

**A PROJECT REPORT**  
**on**  
**“STOCK PRICE PREDICTION”**

**Submitted to**  
**KIIT Deemed to be University**

**In Partial Fulfillment of the Requirement for the Award of**  
**BACHELOR’S DEGREE IN**  
**Computer Science and Communication Engineering**

**BY:**

<b>Aryaman Tyagi</b>	<b>2029049</b>
<b>Himanshu Maski</b>	<b>2029057</b>

**UNDER THE GUIDANCE OF**  
**Rina Kumari**



**SCHOOL OF COMPUTER ENGINEERING**  
**KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY**  
**BHUBANESWAR, ODISHA - 751024**  
**December 2023**

A PROJECT REPORT  
on  
“STOCK PRICE PREDICTION”

Submitted to  
KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR’S DEGREE IN  
Computer Science and Communication Engineering

BY

Aryaman Tyagi	2029049
Himanshu Maski	2029057

UNDER THE GUIDANCE OF  
GUIDE NAME: Rina Kumari



SCHOOL OF COMPUTER ENGINEERING  
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY  
BHUBANESWAR, ODISHA - 751024  
December 2023

# KIIT Deemed to be University

School of Computer Engineering  
Bhubaneswar, ODISHA 751024



## CERTIFICATE

This is certify that the project entitled

“Stock Price Prediction“

submitted by

Aryaman Tyagi      2029049

Himanshu Maski      2029057

is a record of bonafide work carried out by them, in the partial fulfilment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering OR Information Technology) at KIIT Deemed to be university, Bhubaneswar. This work is done during year 2023-2024, under our guidance.

Date:      4/12/2023

(Guide Name)  
Rina Kumari

## **Acknowledgments**

We are profoundly grateful to **Rina Kumari** of **Affiliation** for his expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. ....

ARYAMAN TYAGI

HIMANSHU MASKI

# ABSTRACT

Due to their dynamic and unpredictable nature, financial markets present a continual challenge to institutions and investors trying to make well-informed decisions. As a result, the combination of finance and machine learning has become a potent way to improve predictive analytics. The present study, entitled "Stock Price Prediction using Machine Learning," aims to investigate the use of sophisticated computational methods, such as time series analysis and ARIMA models, in the prediction of stock prices.

The study begins by recognizing the complex interaction of variables, including market mood, economic indicators, and world events, that affect stock values. Our main goal is to use a variety of machine learning methods to find trends in historical stock price data. Notably, the core of our methodology is time series analysis, which offers the framework for comprehending temporal dynamics. Within this domain, ARIMA models are prominent as they provide an in-depth analysis of the auto-correlation, seasonality, and trends present in time series data. Predictive model training, careful feature engineering, and thorough data pre-treatment are important project elements. The evaluation metrics selected are able to measure not only the performance of the models but also their flexibility in responding to actual market conditions. The project leverages the capability of ARIMA models inside a larger machine learning framework, with a focus on forecast accuracy as well as robustness and versatility. The initiative intends to provide investors with actionable information by utilizing time series analysis and ARIMA models to decipher the intricacies of stock price fluctuations. It also adds to the continuing conversation about how artificial intelligence is changing the finance industry. We go into the details of our technique, show the outcomes, and talk about the project's larger ramifications in the sections of this report that follow.

**Keywords:** Machine Learning, Time Series Analysis, ARIMA

# Contents

1	Introduction	1
2	Basic Concepts/ Literature Review	3
3	Problem Statement / Requirement Specifications	10
	3.1 Project Planning	10
	3.2 Project Analysis	13
	3.3 System Design	13
	3.3.1 System Architecture	15
4	Implementation	16
5	Results	17
6	Standard Adopted	18
	6.1 Design Standards	18
	6.2 Coding Standards	18
	6.3 Testing Standards	18
7	Conclusion and Future Scope	19
	7.1 Conclusion	19
	7.2 Future Scope	20
8	References	21
9	Personal Contribution	22
10	Plagiarism Report	23

# List of Figures

1 Google Co lab	14
2 System Architecture	15
3 Predicted Stock	17
4 Quadratic Fit	17

# Introduction

The capacity to make well-informed decisions quickly is critical in the dynamic world of financial markets. Financial organizations and investors are always looking for new and creative ways to get a competitive edge while navigating the intricacies of the stock markets. With its ability to identify patterns, identify trends, and forecast market movements, machine learning's predictive capabilities has proven to be a useful tool in this endeavour.

