D214 Data Analytics Graduate Capstone

Task 2: Data Analytics Report and Executive Summary

Jonathon “Jon” Fryman

Data Analytics 01/01/2022

Student ID 00974544

Program Mentor: Lea Yoakem

C: 419-206-6989

Table of Contents

[Part I: Research Question 4](#_Toc129794837)

[A. Summary of Original Research Question 4](#_Toc129794838)

[A1. Original Research Question 4](#_Toc129794839)

[A2. Research Question Justification 4](#_Toc129794840)

[A3. Description of Context for Research Question 4](#_Toc129794841)

[A4. Hypothesis Discussion 5](#_Toc129794842)

[Part II: Data Collection 5](#_Toc129794843)

[B. Report on Data Collection Process 5](#_Toc129794844)

[B1. Advantages of Data-Gathering Methodology 6](#_Toc129794845)

[B2. Disadvantages of Data-Gathering Methodology 6](#_Toc129794846)

[B3. Challenges 7](#_Toc129794847)

[Part III: Data Extraction and Preparation 8](#_Toc129794848)

[C. Describe Data Extraction and Preparation Process 8](#_Toc129794849)

[C1. Import Python Packages 8](#_Toc129794850)

[C2. Read in Existing Data 9](#_Toc129794851)

[C3. Join the DataFrames using Common Key 11](#_Toc129794852)

[C4. Review Data Types Contained within DataFrame 12](#_Toc129794853)

[C5. Review For Missing and Null Values 13](#_Toc129794854)

[C6. Review Statistical Summary and Plot Distribution 14](#_Toc129794855)

[C7. Tools and Techniques 16](#_Toc129794856)

[C8. Tool and Technique Justification 16](#_Toc129794857)

[C9. Tool and Technique Advantages 17](#_Toc129794858)

[C10. Tools and Techniques Disadvantages 17](#_Toc129794859)

[Part IV: Analysis 18](#_Toc129794860)

[D. Report on Data Analysis Process 18](#_Toc129794861)

[D1. Description of Analysis Technique 18](#_Toc129794862)

[D2. Exploratory Data Analysis 18](#_Toc129794863)

[D3. Statistical Testing 23](#_Toc129794864)

[D4. Creation of Time Series Model 27](#_Toc129794865)

[D5. Calculations Performed 32](#_Toc129794866)

[D6. Analysis Technique Justification 33](#_Toc129794867)

[D7. Analysis Technique Advantages 33](#_Toc129794868)

[D8. Analysis Technique Disadvantages 33](#_Toc129794869)

[Part V: Data Summary and Implications 34](#_Toc129794870)

[E. Summary of Data Analysis Implications 34](#_Toc129794871)

[E1. Data Analysis Implications 34](#_Toc129794872)

[E2. Data Analysis Results 34](#_Toc129794873)

[E3. Data Analysis Limitations 36](#_Toc129794874)

[E4. Recommended Course of Action 36](#_Toc129794875)

[E5. Future Data Study Directions 37](#_Toc129794876)

[F. Sources 37](#_Toc129794877)

[Figure 1: Python Imports 9](#_Toc130119643)

[Figure 2: Loading the individual .csv files 10](#_Toc130119644)

[Figure 3: Combining data into single DataFrame. 11](#_Toc130119645)

[Figure 4: Using .info() function to review DataFrame column datatypes. 12](#_Toc130119646)

[Figure 5: Reviewing DataFrame for null/missing values. 13](#_Toc130119647)

[Figure 6: Reviewing DataFrame statistics with .describe() function. 14](#_Toc130119648)

[Figure 7: Historical Distribution Plots 15](#_Toc130119649)

[Figure 8: Tableau Graphs Reflecting Closing Price 19](#_Toc130119650)

[Figure 9: Correlation Heatmap 20](#_Toc130119651)

[Figure 10: Autocorrelation Plots 21](#_Toc130119652)

[Figure 11: Seasonal Decomposition Plots 1 22](#_Toc130119653)

[Figure 11: Seasonal Decomposition Plots 2 23](#_Toc130119654)

[Figure 12: adfuller test code 24](#_Toc130119655)

[Figure 13: adfuller test results 25](#_Toc130119656)

[Figure 14: Training/Test Data Split Plot Code 25](#_Toc130119657)

[Figure 15: Training/Test Data Split Plots 1 26](#_Toc130119658)

[Figure 16: Training/Test Data Split Plots 2 27](#_Toc130119659)

[Figure 17: Auto Arima Function 28](#_Toc130119660)

[Figure 18: Auto Arima Output 1 29](#_Toc130119661)

[Figure 19: Auto Arima Output 2 29](#_Toc130119662)

[Figure 20: Auto Arima Output 3 30](#_Toc130119663)

[Figure 21: Auto Arima Output 4 31](#_Toc130119664)

[Figure 22: Auto Arima p, q, and d values 32](#_Toc130119665)

[Figure 23: ARIMA Metric Check 32](#_Toc130119666)

[Figure 24: Accuracy Metric Output 34](#_Toc130119667)

[Figure 25: MAPE Accuracy Metric Calculation 35](#_Toc130119668)

[Figure 26: MAPE Metric Output for APPL 35](#_Toc130119669)

[Figure 27: MAPE Scores 36](#_Toc130119670)

# Part I: Research Question

# Summary of Original Research Question

## A1. Original Research Question

To what extent can a company’s future daily per-share closing stock dollar value be accurately predicted?

With a multitude of options for perspective investors to choose from, it can often be difficult to identify the best investment opportunities as well as the most prudent time to invest. To explore these questions, this author will analyze data associated with value and volume for 12 of the largest American based tech companies.

## 

## A2. Research Question Justification

The justification for the outlined research question is the fiscal responsibility required when investing large sums of money. The outlined research aims to provide a deeper comprehension regarding the market performance for a subset of American based tech companies through the review and analysis of historical data related to their price and volume data dating back to January 2010.

This will facilitate an understanding regarding the existence of any seasonality, trends associated with volume/price change for the individual stocks, and general market fluctuations, and better equip investors when the information required to make prudent investments.

## A3. Description of Context for Research Question

The stock market can often be riddled with uncertainty and filled with great fluctuation. To facilitate educated investing, it is essential to ascertain trends and seasonality that exists within a given stocks performance prior to making a large investment.

