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 7](#_Toc129794851)

[C3. Join the DataFrames using Common Key 8](#_Toc129794852)

[C4. Review Data Types Contained within DataFrame 8](#_Toc129794853)

[C5. Review For Missing and Null Values 8](#_Toc129794854)

[C6. Review Statistical Summary and Plot Distribution 9](#_Toc129794855)

[C7. Tools and Techniques 10](#_Toc129794856)

[C8. Tool and Technique Justification 10](#_Toc129794857)

[C9. Tool and Technique Advantages 10](#_Toc129794858)

[C10. Tools and Techniques Disadvantages 10](#_Toc129794859)

[Part IV: Analysis 11](#_Toc129794860)

[D. Report on Data Analysis Process 11](#_Toc129794861)

[D1. Description of Analysis Technique 11](#_Toc129794862)

[D2. Exploratory Data Analysis 11](#_Toc129794863)

[D3. Statistical Testing 12](#_Toc129794864)

[D4. Creation of Time Series Model 12](#_Toc129794865)

[D5. Calculations Performed 12](#_Toc129794866)

[D6. Analysis Technique Justification 12](#_Toc129794867)

[D7. Analysis Technique Advantages 13](#_Toc129794868)

[D8. Analysis Technique Disadvantages 13](#_Toc129794869)

[Part V: Data Summary and Implications 14](#_Toc129794870)

[E. Summary of Data Analysis Implications 14](#_Toc129794871)

[E1. Data Analysis Implications 14](#_Toc129794872)

[E2. Data Analysis Results 14](#_Toc129794873)

[E3. Data Analysis Limitations 14](#_Toc129794874)

[E4. Recommended Course of Action 14](#_Toc129794875)

[E5. Future Data Study Directions 14](#_Toc129794876)

[F. Sources 15](#_Toc129794877)

# 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 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. In an attempt to explore these questions, this author will analyze data associated value and volume for 14 of the largest American based tech companies.

To assist with this analysis, several data analytics tools and techniques willl be utilized.

## A2. Research Question Justification

The justification for the outlined research question is to provide a better understanding in relation to the price and volume for the given companies at different points in time throughout the last 13 years.

This will facilitate an understanding regarding the existence of seasonality, trends associated with volume, and market fluctuations.

## A3. Description of Context for Research Question

As is commonly known, the stock market can often times be riddled with uncertainty and great fluctuation. In an effort to facilitate educated investing, it is prudent to attempt to ascertain trends and seasonality that exists within a given stock prior to making a large investment.

As outlined in the preceding sections of this report, 14 of the largest tech companies will have their market performance analyzed for the previous 13 years of available data in an effort to forecast expected future performance for the respective company's stock price. Per the New York Stock Exchanges available data, an average of $18.9 billion is traded on the stock market per day. With such a massive amount of money invested, an investor stands to lose a significant amount of funding if poor investments are made based on a 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 comfort for investors prior to committing large amounts of funding.

## A4. Hypothesis Discussion

The hypothesis of this research project is that the performance of a specific subset of 14 tech companies market performance can be accurately forecast by training on historical performance and validating the accuracy of the forecast by generating a forecast and comparing to a subset of the historical dataset aside for validation.

# Part II: Data Collection

# Report on Data Collection Process

The data used to further analysis the previously proposed research question is publicly available at 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 comanies included for analysis are:

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

These columns contained within each of the individual csv files are:

* Date
* Open
* High
* Low
* Close
* Adj Close
* Volume

The data being 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 consists of examining existing data. For this specific analysis, this includes examining existing records related to historical stock prices for 14 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 period for 14 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

Similar to 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 period, 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 lead to harm to the companies bottom-line and public reputation.

## B3. Challenges

One of the primary challenges with 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 in the direction needed.

For the current analysis, time was spent to ensure a proper understanding of what information was contained within the existing data 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 over a time period often dating back to 2010. It required time to be spent to determine if all companies had data available for the same time period, or if some of the companies included only included information for an abbreviated period in comparison to others.

The steps to clean and analyze this dataset are outlined in other 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 standard techniques were used utilizing common Python packages such as Pandas, Numpy, and MatPlotLib.

The primary environment used 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 following sections.

An explanation, justification, advantages, and disadvantages for each step of the preparation process have been included in the subsequent sections for review.

## C1. Import Python Packages

* **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

* **Explanation**: Through the use of the pandas 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 dataframes 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 a 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 of the days included for review.

* **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 period 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

* **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 print out 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.

* **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. These statistics are:

* **count**
* **mean**
* **std**
* **min**
* **25%**
* **50%**
* **75%**
* **max**
* **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 provide that provide calculations associated with each columns 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 where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

## 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

## 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.

## 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 in size.
* 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.

# 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

## D3. Statistical Testing

## 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 actually 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.

## D5. Calculations Performed

## 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 implications 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 results of the performed analysis indicate that a daily stock price can be predicted with an accuracy generally above 85%.

## 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 more closely align with the actual price 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 of time, the model may lag behind 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