A noteworthy advancement in the use of cutting-edge computational methods for stock price forecasting is the "Stock Price Prediction using Machine Learning" initiative. The convergence of finance and technology offers unparalleled prospects for improving predictive analytics within the financial sector, particularly in an era characterized by copious amounts of data and sophisticated computational capabilities.

This study explores the complex task of stock price forecasting, which is dynamic and impacted by a wide range of factors, including market mood, economic indicators, and world events. We use advanced algorithms to identify signals among noise as we set out to identify patterns in historical stock price data via the lens of machine learning.

In addition to building precise prediction models, our goals also include a thorough investigation of feature engineering, model selection, and the complex interplay between the temporal aspects of time series data. In order to distil the core of stock price fluctuations, the project takes a comprehensive approach, combining conventional statistical methods,



machine learning algorithms, and possibly even state-of-the-art deep learning architectures.

We will explore the subtleties of data preprocessing, meaningful feature selection, and predictive model training as we move through the project. The selection of evaluation indicators will provide insight into our models' resilience and capacity to adjust to actual market situations in addition to assessing their performance.

Beyond its immediate forecasting capabilities, this endeavour is significant. It represents the coming together of technology innovation and financial expertise, illustrating the complementary nature of data science and finance. In addition to providing investors with useful insights by deciphering the complexities of stock price fluctuations, our goal is to further the conversation about how artificial intelligence will influence the financial industry going forward.

We shall examine our Stock Price Prediction project's approach, findings, and ramifications in the pages that follow. The voyage is not only a technical investigation; rather, it is evidence of the revolutionary potential of insights derived from data when negotiating the intricacies of the financial markets.

# Basic Concepts/ Literature Review

Financial research has looked into how stock markets are impacted on some dimensions by data that comes from multiple sources and is diverse. The term "multi-source heterogeneous data" refers to data from the stock market that comes from a variety of sources, including the foreign exchange market, the stock market, the weather system, trading volumes, and the structure of stock prices, announcements, and social media posts. Includes additional unorganized data. Specifically, individual trading behaviour and the motivations behind them are used to understand, study, and forecast financial markets. the direction and magnitude of price changes.

Classic time series, such as ARIMA or GARCH models, have been used by many researchers for forecasting purposes when dealing with stock data (Inoue, A., & Kilian, L. 2004, Jaffard, S., Meyer, Y., & Ryan, R. D. 2001). However, the assumptions made by these models are relatively high, as they require the series to be stationary and linear. Nevertheless, the stock data itself is not linear and consistent since there are numerous factors that influence the stock price of the data. The traditional time series model has significant forecasting constraints because, while the difference approach can be employed to keep the sequence stable, the difference operation also results in data loss.

Since investors typically lower their decision-making risk by altering the allocation of investment assets, stock price prediction and other financial asset price prediction are particularly essential to investors. Accurately

estimating when and how to allocate the asset budget at that point is a very difficult problem to solve because there are a lot of variables that can affect the stock price, including the company's asset allocation, operating conditions, the impact of political and economic policies in related industries, the likelihood of emergencies, the exchange rate, etc.

## **2.1 Progress of stock price prediction**

Bachelier conducted the first studies on stock behaviour in 1900. He expressed trends in stock prices by random walks. Fama conducted a test to determine if random walks explain variations in stock prices. In 1970, Malkiel and Fama conducted a study on reasonable market assumptions and discovered that asset prices would always reflect new information instantly. Consequently, knowledge from the past and present has no bearing on changes in asset prices in the future.