As outlined in the preceding sections of this report, 12 of the largest American tech companies will have their market performance analyzed using the previous 13 years of available data to forecast expected future performance for the respective company's stock price.

Per the New York Stock Exchange’s available data, an average of $18.9 billion is traded on the stock market each day. With such a tremendous amount of money being invested, perspective investors stand to lose a substantial amount of funding if poor investments are made due to lack of research into the available data for a given stock.

Utilizing the historical stock information for the companies being analyzed, the objective is to develop a model capable of forecasting future stock performance with an accuracy that can provide a greater level of confidence for investors prior to committing large amounts of funding.

## A4. Hypothesis Discussion

The hypothesis being tested throughout this research project is based upon the performance of 12 American tech companies market performance, and if it can be accurately forecast. This will be achieved by training a prediction model on a stock’s historical performance, validating accuracy of a prediction model through comparison against a subset of the historical dataset that’s been excluded from training for validation purposes, and assessing the Mean Absolute Percentage Error(MAPE) of the predictions.

The following passage will further expand upon the null and alternate hypotheses for the current data analysis project in greater detail:

**Hypothesis**: The hypothesis of the current data analysis is that the Mean Absolute Percentage Error(MAPE) score associated with each predicted stock closing price will be below 20%. A MAPE score below 20% indicates the difference between the dollar value for the predicted and actual stock closing price is smaller than 20% which is a generally considered good MAPE score.   
  
**Null hypothesis**: The null hypothesis is that the MAPE score is 20% or greater. This would indicate less than 80% similarity on the predicted dollar value when compared to the actual stock closing price.  
  
**Alternate Hypothesis**: The alternate hypothesis is the predictions associated MAPE scores are less than 20%. A MAPE score in this range would indicate a greater than 80% similarity on the predicted dollar value in comparison to the actual stock closing price.

# Part II: Data Collection

# Report on Data Collection Process

The dataset used for the current analysis is a publicly available dataset obtained from <https://www.kaggle.com/datasets/evangower/big-tech-stock-prices>.

The full dataset originally consisted of 14 unique .csv files. Each file is associated with a specific tech companies stock performance for trading days beginning as early as January 2010. The companies included within the original dataset are:

* Adobe
* Amazon
* Apple
* Cisco
* Google
* IBM
* Intel
* Meta
* Microsoft
* Netflix
* Nvidia
* Oracle
* Salesforce
* Tesla

After a review of the data within the individual files, it was determined the Meta and Tesla datasets contained a smaller set of historical trading days in comparison to the remaining 12 companies. As a result, the data for these Meta and Tesla was excluded from the current analysis. The process used to determine the files associated with Meta and Tesla contained an inconsistent number of trading days in comparison to the others is reflected within section C of this report.

These columns contained within each of the original 14 .csv files are:

* **Date**: The specific date of a given trading day.
* **Open**: The starting period of trading on a securities exchange or organized over-the-counter market.
* **High**: The highest price a stock traded during the trading day.
* **Low**: The lowest price a stock traded during the trading day.
* **Close**: The closing price a stock traded during the trading day.
* **Adj Close**: The closing price after adjustments for all applicable splits and dividend distributions.
* **Volume**: The number of shares traded in a particular stock, index, or other investment over a specific period of time.

The dataset used for this analysis is provided under a CC0 1.0 Universal Public Domain Dedication license that allows users to share and adapt the data with proper credit given to the original data provider.

**Data Gathering**: The data-gathering methodology to be used for the current analysis is documents and records. This methodology makes use of existing data for gathering information. For this specific analysis, this includes examining existing records related to historical market performance for 12 tech companies over a period beginning in January 2010.

## B1. Advantages of Data-Gathering Methodology

Utilizing a CC0 1.0 Public Licensed Dataset has a multitude of advantages. Utilizing the Python package yfinance was initially considered to independently gather the data required for analysis. However, it was determined taking this step would take considerable time and effort to ensure the collected data was consistent across the subset of companies being analyzed.

Additionally, when utilizing a public dataset, it allows any number of data analysts to draw conclusions using the same core data. This can facilitate shared learning, and expand the insight drawn that may not have been otherwise observed when individual collecting data.

## B2. Disadvantages of Data-Gathering Methodology

Similar to the previously outlined advantages of utilizing a public dataset, there can similarly be disadvantages associated with this gathering methodology.

A primary disadvantage is a lack of control over the selection of included companies, time, or variables within the dataset.

Similarly, it also requires trusting the original data gatherer had properly gathered the information contained within the data. It provides little benefit to an individual or company to be performing analysis using inaccurate data.

In fact, using inaccurate data could directly lead to complications for a company. If a recommended course of action is developed based upon a set of data that inaccurately portrays the subject matter being analyzed, it could potentially harm a company’s revenue and public reputation.

## B3. Challenges

One of the primary challenges when analyzing a public dataset is ensuring the data is properly cleaned and able to adequately align with the selected research topic.

In a typical business environment, a specific question is posed that requires analysis to be completed. It would normally be at this time a data analyst would determine what data is required to perform the outlined analysis and that it be gathered in a way that properly aligns with the goals of the outlined analysis.

When utilizing a publicly available dataset, the data is often compiled independent of any specific research question. This could lead to a business needing to devote additional time and resources to ensure the data has been catered to the specific data analysis request.

For the current project, a considerable amount of time was spent ensuring a proper understanding of what information was contained within the dataset had been obtained. This also required performing exploratory analysis to determine what, if any, cleaning, and reconfiguration of the data was required.

For the currently used dataset, it initially contained a folder of 14 unique .csv files, each associated with a specific tech companies stock performance for a period often dating back to January 2010.

After the effort was made to determine if all the included companies had data available for the identical period, it was determined two of the companies, Meta and Tesla, had incomplete data for the complete time period.

Due to the remaining 12 companies contained within the dataset having data available for the same historical range of dates, it was decided to exclude Meta and Tesla from the current analysis.

The specific steps taken to analyze and clean the datasets selected for analysis have been outlined in the subsequent sections of this report.

# Part III: Data Extraction and Preparation

# Describe Data Extraction and Preparation Process

The process of extracting, cleaning, and preparing the data utilized a set of often Python packages such as Pandas, Numpy, and MatPlotLib. The details regarding the specifics of the packages are discussed in further detail within section C1 of this report.

The primary environment used throughout the preparation and analysis process was Jupyter Notebooks in partnership with Python version 3.8.

The individual steps taken to prepare the Jupyter Notebook environment, import the required libraries, and clean the original data are outlined in detail within the subsections of section C of this report.

To provide additional context into the specifics of each stage of the preparation process when possible, an explanation, justification, advantage, and disadvantage are included within each of the following subsections for review.

## C1. Import Python Packages

The packages utilized for the analysis performed within this report include:

* **MatPlotLib**: A comprehensive library for creating static, animated, and interactive visualizations in Python.
* **Numpy**: Offers comprehensive mathematical functions, random number generators, linear algebra routines, and more.
* **Pandas**: A library for data manipulation and analysis.
* **Seaborn**: A data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.
* **Sklearn**: Built upon NumPy, SciPy, and matplotlib. Provides simple and efficient tools for predictive data analysis.
* **statsmodels**: A package that allows users to explore data, estimate statistical models, and perform statistical tests.
* **pdarima**: A statistical library designed to fill the void in Pythons time series analysis capabilities.
* **datetime**: Supplies classes for manipulating dates and times.

As shown in the following image, the libraries required were collectively imported in the first cell of the Jupyter Notebook.

Graphical user interface, text, application

Description automatically generated

Figure : Python Imports

* **Explanation**: This step imports the packages deemed essential for the overall analysis to be performed.
* **Justification**: Utilizing the existing packages and modules removes the requirement of having to manually create similar functions to facilitate the common step of data preparation.
* **Advantage**: This step saves a significant amount of time and utilizes existing trial and error by the Python Package developer to ensure as many issues as possible have already been resolved.
* **Disadvantage**: There can be a slight learning curve in relation to determining what the existing package contains and is capable of. However, this is essential to ensure the available functionality is properly utilized with the proper arguments passed at runtime.

## C2. Read in Existing Data

As outlined on the image shown below, each of the original 14 .csv files that corresponded to a single company’s historical market performance was independently read in as a .csv file using the Pandas .read\_csv() function.

Also shown within this step was the determination of how many total days of historical data were available within each of the files. As can be seen within the output, the META and TSLA files did not contain a consistent number of historical datapoints when compared to the other 12 files. As a result, these two files were excluded from the subsequent stages of data cleaning and preparation.

Table

Description automatically generated

Figure : Loading the individual .csv files

* **Explanation**: Using the panda’s .read\_csv() function, a csv file can be converted into a DataFrame. Similarly, using the pandas .value\_counts() function makes the determination that specific files within the dataset did not contain a uniform number of values.
* **Justification**: The original dataset consisted of 14 unique csv files, making pandas the logical choice of package to read in the data for analysis. This is primarily due to being a well-used package with significant documentation available to assist in the process of reading in the data and determining the value count for each file.
* **Advantage**: Pandas is a widely adapted package with a significant amount of documentation available. This facilitates an ease of use few other Python packages have available.
* **Disadvantage**: In comparison to directly opening a .csv file with an application like Microsoft Excel, pandas requires additional steps be taken before being able to fully view the information contained within the data.

## C3. Join the DataFrames using Common Key

As reflected in the image shown below, the original DataFrame associated with each of the 12 tech companies were joined into a single DataFrame.

This was facilitated by creating an empty DataFrame with an index column associated with the dates found within the original .csv files. A column was then created for each of the companies was populated with the Close value associated with the specific trading day found in the index for each individual company.

The index of the new DataFrame was set with 3,271 periods beginning from 01/04/2010 to match the count found for the Date column within the preceding stage of preparation. A frequency of ‘B’ was used to signify the date range should only contain business days.

Graphical user interface, text, application, email

Description automatically generated

Figure : Combining data into single DataFrame.

* **Explanation**: Each of the 12 companies selected for analysis have their stock information contained within individual .csv files. To facilitate an easier process for training/testing future prediction models. these steps were performed to combine the relevant ‘Close’ column for all companies into a single DataFrame.
* **Justification**: Due to challenges associated with forecasting from 14 unique files, it was determined that combining each of the companies closing price for the specific trading days under analysis would be a more practical solution. This was facilitated by creating an empty DataFrame using the date as an index followed by adding a column for each of the companies labelled with their stock symbol. The column associated with each individual company was then populated with their closing price that associated with the trading day reflected in the row index.
* **Advantage**: It was determined to be substantially more practical and efficient to analyze a single DataFrame containing only the columns required for the current analysis. It would take considerably more time and effort to forecast across12 different files/DataFrames which also contained data not essential to the current research.
* **Disadvantage**: To prevent the single DataFrame from containing to much information, all columns besides the company's closing price data was excluded from the new DataFrame. If future review of data contained within the excluded columns is required, it will be necessary to either use the original 12 DataFrames or to create a new DataFrame at that time containing the relevant information.

## C4. Review Data Types Contained within DataFrame

Utilizing the pandas .info() method, the datatype for each column of a DataFrame can easily be reviewed. As shown in the image below, all columns are a float64 datatype except the Date column. The data column had earlier been converted into the appropriate datetime format.

Table

Description automatically generated

Figure : Using .info() function to review DataFrame column datatypes.

* **Explanation**: The preceding section contains the step taken to review the data type associated with each variable contained within the newly created DataFrame.
* **Justification**: To ensure the data contained within the DataFrame is in an appropriate form to be used for analysis, it is imperative to review the datatype of each variable contained within the DataFrame prior to performing analysis upon the data.
* **Advantage**: This step allows corrective action to be taken if it is determined a variable or variables required for the specified analysis are not of an appropriate data type.
* **Disadvantage**: If a variable is determined to have the incorrect datatype associated, additional corrective action will need to be performed. While performing the correction is not a disadvantage to the overall research, it would require additional steps to be taken that the .info() method is not equipped to perform.

## C5. Review For Missing and Null Values

Although there are a multitude of methodologies for determining if any columns within a DataFrame are absent of values, the following section shows two distinct methods that can be used to obtain this information.

First. a printout for each of the columns associated to a specific stock to reflect a total number of values contained within the respective column was created. This is achieved using the pandas .value\_counts() function.

Second, similar to the approach taken within section C2, utilizing the pandas .isna() function in partnership with .sum() provides confirmation that no rows within the DataFrame are absent of a value.

Table

Description automatically generated

Figure : Reviewing DataFrame for null/missing values.

* **Explanation**: The preceding section reviews the DataFrame to ensure there are an equal number of values for all columns contained.
* **Justification**: Prior to completing analysis, it is important to ensure there are no gaps in the available data that could skew the results or predictions. It is also important to understand any reasons a given set of data would be missing values.
* **Advantage**: The primary advantage associated with reviewing for missing data is to pre-emptively position yourself to address any missing values prior to allowing the missing data to skew the results of any performed analysis or predictions.
* **Disadvantage**: It can be difficult to discern the reasoning a given datapoint is absent just from the identification that there are an inequal count of values across the fields of a DataFrame.

## C6. Review Statistical Summary and Plot Distribution

As seen in the following image, the use of the .describe() function provides several statistics related to numeric columns contained within a DataFrame.

**Graphical user interface, application, table

Description automatically generated**

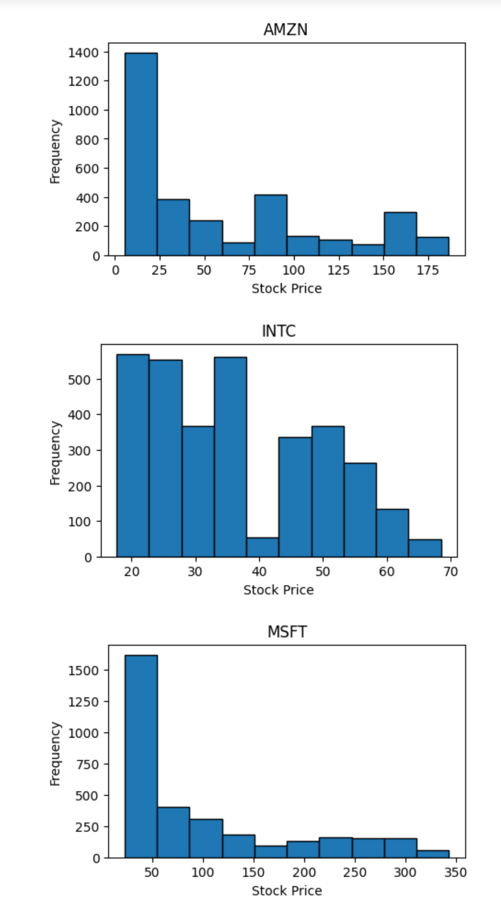
Figure : Reviewing DataFrame statistics with .describe() function.

These statistics are:

* **count:** A count of the total values included within the column.
* **mean:** The average amount contained within the column.
* **std:** The amount of variation or dispersion of a set of values.
* **min:** The minimum value contained within the column.
* **25%:** Represents results in the 25th percentile.
* **50%:** Represents results in the 50th percentile.
* **75%:** Represents results in the 75th percentile.
* **Max:** The maximum value contained within the column.

The following visualizations are the historical distribution plots for each of the companies included for analysis.

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated**

Figure : Historical Distribution Plots

* **Explanation**: The actions performed within this step provide a summary of the values found within each of the columns of the DataFrame. As outlined in the initial portion of this section, a multitude of statistics are provided that provide insight into the range of values associated with a specific column.
* **Justification**: It is important to understand the data to be analyzed. This step provides clear insight into the minimum, maximum, and average values for the closing stock prices associated with each individual company.
* **Advantage**: The insight gained can assist in identifying if potential outlier datapoints exist within the data that could skew the future analysis to be performed.
* **Disadvantage**: While this stage does provide statistics useful in detection of outlier data points, it does not provide information regarding the reasoning. An outlier value could be due to a datapoint collected in error or simply had an incorrect value entered when the data was created.

## C7. Tools and Techniques

The tools used for this analysis are as follows:

* **Jupyter Notebooks**: Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations, and narrative text. Its uses include data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more(K, 2020).
* **Python**: Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently whereas other languages use punctuation, and it has fewer syntactical constructions than other languages.
* **Tableau**: A powerful data visualization tool used in the business intelligence industry.

The primary technique used for this analysis was time series analysis. This is a technique in statistics that deals with time series data and trend analysis. Time series data follows periodic time intervals that have been measured in regular time intervals or collected in particular time intervals.

## C8. Tool and Technique Justification

The justification for the utilized tools and techniques are as follows:

* **Jupyter Notebooks**: As is outlined below in the advantages section of this report, Jupyter Notebooks is an excellent tool to use when having the results of executed code appear directly with the code itself is of benefit to the user.
* **Python**: Due to Python's support for an extensive set of libraries that allow easy exploration of data, the libraries facilitating time-series model creation, and my general familiarity with the language across a wide spectrum of uses, Python was determined to be the optimal programming language utilized for this analysis
* **Tableau**: Tableau makes it incredibly easy to create data visualizations directly in a user interface that allows creating filters and creating calculated fields to derive insight that would not otherwise be available.
* **Time Series Analysis**: Time series analysis is used when seeking to analyze the past to forecast the future. Due to the subject nature of the current research, this was found to be an optimal technique for forecasting future stock performance.

## C9. Tool and Technique Advantages

The advantages of the utilized tools and techniques are as follows:

* Jupyter Notebooks is great for showcasing your work/analysis. This is a result of both the code and results easily within the same cell of the notebook.
* The Python programming language is easy to read, learn, and write in comparison to many other programming languages.
* Python is a free and open-sourced language.
* Python has a vast, extensive set of libraries to facilitate nearly any objective a programmer/data analyst would be pursuing.
* Tableau quickly creates interactive visualizations that can be adjusted in the user interface. It provides easy selection of multiple visualization types depending on the type of data being provided for the specific worksheet.
* Time series analysis is specifically useful when attempting to identify patterns and/or forecast the future. This aligns perfectly for the subject matter of this project.

## C10. Tools and Techniques Disadvantages

The disadvantages of the utilized tools and techniques are as follows:

* When creating code in Jupyter Notebooks, it is very easy to end of with duplicate code rather than the standard creation of functions, classes, and objects. This can become difficult to maintain as your notebook grows.
* As Python is an interpreted language, it can be slow in comparison to languages like C/C++ or Java.
* Due to Python being a dynamically typed language, it can often lead to run-time errors and require more testing when compared to other programming languages.
* Tableau has a high cost associated with licensing that may make it impractical to use in many circumstances depending on the environment the data analysis is being performed(example: working for a large company or enrolled in an institution of higher learning).
* A primary disadvantage of time series analysis is it can sometimes struggle to accurately identify the correct model to represent data.

