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 4](#_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 6](#_Toc129794847)

[Part III: Data Extraction and Preparation 7](#_Toc129794848)

[C. Describe Data Extraction and Preparation Process 7](#_Toc129794849)

[C1. Import Python Packages 7](#_Toc129794850)

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

[C3. Join the DataFrames using Common Key 9](#_Toc129794852)

[C4. Review Data Types Contained within DataFrame 10](#_Toc129794853)

[C5. Review For Missing and Null Values 11](#_Toc129794854)

[C6. Review Statistical Summary and Plot Distribution 12](#_Toc129794855)

[C7. Tools and Techniques 15](#_Toc129794856)

[C8. Tool and Technique Justification 15](#_Toc129794857)

[C9. Tool and Technique Advantages 16](#_Toc129794858)

[C10. Tools and Techniques Disadvantages 16](#_Toc129794859)

[Part IV: Analysis 16](#_Toc129794860)

[D. Report on Data Analysis Process 16](#_Toc129794861)

[D1. Description of Analysis Technique 16](#_Toc129794862)

[D2. Exploratory Data Analysis 17](#_Toc129794863)

[D3. Statistical Testing 21](#_Toc129794864)

[D4. Creation of Time Series Model 25](#_Toc129794865)

[D5. Calculations Performed 30](#_Toc129794866)

[D6. Analysis Technique Justification 31](#_Toc129794867)

[D7. Analysis Technique Advantages 31](#_Toc129794868)

[D8. Analysis Technique Disadvantages 31](#_Toc129794869)

[Part V: Data Summary and Implications 32](#_Toc129794870)

[E. Summary of Data Analysis Implications 32](#_Toc129794871)

[E1. Data Analysis Implications 32](#_Toc129794872)

[E2. Data Analysis Results 32](#_Toc129794873)

[E3. Data Analysis Limitations 34](#_Toc129794874)

[E4. Recommended Course of Action 34](#_Toc129794875)

[E5. Future Data Study Directions 34](#_Toc129794876)

[F. Sources 34](#_Toc129794877)

[Figure 1: Python Imports 8](#_Toc129962315)

[Figure 2: Loading the individual .csv files 9](#_Toc129962316)

[Figure 3: Combining data into single DataFrame 10](#_Toc129962317)

[Figure 4: Using .info() function to review DataFrame column datatypes 11](#_Toc129962318)

[Figure 5: Reviewing DataFrame for null/missing values 12](#_Toc129962319)

[Figure 6: Reviewing DataFrame statistics with .describe() function 13](#_Toc129962320)

[Figure 7: Historical Distribution Plots 14](#_Toc129962321)

[Figure 8: Tableau Graphs Reflecting Closing Price 17](#_Toc129962322)

[Figure 9: Correlation Heatmap 18](#_Toc129962323)

[Figure 10: Autocorrelation Plots 19](#_Toc129962324)

[Figure 11: Seasonal Decomposition Plots 1 20](#_Toc129962325)

[Figure 11: Seasonal Decomposition Plots 2 21](#_Toc129962326)

[Figure 12: adfuller test code 22](#_Toc129962327)

[Figure 13: adfuller test results 23](#_Toc129962328)

[Figure 14: Training/Test Data Split Plot Code 23](#_Toc129962329)

[Figure 15: Training/Test Data Split Plots 1 24](#_Toc129962330)

[Figure 16: Training/Test Data Split Plots 2 25](#_Toc129962331)

[Figure 17: Auto Arima Function 26](#_Toc129962332)

[Figure 18: Auto Arima Output 1 26](#_Toc129962333)

[Figure 19: Auto Arima Output 2 27](#_Toc129962334)

[Figure 20: Auto Arima Output 3 28](#_Toc129962335)

[Figure 21: Auto Arima Output 4 29](#_Toc129962336)

[Figure 22: Auto Arima p, q, and d values 30](#_Toc129962337)

[Figure 23: ARIMA Metric Check 30](#_Toc129962338)

[Figure 24: Accuracy Metric Output 32](#_Toc129962339)

[Figure 25: MAPE Accuracy Metric Calculation 33](#_Toc129962340)

[Figure 26: MAPE Metric Output for APPL 33](#_Toc129962341)

# 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 forecast?

With so many stocks for perspective investors to choose from, it is often difficult for investors to know the best companies to invest in as well as the best time to invest in them. 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 to provide better comprehension of the price and volume for the analyzed companies market performance at various points in time since 2010.

This will facilitate an understanding regarding the existence of seasonality, any trends associated with volume/price change, and general market fluctuations.

## A3. Description of Context for Research Question

The stock market can often be riddled with uncertainty and great fluctuation. To facilitate educated investing, it is prudent to attempt 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 for 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 massive 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 of this research project is that the performance of 12 American tech companies market performance can be accurately forecast. This is to be achieved by training 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.

The following passage will further expand upon the null and alternate hypothesis’ 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 that the MAPE score is less than 20%. This would indicate a greater than 80% similarity on the predicted dollar value when compared 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 from <https://www.kaggle.com/datasets/evangower/big-tech-stock-prices>.

The full dataset consists of 14 unique csv files. Each file is associated with a specific tech companies stock performance for trading days beginning in 2010. The companies included for analysis are:

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

Due to the Meta and Tesla datasets containing a smaller set of historical trading days in comparison to the others, the data for these two companies was excluded from the current analysis.

These columns contained within each of the individual .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 this 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 stock prices for 12 tech companies over a period beginning in 2010. This includes historical open and close prices, and overall volume.

## B1. Advantages of Data-Gathering Methodology

Utilizing a CC0 1.0 Public Licensed Dataset has a multitude of advantages. While not impossible to independently ascertain the data included within this dataset, it would take considerable time and effort to gather such an extensive time for 12 companies.

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

## B2. Disadvantages of Data-Gathering Methodology

Like a public dataset providing many advantages to the analysts utilizing it, there are often disadvantages associated with this method of data gathering as well.

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

Another disadvantage is trusting that the included datapoints have been correctly gathered and properly reflect the information to be analyzed. It does not serve a data analyst well to perform analysis on inaccurate statistics.

In fact, it could lead to complications for a company if recommended course of action is developed based upon a set of data that inaccurately portrays the subject matter being analyzed. This could potentially harm to the 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 setting, a specific question is posed that requires analysis to be completed. It would be at this time that the data required to perform the analysis is gathered in a way that best aligns with the specific goals of the analysis outlined.

When utilizing publicly available datasets, the data is compiled independent of any specific research question which requires time and energy to cater the data to the specific data analysis being performed.

For the current analysis, time was spent to ensure a proper understanding of what information was contained within the existing dataset and performing analysis to determine what, if any, cleaning, and reconfiguration of the data would be required.

This specific dataset 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 effort was made to determine if all companies had data available for the identical period, it was determined two of the companies, Meta and Tesla, had incomplete data for the time period. Due to the remaining 12 companies having historical data available for the same historical range of dates, it was decided to exclude Meta and Tesla from the currently analysis.

The individual steps taken to analyze and clean the remaining datasets are outlined in the following sections of this report.

# Part III: Data Extraction and Preparation

# Describe Data Extraction and Preparation Process

To extract, clean, and prepare the data, a set of industry standard techniques were performed utilizing common Python packages such as Pandas, Numpy, and MatPlotLib.

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

The steps required to properly prepare the Jupyter Notebook environment, import the required libraries, and clean the original data are outlined in the subsequent sections of this report.

An explanation, justification, advantages, and disadvantages for each step of this process have been included for review.

## C1. Import Python Packages

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

The packages utilized for this analysis included:

* **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.

Graphical user interface, text, application

Description automatically generated

Figure : Python Imports

* **Explanation**: This step imports the packages required to load the existing code. It also facilitates the cleaning and preparation steps of the analysis process.
* **Justification**: Utilizing the existing packages and modules removes the requirement of having to manually code 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 having to learn what exists within the existing packages to ensure the functionality is properly utilized.

## C2. Read in Existing Data

As shown on the image below, each individual .csv file that corresponds to a single company that’s been selected for analysis was independently read in as a .csv file using the Pandas library.

Text

Description automatically generated

Figure : Loading the individual .csv files

* **Explanation**: Using the panda’s library, a csv file can be converted into a DataFrame.
* **Justification**: The original dataset consisted of 14 unique csv files, making pandas the logical choice of package to read in the data for analysis
* **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 opening a .csv file in an application like Microsoft Excel, pandas does require additional steps be taken before being able to fully view the information contained within the data.

## C3. Join the DataFrames using Common Key

As shown in the following section, the original DataFrame associated with each of the tech companies included for review have been joined into a single DataFrame.

This was facilitated by creating an empty DataFrame with an index column associated with the dates included within the original .csv files. A column was then created for each of the companies that will contain the companies closing price for their stock for all the days included for review.