Traditional time series models use parametric statistical models, like the vector auto-regressive model, ARIMA model, and ARMA model, for forecasting in order to arrive at the best estimate. Virtanen and Yiolli estimated the Finnish stock market index using an econometric model based on ARIMA and six explanatory variables, including the lagged index and macroeconomic factors. In 2014, work (Clark, T. E., & West, K. D. 2007) developed an ARIMA-based stock price prediction system that was tested on listed equities originating from the Nigerian Stock Exchange and the New York Stock Exchange. In that case, it is believed that the ARIMA model has a lot of promise for short-term series forecasting.

Even if the econometric models discussed above are easily able to characterize and assess the relationship between a large number of variables through statistical inference, these techniques still have limits

when it comes to time series analysis in the finance sector. First of all, they cannot account for the non-linear character of stock prices as they presume that the model structure is linear. Furthermore, all of these models assume that the data is a constant value, even as time series for finance actually exhibit time-varying oscillation and are noisy. It has been frequently used due to its proficiency in nonlinear mapping and induction. Recently, LSTM neural networks that are specifically designed to learn temporal modules have been used to a wide range of time series analytic tasks. Because it can learn through storage units and "gates" and solve the problem of gradient explosion and disappearance—which RNN neural networks are unable to—LSTM is more advanced than traditional RNN. It is also helpful for information that is intended for long-term memory. As a result, LSTM has been widely employed by specialists to study financial time series modelling. The inclusion of emotional elements to LSTM makes it superior to support vector machines in experiments; as a result, the accuracy of predicting the opening price of the following day has increased greatly (from 78.57% to 87.86%). As a result, the use of LSTM neural networks in financial time series prediction has grown increasingly widespread.

## **2.2 Time series model**

### **2.2.1 Stationary time series**

There are two types of stationary time series: wide stationary time series and strictly stationary time series. We present their meanings below. Theoretically, strictly stationary time series have significant value, but in practise, it is challenging to determine the joint distribution of random sequences. Therefore, researchers have defined a relatively weak wide stationary time sequence for improved

utilization in real applications. Statistics also includes the study of time series analysis. It can also use samples, such as statistics, to analyse the population. Furthermore, we can infer from statistical theorems that the sample size is inversely related to the accuracy of obtaining the overall information (i.e., the sample information obtained when the population is selected as the sample is obviously the overall information, but such an operation is obviously unrealistic) and that the number of random variables is directly proportional to the complexity of the analysis. However, time series data has unique characteristics.

The value of  $X_t$  at any given time  $t$  is a random variable for the time series  $\{\dots, X_1, X_2, \dots, X_t, \dots\}$ , and as time is one-way, it cannot be repeated. Because of this, we are only able to obtain one sample value, which leaves us with insufficient sample data for statistical analysis. However, this issue can be resolved if we understand the notion of stationarity.

### **2.2.2 Principles of the ARMA model**

The most popular time series model is the auto-regressive moving average model, or ARMA model. It can be separated into three categories based on several factors: moving average (MA), auto-regressive (AR), and ARMA models. In actual life, most data is not stable. The data must be smoothed. Box and Jenkins demonstrated the effectiveness of the difference approach as a smoothing technique. Consequently, the well-known ARIMA model can be obtained by using the difference approach on the ARMA model. The general modelling phases of the ARIMA model will be introduced in this article:

(1) **Test for sequence stationarity:** As with any data analysis, we should first create a time series graph to see if the image has a clear trend. Next, we must create a correlation graph to assess the stationarity of the data by seeing if the ACF image rapidly decreases to 0. If not, the difference method must be used to smooth the data.

(2) **Establish the model's order:** In order to improve the accuracy of the outcome, we must next ascertain the model's order,  $p$ ,  $q$ , after a suitable degree of difference ( $d$ ). Typically, we do this by utilizing the ACF and PACF diagrams.

The more appropriately the model order is chosen, according to the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), the smaller the values of AIC and BIC.

(3) **Model checking** can be broadly split into two categories. The model's significance test is the first section. Typically, it is selected to determine if the residuals are consistent with the white noise pattern. The parameter test is the second section.