# Part IV: Analysis

# Report on Data Analysis Process

## D1. Description of Analysis Technique

Over the course of the performed analysis, several techniques were utilized. These techniques include:

* **Exploratory Data Analysis**
* **Statistical Testing**
* **Creation of Time Series Model**

Each of the three analysis techniques used will be further overviewed below.

* **Exploratory Data Analysis (EDA):** EDA refers to the critical process of performing initial investigations on data in an attempt at discovering patterns, spotting anomalies, and checking assumptions with the help of summary statistics and graphical representations(Patil, 2022).
* **Statistical Testing:** Statistic hypothesis testing is a method of statistical inference used to decide whether the data at hand sufficiently supports a particular hypothesis. Hypothesis testing allows the creation of probabilistic statements about population parameters(Wikipedia, 2023).
* **Creation of Time Series Model:** A Time Series is a collection of data points that are stored with respect to their time. Mathematical and statistical analysis is performed on this type of data to find hidden patterns and meaningful insights, which is referred to as *time-series analysis*. Time-series modeling techniques are used to understand past patterns from the collected data to try and forecast future horizons(Vishwas & Patel, 2020).

## D2. Exploratory Data Analysis

In partnership with the exploratory data analysis(EDA) performed in section C of this report, the following EDA was performed for the current dataset.

As shown below. a visualization was created using Tableau that reflects the average closing price by year broken down for each individual companies trading symbol. This specific Tableau dashboard features a box and whiskers plot and a line plot.

Chart, box and whisker chart

Description automatically generated

Figure : Tableau Graphs Reflecting Closing Price

Additionally, a correlation heatmap, autocorrelation plot, and seasonal\_decompose for all the companies included within the dataset was created.

The following image reflects the generated correlation heatmap. This heatmap reflects the correlation between stocks included within the created DataFrame.

## Table Description automatically generated with medium confidence

Figure : Correlation Heatmap

Next, the following image reflects autocorrelation plots that were generated for each company included in the analysis. The plots shown in the output were generated using the lag\_plot() function.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure : Autocorrelation Plots

Finally, the statsmodels seasonal\_decompose() function was used to generate seasonal decomposition plots for each of the analyzed companies. The seasonal\_decompose() function provides output reflecting trend, seasonality, and residuals. The seasonal\_decompose() output can be found below.

A picture containing application

Description automatically generated A picture containing graphical user interface

Description automatically generated

Figure : Seasonal Decomposition Plots 1

Graphical user interface

Description automatically generated with medium confidenceGraphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated with medium confidenceGraphical user interface, application

Description automatically generated

Figure 11: Seasonal Decomposition Plots 2

## 

## D3. Statistical Testing

The first statistical tests performed was to perform the augmented Dickey-Fuller test utilizing the adfuller() function for each column in the DataFrame. The adfuller test is a method of testing stationarity within a time series. As shown in Figure 13, none of the associated columns had a p-value < 0.05. This reflects that the data is non-stationary.

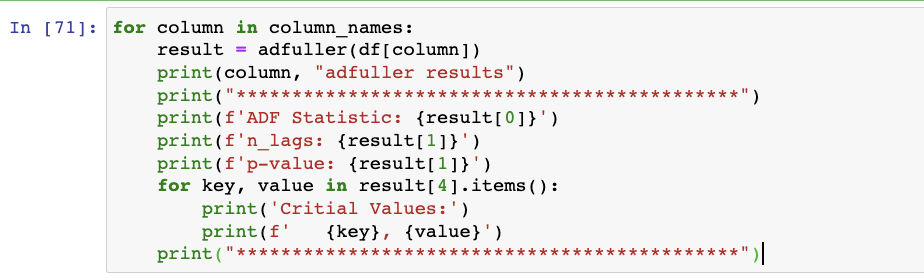


Figure 12: adfuller test code

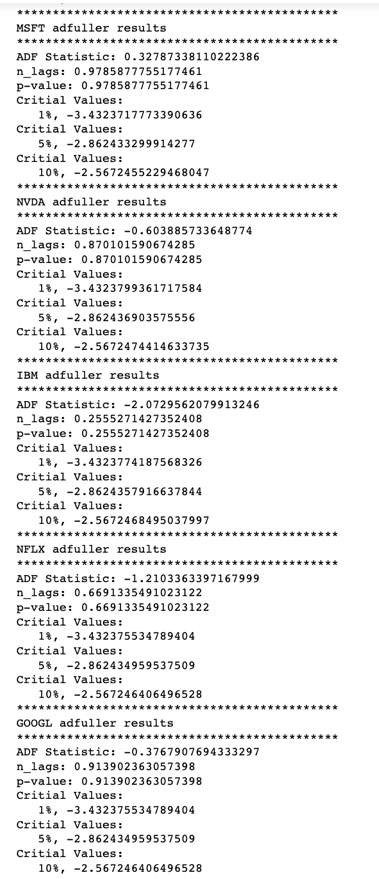
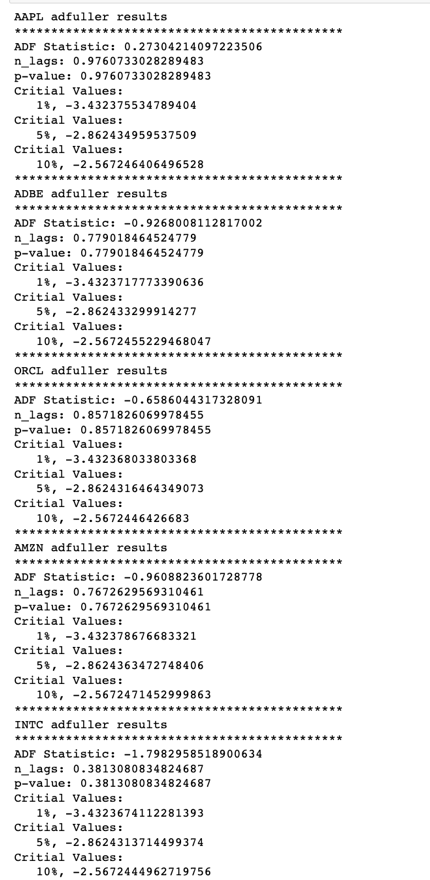
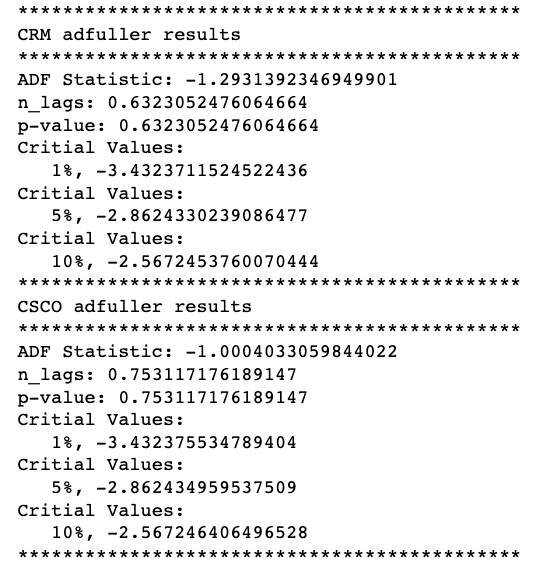
 

