D214 - 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
jfryma1@wgu.edu

Research Question

A. Summarize the original real-data research question you identified in task 1. Your summary should include justification for the research question you identified in task 1, a description of the context in which the research question exists, and a discussion of your hypothesis.

Original Research Question

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

Research Question Justification

Description of Context for Research Question

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 forcast 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.

Data Collection

B. Report on your data-collection process by describing the relevant data you collected, discussing one advantage and one disadvantage of the data-gathering methodology you used, and discussing how you overcame any challenges you encountered during the process of collecting your data.

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 (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.

Advantage of Data-Gathering Methodology

Utilizing a CC0 1.0 Public Licensed Dataset has a multitude of advantages.

Disadvantage of Data-Gathering Methodology

A general disadvantage for this analysis was utilizing a public dataset that had preselected variables and observation period.

Challenges

One of the challenges with using the previously mentioned public dataset was ensuring the data was properly cleaned and prepared to facilitate as complete an answer to the stated research question.

Data Extraction and Preparation

C. Describe your data-extraction and -preparation process and provide screenshots to illustrate each step. Explain the tools and techniques you used for data extraction and data preparation, including how these tools and techniques were used on the data. Justify why you used these particular tools and techniques, including one advantage and one disadvantage when they are used with your data-extraction and -preparation methods.

For this report, the primary data analytic technique to be used is a time series analysis. Time series analysis is a method of analyzing a sequence of data points collected over consistent internals of time. This analysis technique allows insight into how specific variables change over time.

Time Series analysis requires many data points to ensure consistency and reliability. Ensuring that there is sufficient data helps ensure that any trends or patterns identified are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting/predicting future data based on historical data.

Import Python Packages

In [1]:
```
from statsmodels.graphics.tsaplots import plot_pacf
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.tsa.statespace.sarimax import SARIMAX
from statsmodels.tsa.holtwinters import ExponentialSmoothing
from statsmodels.tsa.stattools import adfuller
from sklearn import preprocessing
import matplotlib.pyplot as plt
from tqdm import tqdm_notebook
import numpy as np
import pandas as pd
from itertools import product
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

- **Explanation**: This step imports the packages required to load the existing code. It also facilitates the cleaning and preperation 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 preperation.
- **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.

Read in the existing data

In [2]:
```
CSCO = pd.read_csv("Datasets/CSCO.csv")
ADBE = pd.read_csv("Datasets/ADBE.csv")
ORCL = pd.read_csv("Datasets/ORCL.csv")
AMZN = pd.read_csv("Datasets/AMZN.csv")
INTC = pd.read_csv("Datasets/INTC.csv")
MSFT = pd.read_csv("Datasets/MSFT.csv")
NVDA = pd.read_csv("Datasets/NVDA.csv")
IBM = pd.read_csv("Datasets/IBM.csv")
NFLX = pd.read_csv("Datasets/NFLX.csv")
TSLA = pd.read_csv("Datasets/TSLA.csv")
GOOGL = pd.read_csv("Datasets/GOOGL.csv")
META = pd.read_csv("Datasets/META.csv")
AAPL = pd.read_csv("Datasets/AAPL.csv")
CRM = pd.read_csv("Datasets/CRM.csv")
```

In [3]:
```
AAPL['Symbol'] = "AAPL"
ADBE['Symbol'] = "ADBE"
ORCL['Symbol'] = "ORCL"
AMZN['Symbol'] = "AMZN"
INTC['Symbol'] = "INTC"
MSFT['Symbol'] = "MSFT"
NVDA['Symbol'] = "NVDA"
IBM['Symbol'] = "IBM"
NFLX['Symbol'] = "NFLX"
TSLA['Symbol'] = "TSLA"
GOOGL['Symbol'] = "GOOGL"
META['Symbol'] = "META"
CRM['Symbol'] = "CRM"
```

- **Explanation**:
- **Justification**:
- **Advantage**:
- **Disadvantage**:

Join the DataFrames together using a common key

In [4]:
```
df = pd.concat([AAPL, ADBE, ORCL, AMZN, INTC, MSFT, NVDA, IBM, NFLX, TSl
df = df.sort_values(by=['Date'])
```

In [5]:
```
df = df.reindex(columns=["Date", "Symbol", "Open", "High", "Low", "Close
df.head(2)
```

Out[5]:

	Date	Symbol	Open	High	Low	Close	Adj Close	Volume
0	2010-01-04	AAPL	7.6225	7.660714	7.5850	7.643214	6.515213	493729600
0	2010-01-04	NVDA	4.6275	4.655000	4.5275	4.622500	4.242311	80020400

In [6]:
```
df.to_csv('df.csv')
```

- **Explanation**:
- **Justification**:
- **Advantage**:
- **Disadvantage**:

Review the datatypes contained within DataFrame

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 41817 entries, 0 to 2687
Data columns (total 8 columns):
 #   Column     Non-Null Count  Dtype  
---  ------     --------------  -----  
 0   Date       41817 non-null  object 
 1   Symbol     41817 non-null  object 
 2   Open       41817 non-null  float64
 3   High       41817 non-null  float64
 4   Low        41817 non-null  float64
 5   Close      41817 non-null  float64
 6   Adj Close  41817 non-null  float64
 7   Volume     41817 non-null  int64  
dtypes: float64(5), int64(1), object(2)
memory usage: 2.9+ MB
```