Graphical user interface, text, email

Description automatically generated

Figure : Combining data into single DataFrame.

* **Explanation**: As each of the 14 original tech companies included have their stock information included in separate .csv files, steps were performed to combine the relevant information for all companies into a single DataFrame.
* **Justification**: Due to challenges of analyzing 14 files within a single Jupyter Notebook, it was determined that combining each of the companies closing price for each of the trading days included within the time analyzed would be ideal. This was facilitated by creating an empty DataFrame with the date as the index and adding a column for each of the companies. 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 is significantly easier to analyze a single DataFrame in comparison to performing analysis across 14 different files/DataFrames.
* **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.

## 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 of each variable contained within the 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.
* **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**: There are no disadvantages of performing this step.

## 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 a printout for each of the columns associated with a specific stock to reflect a total number of values contained within the respective column. Another method is using the .info() function contained within the previous section which includes the number of non-null records for each of the columns contained within the DataFrame.

Text

Description automatically generated with medium confidence

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 analysis performed. It is also important to understand any reasons a given set of data would be missing.
* **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 any analysis performed.
* **Disadvantage**: It can often be difficult to discern the reason a given set of data may be 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 section, 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**

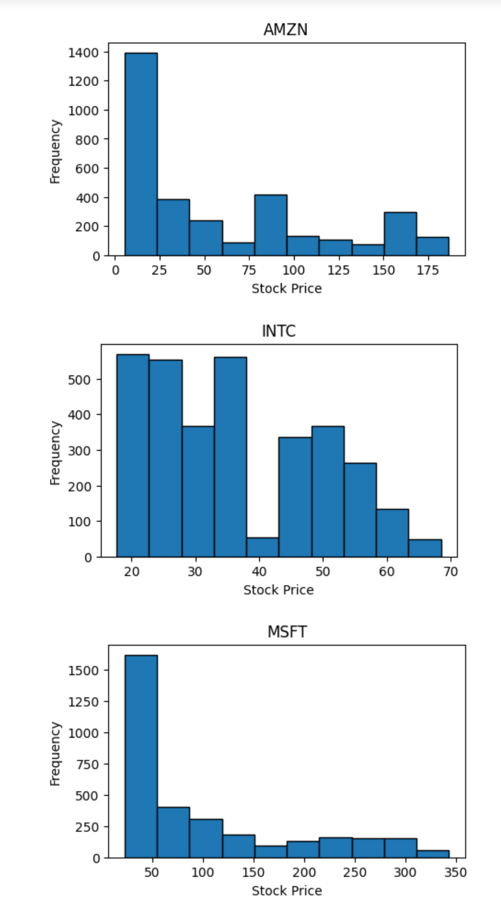
Figure : Reviewing DataFrame statistics with .describe() function.

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 calculations associated with each column’s values.
* **Justification**: It is important to understand the data that is to be analyzed. This step provides a clear insight into the minimum, maximum, and average values found associated with closing stock prices for all the companies being analyzed.
* **Advantage**: The insight gained can assist with determining if outliers exist within the data that could be skewing the future analysis to be performed.
* **Disadvantage**: While this stage does provide the statistics that can be used to detect outlier data points, it does not provide any information regarding if that datapoint may have been collected in error or any other information that provides clarity into the methods used to collect it.

## C7. Tools and Techniques

The tools and techniques used for this analysis are as follows:

* **Jupyter Notebooks**: The 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.

## 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.

## 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.

## 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).

# 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 refers to the critical process of performing initial investigations on data in an attempt at discovering patterns, spotting anomalies, testing hypothesis, 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 addition to the exploratory data analysis(EDA) performed in section C, the following EDA was performed for the current dataset.

The first visualization was created within Tableau and reflects the average closing price by year broken down by each individual companies trading symbol.

Chart, box and whisker chart

Description automatically generated

Figure : Tableau Graphs Reflecting Closing Price

This includes generating a correlation heatmap, an autocorrelation plot, and a seasonal\_decompose for all the companies included within the dataset.

The first image is a correlation heatmap reflecting the correlation between the different company stocks.

## Table Description automatically generated with medium confidence

Figure : Correlation Heatmap

Next, an autocorrelation plot was included for each of the analyzed companies 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

The seasonal\_decompose() function was also performed for each of the analyzed companies. The output for this action can be seen in the following visualizations.

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 the adfuller test for each column in the DataFrame. This test is a method of testing stationarity within a time series.

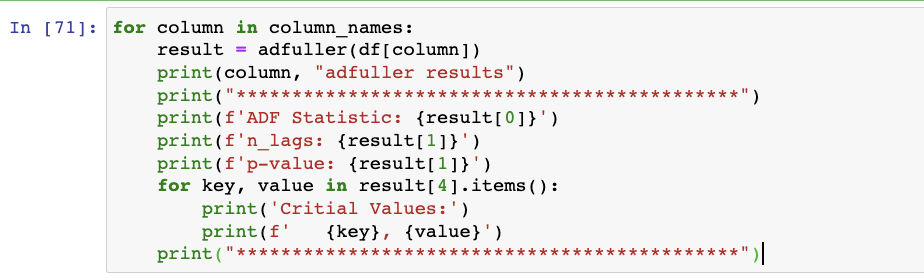


Figure 12: adfuller test code

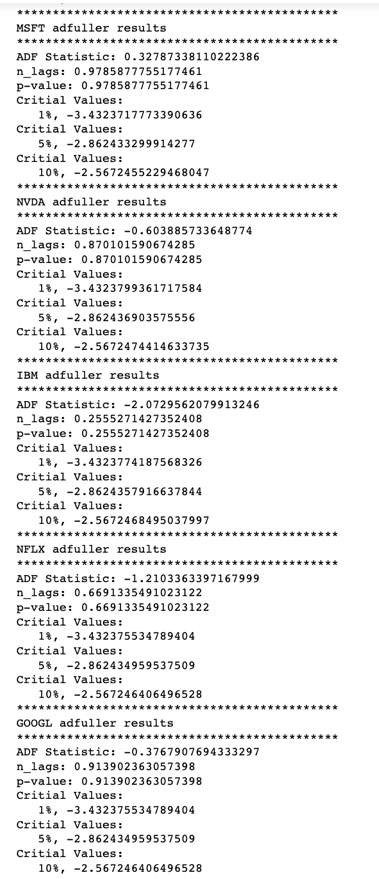
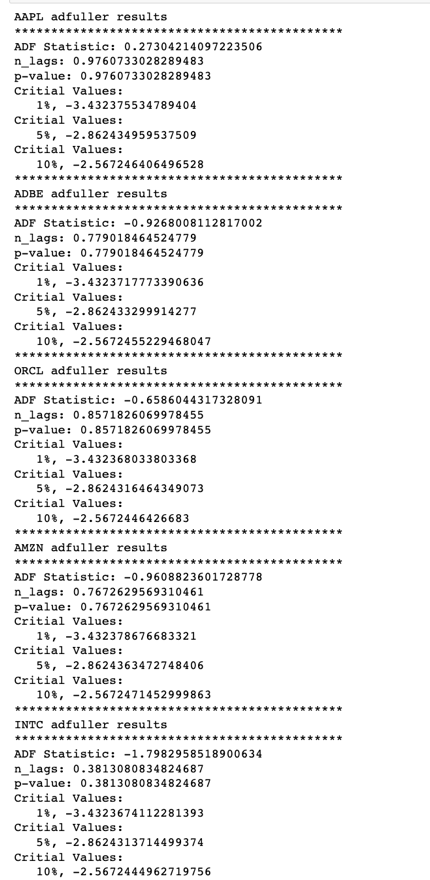
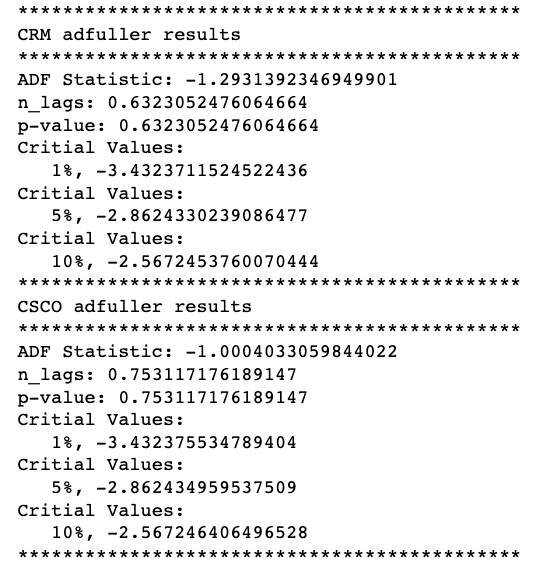
 

Figure 13: adfuller test results

The following visualizations exhibit the breakdown of where the training and test data split was made for each of the analyzed companies as outlined by the change in color from green to blue in the graphs.

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.

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. 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.

Text

Description automatically generated

Figure 17: Auto Arima Function

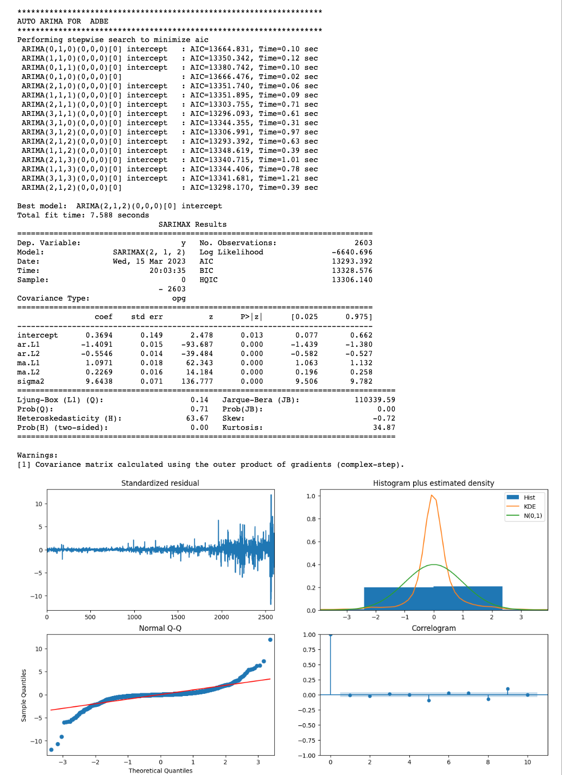
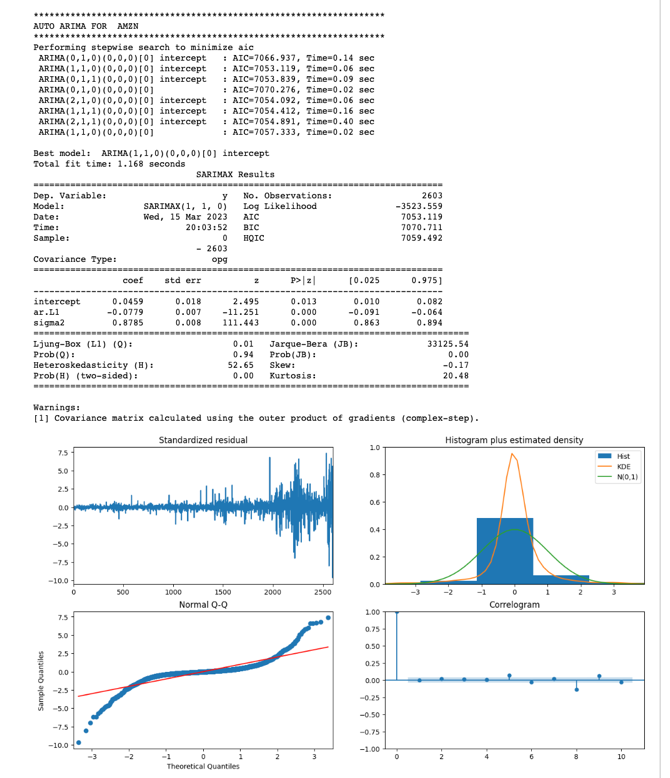
 

Figure 18: Auto Arima Output 1

## 

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 data.

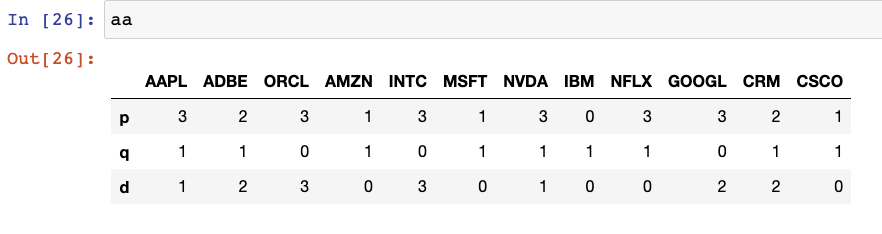


Figure 22: Auto Arima p, q, and d values

## D5. Calculations Performed

The following section reviews the calculations performed to assess the models’ capabilities once fitting of the model had been completed.

As can be seen 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:** 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:** 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 includes that it can be easy to misuse the tests themselves. If the tests used are not carefully constructed, the results can be skewed incorrectly.
* **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.

## 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%.

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

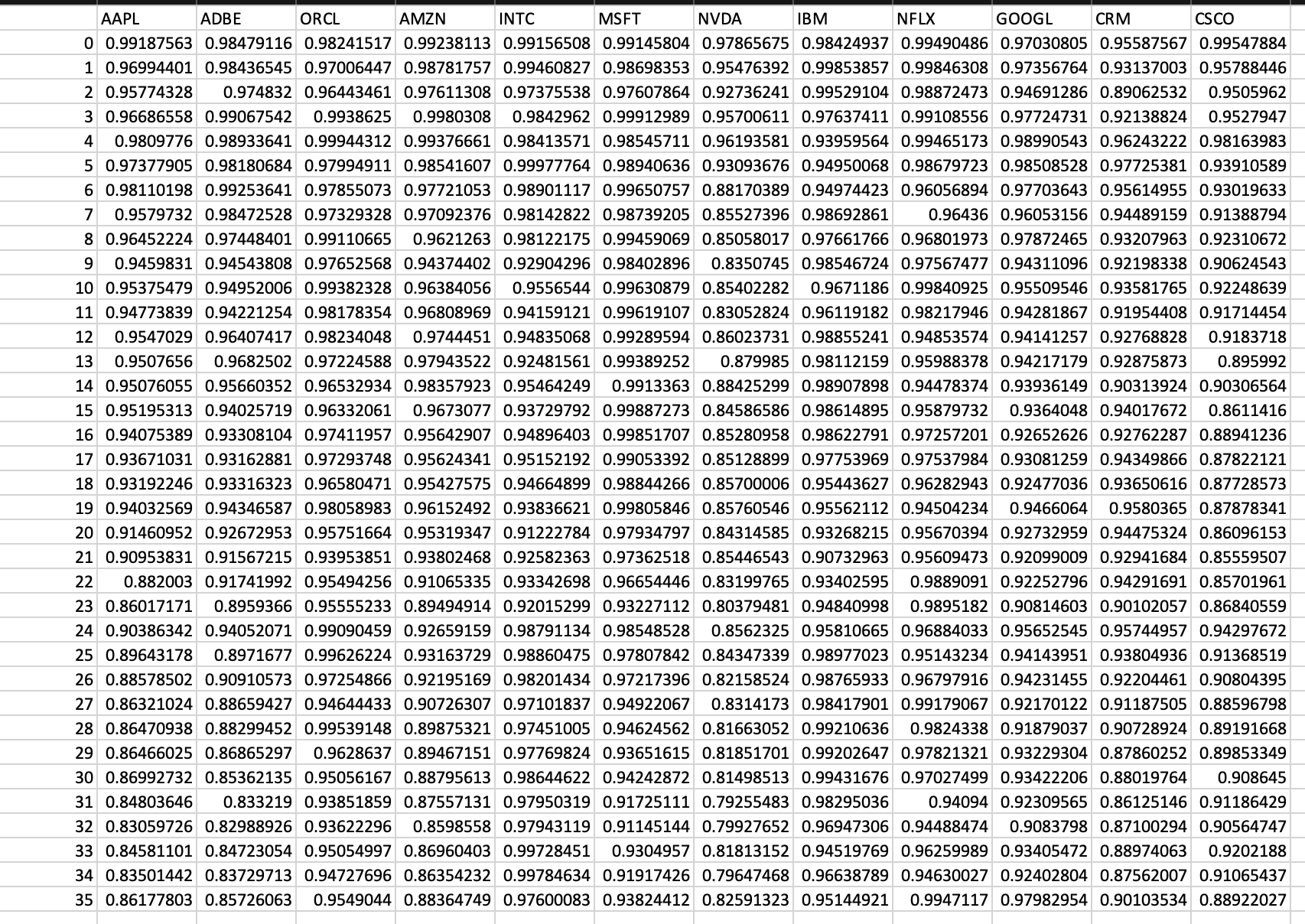


Figure 24: Accuracy Metric Output

In addition to the accuract 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

## 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, it warrants more fine tuning to the model to determine if performance can be fine-tuned to 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 market performance.

However, as stated in the previous section, it would be prudent to attempt fine-tuning the model to correct the 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

1. It would be prudent to facilitate the model being able to regularly update the historical data to allow predictions to continue being made for trading days in the future based upon historical data that is not yet currently available.
2. Improving the model to collect data for additional companies would allow predictions to be generated for any company with 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