Figure 13: adfuller test results

The following visualizations exhibit the breakdown of how training and test data was created for each of the analyzed companies. The separation of training and test data can be identified by the change from blue to green in Figure 15-16.

Graphical user interface, text, application

Description automatically generated

Figure 14: Training/Test Data Split Plot Code

Chart, histogram

Description automatically generated Graphical user interface, chart, histogram

Description automatically generated

Figure 15: Training/Test Data Split Plots 1

Graphical user interface, chart

Description automatically generated with medium confidenceGraphical user interface, chart, histogram

Description automatically generated

Figure 16: Training/Test Data Split Plots 2

## D4. Creation of Time Series Model

The Time Series model used for this analysis was an ARIMA model. ARIMA, short for 'Auto Regressive Integrated Moving Average' is a class of models that 'explains' a given time series based on its own past values. This is useful in forecasting future values.

Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modeled by ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, and q. **p** is the order of the AR term, **q** is the order of the MA term, and **d** is the number of differencing required to make the time series stationary.

To provide additional context, **p** is the order of the 'Auto Regressive term. This refers to the number of lags of Y to be used as predictors. **q** is the order of the 'Moving Average (MA). This refers to the number of lagged forecast errors that should go into the ARIMA model. **d** is the minimum number of differencing needed to make the series stationary. If the series is already stationary, then d = 0 (Sosna, 2021).

When it comes to making the appropriate decision for the level of decisioning to be completed, the correct differencing level is whatever is required to get a near stationary series which roams around the defined mean and the ACF plot reaches zero quickly.

The first step in appropriately determining if **d** should be set as zero is to test if the series is stationary by performing the Augmented Dickey Fuller Test. This step can be found earlier within this section of the report.

If the calculated p-value is less than the significance level (0.05), it can be inferred that the time series is stationary. If the p-value is greater than 0.05, the appropriate order of differencing will need to be determined.

The following images outline the process used to identify the appropriate values for **p**, **q**, and **d** for each of the companies. This was achieved using pmdarima’s auto\_arima() function. As shown in the output, the auto\_arima function determines the correct values to be used for p, d, and q. Once this step has been completed, it fits the model and generates predictions. This can be observed in Figures 17-21 shown below.

Text

Description automatically generated

Figure 17: Auto Arima Function

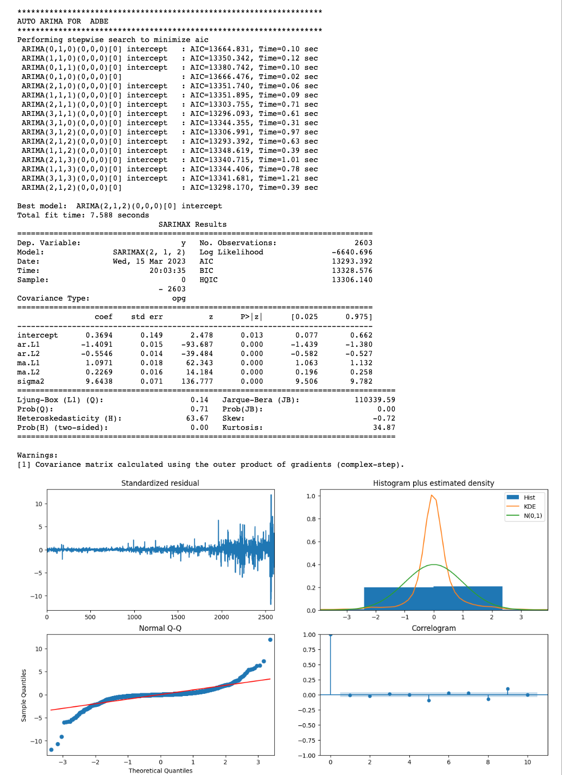
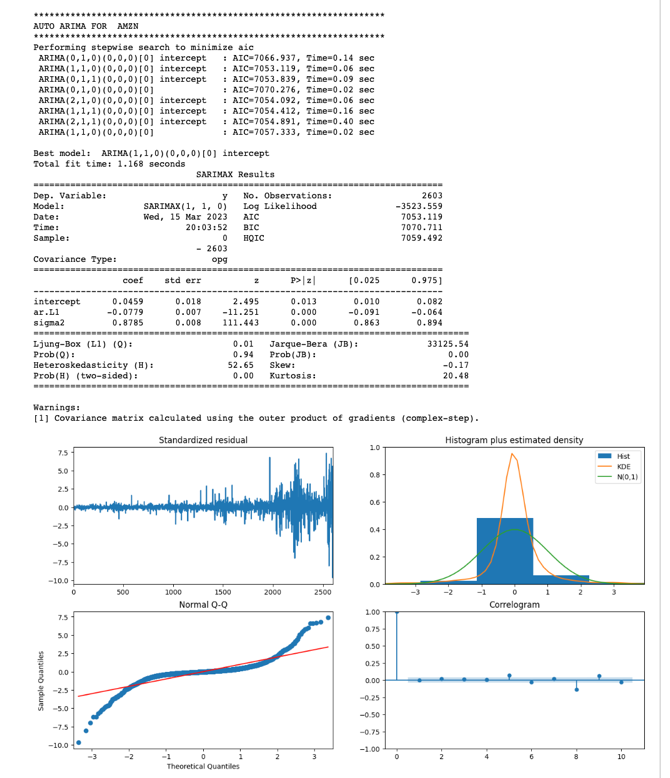
 

Figure 18: Auto Arima Output 1

## 

Figure 19: Auto Arima Output 2

## 

Figure 20: Auto Arima Output 3

## 

Figure 21: Auto Arima Output 4

The following image outlines the p, q, and d value determined as the best fit by the auto\_arima() function for each of the individual company’s being analyzed.

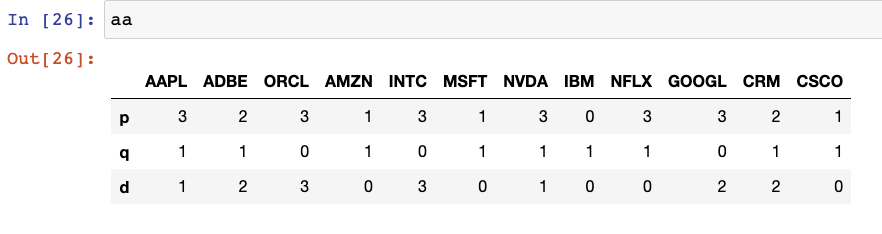


Figure 22: Auto Arima p, q, and d values

## D5. Calculations Performed

The following section reviews the calculations performed to assess the models’ capabilities once fitting of the model was complete.

As shown in the following image, four distinct series/DataFrames were created to hold the actual result, forecasted result, the difference between actual/forecast, and the accuracy for the tested training day.

Additionally, a printout in the console was completed for each tested day with metrics provided for the four previously mentioned metrics.



Figure 23: ARIMA Metric Check

## D6. Analysis Technique Justification

The three techniques overviewed above all play an essential role in ensuring the performed data analysis is as reliable as possible.

The **exploratory data analysis** step is a technique used to review the existing data allowing better familiarity with what is contained within the data.

The **statistical testing** stage is used to provide a mechanism for making quantitative decisions about a process or processes. The intent is to determine if enough evidence exists to "reject" a conjecture or hypothesis about a given process. The conjecture is also commonly known as the null hypothesis.

The creation of the **time series model** stage was used to determine if accurate predictions could be generated for the stock closing price for the included companies.

## D7. Analysis Technique Advantages

The advantages for each of the three techniques used will be outlined below:

* **Exploratory Data Analysis (EDA):** The advantages of the EDA stage are that it can help identify obvious errors within the data, facilitate better understanding of patterns within the data, assist with detecting outliers and/or anomalous events, and find interesting relationships among variables.
* **Statistical Testing:** The advantages of the statistical testing stage include providing a mechanism for making quantitative decisions regarding a process of processes.
* **Creation of Time Series Model:** Some of the advantages of the Time Series Modeling include facilitating a better understanding of underlying causes of trends or systematic patterns over time. This includes identification of seasonal trends and an understanding of why any such trends occur.

## D8. Analysis Technique Disadvantages

The disadvantages for each of the three techniques used will be outlined below:

* **Exploratory Data Analysis (EDA):** Some disadvantages of the EDA stage include providing inconclusive results, a lack of standardized analysis, a small sample population, and/or outdated information that can adversely affect the authenticity of information.
* **Statistical Testing:** The disadvantages of the statistical testing stage include the ability of misusing the tests themselves. If the tests performed are not carefully constructed, the results can be incorrectly skewed.
* **Creation of Time Series Model:** Some of the disadvantages of the Time Series Modeling include issues with generalization from a single study, difficulty in obtaining appropriate measures, and problems with identifying the correct model to represent the data.

# Part V: Data Summary and Implications

# Summary of Data Analysis Implications

## E1. Data Analysis Implications

The implication of the performed analysis is that future stock performance can be reasonably predicted with a relatively high accuracy. Utilizing the forecasted stock prices, an investor can gain a level of comfort with how the stock is estimated to perform based upon its past performance prior to making a substantial investment.

## E2. Data Analysis Results

The original hypothesis for the performed analysis was to determine if a stock closing price could be predicted with greater than or equal to 80% accuracy. Based upon the results of the performed analysis, the null hypothesis can be rejected as the results of the performed analysis indicated the closing price can be predicted with an accuracy ranging from 80%-99%.

While the Closing Price for Nvidia did have three predictions with an accuracy metric reflecting 79.2%, 79.6%, and 79.9%, the remaining 429 predictions for all companies were well above the 80% accuracy metric threshold outlined for the null hypothesis.

The output shown below reflects the accuracy metric for each of the tested days separated by company.

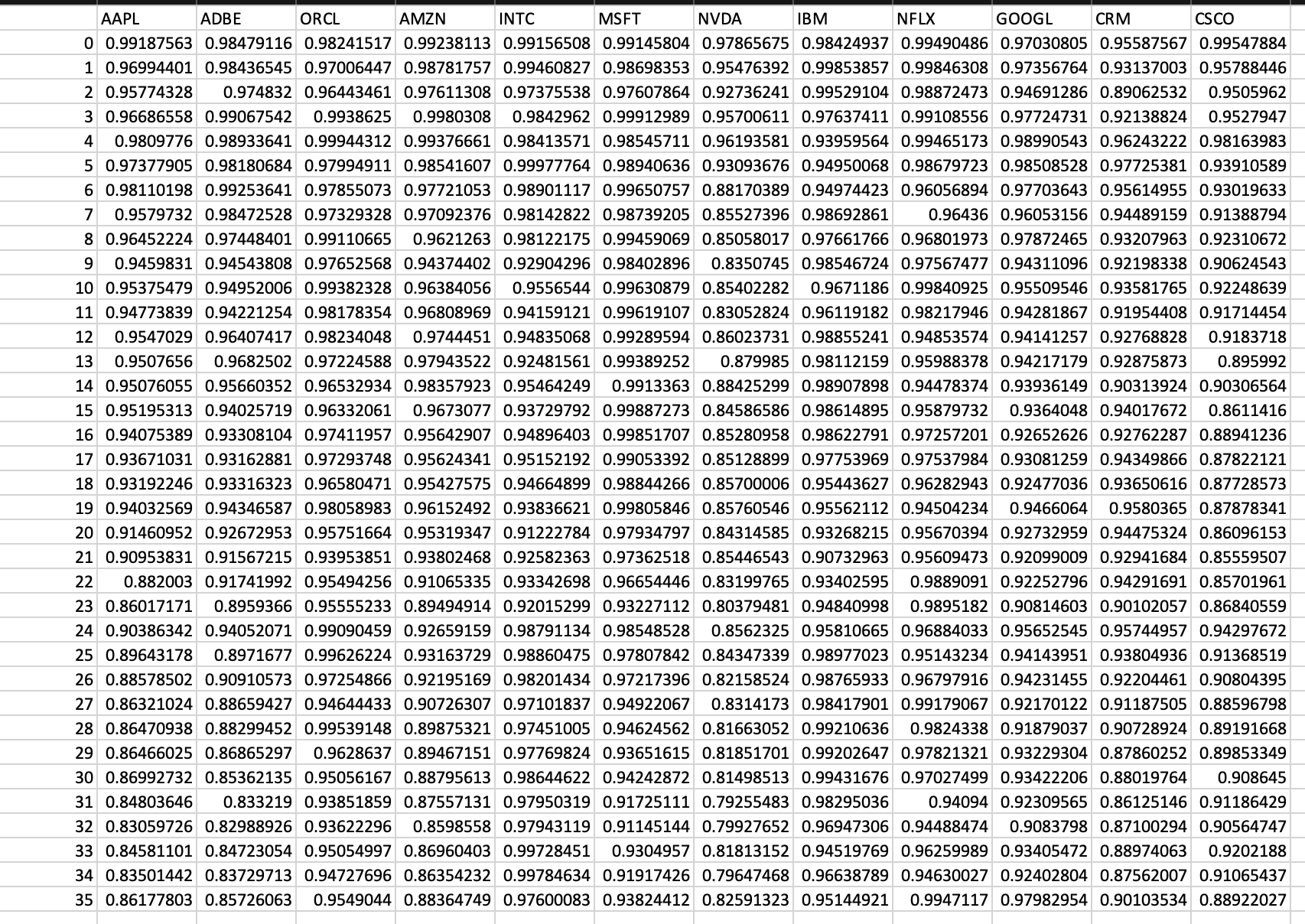


Figure 24: Accuracy Metric Output

In addition to the accuracy metric referenced above, a separate MAPE score was calculated for each companies performance on each individual trading day. This metric provided a screen print-out that included the raw MAPE score, as well as a MAPE score that had been formatted to be in a percentage format.

Table

Description automatically generated with low confidence

Figure 25: MAPE Accuracy Metric Calculation

Table

Description automatically generated Table

Description automatically generated with low confidence

Figure 26: MAPE Metric Output for APPL

The following image reflects the MAPE score associated with each company for the 36 days predictions were forecast. As shown in the Figure 27, no MAPE score met the 20% threshold required to accept the null hypothesis.



Figure 27: MAPE Scores

## 

## E3. Data Analysis Limitations

## The limitations of the performed analysis were that the predicted estimates were consistently lower than the stocks actual market price for a given day. In this case, it is beneficial to error on the side of caution and have the prediction be below an actual price as opposed to above. However, it would warrant additional fine tuning of the model to determine if performance can be improved to better align with the actual price more closely on a given day.

## E4. Recommended Course of Action

The recommended course of action based upon the completed analysis is that utilizing the developed model can assist with investment decisions if a company is seeking insight into an expected company's future market performance.

However, as stated in the previous section, it would be prudent to attempt fine-tuning of the model to improve predictions consistently being below the actual price. If a company’s market value significantly increases for a given period, the model may lag in reflecting accurate predictions at the new market value. This could cause an investor to underestimate any gains from their investment.

## E5. Future Data Study Directions

In regard to future data study, there are two key points that would be essential to maintaining the accuracy of the current prediction model and improving it for additional use.

1. It would be prudent to facilitate the model being able to regularly update the collected historical data. This would allow predictions to continue being made for trading days in the future based upon historical data that is not yet currently available.
2. It would create additional opportunity to implement Improvement to the current prediction model that allows the collection of data for additional companies. This would facilitate predictions being generated for any company trading on the stock market that has historical data available for collection.

# Sources

* 1. Creative Commons License Deed. Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International - CC BY-NC-SA 4.0. (n.d.). Retrieved February 23, 2023, fromhttps://creativecommons.org/licenses/by-nc-sa/4.0/
  2. Gower, E. (2023, January 30). Big Tech Stock prices. Kaggle. Retrieved February 23, 2023, fromhttps://www.kaggle.com/datasets/evangower/big-tech-stock-prices
  3. Time series analysis: Definition, types, techniques, and when it's used. Tableau. (n.d.). Retrieved February 23, 2023, from https://www.tableau.com/learn/articles/time-series-analysis#definition
  4. Li, S. (2018, September 5). An end-to-end project on time series analysis and forecasting with python. Medium. Retrieved February 23, 2023, from https://towardsdatascience.com/an-end-to-end-project-on-time-series-analysis-and-forecasting-with-python-4835e6bf050b
  5. Stephen Allwright. (2022, December 6). What is a good MAPE score? (simply explained). Stephen Allwright. Retrieved February 24, 2023, from https://stephenallwright.com/good-mape-score/
  6. Patil, P. (2022, May 30). What is exploratory data analysis? Medium. Retrieved March 10, 2023, from https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
  7. Wikimedia Foundation. (2023, March 3). Statistical hypothesis testing. Wikipedia. Retrieved March 10, 2023, from https://en.wikipedia.org/wiki/Statistical\_hypothesis\_testing
  8. Vishwas, B. V., & Patel, A. (2020). Hands-on time series analysis with python: From basics to bleeding edge techniques. Apress.
  9. K, B. (2020, December 22). Everything you need to know about Jupyter notebooks! Medium. Retrieved March 10, 2023, from https://towardsdatascience.com/everything-you-need-to-know-about-jupyter-notebooks-10770719952b
  10. Python - overview. Tutorials Point. (n.d.). Retrieved March 10, 2023, from https://www.tutorialspoint.com/python/python\_overview.htm
  11. Yfinance. PyPI. (n.d.). Retrieved March 18, 2023, from https://pypi.org/project/yfinance/
  12. What is time series analysis? TIBCO Software. (n.d.). Retrieved March 19, 2023, from https://www.tibco.com/reference-center/what-is-time-series-analysis
  13. Sosna, M. (2021, August 19). *A deep dive on Arima models*. Medium. Retrieved March 19, 2023, from https://towardsdatascience.com/a-deep-dive-on-arima-models-8900c199ccf