- **Explanation**:
- **Justification**:
- **Advantage**:
- **Disadvantage**:

Review value counts to ensure no missing values

In [8]:
```
print("Date Column Values", df['Date'].value_counts().sum())
print("Symbol Column Values", df['Symbol'].value_counts().sum())
print("Open Column Values", df['Open'].value_counts().sum())
print("High Column Values", df['High'].value_counts().sum())
print("Low Column Values", df['Low'].value_counts().sum())
print("Close Column Values", df['Close'].value_counts().sum())
print("Adj Close Column Values", df['Adj Close'].value_counts().sum())
print("Volume Column Values", df['Volume'].value_counts().sum())
```

```
Date Column Values 41817
Symbol Column Values 41817
Open Column Values 41817
High Column Values 41817
Low Column Values 41817
Close Column Values 41817
Adj Close Column Values 41817
Volume Column Values 41817
```

- **Explanation**:
- **Justification**:
- **Advantage**:
- **Disadvantage**:

Review a statistical summary of each variable and plot the distribution

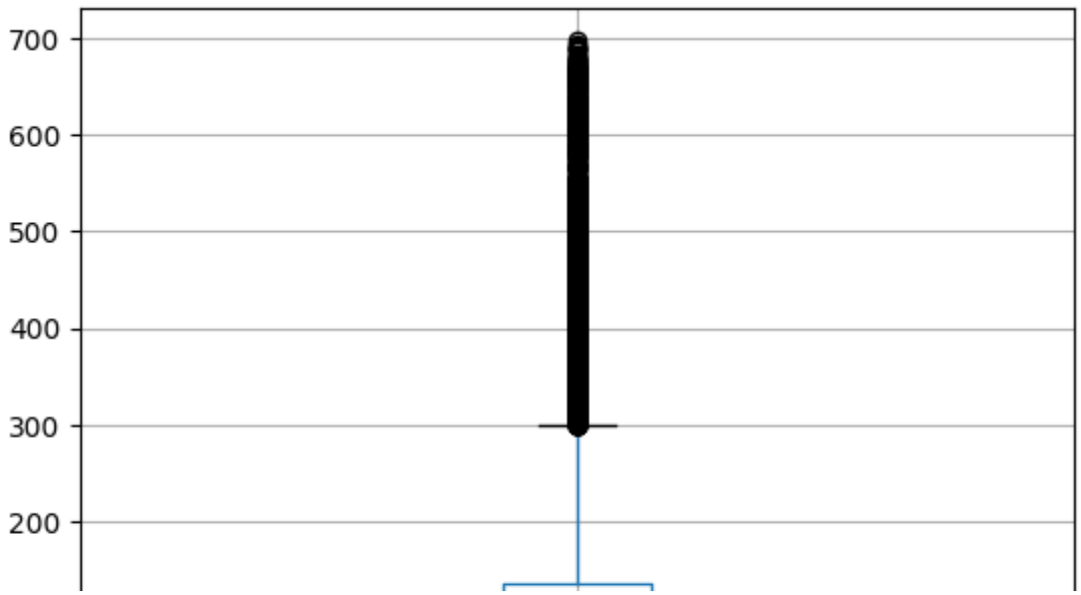
In [9]:
```
df.describe()
```

Out[9]:

