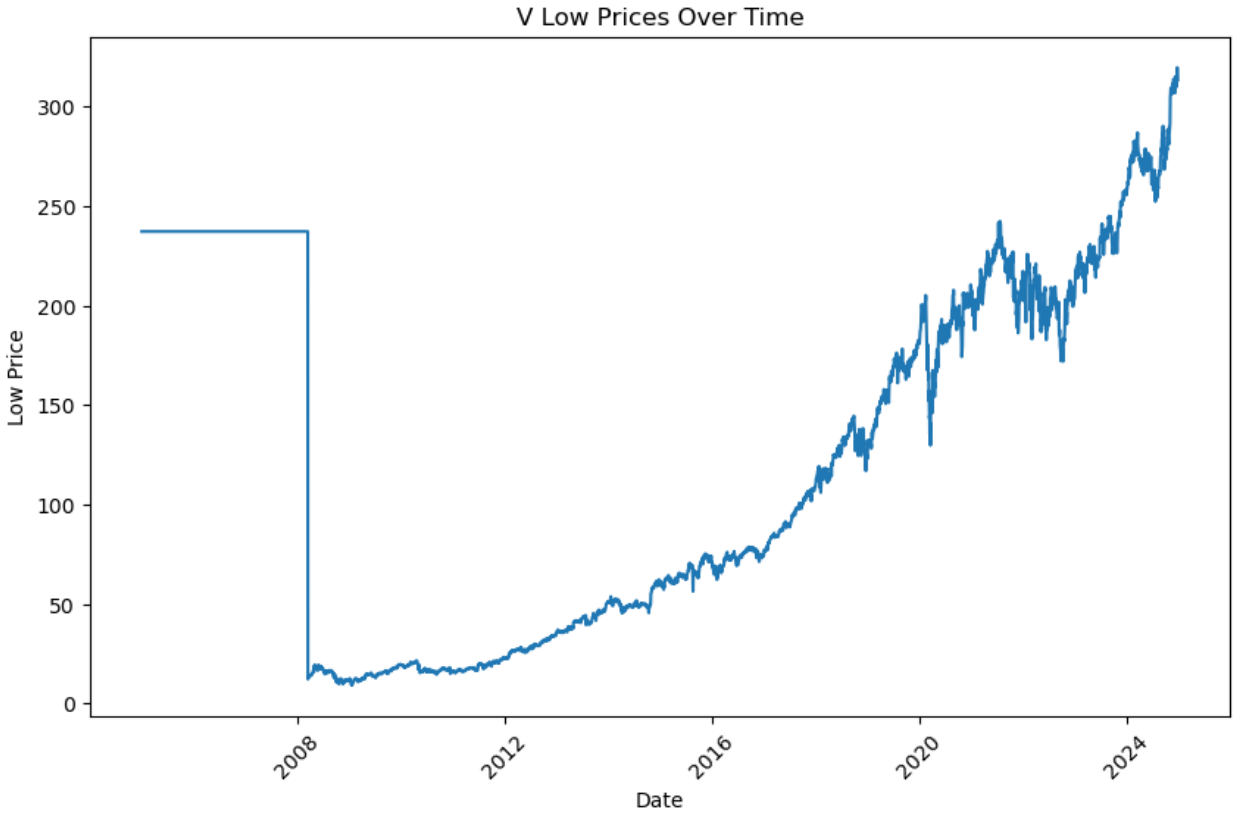

TSLA Low Prices Over Time
Low Price
0
100
200
300
400
2008
2012
2016
2020
2024
Date

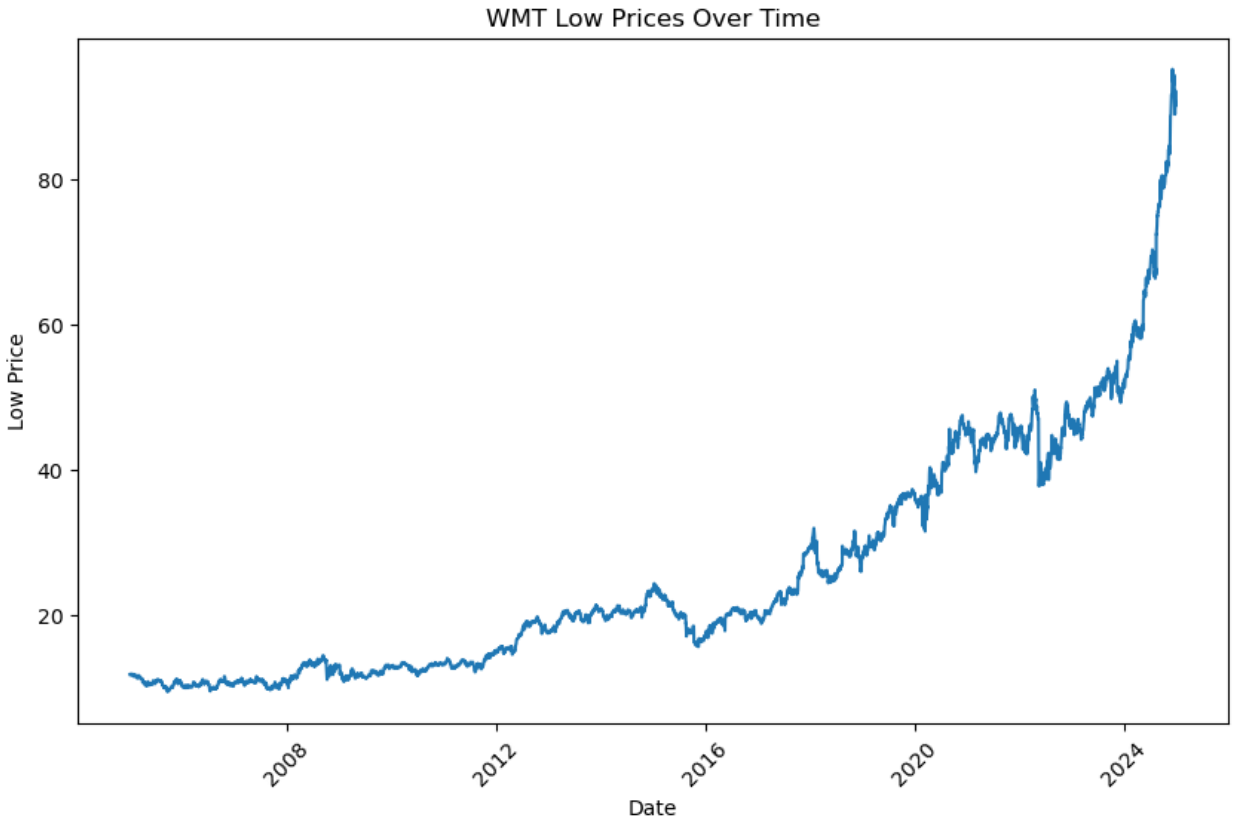

META Low Prices Over Time
Low Price
0
100
200
300
400
500
600
2008
2012
2016
2020
2024
Date

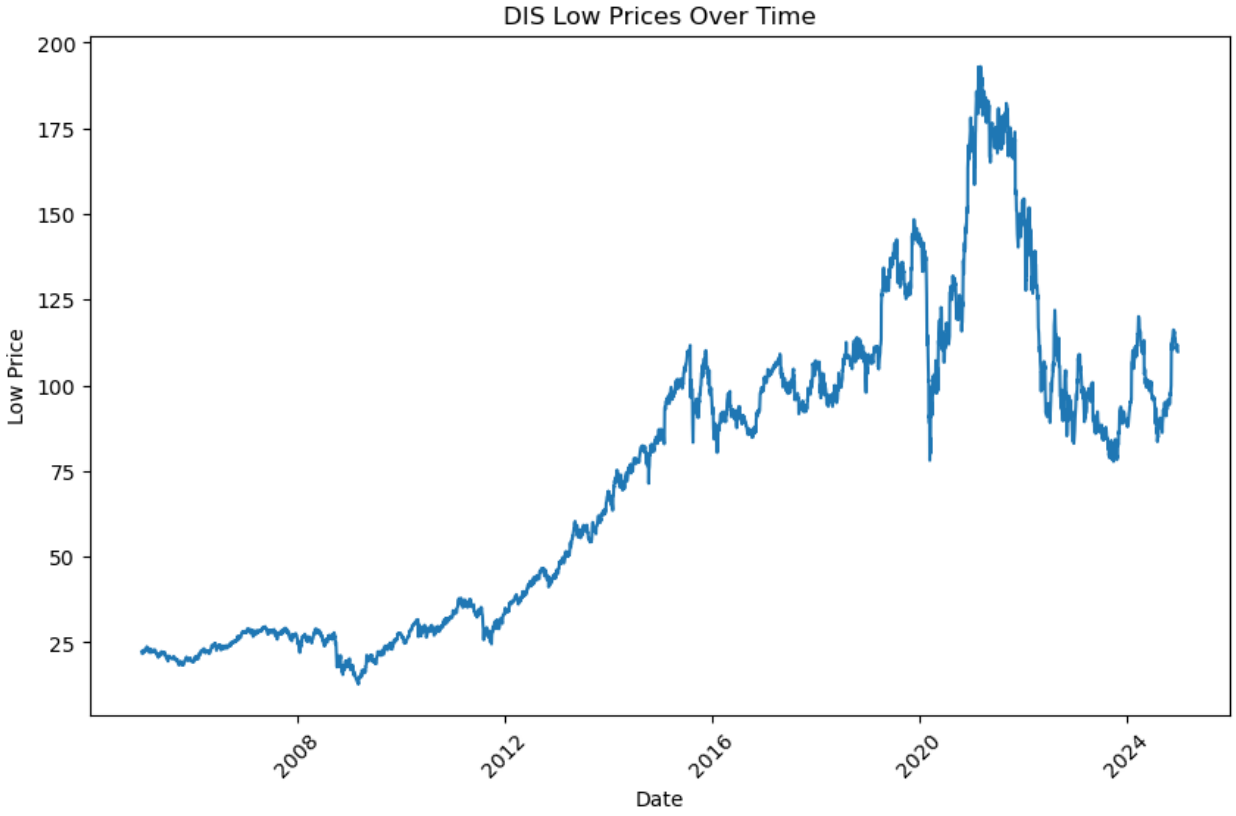

NVDA Low Prices Over Time
Low Price
0
20
40
60
80
100
120
140
2008
2012
2016
2020
2024
Date

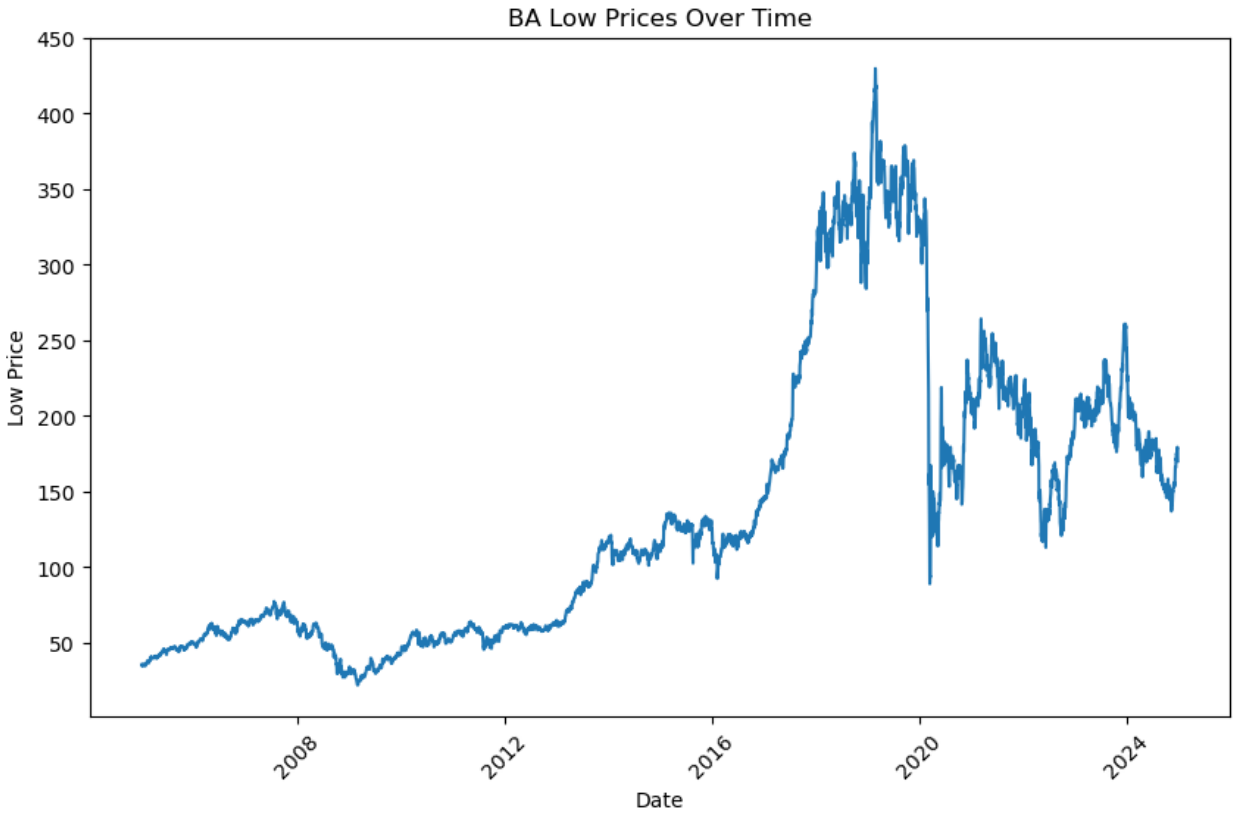

BRK-B Low Prices Over Time
Low Price
100
200
300
400
500
2008
2012
2016
2020
2024
Date


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
450
400
350
300
250
200
150
100
50
2008
2012
2016
2020
2024
Date

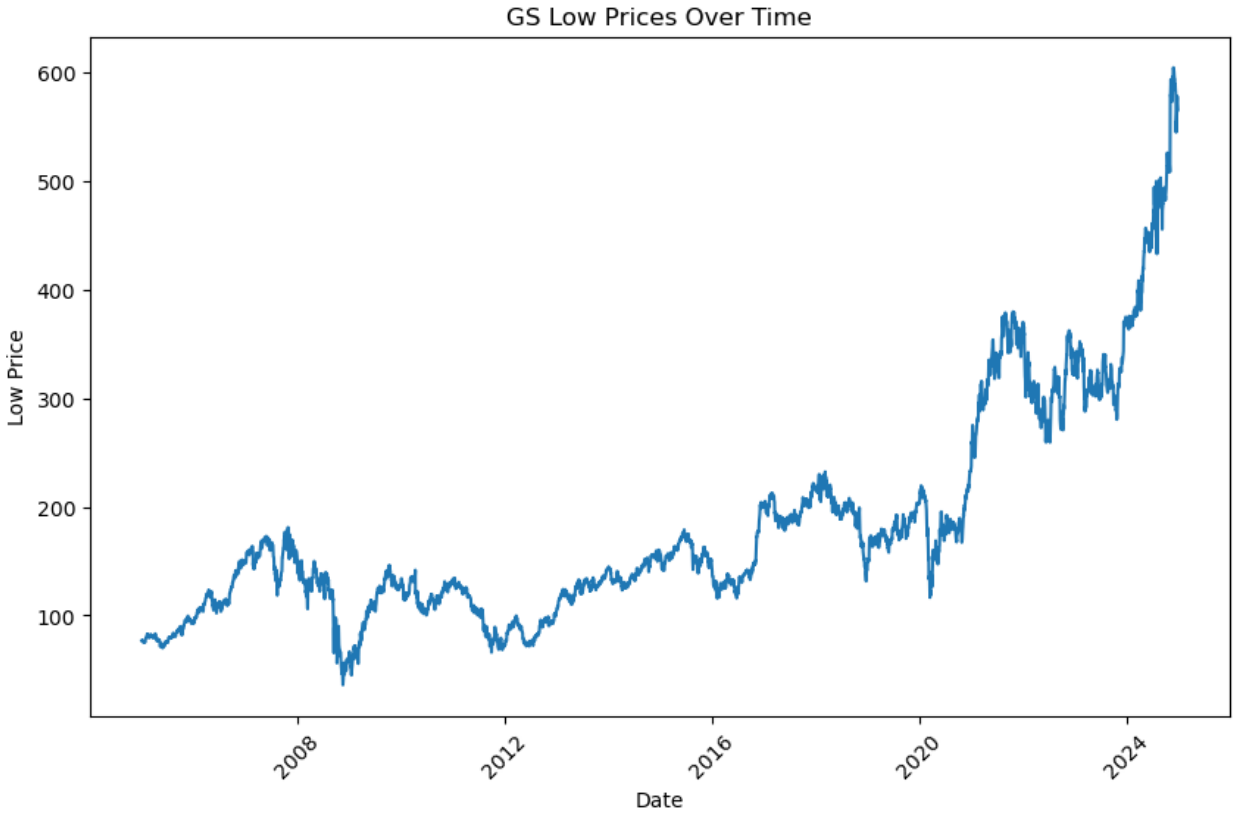
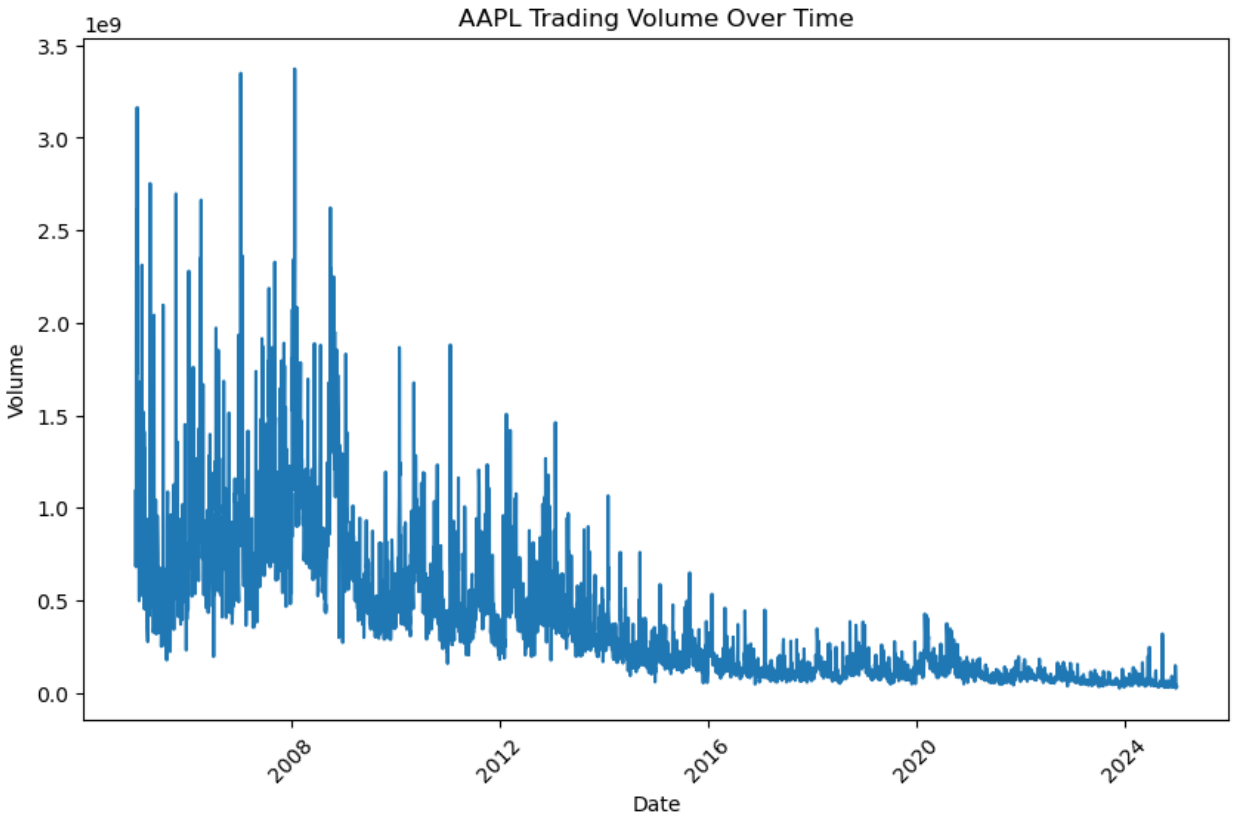
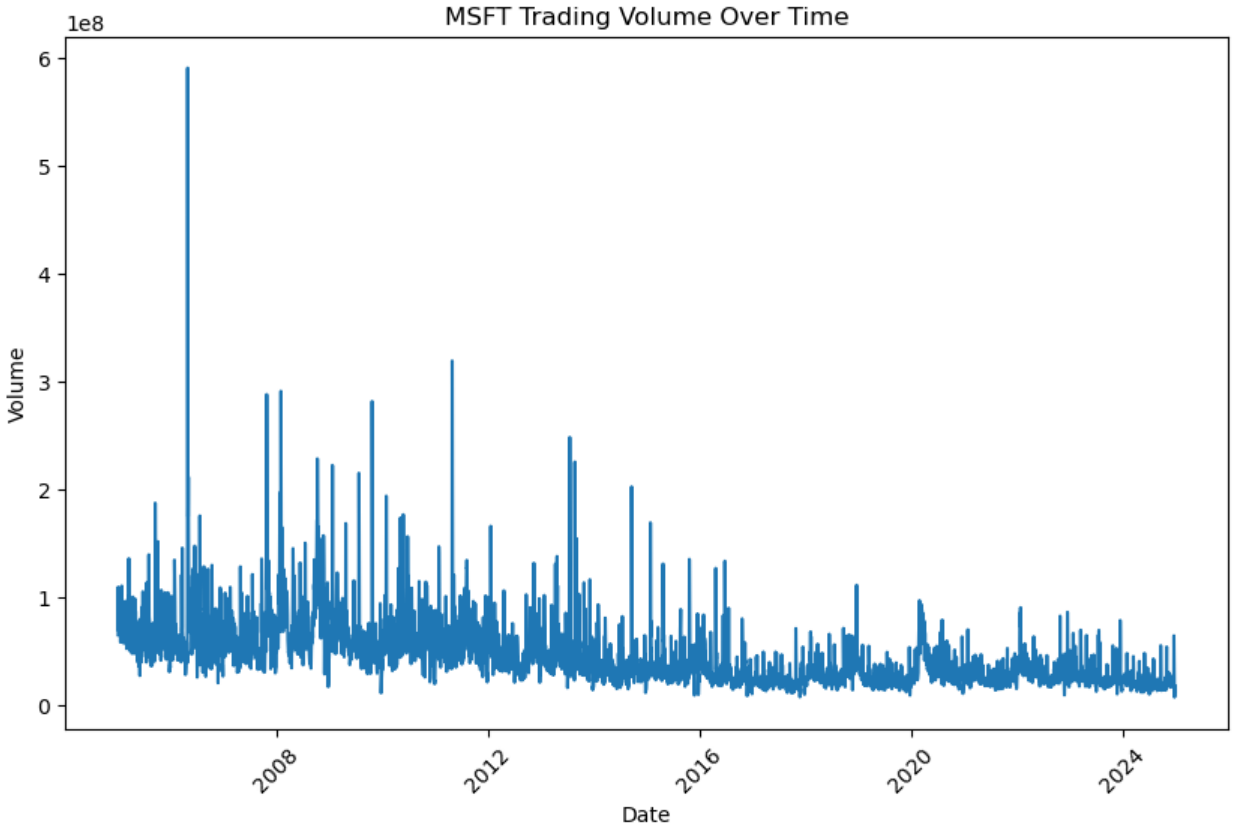

GS Low Prices Over Time
Low Price
600
500
400
300
200
100
2008
2012
2016
2020
2024
Date

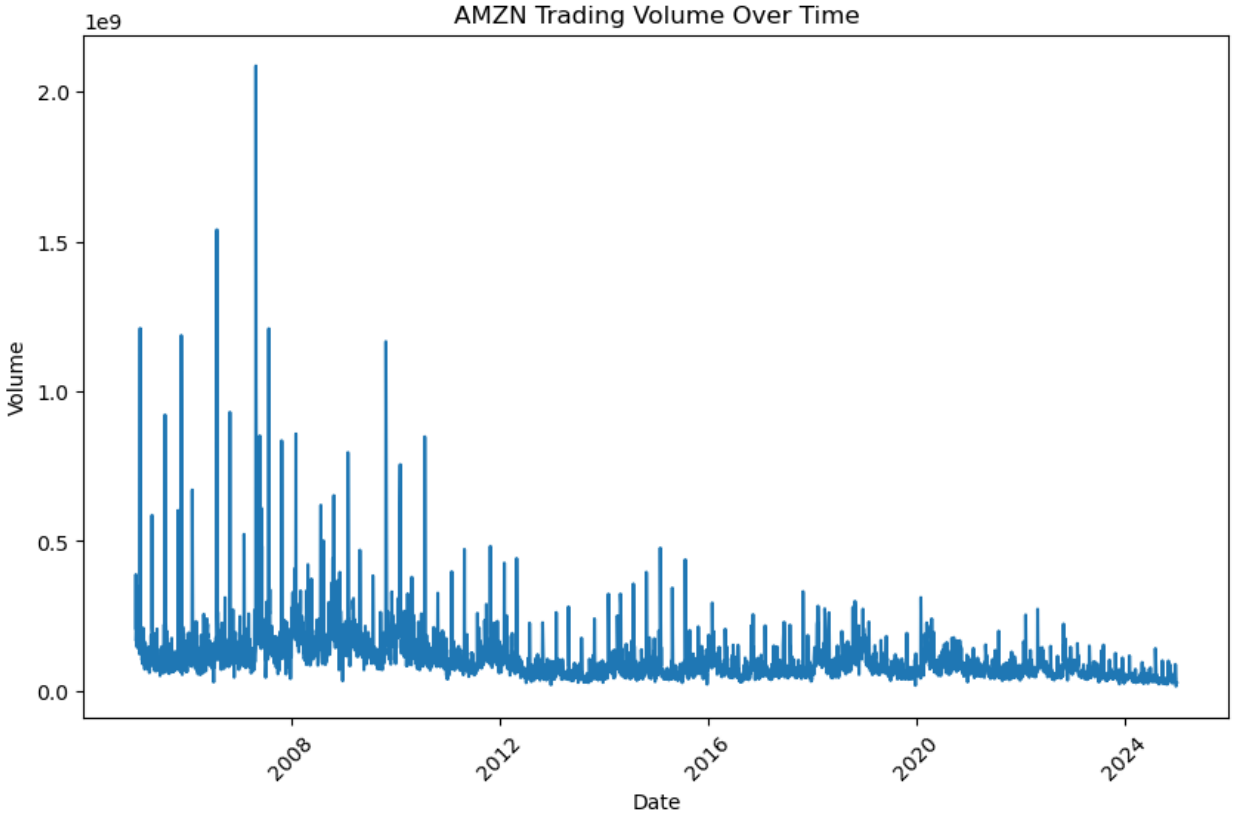
Figure 4.4: Line plots for low prices of 15 companies