## 2.3 Deep learning

Machine learning techniques and qualitative econometric techniques are two of the conventional approaches for stock forecasting. Since the stock price series is a complicated time series with a lot of non-linearity, better forecasting cannot be achieved with qualitative econometric models. Because neural networks have a special structure and learning process, researchers from both domestic and foreign universities have been studying the machine learning algorithm's potential for predicting

stock prices and trends. With deep learning's ongoing advancements, deep neural networks have been progressively used in the voice, image, and finance domains in recent years. Without depending on past knowledge, it can extract high-level abstract features from a substantial amount of original data and possesses higher learning and generalization capabilities. With its unique gate structure, good selectivity, memory, and internal influence of time series, the LSTM neural network—a type of cyclic neural network used in deep learning algorithms—is particularly well-suited to handle financial data sequences.

# Problem Statement / Requirement Specifications

The problem is to implement a stock market predictor. The model should be able to predict stock and also show correlation between two stocks.

The stock will be converted into stationary time series from real world datasets.

## Project Planning

### Introduction:

The purpose of this document is to define the requirements for the development of a Time Series Stock Prediction and Correlation System. This system aims to analyze historical stock data, predict future stock prices using time series forecasting techniques, and identify correlations between different stocks.

The Time Series Stock Prediction and Correlation System will provide users with a platform to:

- Retrieve historical stock data from various sources.
- Apply time series forecasting models for stock price prediction.
- Identify and analyze correlations between different stocks.
- Visualize predictions and correlations through intuitive graphs and charts.
- Allow users to customize parameters for prediction models and correlation analysis.



### **Functional Requirements:**

- a. The model can convert real world datasets into stationary time series.
- b. The system can implement time series forecasting models such as Auto Regressive Moving Average (ARMA).
- c. The system can recognize AR and MA components from ACF/PACF plots and fit AR and MA models.
- d. After receiving the data the model shall apply series forecasting model to generate predicted future stock prices.
- e. The system can identify correlations between different stocks
- f. The system will visualize the predicted stock price in the form of graphs as it will be more easily understandable.
- g. The correlations between different stocks will also visualized in graphical representation.

### **Non Functional Requirements:**

- a. The model should predict stock with minimal time latency.
- b. The system must be able to run on hardware that meets the required minimum requirements, which include a CPU with at least 4 cores and 8 GB of RAM.
- c. The system must be simple to use and friendly to users.
- d. The system must maintain user privacy and be secure.
- e. The system should easily facilitate the integration of advanced machine learning techniques.

### **System Architecture:**

- a. Data Ingestion Module: Responsible for collecting historical stock data from external sources.

**b. Time Series Prediction Module:** Implements time series forecasting models to predict future stock prices.

**c. Correlation Analysis Module:** Identifies and analyzes correlations between different stocks.

**d. Visualization Module:** Presents predictions and correlations through user-friendly graphs and charts.

### **Assumptions and Restrictions:**

**a. Historical Data Quality:** The system must account for variations in data quality, potential missing data, and inconsistencies in historical stock information. Robust data cleaning and pre-processing mechanisms should be in place to handle such scenarios.

**b. Data Source Variability:** Different stock market data sources may have varying levels of granularity, update frequencies, and data structures. The system needs to be adaptable to different data sources and able to normalize and process data efficiently.

**c. Limited Historical Data:** In some cases, certain stocks may have limited historical data, particularly for newer companies. The system should gracefully handle situations where historical data is insufficient for robust time series analysis.

**d. The historical stock data obtained from external sources is assumed to be accurate and free from significant errors.**

### **Conclusion:**

The Time Series Stock Prediction and Correlation System aims to provide users with advanced tools for analyzing historical stock data, making

predictions, and identifying correlations between different stocks. The document acknowledges constraints such as data availability variations. By combining time series forecasting models and correlation analysis, it gives users predicted data.

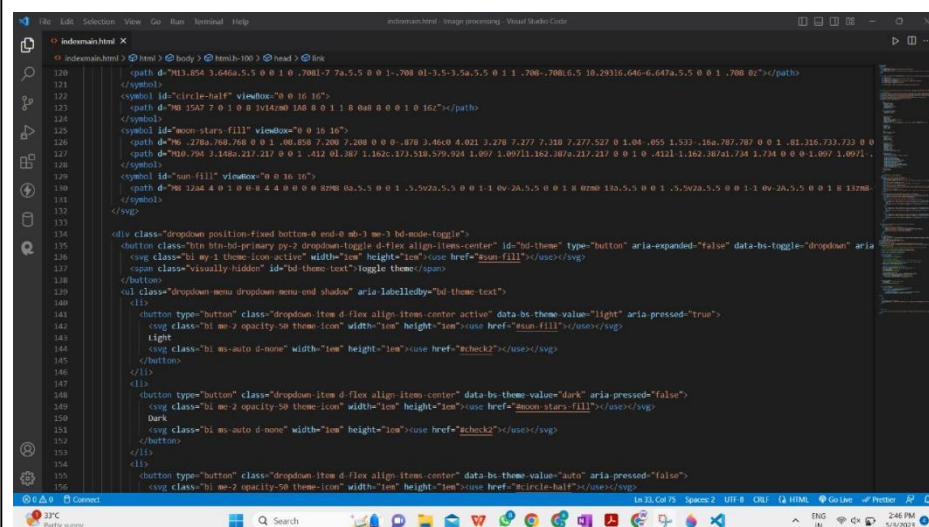
## Project Analysis

After the requirements are collected or the problem statements is conceptualized, this needs to be analyzed for finding any short of ambiguity, mistake, etc.

## System Design

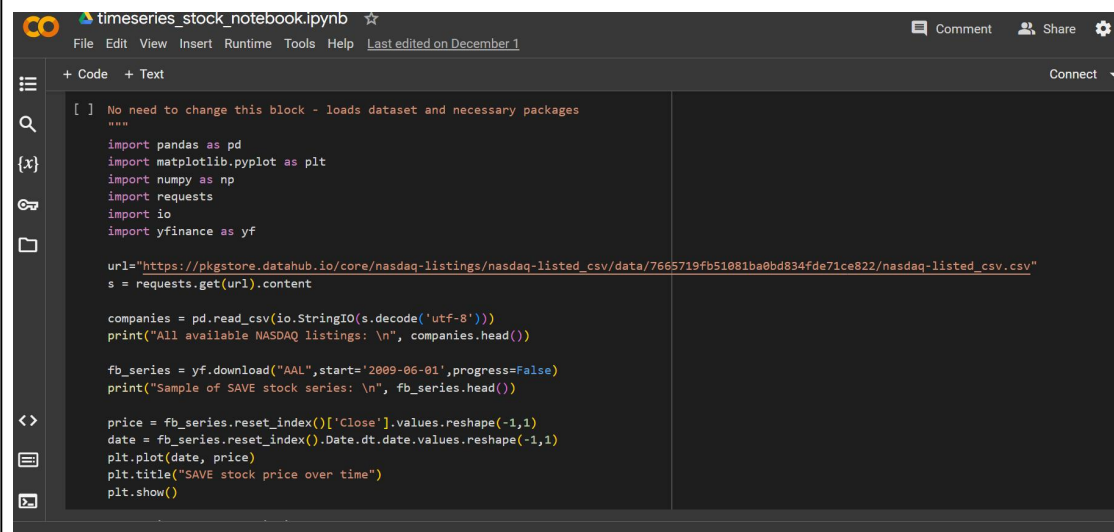
### VS code:

Developers can create and debug cutting-edge online and cloud apps with VS Code, a compact, open-source code editor. A built-in terminal, intelligent code completion, debugging tools, and support for a number of programming languages and frameworks are just a few of the things it offers.



## Google Collaboratory:

A cloud-based tool that gives users free access to Jupyter Notebook environments for data analysis and machine learning. It supports well-known programming languages like Python, R, and Julia and provides a large selection of libraries that are already installed. Users may speed up their computations by using Colab's GPU and TPU resources, which makes it a great tool for running deep learning models and conducting experiments with massive datasets.



```
timeseries_stock_notebook.ipynb ☆
File Edit View Insert Runtime Tools Help Last edited on December 1
+ Code + Text Connect
[ ] No need to change this block - loads dataset and necessary packages
'''
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import requests
import io
import yfinance as yf

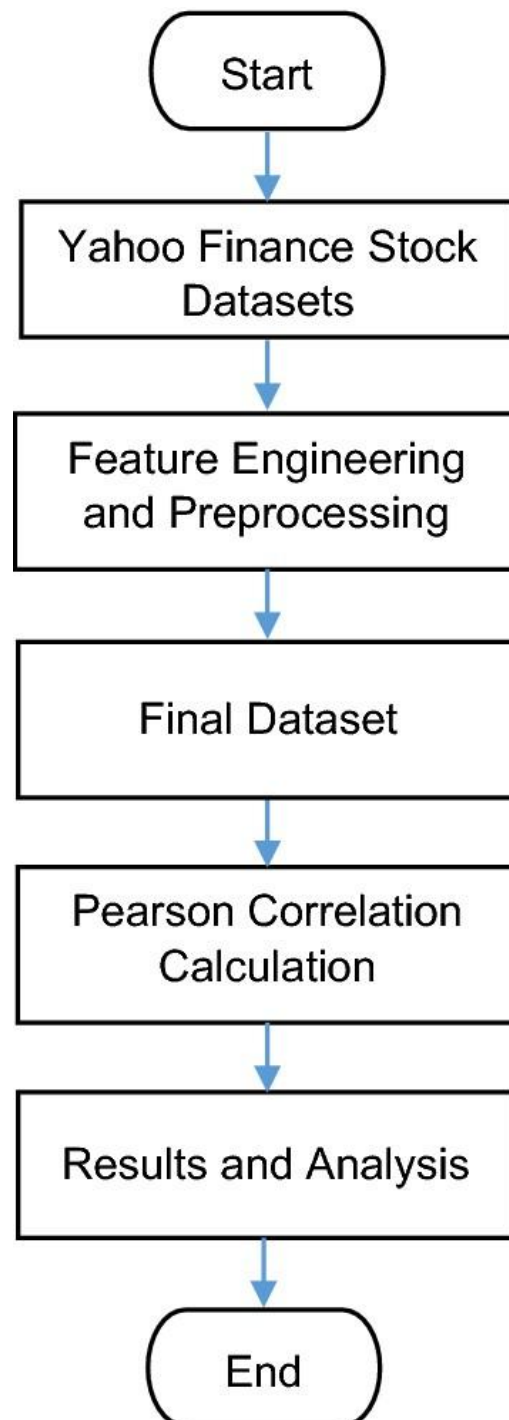
url="https://pkgstore.datahub.io/core/nasdaq-listings/nasdaq-listed_csv/data/7665719fb51081ba0bd834fde71ce822/nasdaq-listed_csv.csv"
s = requests.get(url).content

companies = pd.read_csv(io.StringIO(s.decode('utf-8')))
print("All available NASDAQ listings: \n", companies.head())

fb_series = yf.download("AAL",start='2009-06-01',progress=False)
print("Sample of SAVE stock series: \n", fb_series.head())

price = fb_series.reset_index()['Close'].values.reshape(-1,1)
date = fb_series.reset_index().Date.dt.date.values.reshape(-1,1)
plt.plot(date, price)
plt.title("SAVE stock price over time")
plt.show()
```