	Open	High	Low	Close	Adj Close	Volume
count	41817.000000	41817.000000	41817.000000	41817.000000	41817.000000	4.181700e+04
mean	93.629224	94.794504	92.408277	93.633660	89.635823	5.456459e+07
std	104.216496	105.624383	102.676037	104.180845	103.506444	9.637860e+07
min	1.076000	1.108667	0.998667	1.053333	1.053333	5.892000e+05
25%	26.320000	26.629999	25.995714	26.305000	23.390303	8.519300e+06
50%	51.500000	52.139999	50.915001	51.572498	49.169998	2.662780e+07
75%	135.559998	136.892929	133.957932	135.449326	119.370003	6.117400e+07
max	696.280029	700.989990	686.090027	691.690002	691.690002	1.880998e+09

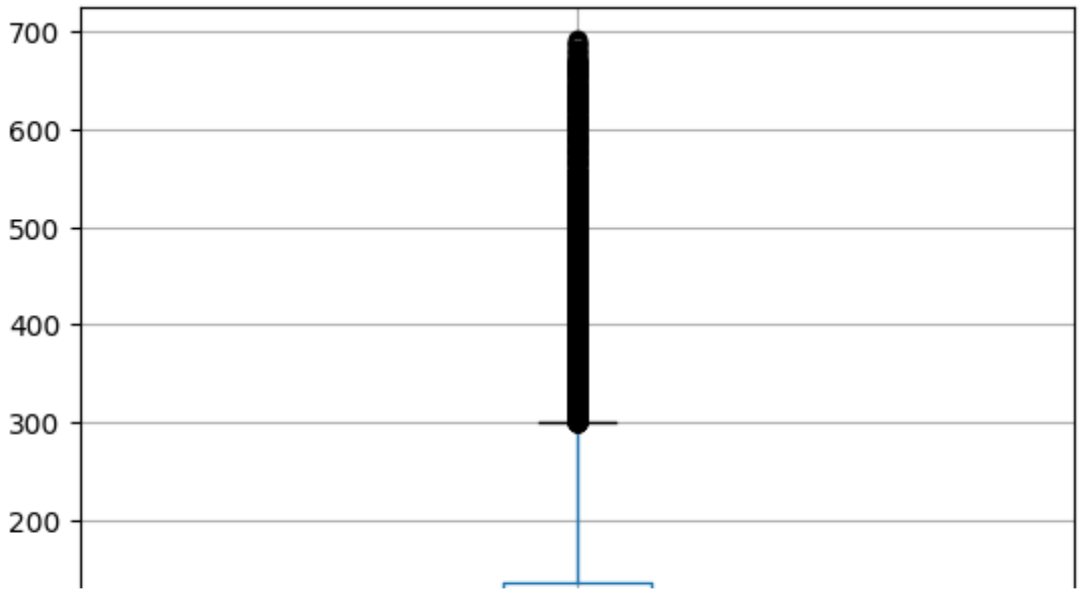
In [10]:
```
print(df.boxplot(column=['Open']))
```

```
AxesSubplot(0.125,0.11;0.775x0.77)
```