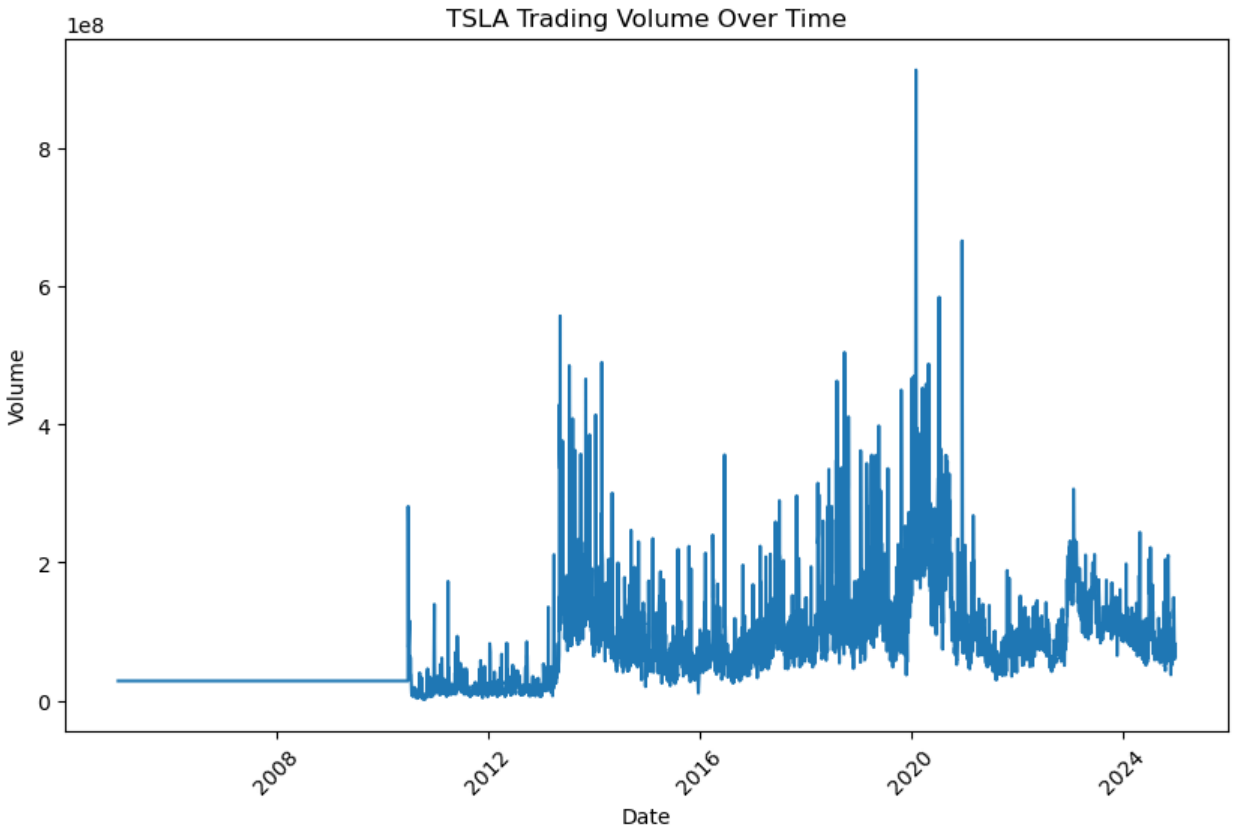
e. Time Plots for Trading Volume:

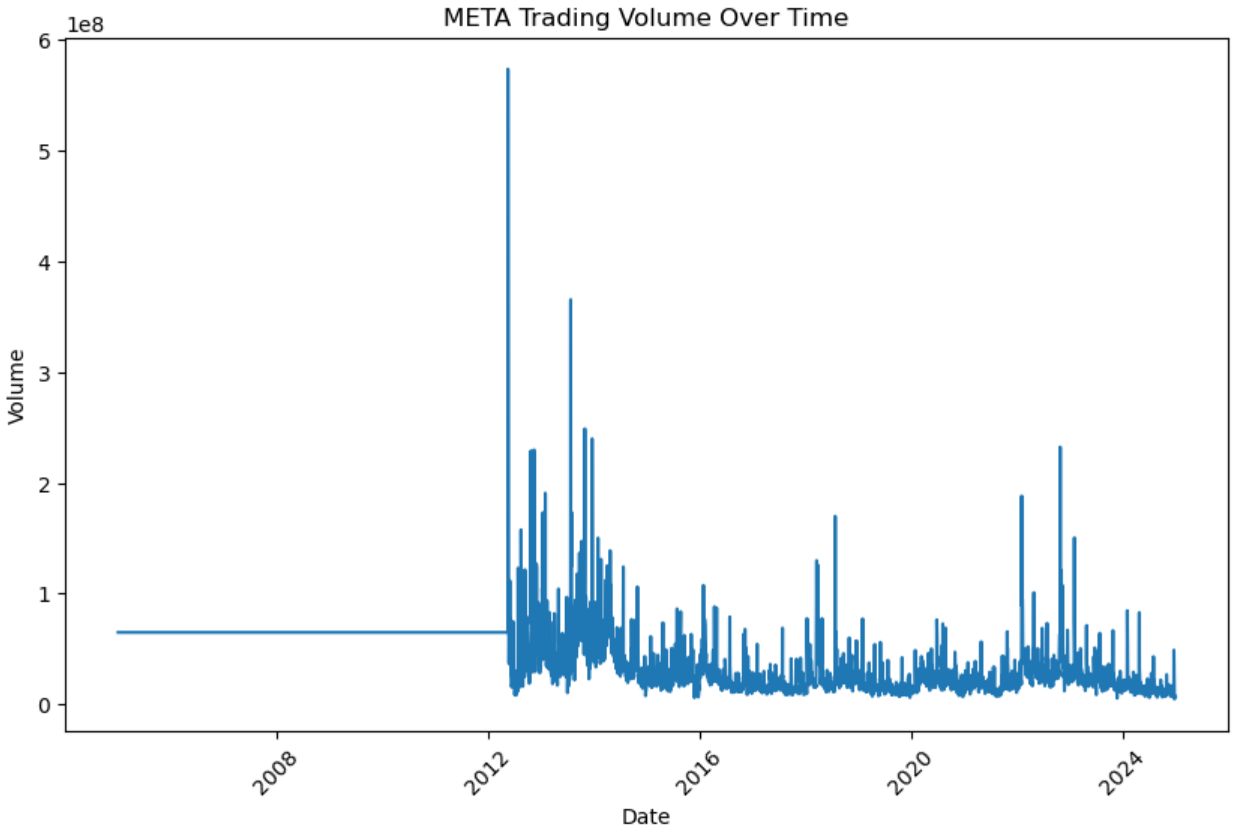
This graph is designed to show the time series for each company for trading volume. The number of shares traded on a particular day can be thought of as a broadcasting unit of market attraction and liquidity, with heftier periods of volume correlating with meaningful market events or company-specific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

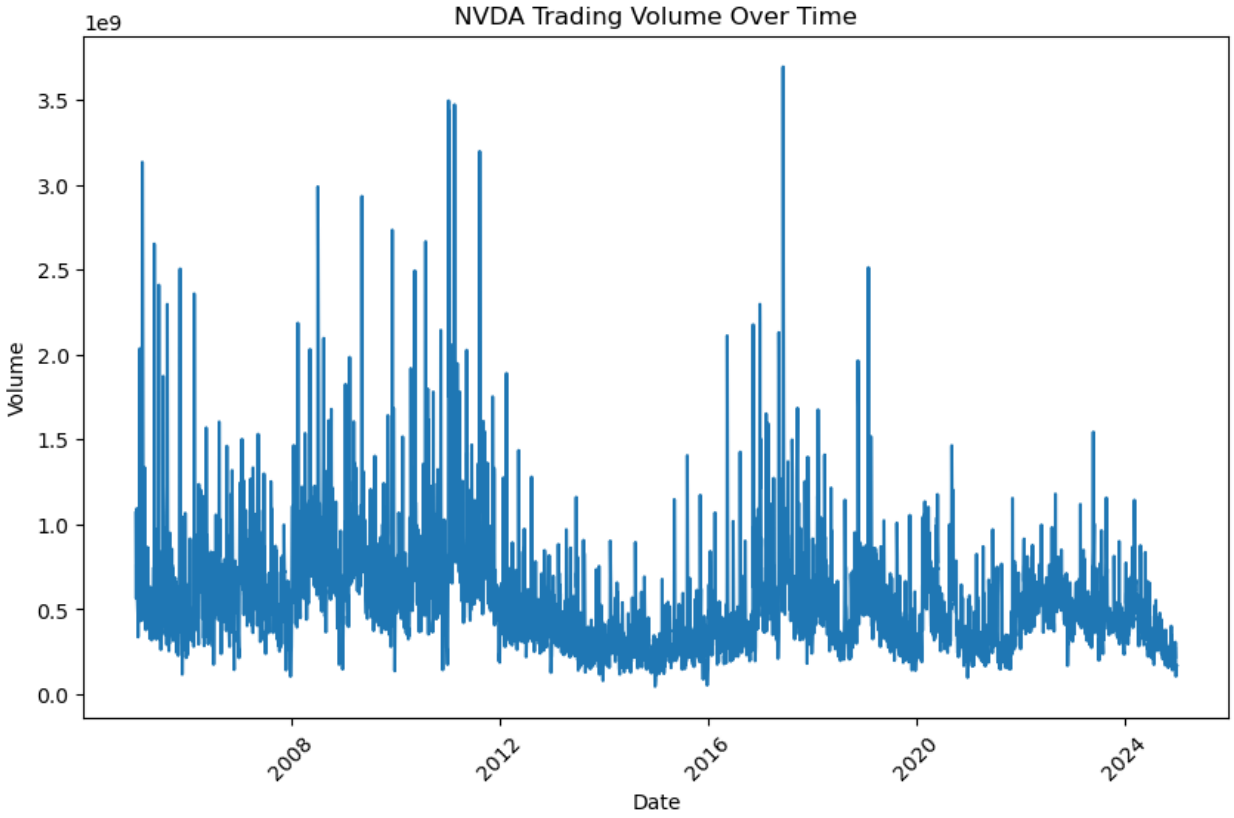


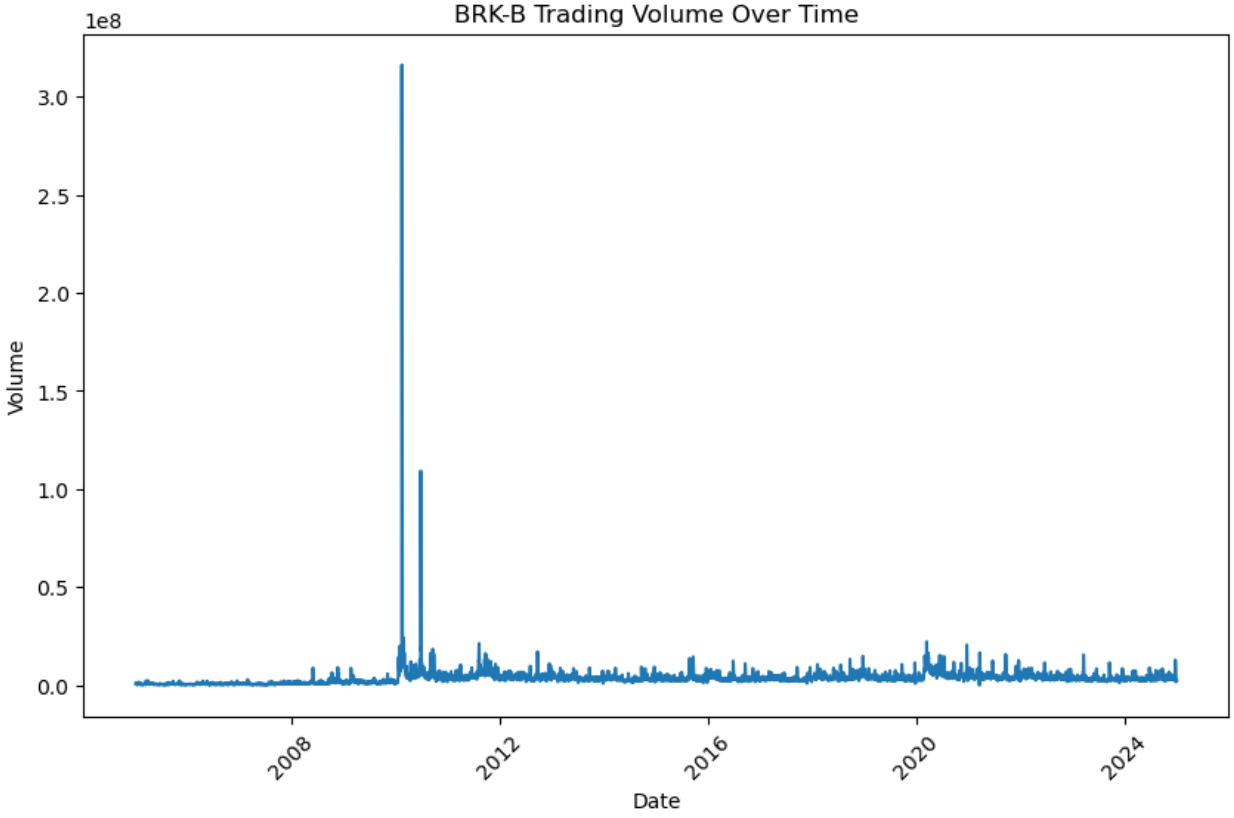

MSFT Trading Volume Over Time
1e8
6
5
4
3
2
1
0
Volume
2008
2012
2016
2020
2024
Date

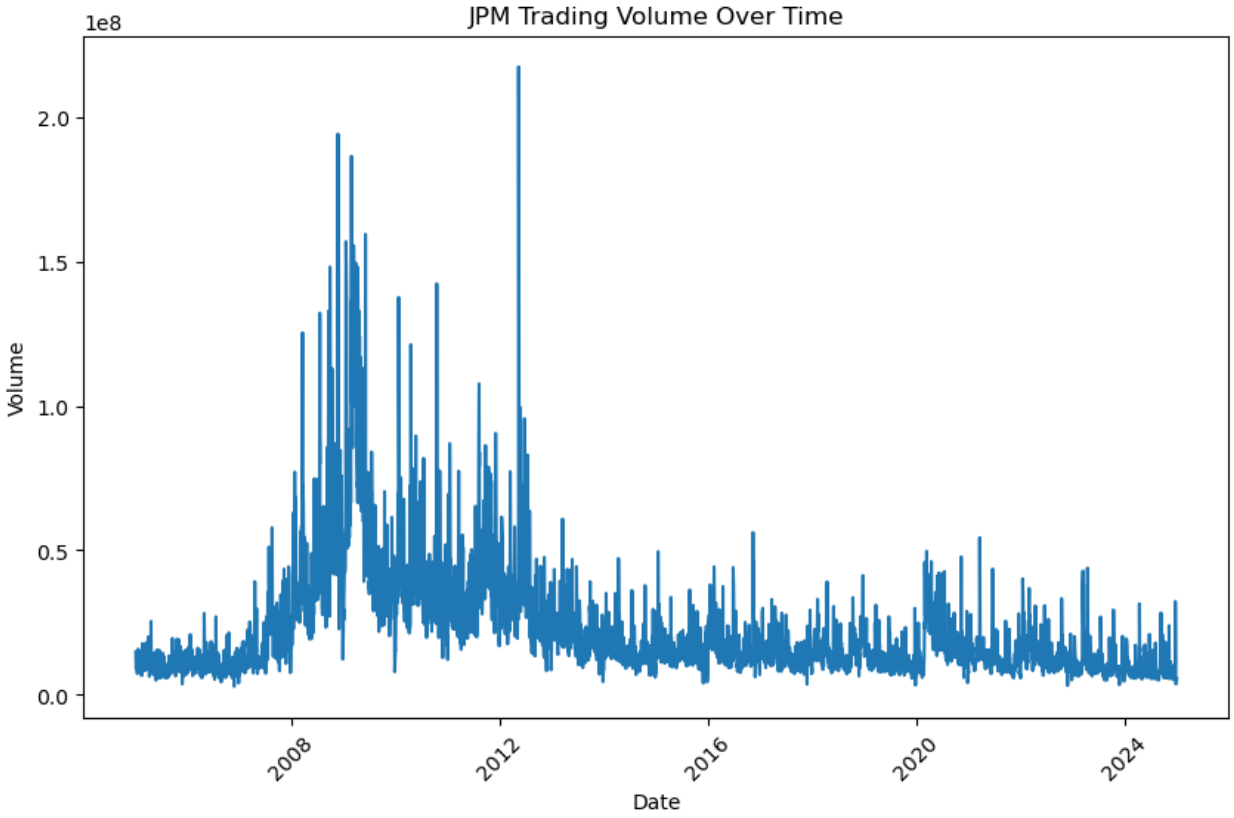

AMZN Trading Volume Over Time
1e9
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date

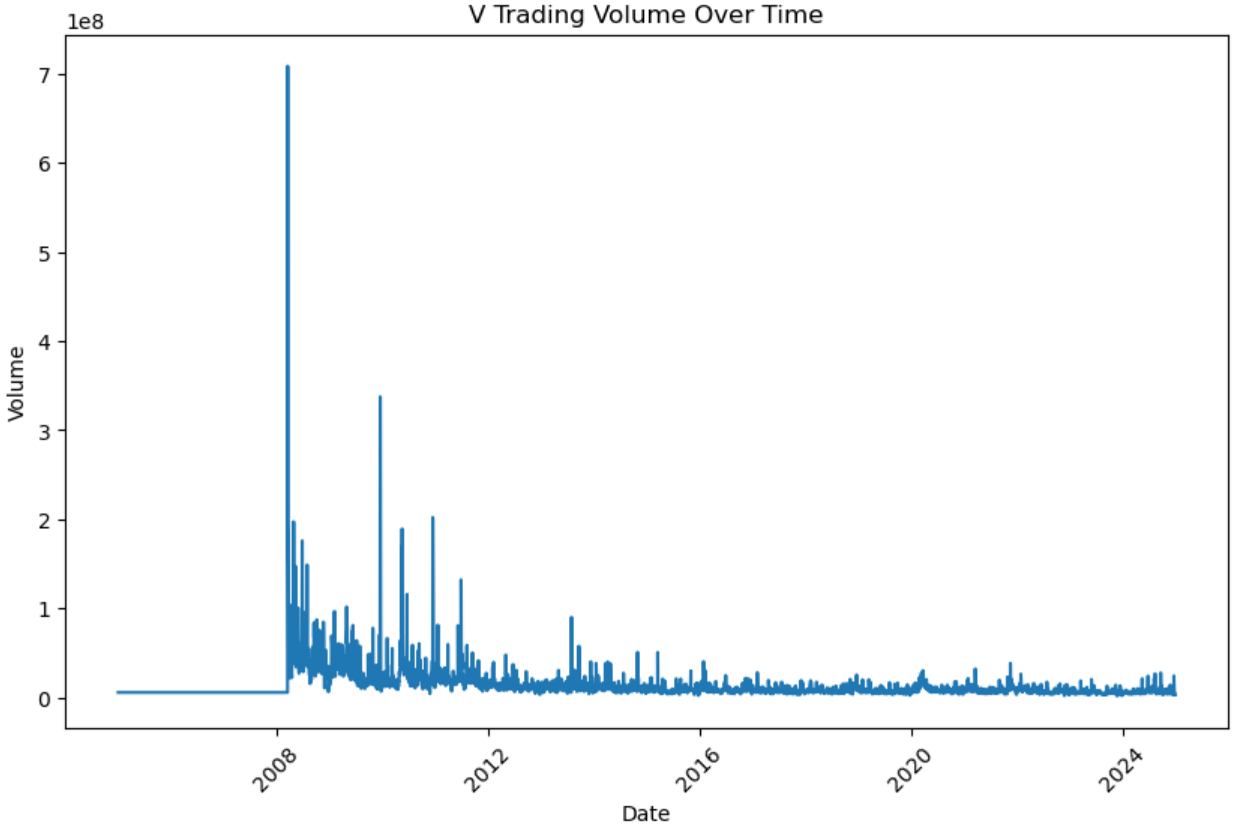

TSLA Trading Volume Over Time
1e8
8
6
4
2
0
Volume
2008
2012
2016
2020
2024
Date

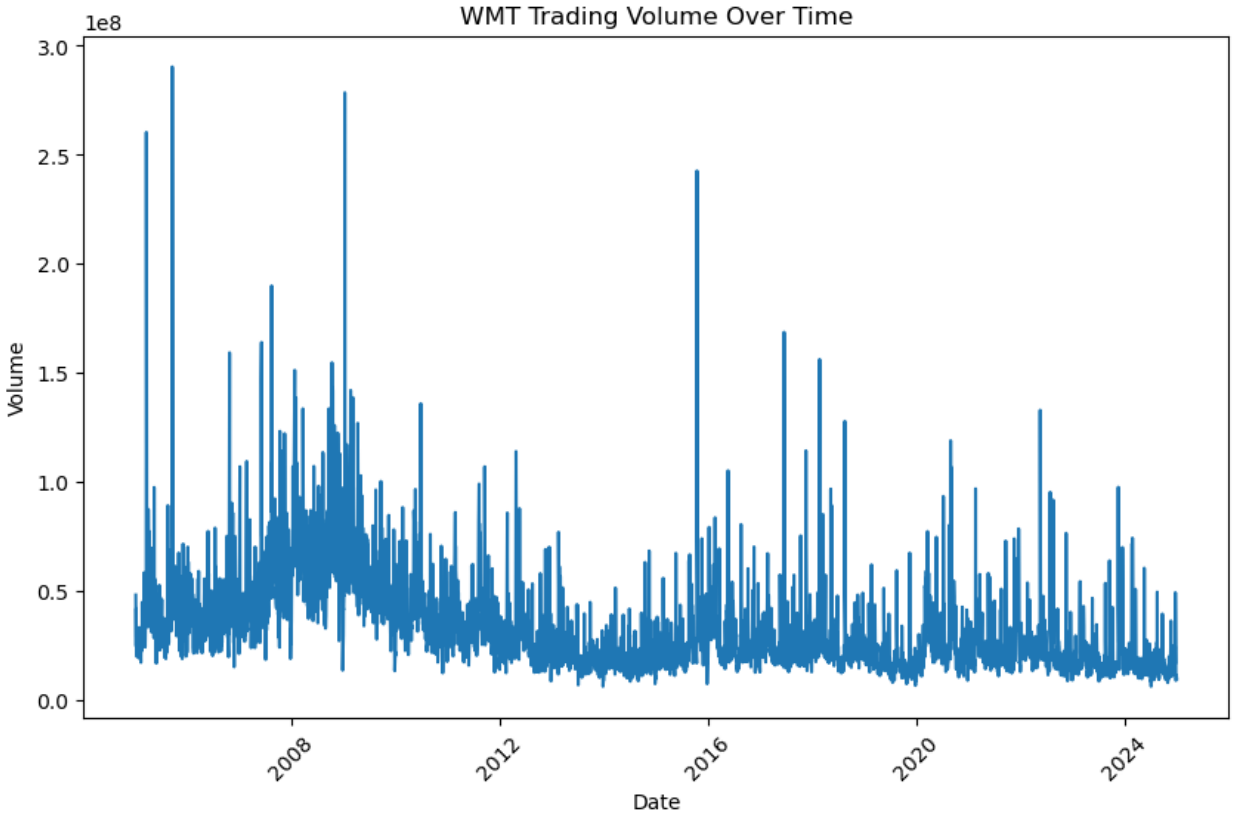

META Trading Volume Over Time
1e8
6
5
4
3
2
1
0
Volume
2008
2012
2016
2020
2024
Date

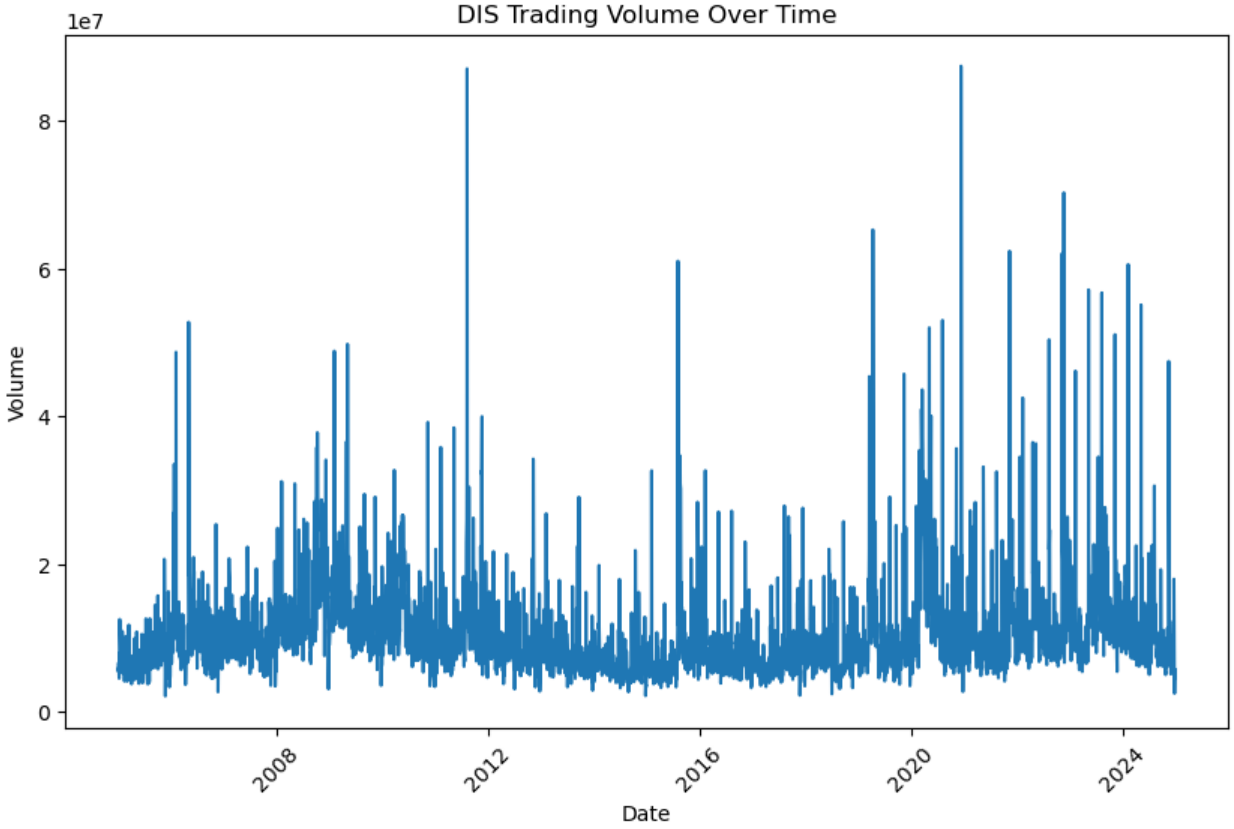

NVDA Trading Volume Over Time
1e9
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date

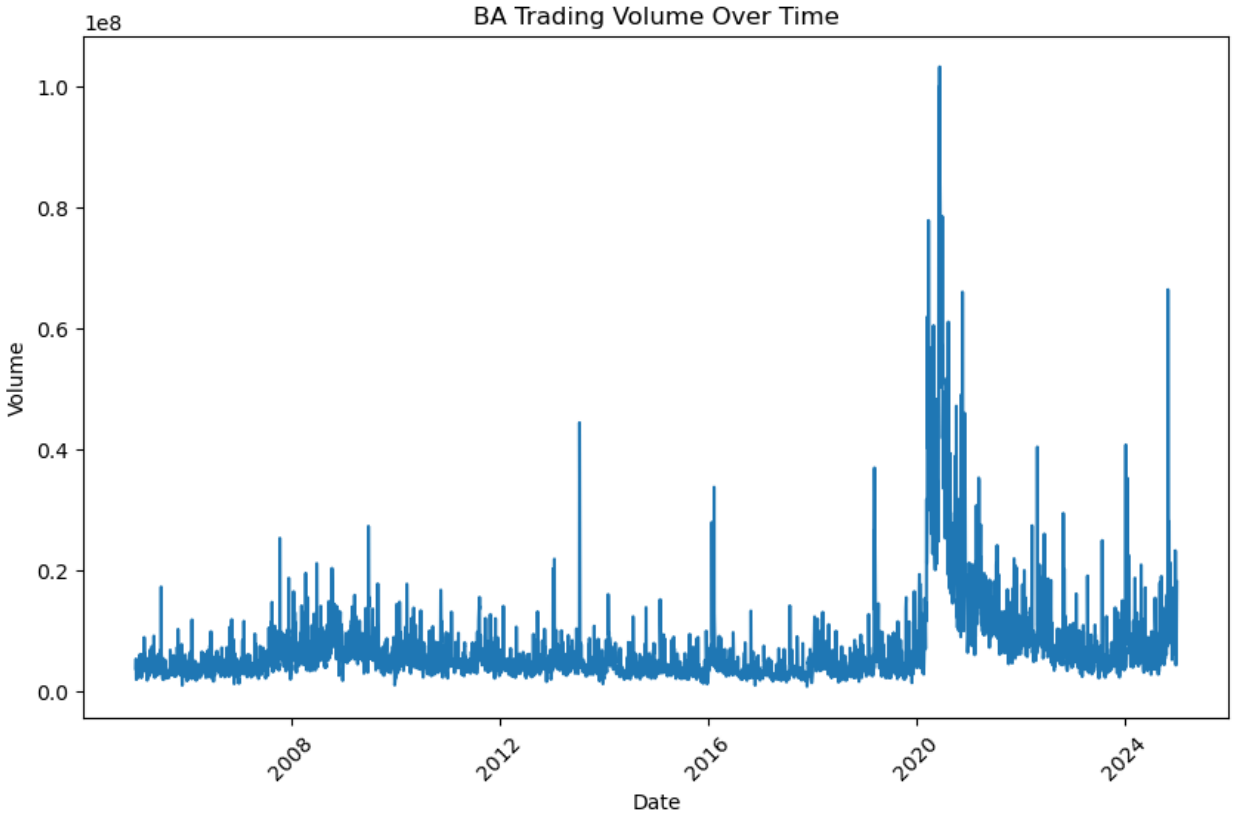

BRK-B Trading Volume Over Time
1e8
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date

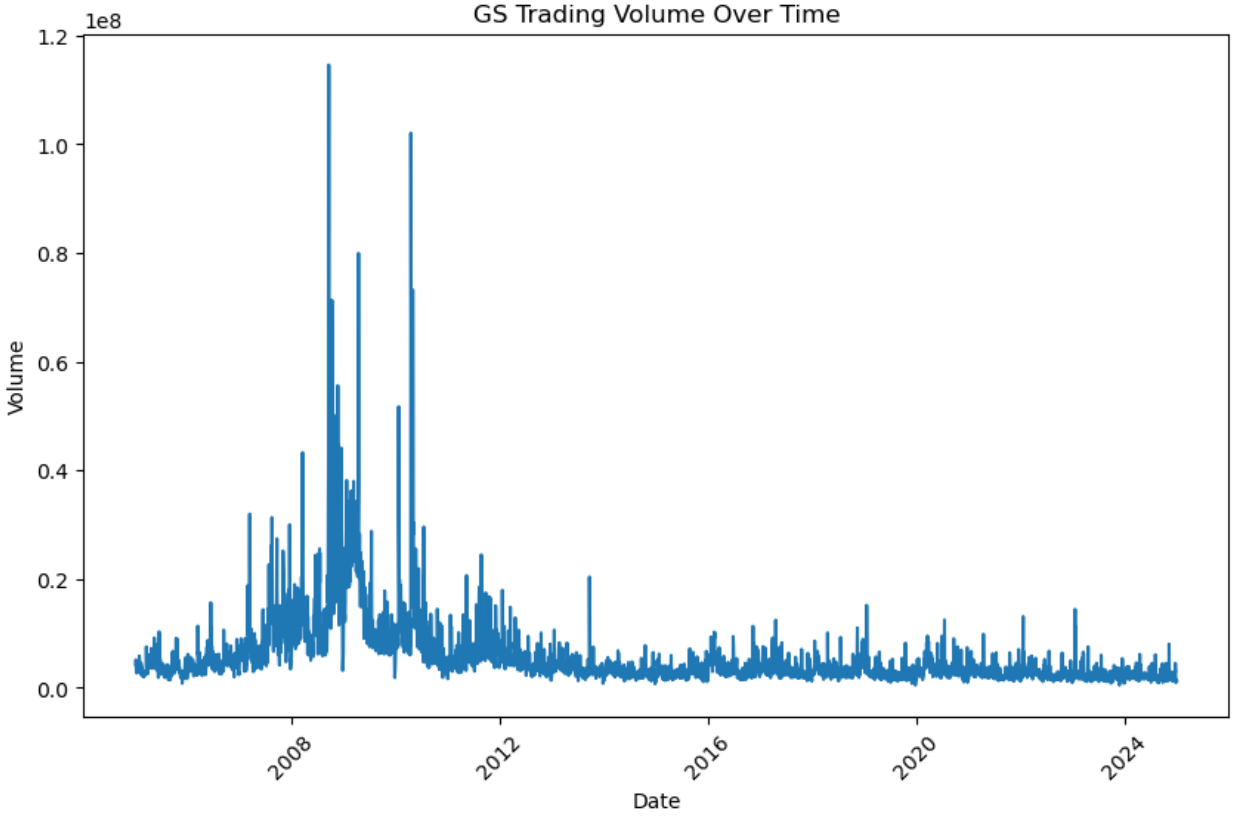

JPM Trading Volume Over Time
1e8
2.0
1.5
1.0
0.5
0.0
Volume
2008
2012
2016
2020
2024
Date


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

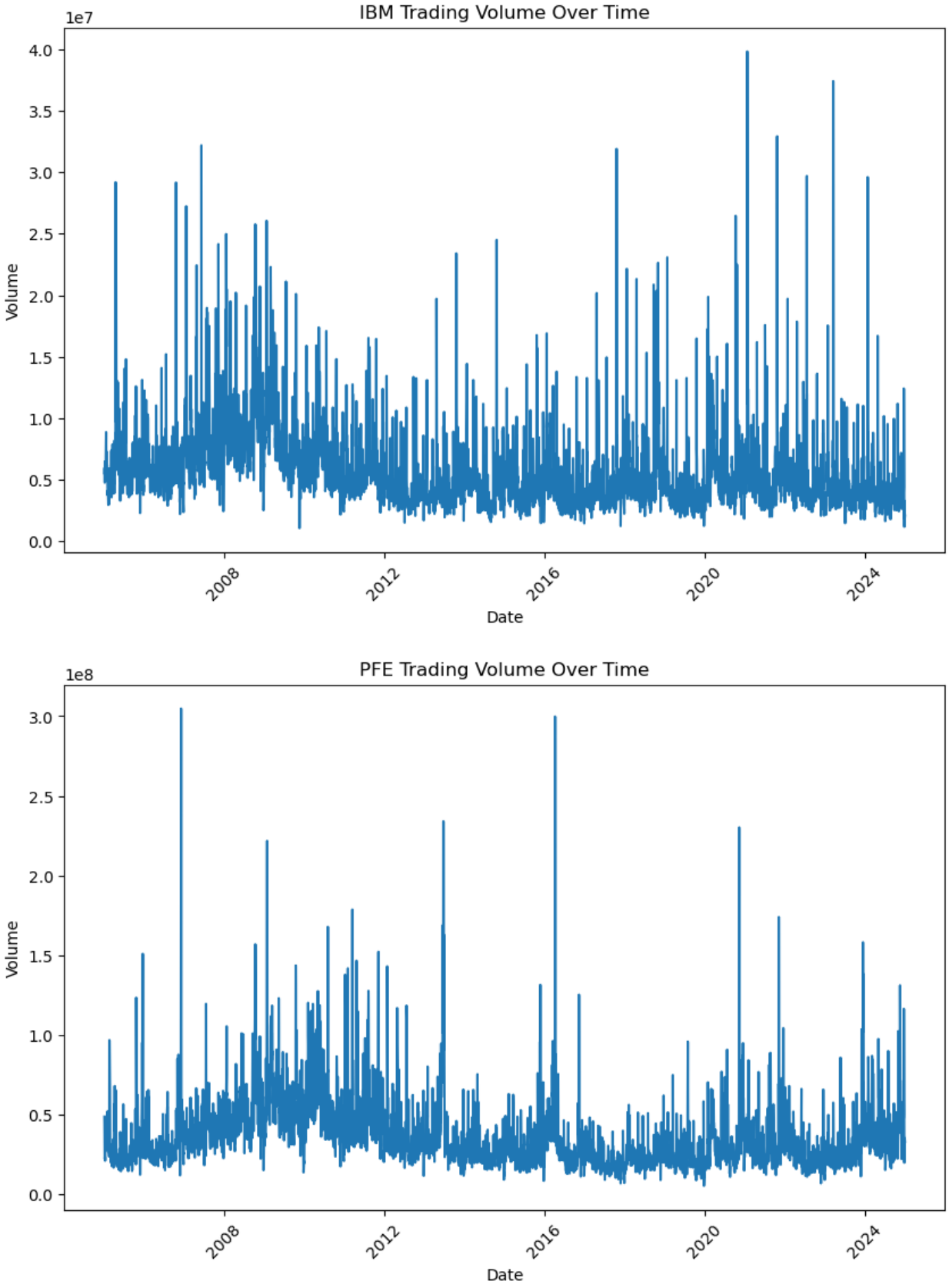


Figure 4.5: Trading volume of 15 companies over time

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.

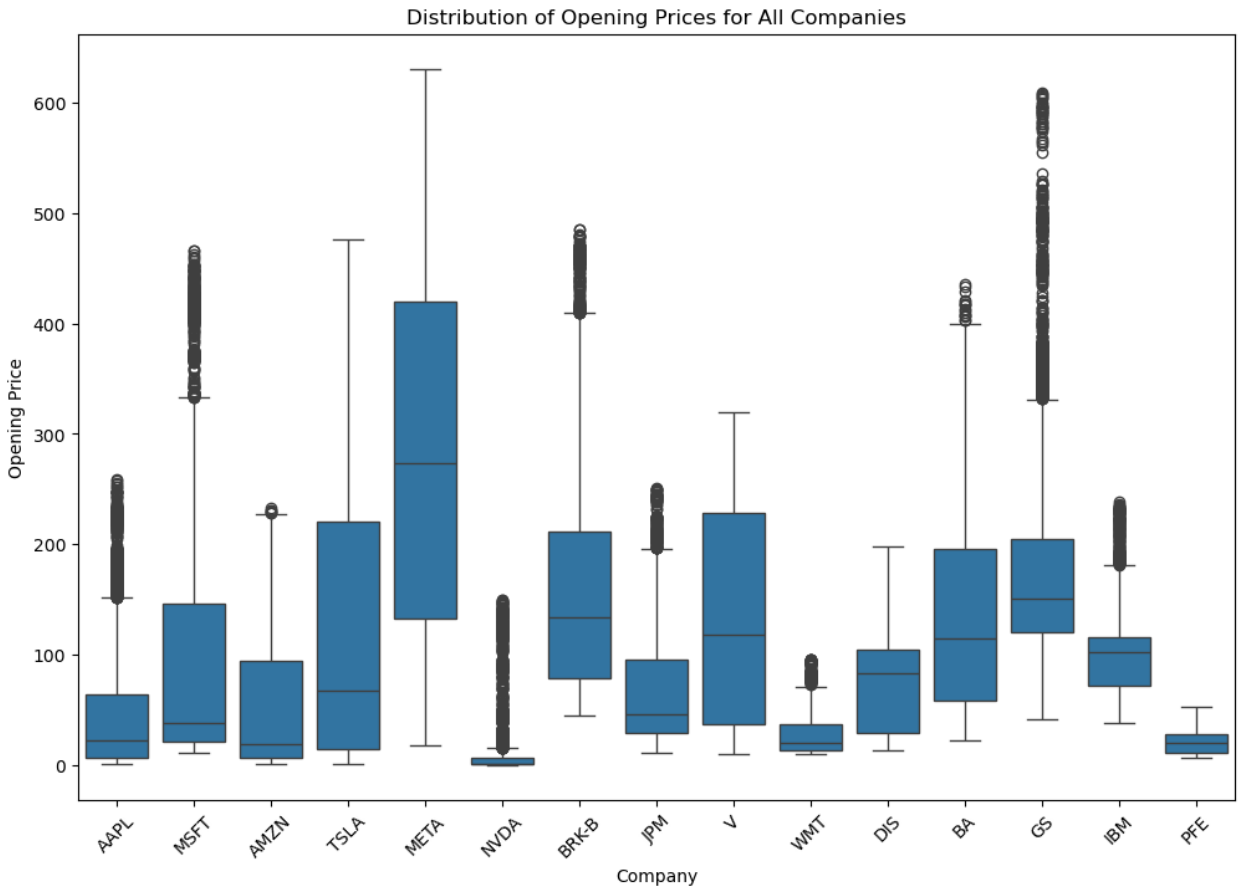


Figure 4.6: Box plot for distribution of opening price of 15 companies

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.

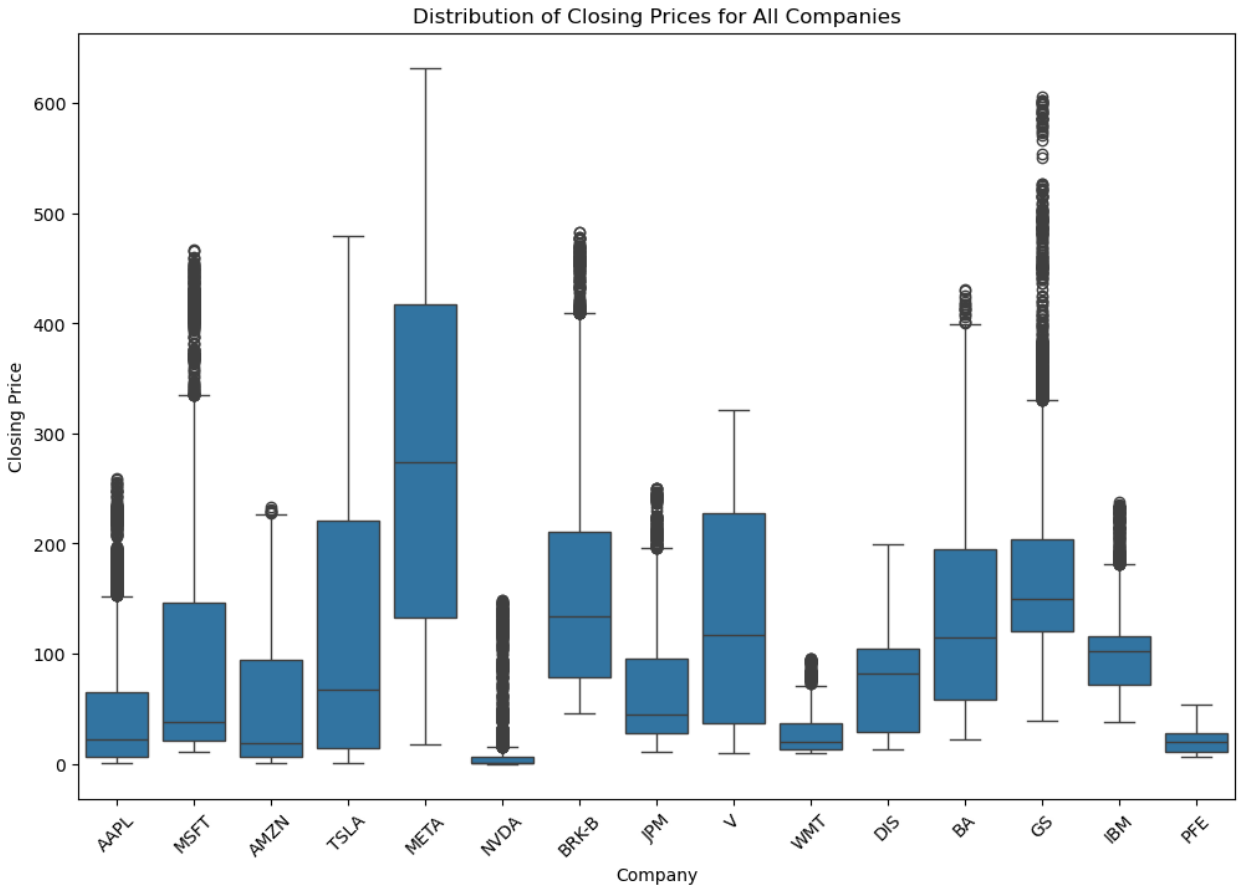


Figure 4.7: Box plot for distribution of closing price of 15 companies

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.

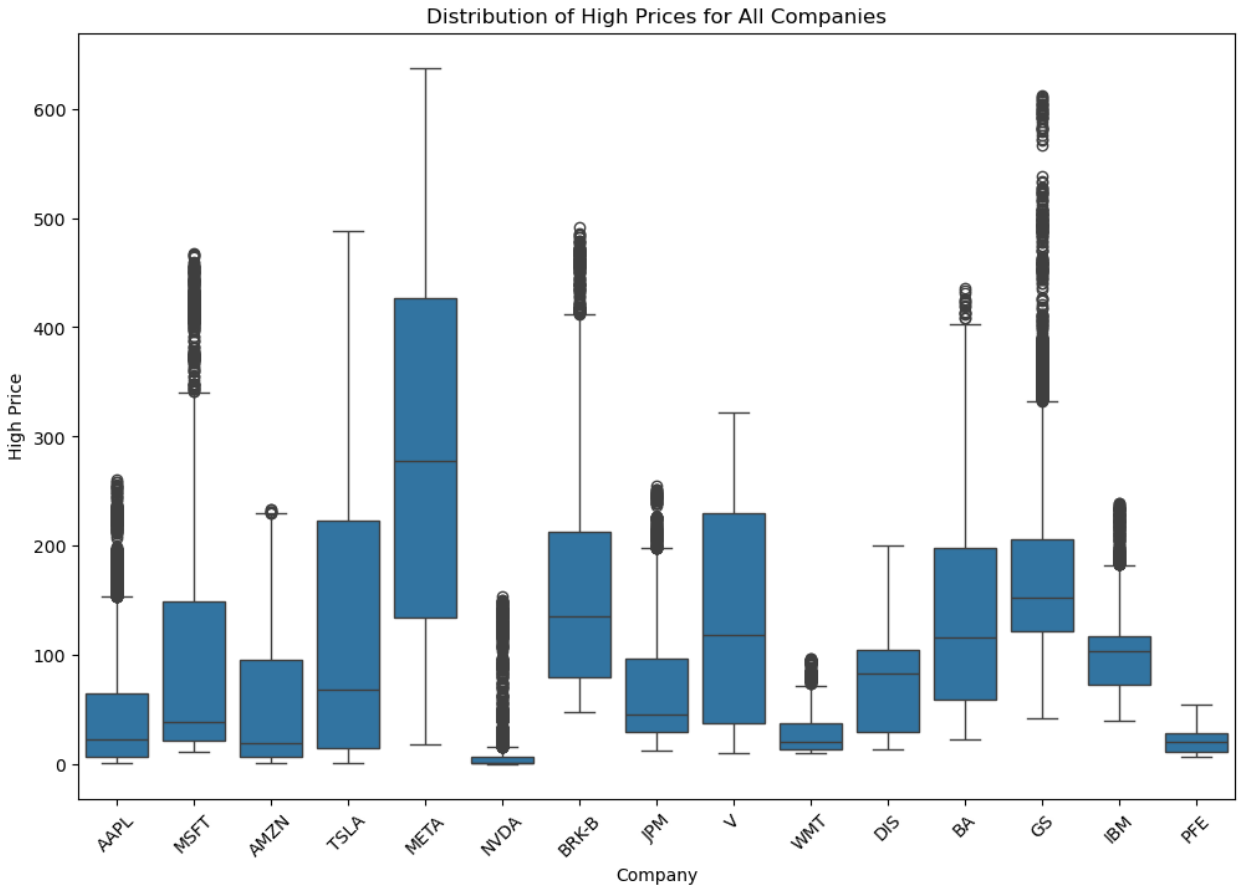


Figure 4.8: Box plot for distribution of high price of 15 companies

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.

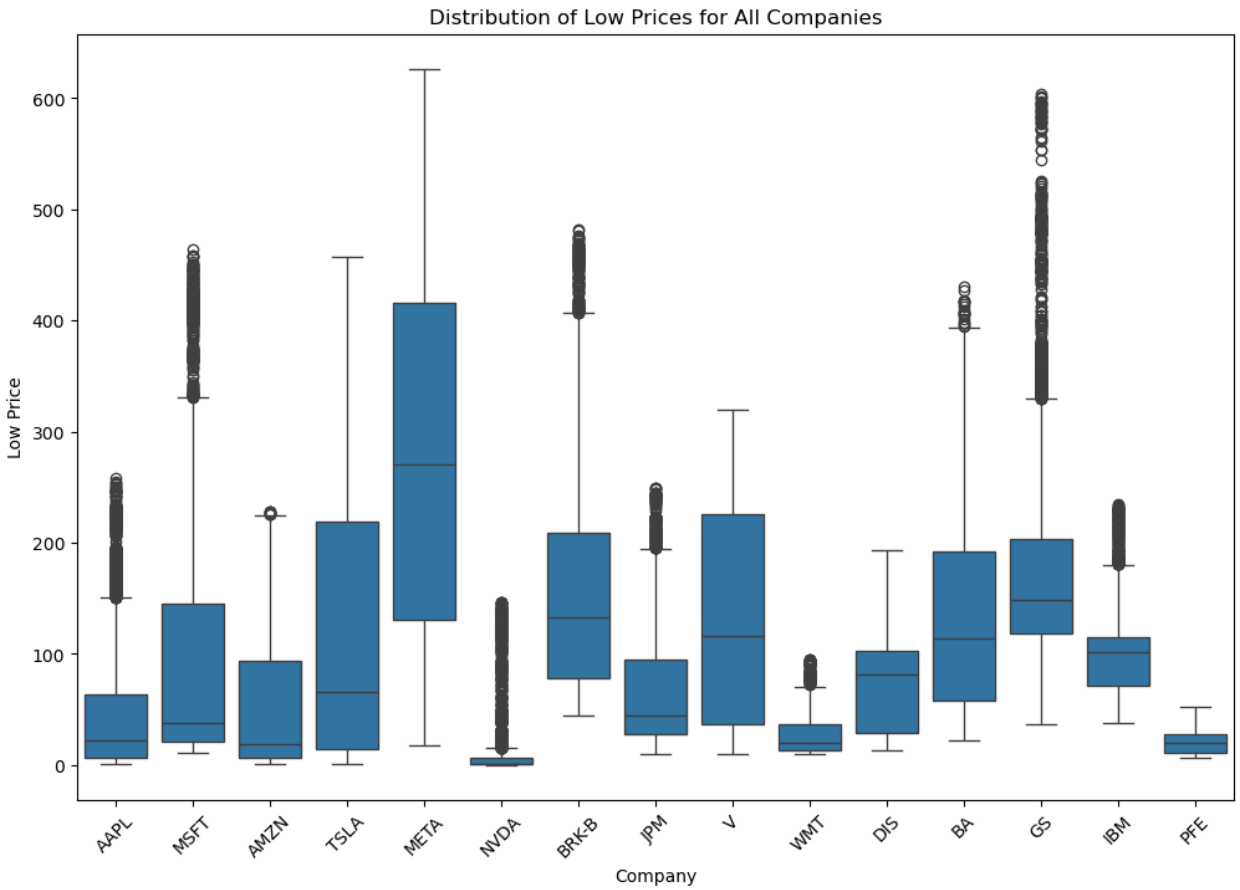


Figure 4.9: Box plot for distribution of low price of 15 companies

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.

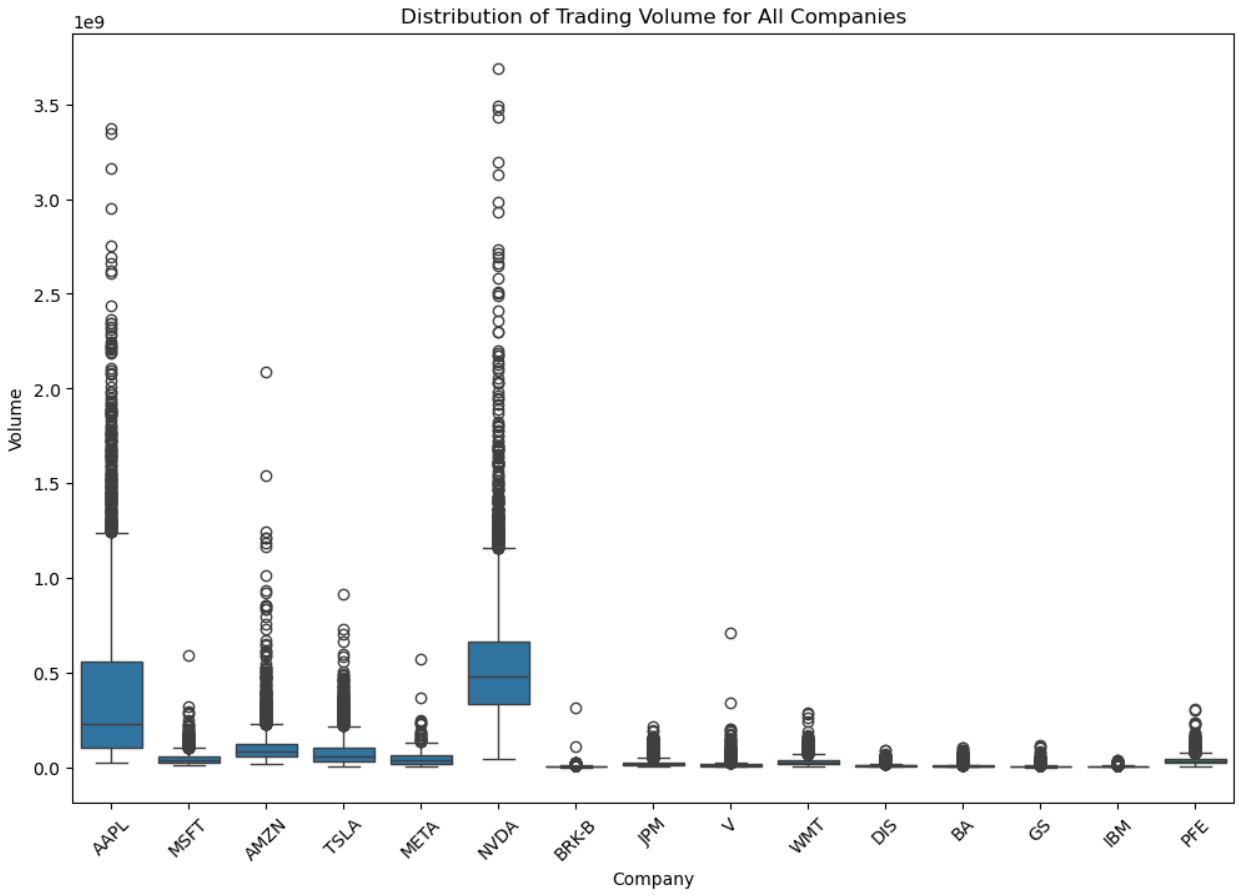


Figure 4.10: Box plot for distribution of trading volume of 15 companies

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.

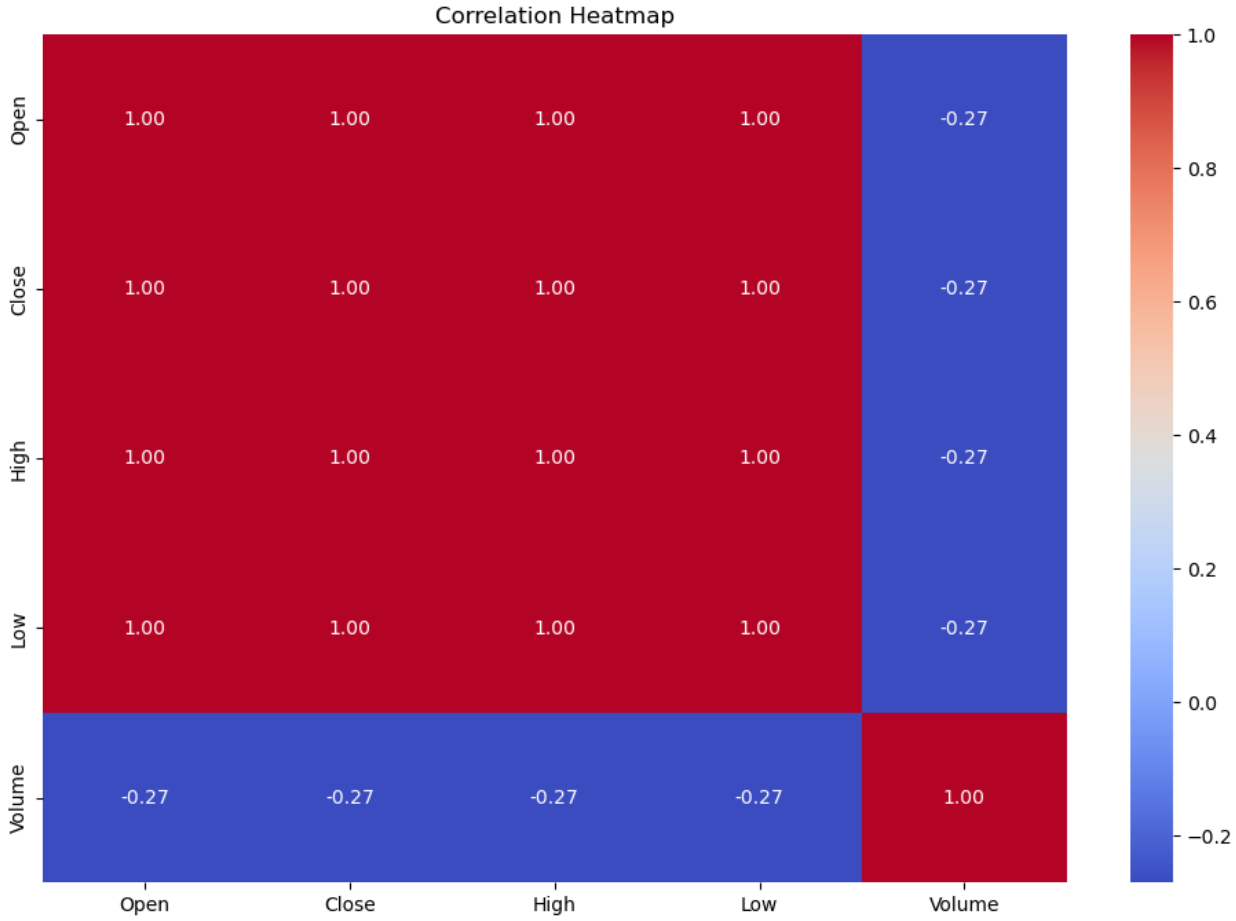


Figure 4.11: correlation heatmap between various prices

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.

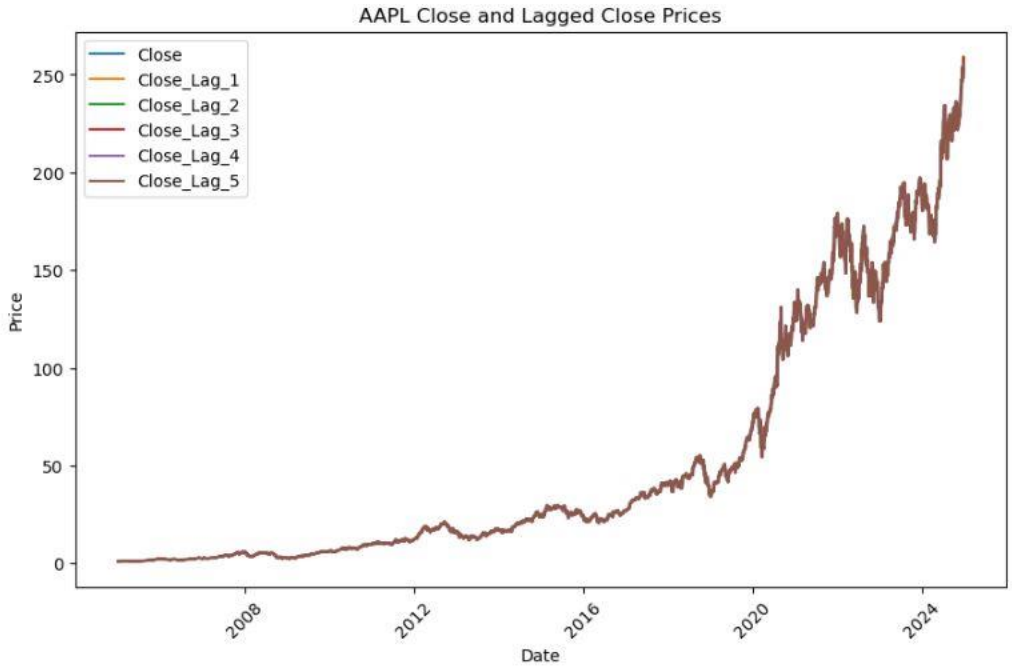


Figure 4.12: Visualization of Lagged Features for a sample company (AAPL)

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.

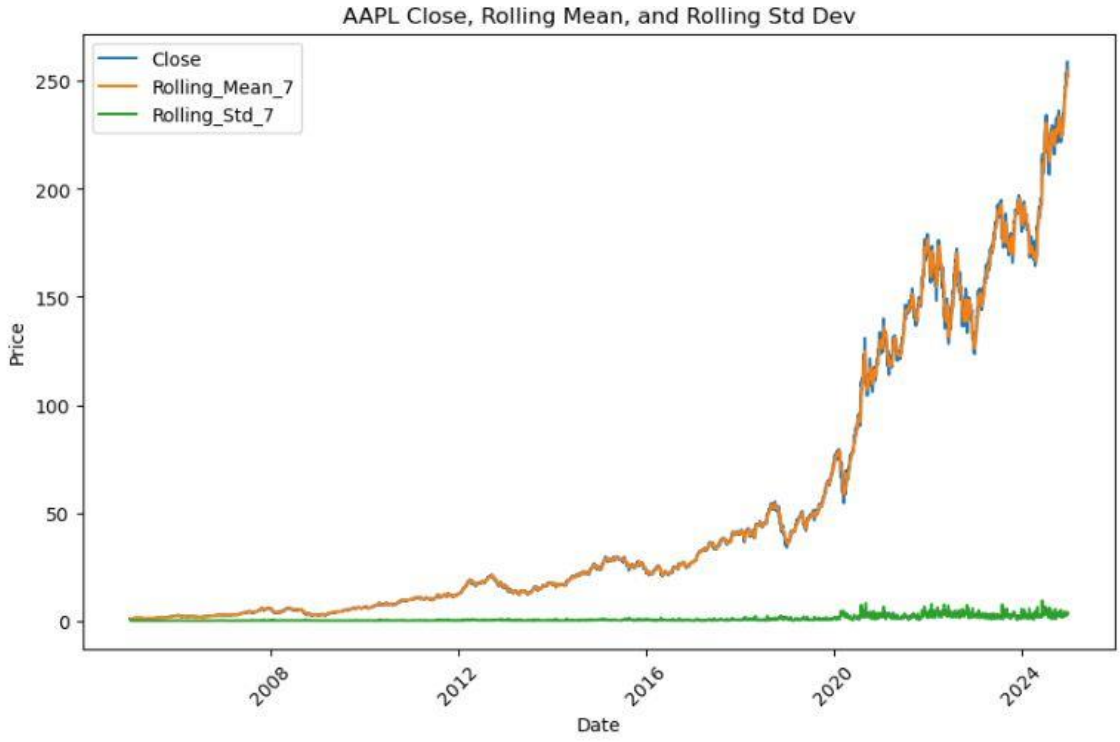


Figure 4.13: closing prices, rolling mean, and rolling standard deviation for AAPL

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.

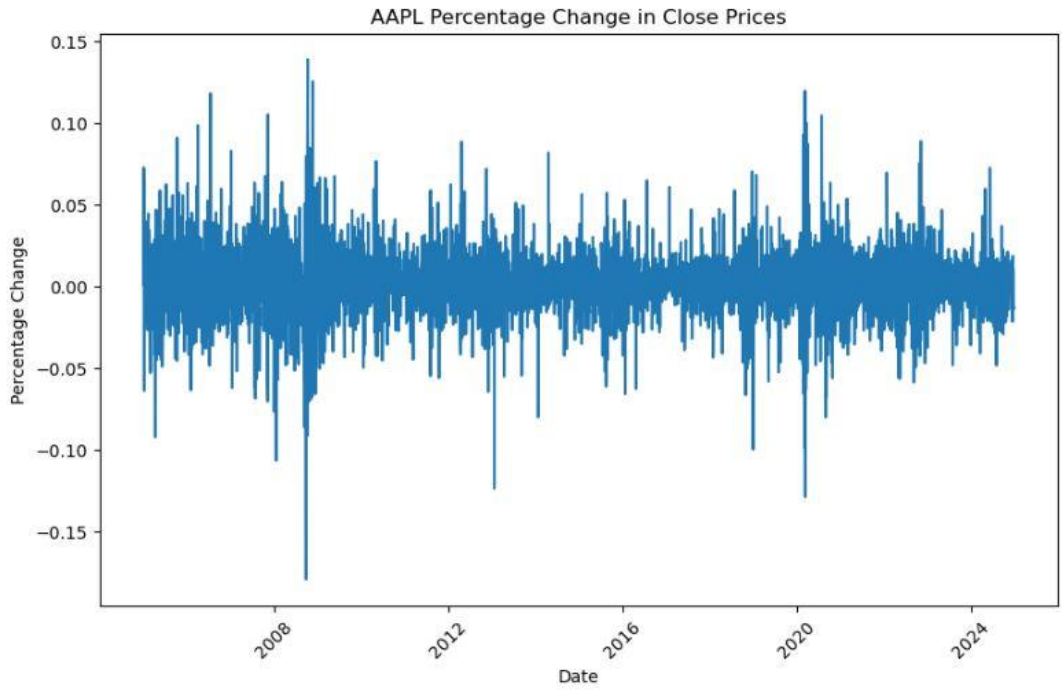


Figure 4.14: The percentage change in closing prices for AAPL

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.

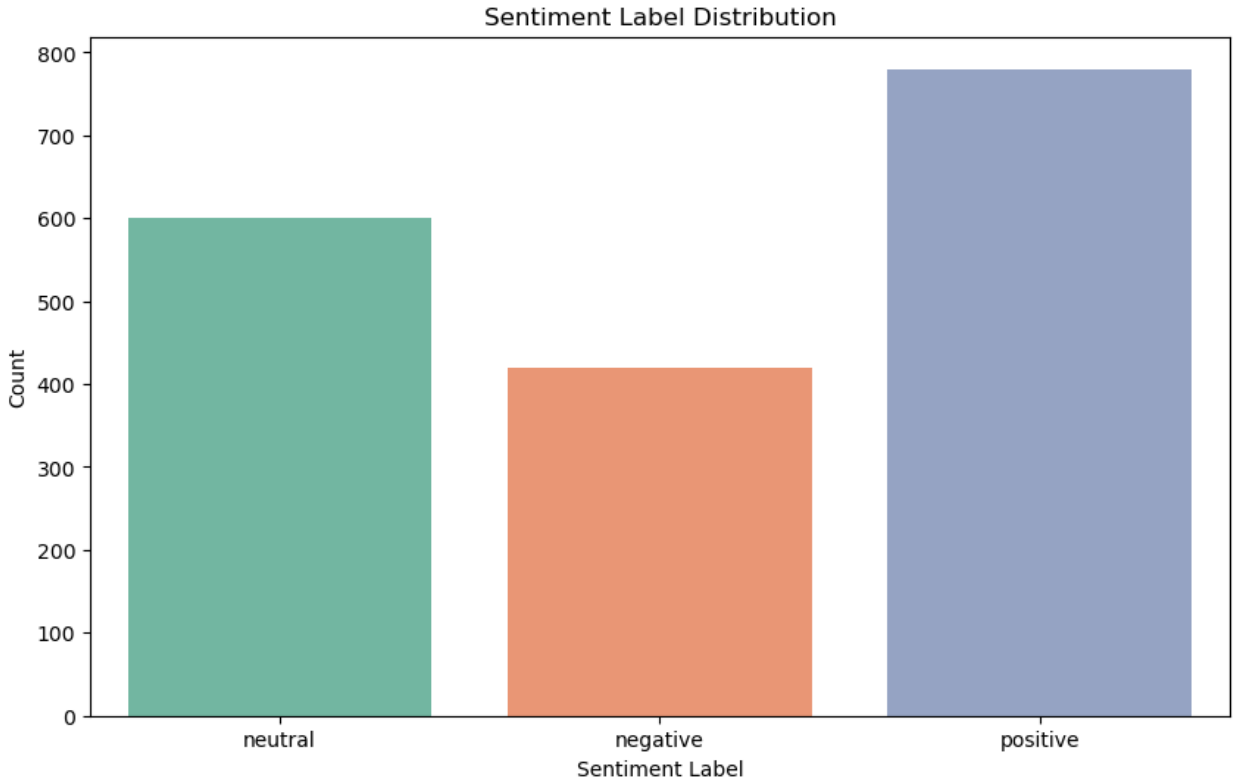


Figure 4.15: Sentiment distribution

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.

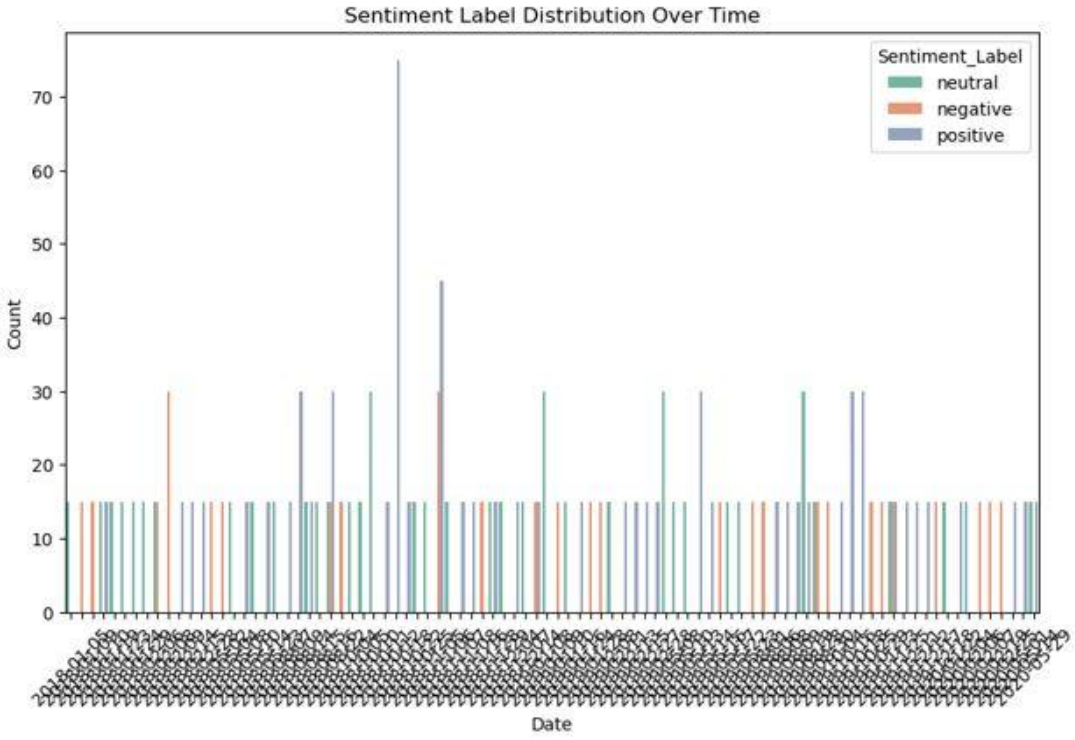


Figure 4.16: Sentiment distribution over time

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.

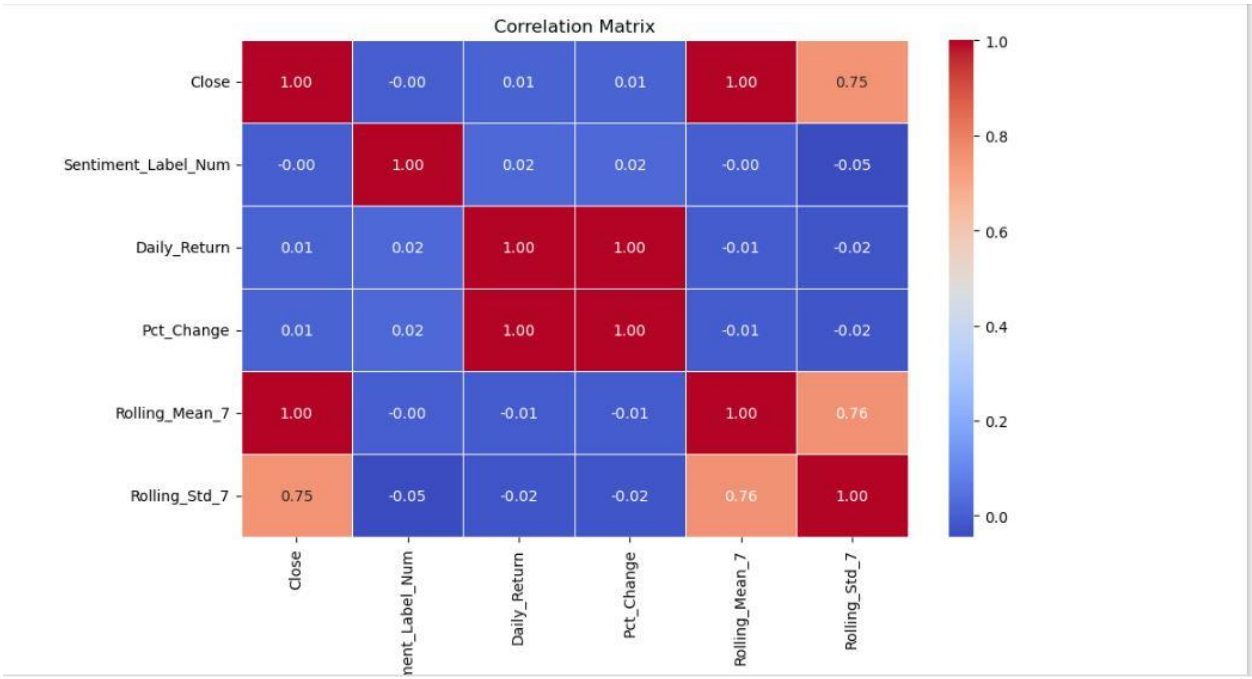


Figure 4.17: Correlation between sentiment score and stock price

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.

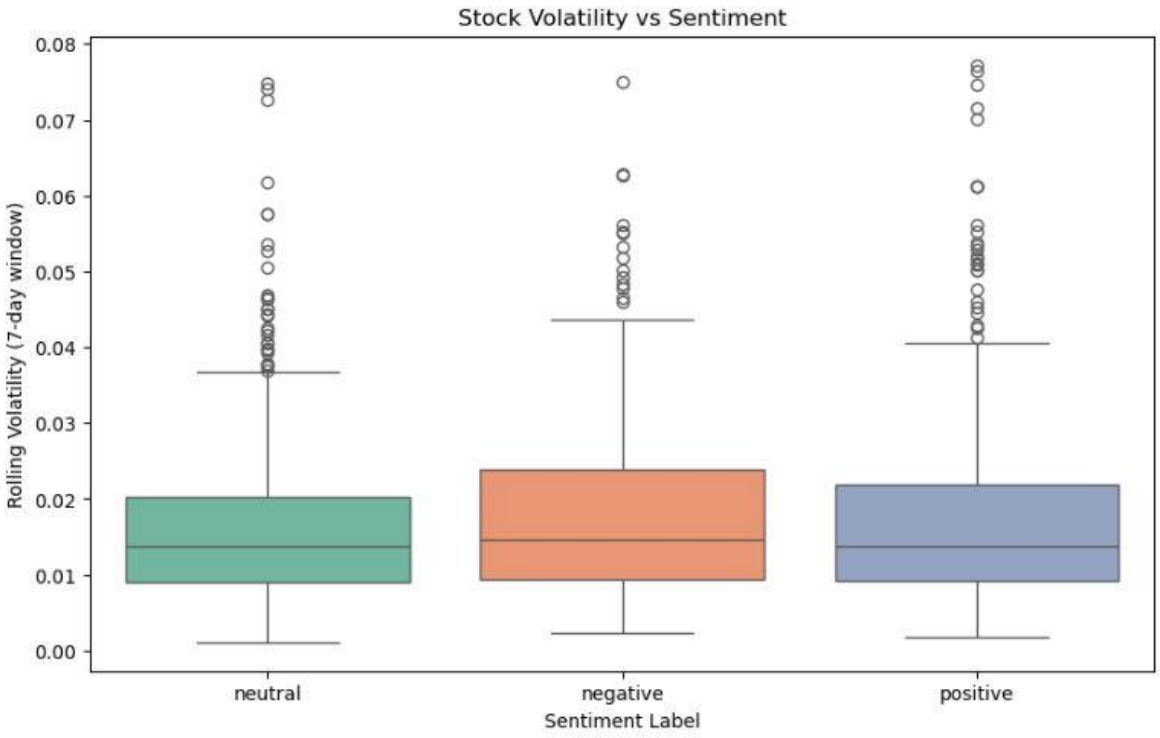


Figure 4.18: Stock volatility vs Sentiment

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.

CHAPTER 5

DISCUSSION AND FUTURE WORK

5.1 Introduction

This chapter outlines the primary findings and conclusions of the study conducted on sentiment analysis and stock market prediction. The objective was to discover a connection between the financial news sentiment and stock price dynamics and to establish a predictive model between sentiment and stock price. This study thus showed the power of machine learning and natural language processing (NLP) techniques for use in forecasting applications through FinBERT for sentiment classification and Random Forest for stock price prediction.

This chapter does not just summarize the contributions made by the research but also goes ahead to suggest future research directions and areas within which improvements could be made. The enhanced dataset, improved sentiment classification, incorporation of deep learning models, and better explainability, among others, are promoting future development in this field of study.

5.2 Summary

Foremostly, this study has developed a prediction model that uses financial news sentiment to predict stock price movements integrated with different stock market features. A systematic approach was developed for the research, which consisted of data collection, preprocessing, exploratory data analysis (EDA), sentiment classification through FinBERT, feature engineering, and application of Random Forest in machine learning modeling.

In the sentiment analysis section, financial news was classified to three categories: positive, neutral, or negative. This classification allows understanding of

the extent to which investor sentiments play a role in the behavior of the stock market. Results show that positive sentiment usually translates into an increase in stock price, while negative sentiment usually leads to decreased stock prices. Sentiment thus gives extra dimensions to market insight that can only be derived from stock price analysis.

For the machine learning experiment, a Random Forest classifier is created that predicts stocks' price movements based on a combination of market parameters and sentiment-based features. The model learns the patterns and trends effectively and establishes that the inclusion of sentiment analysis improves prediction accuracy. Key features of improved model performance through feature engineering include lagged stock prices, rolling statistics, percentage changes, and sentiment scores.

5.3 Key Findings

The major findings of this research can be summarized as follows:

a. Sentiment has a measurable impact on stock prices

Statistical analyses of the results indicated that the market sentiment obtained from financial news would move in one direction with stock price trends. Positive sentiment or happy news is generally followed by an upward movement of stock price, whereas adverse news tends to decrease prices before the actual price declines. The results favor the view that financial news sentiment can act as a leading indicator of market trends.

b. Feature engineering enhances prediction accuracy

Integrating additional features like lagged prices, rolling means, and volatility indicators-with sentiment features greatly improved its performance. This incorporation of prior stock price movements with market sentiment trends allowed the model to better understand short-term price fluctuations.

c. Random Forest is effective for sentiment-driven stock price prediction

Random Forest Classifier showed a remarkably good capability to predict stock price movement patterns based on sentiment and market features; the ensemble learning concept deals well with the non-linear relationships between the variables and is very suitable for predicting stock market behavior. Future work highlights that deep learning models may enhance the predictive performance even more.

d. Sentiment trends influence market volatility

The study also pointed out the relationship between trends in sentiment and volatility in stock market prices. Spells of extreme positive and negative sentiment were highly correlated with greater market fluctuations, and hence using sentiment analysis would likely be the case for an early warning indicator of impending volatile market conditions.

This is a successful study into the importance of sentiment analysis in predicting stock markets. Ultimately, readers and clients would appreciate this work as it combines historical stock data with the result of financial news sentiment for better insight in decision-making for profitable trades by investors, analysts, and financial institutions.

5.4 Future Work

Although this research provides a solid ground on which studies on sentiment-based stock market prediction may be built, it still has quite a lot to present in terms of further research and improvements. Some of these include:

a. Expanding data source

Currently, it restricts its scope: sentiment data is drawn exclusively from financial news. Stock market sentiment is affected by all possible things such as social media and investors' opinions and earnings reports and macroeconomic indicators. Possible great improvements in sentiment analysis in the future might have Social media platforms (e.g,. Twitter, Reddit, StockTwits) for capturing real-time

investor sentiment. This additional data from the identified sources can provide a complete understanding of sentiment-driven market trends.

b. Incorporating Advanced Deep Learning Models

While Random Forest delivered a solid benchmark for stock price prediction, there is still hope for further improvement by the use of the following: deep learning architectures like LSTM networks or even transformer-based models (such as GPT-4 or T5). Since LSTM networks are very much time-series-prediction sensors, they can also be chosen to represent the longer trends of stock prices considering their long memory and capability to be trained on time-frame predictions only. Sentiment analysis within transformer models (e.g., FinBERT, GPT-4) can perform better because they capture the contextual meaning and nuanced sentiment of financial text. This transient expectancy may yield more effective prediction models with the combination of LSTM and transformer models.

c. Refining Sentiment Analysis Techniques

Presently, the mood classification mechanism of news identifies it as either positive, neutral, or negative. But when it comes to financial markets, emotions get different. Hence, the future requirements of sentiment classification research can be as follows: Aspect-Based Sentiment Analysis (ABSA) for identifying the opinions that are related specifically to a company versus general market sentiment. Using financial-specific sentiment lexicons to boost classification accuracy. Detecting multi-sentiment tones from a single article through multi-label classification. By improving sentiment analysis techniques, researchers can capture different micro-sentiment trends, thus improving prediction accuracy.

5.5 Conclusion

Through this research, it is evident that financial news sentiment plays such an important role in stock price movements that a machine learning model can efficiently predict stock market movement using sentiment data alone. Using FinBERT for sentiment analysis and Random Forest for prediction, existing work would greatly contribute to the foundation for further research in sentiment-driven financial forecasting approaches.

Further research should aim toward broadened data sources, deep learning techniques, improved sentiment classification, and a real-time trading model. These developments will further enhance accuracy, interpretation, and real-world applicability of sentiment models in stock prediction.

Through this research, field would now contribute to a developing course on increasing sentiment-aware financial modeling, while it demonstrates that sentiment indeed renders stock market forecasting easier and allows a better outcome for investments.

References

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157–166.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8.

Box, G. E. P., & Jenkins, G. M. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2), 15-21.

Daas, P. J., Puts, M. J., Buelens, B., & van den Hurk, P. A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31(2), 249-262.

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, 270(2), 654-669.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399.

Kim, H., & Won, J. (2020). Hybrid models combining sentiment analysis and LSTM for stock price prediction. *Expert Systems with Applications*, 143, 113085.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.

Mittal, A., & Goel, A. (2012). Stock prediction using Twitter sentiment analysis. *Stanford University Research Paper*, 15(2), 1-5.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. *Expert Systems with Applications*, 41(16), 7653-7670.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3), 425-442.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139-1168.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267-307.

Tsay, R. S. (2005). *Analysis of Financial Time Series*. John Wiley & Sons.

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Araci, D. (2019). FinBERT: A pre-trained language model for financial communications. *arXiv preprint arXiv:1908.10063*.