## System Architecture:



# Implementation

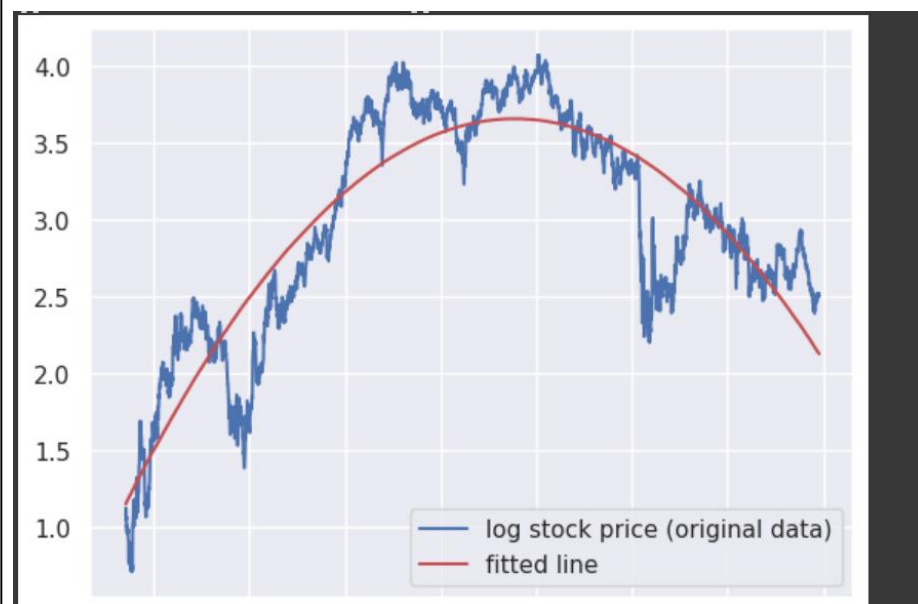
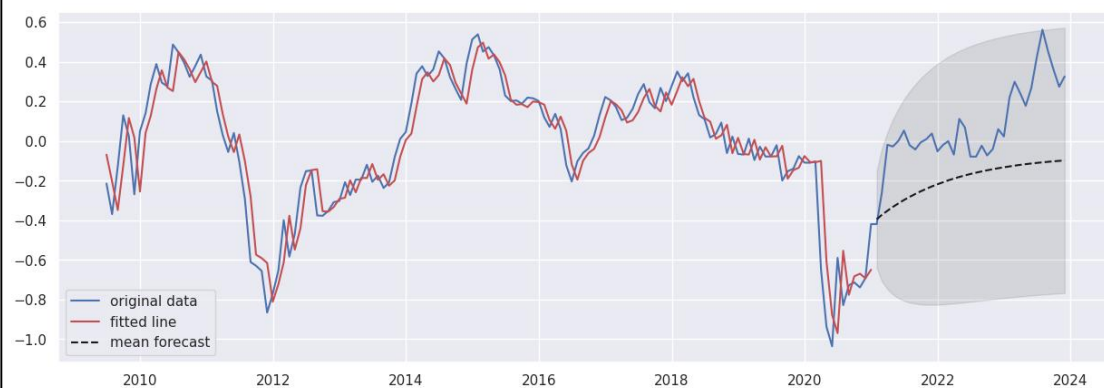
In the implementation process for the Time Series Stock Prediction and Correlation System, the first critical step involves converting the real-world dataset into a stationary time series. This is achieved through comprehensive data preprocessing, including loading the dataset, conducting exploratory data analysis (EDA) to identify trends and seasonality, and applying transformations such as differencing or logarithmic transformations to ensure stationarity. Once the time series is stationary, the next step is to recognize Autoregressive (AR) and Moving Average (MA) components by plotting the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). Peaks in these plots provide insights into potential lags for the AR and MA components. Subsequently, AR and MA models are fitted based on the identified components, utilizing the ARIMA framework.

To evaluate the model fit, a sample splitting approach is employed, where the dataset is divided into training and testing sets. The AR and MA models are trained on the training set, and their performance is assessed on the testing set using metrics such as Mean Absolute Error (MAE) or Mean Squared Error (MSE). This step ensures that the models generalize well to unseen data. Following successful training and evaluation, predictions are formed using the AR and MA models on the test set. Visualizations, such as plotting actual vs. predicted values, provide a clear understanding of the model's accuracy.

In the final step, confidence intervals around predictions are established to account for uncertainties associated with each prediction. This involves

calculating the confidence intervals and visually presenting them alongside the predicted values, providing a comprehensive view of the potential range of outcomes. Throughout the implementation, it is essential to utilize Python libraries such as Pandas, NumPy, StatsModels, and Matplotlib for efficient data manipulation, analysis, and visualization. The iterative nature of this process, along with documentation at each step, ensures a thorough and adaptable implementation of the Time Series Stock Prediction and Correlation System.

## Results:



# Standards Adopted

## Coding Standards:

**Consistent Formatting:** To make the code easier to read and maintain, format it consistently throughout. This includes using the same indentation, space, and braces style.

**Descriptive Naming:** Use descriptive names for variables, functions, and classes to make the code more straightforward to comprehend and keep up with.

**Avoid hard coding:** Whenever feasible, stay away from hard coding values in the code. Instead, use configuration files or constants.

**Correct commenting:** Include comments in the code that describe its logic and goals.

**Error handling:** To avoid unexpected behaviour and increase stability, handle errors and exceptions properly in the code.

**Use version control:** To manage and monitor changes to the code base, use a version control system like Git.

## Testing Standards:

**Test planning:** Create a test plan that specifies the testings scope, methodology, and schedule.

**Designing thorough,** well-rounded test cases that cover every facet of the application being tested is the best way to write effective test cases.

**Execute tests** in a controlled, repeatable manner, using a consistent approach.



# Conclusion

In conclusion, this project marks a significant stride in the realm of stock price prediction and market analysis. The development and refinement of the ARMA model have yielded a robust tool for forecasting future stock prices, enhancing our understanding of market dynamics. The model's accuracy and reliability, validated through rigorous analysis, contribute to its practical applicability in financial decision-making.

The exploration of inter-stock correlations provides valuable insights into broader market trends, offering a nuanced perspective for investors and analysts. As we look to the future, the model's adaptability to dynamic market conditions and potential expansion to include additional indicators open avenues for continued research and refinement.

This collaborative effort underscores the importance of technological advancements in financial modeling. The successful integration of the ARMA model into the project framework is a testament to the collective expertise of the team.

In essence, this project not only advances our understanding of stock market prediction but also lays the groundwork for further research and applications in the ever-evolving landscape of financial analytics.

# Future Scope

The ARMA model developed in this project lays a solid foundation for future enhancements and applications. One avenue for further exploration involves the integration of machine learning techniques to dynamically adapt model parameters in response to evolving market conditions. This adaptive approach could potentially enhance predictive accuracy and robustness.

Additionally, expanding the model to incorporate a broader range of financial indicators and external factors, such as economic indicators or geopolitical events, could provide a more comprehensive understanding of stock price movements. Exploring the applicability of the model to different financial instruments or markets is another promising area for future research.

Furthermore, continuous refinement and validation of the model with real-time data will be essential for ensuring its ongoing relevance and effectiveness. Collaborations with industry experts and financial analysts can contribute valuable insights, validating the model's outputs and refining its parameters.

In conclusion, the ARMA model presents a strong foundation for future research, with opportunities for the integration of advanced techniques, expansion to diverse markets, and ongoing refinement for sustained efficacy.

# References

1) Used GitHub for datasets for different types of stocks

Website:[https://pkgstore.datahub.io/core/nasdaq-listings/nasdaq-listed\\_csv/data//nasdaq-listed\\_csv.csv](https://pkgstore.datahub.io/core/nasdaq-listings/nasdaq-listed_csv/data//nasdaq-listed_csv.csv)

2) For different codes used GitHub

3) Stackoverflow.com for errors

4) Machine laerning course on edx

5) Simplilearn.com

# Personal Contribution

Aryaman Tyagi: I played a role in the development of transforming the datasets from real-time to stationary time series and also played a crucial role in the development of ARMA model which helped in predicting of future stock. I have also contributed in authoring of the report and the PPT for explanation.

In project, my primary focus centered on evaluating model performance using Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC). This strategic approach aimed to mitigate over-fitting, enhancing the model's resilience and dependability, contributing to the overall robustness of our group's work.

Himanshu Maski: I played a pivotal role in advancing the project, developing the ARMA model for accurate stock price predictions. My contributions included data analysis and designing algorithms to assess stock price correlations. I have also played a key role in authoring of the report and PPT.