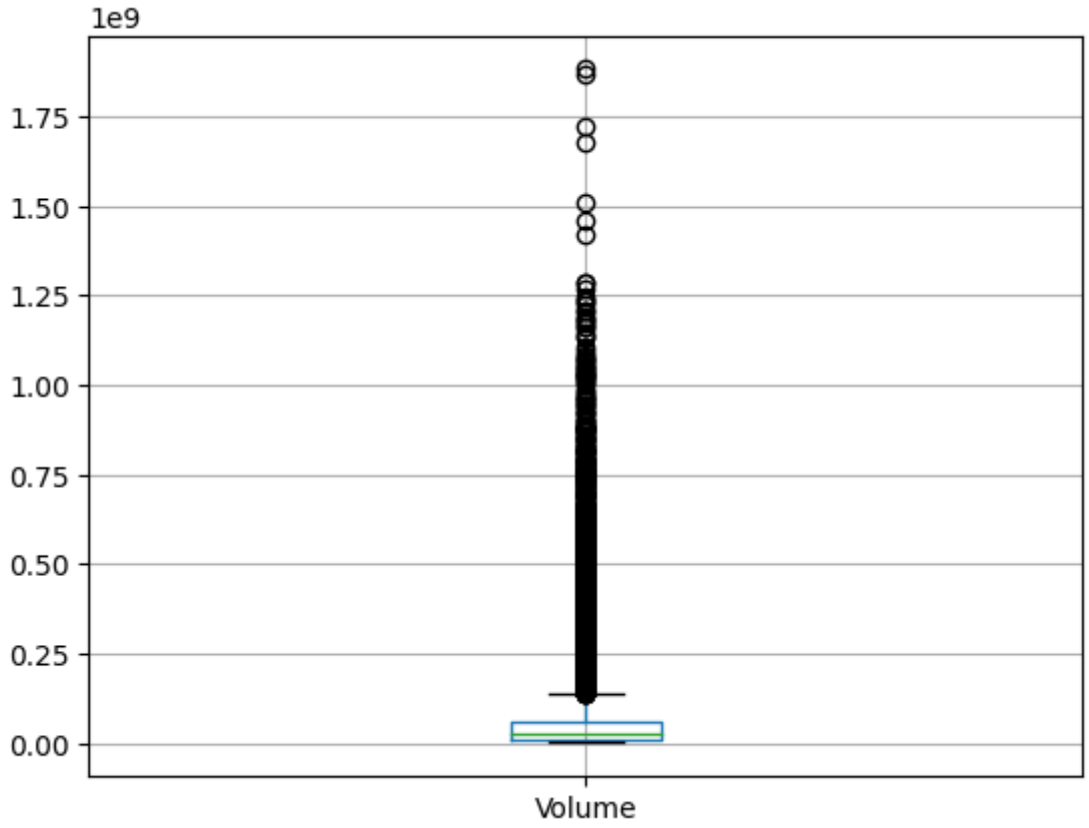
In [11]:
```
print(df.boxplot(column=['Close']))
```

```
AxesSubplot(0.125,0.11;0.775x0.77)
```



In [12]:
```
print(df.boxplot(column=['Volume']))
```

```
AxesSubplot(0.125,0.11;0.775x0.77)
```



In [ ]:

- **Explanation**:
- **Justification**:
- **Advantage**:
- **Disadvantage**:

Check for null values

In [13]:
```
df.isnull().sum()
```

Out[13]:
```
Date         0
Symbol       0
Open         0
High         0
Low          0
Close        0
Adj Close    0
Volume       0
dtype: int64
```

- **Explanation**:
- **Justification**:
- **Advantage**:
- **Disadvantage**:

Create a single target variable for Stock Closing Price

In [14]:
```
df.isnull().sum()
```

Out[14]:
```
Date         0
Symbol       0
Open         0
High         0
Low          0
Close        0
Adj Close    0
Volume       0
dtype: int64
```

- **Explanation**:
- **Justification**:
- **Advantage**:
- **Disadvantage**:

Tools & Techniques

Tool and Technique Justification

Tools and Techniques Advantages

Tools and Techniques Disadvantages

Analysis

D. Report on your data-analysis process by describing the analysis technique(s) you used to appropriately analyze the data. Include the calculations you performed and their outputs. Justify how you selected the analysis technique(s) you used, including one advantage and one disadvantage of these technique(s).

Description of Analysis Techniques

Calculations Performed

In [ ]:

Analysis Technique Justification

Analysis Technique Advantages

Analysis Technique Disdvantages

Data Summary and Implications

E. Summarize the implications of your data analysis by discussing the results of your data analysis in the context of the research question, including one limitation of your analysis. Within the context

of your research question, recommend a course of action based on your results. Then propose two directions or approaches for future study of the data set.

Data Analysis Implications

Data Analysis Results

Data Analysis Limitations

Recommended Course of Action

Future Data Study Directions

1.
2.

F. Acknowledge sources, using in-text citations and references, for content that is quoted.

- 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/
- Gower, E. (2023, January 30). Big Tech Stock prices. Kaggle. Retrieved February 23, 2023, fromhttps://www.kaggle.com/datasets/evangower/big-tech-stock-prices
- 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
- 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
- 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/

In [ ]: