National College of Ireland

Bachelor of Science Honours in Computing

Data Analytics

Academic Year 2020/2021

Umer Iqbal

x17111854

x17111854@student.ncirl.ie

umeriqbal0101@gmail.com

Analyzing and predict stock prices.

Technical Report

Contents

Executive Summary

A company stock price is the highest amount someone is willing to pay for the stock, or the lowest amount that it can be bought for. Technical analysis can be used to predict information on future price movements from historical data.

The project aims to Analyse and predict the historical stock prices of Amazon and Apple Inc etc. In the beginning, the introduction of the project is explained including background, aim, and technology. After that project report briefly discussed the complexity of data, how datasets were acquired/obtained, why datasets are suitable for my project, how datasets are complemented with each other, and characteristics of our datasets, what data visualisations tools were used. Next, we have the KDD methodology section which described a selection of our data, preprocessing/cleaning methods, a transformation of our data, data mining/Machine learning technique (LSTM, ARIMA forecasting, random forest, decision trees, kMeans clustering, hierarchical Clustering, etc.), and evaluation process. Following, project report contains a brief explanation of analysis how datasets were used for pre-processing/cleaning a brief discussion on LSTM, ARIMA forecasting, steps involved during implementation and why these steps were carried out for implementations, characteristics of analysis, advanced statistics (descriptive statistics, Kruskal Wallis test, Mann Whitney test, normality test and Wilcoxon signed-rank test), exploratory data analysis, why did I choose closing price attributes for predicting my stocks values as a predictor in my model. Afterward, all outputs are explained in the results section, describing results tables and figures.

Data visualizations and principal component analysis etc. techniques are used to explore the datasets. Long short-term memory and ARIMA forecasting were used to develop models for the prediction of stock prices. Is it going to increase or decrease in stock prices of Apple Inc and Amazon? Keras, TensorFlow and forecasting packages were used for the smooth development of the prediction model.

1.0 Introduction

1.1. Background

"I will tell you how to become rich. Close the doors. Be fearful when others are greedy. Be greedy when others are fearful." — Warren Buffett. Stock markets were started when the countries in the new world began to trade with each other. As a result, groups of investors pooled their savings and become business partners and co-owners with individual shares in their business to form joint-stock companies. AI in trading allows all the data scientists from all backgrounds to produce algorithm/predicting models that help to solve investment challenges/problems. In the past, technology helps a lot in the stock market to make better decisions and gain a lot of profits. Predicting the stocks has become a very famous trend for all companies/stock's markets, so they can have an idea in advance (based on advanced machine prediction), should they invest in stocks or not.

"It's not whether you're right or wrong that's important, but how much money you make when you're right and how much you lose when you're wrong" By George Soros.

1.2. Aims

The overall aim of this project is to find out future prediction of stock values to have a better understanding for end-users because successfully prediction of future stocks can provide a lot of profit to our end-users. Figure 1 explained the actual and predicted stock prices of Amazon.

Figure 1: Amazon low stock share prices.

The goal will be achieved by analyzing the data from 2010 to 2020 of Apple Inc, Amazon, and the New York Stock market exchange. Whereas historical stock price data contain opening, closing, high and low stock prices of companies. A machine learning model artificial recurrent neural network and ARIMA forecasting will be used to predict stock prices. However, collected data will be imported into Tableau and RStudio for data visualisations.

1.3. Technology

1.3.1 RStudio

RStudio is an IDE for R programming language, for statistical computing and graphics. It is very important for every data analytics project that any data we will use need to be clean e.g., data need to be clean (there should not be any columns that are not important and need to remove), rearranged the columns, changed the names of columns if necessary, deleted some values, deals with missing values, and filled it with means.

Descriptive Statistics has been performed in RStudio which gives me a piece of advanced information on my datasets like, what is the average value, median, mode, quartiles, variance, range, weighted mean, standard deviations, variance, and interquartile range.

ARIMA forecasting is performed in RStudio to predict the next 100 days of stocks and principal component analysis, exploratory data analysis for the transformation of our data.

RStudio has an amazing functionality to perform graphical analysis to give you an understanding of columns in the dataset. Whereas graph has been carried out in RStudio by R programming language like cluster analysis, regression analysis, scatterplot plot, ggplot2, etc.

1.3.2 PyCharm

PyCharm is an IDE used in computer programming language specifically for python. The datasets have been imported in PyCharm via python language. LSTM which is stands for Long short-term memory is an artificial recurrent neural network used in the field of deep learning that is performed in PyCharm for predicting stocks. Further, PyCharm will be used for predicting the stocks of different datasets.

1.3.3 Tableau

Tableau is an interactive data visualization, and it provides advanced visualizations for the datasets. Extracting the dataset as well as different visualization has been performed for the project like candlestick chart, bar graph, scatter graph, etc.

1.3.4 IBM SPSS

IBM SPSS is a platform for advanced statistical analysis tests (normality test, Man Whitney test, Kruskal Wallis test, Wilcoxon Signed ranked test). It is also used to build some graphs for visualizations.

1.3.5 Excel

Excel is used in the project for storing a large amount of data in it as CSV file and visualization.

1.3.6 GitBash

It is an application for Microsoft Windows environments that provide an emulation layer for git cmd. It has been used on regular basis for updating/uploading files on the git hub because it can be download whenever or wherever I want to work on my project.

1.4.Structure

Executive Summary: In this section, a small summary of the report is explained.

Introduction: In the introduction sections background of the project, aims of the project (what are these project goals explained), and briefly discussed technology used in the project.

Data: Briefly explained why data is suitable, how data was acquired, any pre-processing analysis is done and characteristics of data.

Methodology: Explained step-by-step KDD methodology points.

Analysis: Briefly explained how the analysis was done, why I choose these steps for my analysis, why the LSTM model is used in stock price predictions etc.

Results: All diagrams/outputs are added in this section.

Conclusion: Small summary of the report's the advantages/disadvantages of the report were explained in this area.

Further development: In this section, I provide the further part of the project what I will do next like Advanced data mining, predictions, etc.

Reference: Harvard Style reference is provided in this section.

2.0 Data

2.1 Why the datasets are suitable for my project.

Stock market prediction is the act of trying to determine the future value of a company stock prediction. Whereas the successful prediction of stock can lead to a lot of profits for the company. Once, my project idea was ready I decided to find my datasets related to my project idea. Then, I found a dataset of historical stock prices of New York stock exchange from Kaggle.com, Apple and Amazon from yahoo.finance.com and I decided to use it because it contains all the information that I need to predict values for my project like opening price, closing price, Higher stocks, lower stocks, etc. I need datasets for visualizations and through these datasets, I can perform visualizations related to my projects. In my project, I decided to do statistical analysis and machine learning, etc. and these datasets are very complex for the development of models.

2.2 How the datasets are complemented with each other.

Hence, the New York stock exchange dataset which contains attribute like (date, symbol, low, open, close, high, volume) and every column have over nine hundred thousands of rows. On the other hand, Amazon and Apple Inc dataset is taken from https://finance.yahoo.com/ have attribute like (date, open, high low, close, Adj Close, Volume) and every column have over two thousands of rows which provide the history data for both Apple and New York stock exchange. Two of my dataset are complement with each other whereas on New York stock exchange, Amazon and Apple stock exchange datasets which contain completely different information for companies' stocks price.

2.3 How the datasets were acquired/obtain.

Kaggle is an online community for data scientists which allows users to find and publish datasets, explore, and build models in web-based data science environment. It offers a public data platform a cloud-based workbench for data science and artificial intelligence.

As well as New York stock exchange historical price dataset is taken from https://www.kaggle.com/ datasets and Apple and Amazon historical stock price dataset for the last ten years was taken from https://finance.yahoo.com/ yahoo finance provides financial news, data and commentary including stock quotes it also provides some online tools for personal online management.

2.4 Descriptive statistics.

Descriptive statistics is performed on both Apple Inc. and Amazon datasets, for a better understanding of my data insights, describing the basics features of my data in the study, summarising my data in a meaningful full way so pattern might appear.

2.4.1 Descriptive Statistics on Apple datasets.

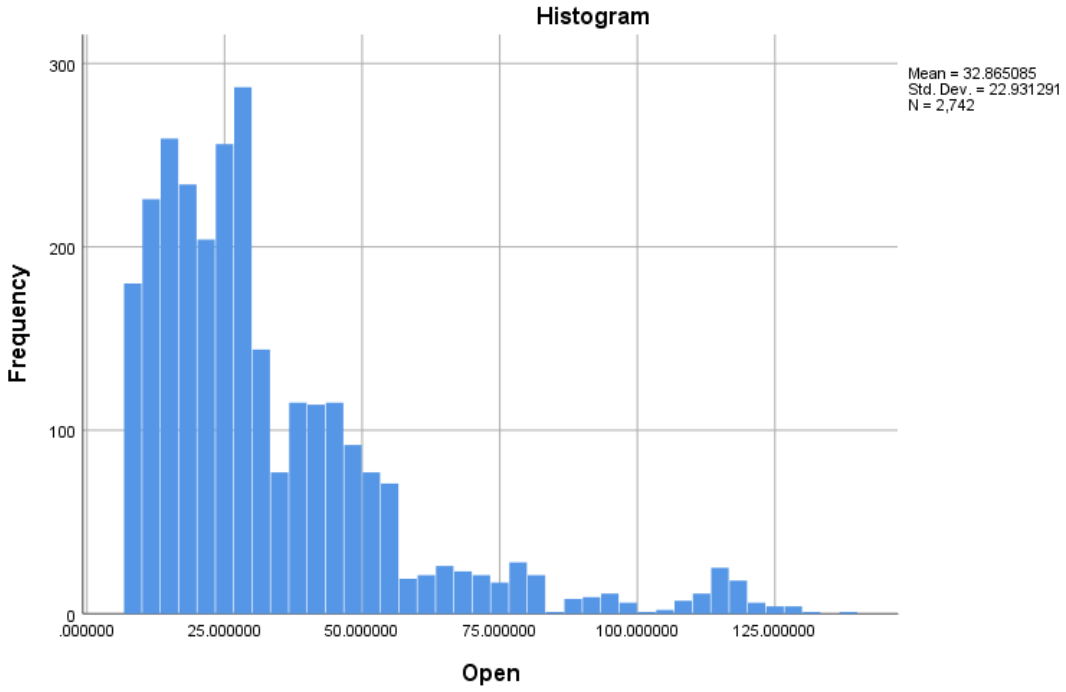


Figure 2: Histogram of Apple Inc opening prices with mean and Std.Dev.

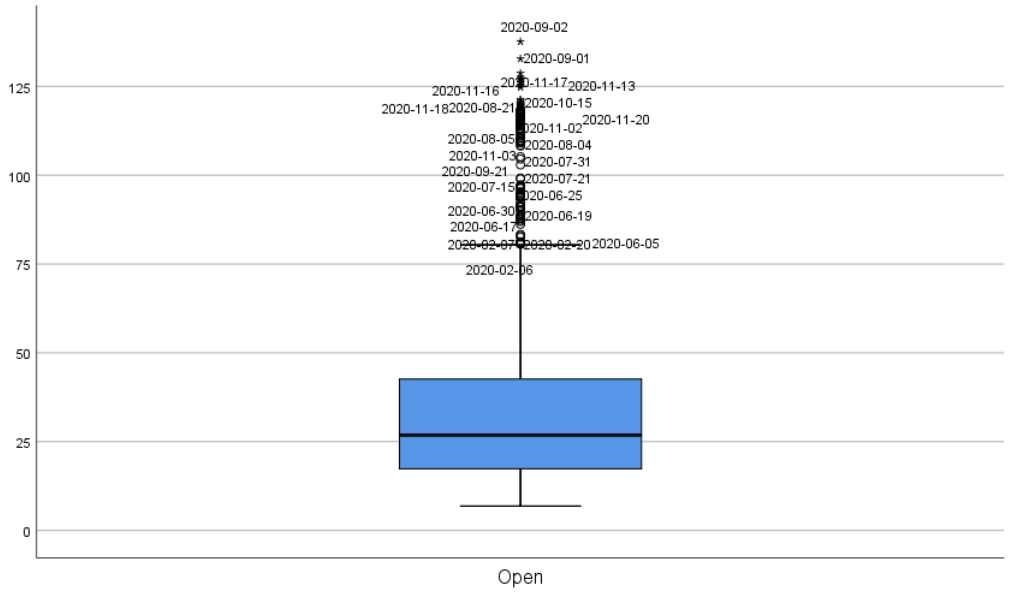


Figure 3: Boxplot of Apple Inc opening prices.

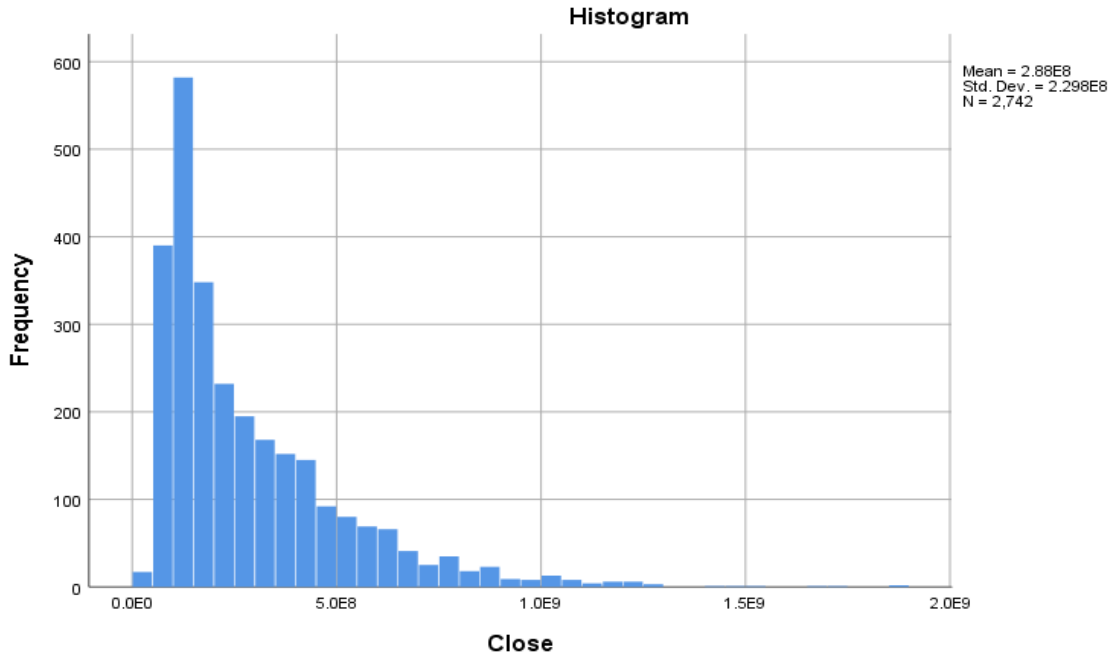


Figure 4: Histogram of Apple Inc closing prices with mean and Std.Dev.

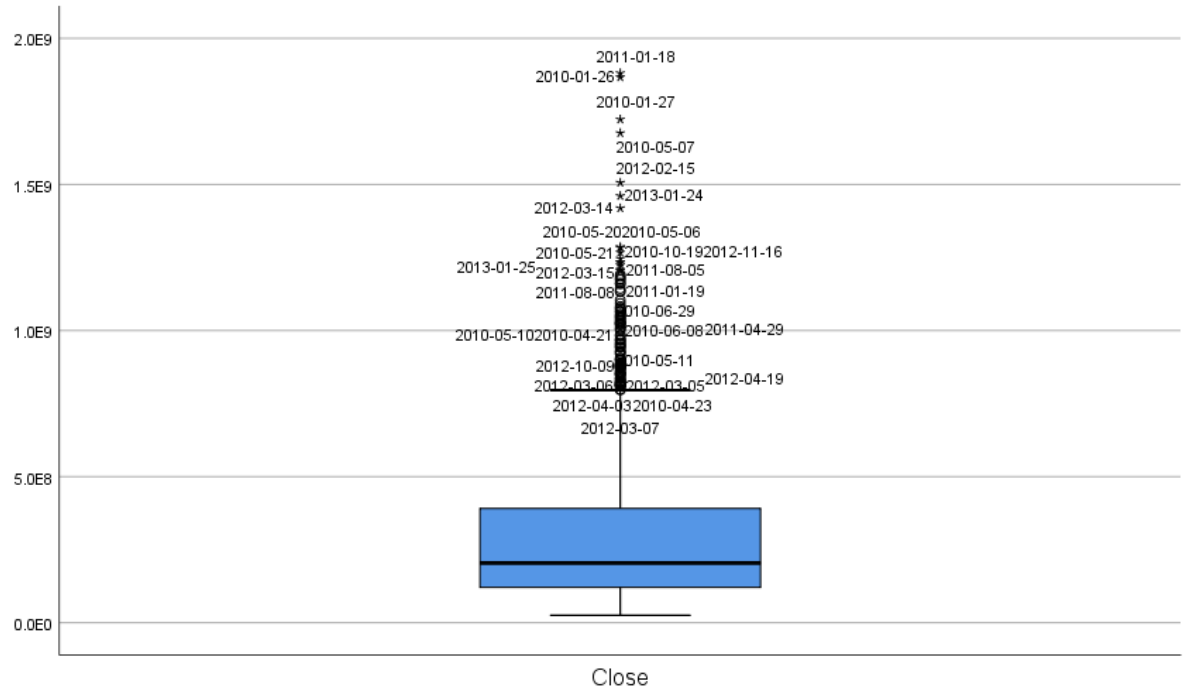


Figure 5: Boxplot of Apple Inc closing prices.

Descriptives

			Statistic	Std. Error
Open	Mean		32.86508533	.437920008
	95% Confidence Interval for Mean	Lower Bound	32.00639871	
		Upper Bound	33.72377195	
	5% Trimmed Mean		30.24182659	
	Median		26.81874950	
	Variance		525.844	
	Std. Deviation		22.93129143	
	Minimum		6.870357	
	Maximum		137.589996	
	Range		130.719639	
	Interquartile Range		25.270537	
	Skewness		1.812	.047
	Kurtosis		3.738	.093
High	Mean		33.20358927	.443942464
	95% Confidence Interval for Mean	Lower Bound	32.33309365	
		Upper Bound	34.07408490	
	5% Trimmed Mean		30.53587566	
	Median		27.03624900	
	Variance		540.407	
	Std. Deviation		23.24665195	
	Minimum		7.000000	
	Maximum		137.979996	
	Range		130.979996	
	Interquartile Range		25.446696	
	Skewness		1.820	.047
	Kurtosis		3.761	.093
Low	Mean		32.52427062	.431664087
	95% Confidence Interval for Mean	Lower Bound	31.67785079	
		Upper Bound	33.37069044	
	5% Trimmed Mean		29.95946877	
	Median		26.56125100	
	Variance		510.928	
	Std. Deviation		22.60370568	
	Minimum		6.794643	
	Maximum		130.529999	
	Range		123.735356	
	Interquartile Range		25.153393	
	Skewness		1.785	.047
	Kurtosis		3.595	.093
Close	Mean		287796340.4	4389216.154
	95% Confidence Interval for Mean	Lower Bound	279189834.4	
		Upper Bound	296402846.4	
	5% Trimmed Mean		262708977.4	
	Median		204481200.0	
	Variance		5.283E+16	
	Std. Deviation		229837396.8	
	Minimum		25432400	
	Maximum		1880998000	
	Range		1855565600	
	Interquartile Range		269948200	
	Skewness		1.867	.047
	Kurtosis		4.953	.093

Figure 6: Descriptive statistics of opening-closing, low and high prices of Apple dataset with mean, median, kurtosis, and range, etc.

2.4.2 Descriptive Statistics on Amazon dataset.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Open	2582	100.0%	0	0.0%	2582	100.0%
Close	2582	100.0%	0	0.0%	2582	100.0%

Figure 7: Processing summary on amazon dataset.

Descriptives

			Statistic	Std. Error
Open	Mean		708.4933384	11.94591265
	95% Confidence Interval for Mean	Lower Bound	685.0687950	
		Upper Bound	731.9178818	
	5% Trimmed Mean		671.5531208	
	Median		385.1750030	
	Variance		368463.868	
	Std. Deviation		607.0122473	
	Minimum		105.930000	
	Maximum		2173.070068	
	Range		2067.140068	
	Interquartile Range		766.825028	
	Skewness		.943	.048
	Kurtosis		-.604	.096
Close	Mean		708.3984779	11.94038077
	95% Confidence Interval for Mean	Lower Bound	684.9847819	
		Upper Bound	731.8121740	
	5% Trimmed Mean		671.5429645	
	Median		385.2399900	
	Variance		368122.693	
	Std. Deviation		606.7311538	
	Minimum		108.610001	
	Maximum		2170.219971	
	Range		2061.609970	
	Interquartile Range		763.915031	
	Skewness		.942	.048
	Kurtosis		-.605	.096

Figure 8: Descriptive statistics of opening-closing prices of Amazon dataset with mean, median, kurtosis, and range, etc.

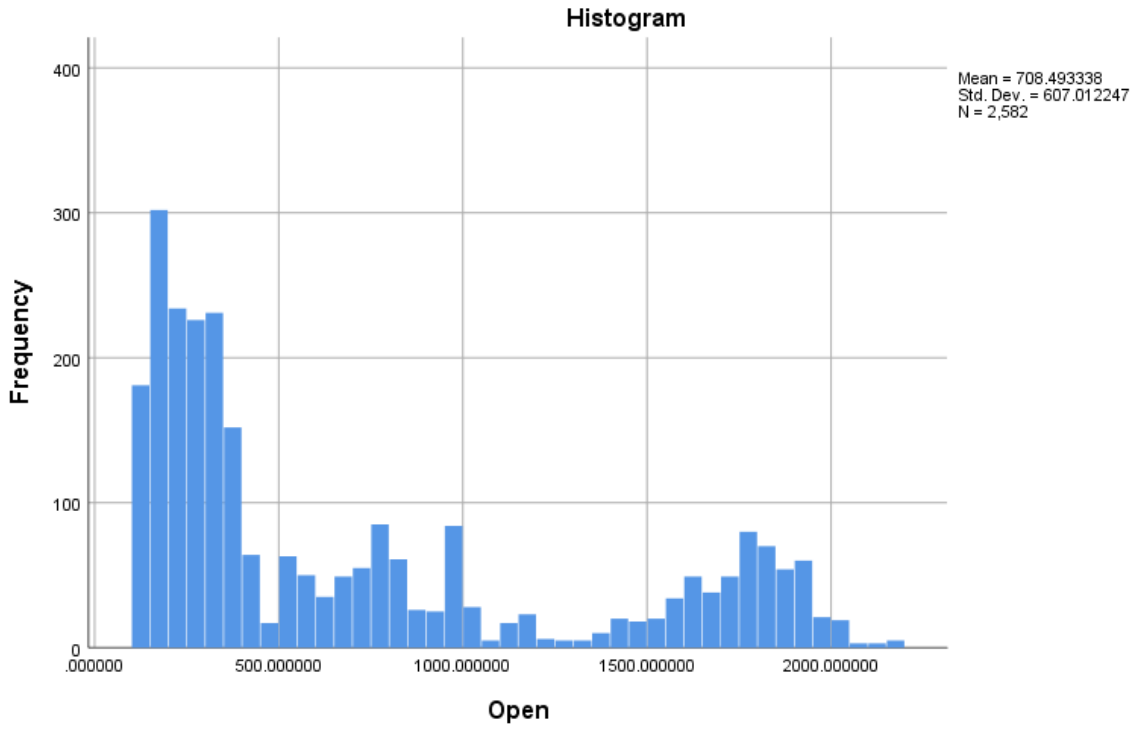


Figure 9: Histogram of Amazon opening stock prices.

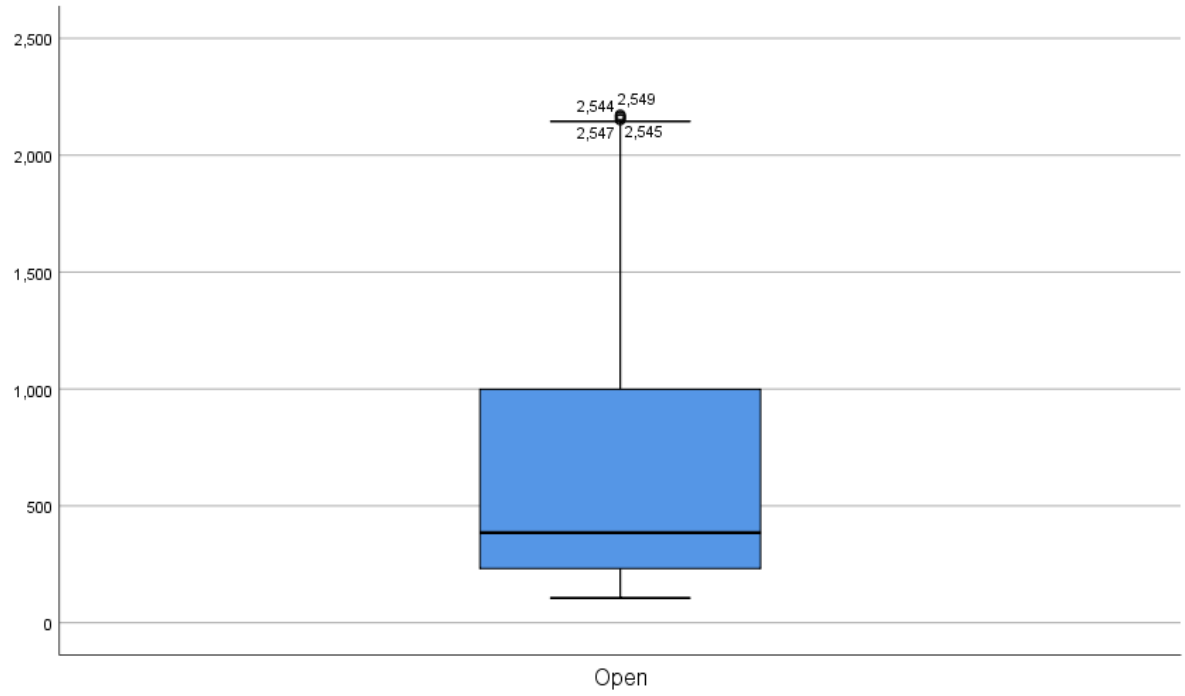


Figure 10: Boxplot of Amazon opening stock prices.

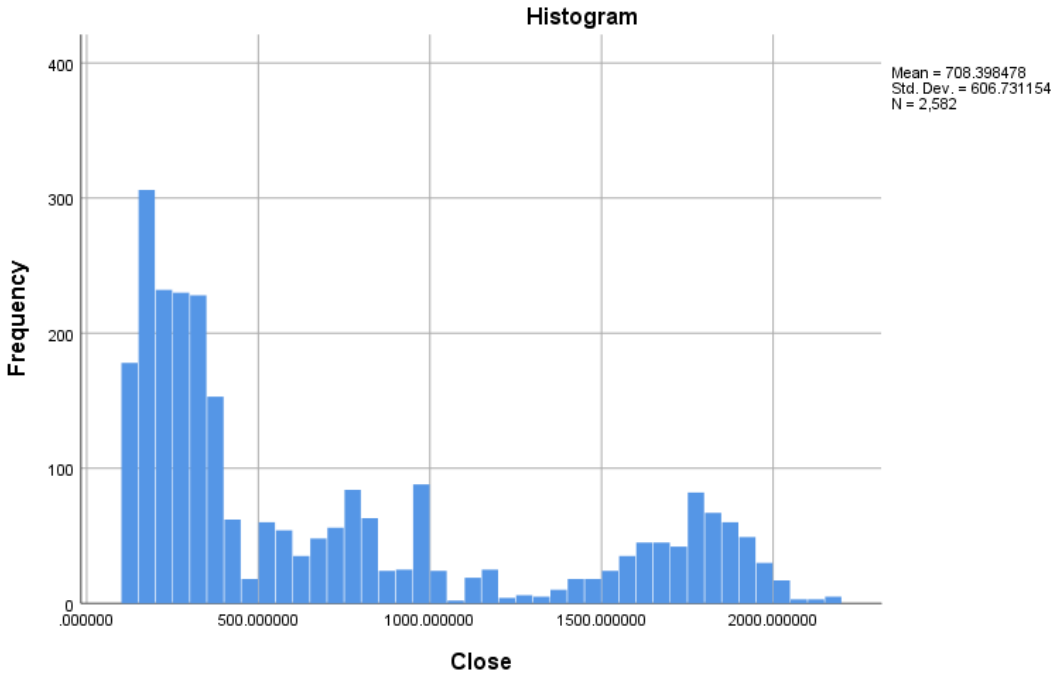


Figure 11: Histogram of Amazon closing stock prices.

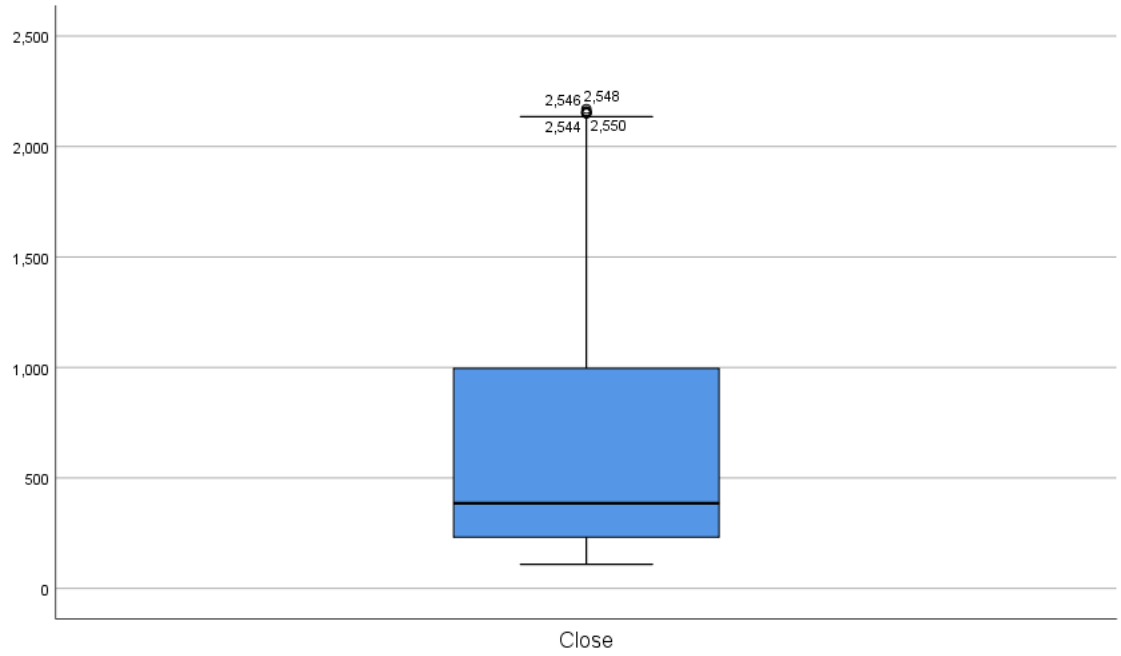


Figure 12: Boxplot of Amazon closing stock prices.

2.5 Main characteristics of Datasets.

Summarize.	New York Stock market exchange Dataset.	Apple Inc	Amazon
File Type.	Excel	Excel	Excel
File Order.	Structure/missing values/ unnecessary columns.	Structure/missing values/ unnecessary columns.	Structure/missing values/ unnecessary columns.
Size of files.	49.2 MB	189 KB	186 KB
The number of instances.	851265	2743	2583
Number of Attributes.	7	7	8
Type of Attributes.	Decimal numbers, Characters, Date.	Decimal/numbers, Characters, Date	Decimal/numbers, Characters, Date

2.6 What data visualisation tools I used?

For visualising my data, I have decided to use Tableau, RStudio, IBM SPSS, excel, and PyCharm and built scatterplot, histogram, ggplot2, line chart, bar chart, and candlestick chart, etc.

Reference:

Sites from New York Stock exchange datasets are taken.

Kaggle.com. 2020. Find Open Datasets and Machine Learning Projects | Kaggle. [online] Available at: https://www.kaggle.com/datasets

[Accessed 20 October 2020].

Sites from Apple Inc stock datasets are taken.

Finance.yahoo.com. 2020. Yahoo Is Now A Part Of Verizon Media. [online] Available at: https://finance.yahoo.com/quote/AAPL/history?p=AAPL

[Accessed 7 December 2020].

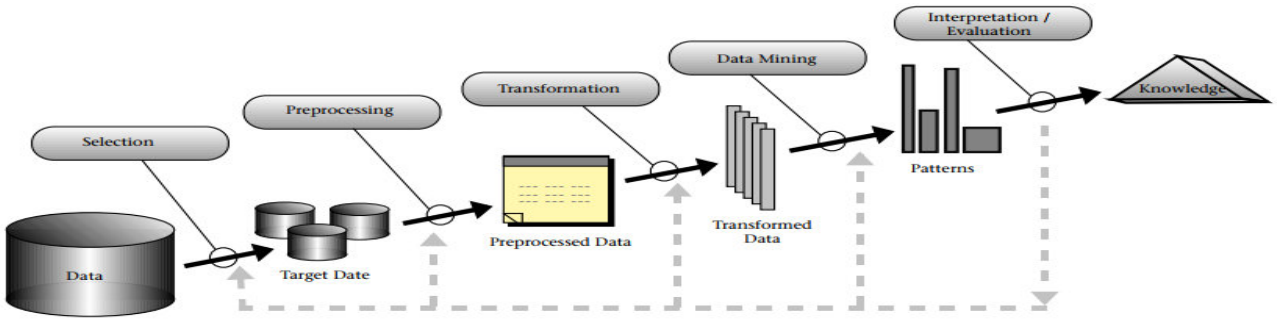
Sites from Amazon Inc stock datasets are taken.

Finance.yahoo.com. 2021. Yahoo is now a part of Verizon Media. [online] Available at: https://finance.yahoo.com/quote/AMZN/history?period1=1262304000&period2=1586217600&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

[Accessed 9 May 2021].

3.0 Methodology

Figure 13: to show steps of the KDD process.



KDD is an application that has its core, the application for specific data mining and pattern discovery.

The term Knowledge Discovery in databases is the process of finding knowledge in data and highlights high-level applications for particular data mining methods. It is useful for those who want to do their research in machine learning, pattern recognition, statistics, artificial intelligence, and data visualizations, etc. It refers to the overall process of discovering useful knowledge from data. It includes the choice of encoding schemes, pre-processing, and sampling of data before data mining steps.

3.1 Data Selection.

In KDD methodology, data selection is the step where we collect our data related to our project for all analytics project dataset need to big for getting some useful information from them. However, I collected the New York stock exchange historical prices dataset which is in excel format and it is providing me all suitable information for my project idea e.g., opening and closing prices of different companies' stocks like WLTW or NUE, etc. whereas this dataset is also providing me high stocks rate and low stocks rate of different companies based on different periods.

After that, apple Inc and Amazon's historical stock price data is collected from https://finance.yahoo.com/ (it provides financial news, data and commentary including stocks quotes). However, these two datasets contain the last ten years of historical stock prices information and are downloaded from the finance.yahoo.com site and contain the following attributes (Date, Open, High, Low, Close, Adj Close, and volume).

3.2 Data pre-processing.

All datasets were imported to RStudio and decided which actions need to take for the cleaning process and then apply machine learning on our datasets.

In New York stock exchange historical prices dataset, few unnecessary columns were not suitable for my project. So, I have to delete it by using R programming language. I checked the null values in New York stock exchange historical prices dataset, if any null values contain my dataset then I filled the null values by taking a mean of that column. In this dataset, the last two

rows 851264 and 851263 were unnecessary/incomplete so I decided to delete them. In this dataset, all column was re-arranged/re-ordered, also changed the names of the column, and changed them to camel case, for my own better understanding. Once my dataset was clean, then I added a few new columns in New York stock exchange historical prices dataset like converting the price of the opening dataset from dollars to euros with help of R programming language.

Apple Inc dataset is also imported in RStudio for the cleaning process. Checked all the columns are there any empty/null values in the dataset, if there are any then it is filled by taking a mean of that column. Re-order all columns and changed the name/spelling of columns with help of R programming language. Apple Inc, the dataset was almost in better condition than other dataset and it did not take much effort in the cleaning processing.

For pre-processing analysis, the amazon data set was imported into RStudio. With help of R programming language, every column was checked and filled empty rows by taking the mean of that column. Re-arranged all columns and changed names of some columns if necessary.

3.3 Data Transformation.

The process which involves changing the value format or structure of data is called data transformation. Process such as data migration, data integration, data warehousing, all involved in the data transformation process.

Data transformation may include constructive adding a new column in datasets, reproducing a new column based on the existing column (changing opening prices from dollars to euros or changing the price from dollars to British pounds). Destructive field deleting unnecessary records from New York stocks exchange and Apple Inc datasets. By adding and deleting the datasets we transform our datasets into new shapes. Data which is transformed make it easy to use for both human and computers.

Data visualisation technique was applied to perform exploratory data analysis. Also, principal component analysis was used on both Apple Inc and Amazon datasets to determine, any correlations between them.

3.4 Data Mining.

The data mining technique is a process of selecting and perform a machine learning algorithm that will be performed on both Apple Inc and Amazon datasets. It is the process of analyzing a massive amount of data to discover business intelligence that helps stock markets to solve problems and take advantage of predicted models.

K-Means clustering intends to partition n objects int *k* cluster in which each object belongs to the cluster with the nearest mean. Where this method produces exactly k different cluster of greatest possible distinction. It is minimized total intra-cluster variance or squared error functions. Whereas kMeans clustering is performed in RStudio for the apple dataset. Hierarchical Clustering is an algorithm that groups similar objects into groups called a cluster. The endpoint of a cluster, where each cluster is separate from the other cluster. It is performed in RStudio with R programming language for the first 40 columns of an open column from the apple dataset. Artificial recurrent neural networks (Long short-term memory) are performed

for predicting stock prices of apple and amazon. ARIMA forecasting is used to predict the next hundred days of stock to check is there a rise or decline in prices. A decision tree algorithm is used in RStudio, which goal is to create a training model that can use to predict attributes (Closing price) by learning simple decision rules from previous data.

3.5 Interpretation/Evaluation.

It is the step where we discover some useful information from our datasets which involves evaluation and interpretation of the pattern to decide the useful data. Here we can decide after getting a response from machine learning (PyCharm), that now we have apple stock predictions based RNN/LSTM model when we should buy or sell stocks and gain profits, or based on ARIMA forecasting we can decide rise or fall in stock prices.

4.0 Analysis

The process of analysis follows the KDD methodology, for producing some useful knowledge from our datasets. At the beginning of every data analytics project, we need to know which domain knowledge is better for our project, what are we expecting to be our outcomes from analysis and KDD is one of the best methodology use in analytics projects for analyzing, pre-processing, transformations, etc. This methodology is consisting of step-by-step processes like data collection, pre-processing, data mining/machine learning, etc. for achieving goals it is better than CRISPDM. My goal of this analysis is to perform exploratory data analysis, principal component analysis (PCA), Artificial Intelligence recurrent neural network (Long short-term memory) to predict stock prices, ARIMA forecasting to predict the next hundred days of rising or falling of stocks, decision tree and random forest.

4.1 Data Selection.

The important step for analyzing data is that data need to exist, whereas I choose Kaggle for collecting my New York stock exchange dataset because it is an online community for data scientists which allows users to find and publish datasets, explore, and build models in the web-based data science community. Although, other data like Apple and Amazon I decided to take from https://finance.yahoo.com/ why I took it from here? because it provides me updated information on historical stocks and I can use it to predict the stocks for apple and amazon for the next few days.

Once my datasets were collected, then I decided to import them in RStudio (it is power ide for R language which provide standard features, environmental variables as well as version control management tools) and PyCharm (it is widely used in big companies for machine learning it can provide support for import like libraries) for pre-processing, predictions of stock, etc.

```
prices <- read.csv("D:/4th Year/Final Year Project/Mid-Implementation/NewYork Stock Exchange/prices.csv")
Apple <- read.csv("D:/4th Year/Final Year Project/Mid-Implementation/NewYork Stock Exchange/Apple.csv")
amazon <- read.csv("D:/4th Year/Final Year Project/Mid-Implementation/NewYork Stock
Exchange/AMZN.csv")
```

Figure 14: It displays codes for importing my datasets in RStudio.

4.2 Data Manipulation.

After importing my Apple, Amazon, and New York stock exchange datasets, in RStudio I performed cleaning approaches/processes because it is one of the important steps before performing any analysis. Removing some rows from the dataset, reordering our columns names of datasets, changing the name of the column in datasets, adding a new column in datasets, finding some empty values in datasets, and filled them by taking the mean of that column.

```
prices[-c(851264, 851263), ]

prices <- prices[-c(851264, 851263), ]

#Reordering column of prices dataset by their position

prices<-prices[, c(2, 1, 3, 4, 5, 6, 7)# All columns in my dataset are re-ordered as I want them to be.

#Reordering column of Apple dataset by their positions

Apple<-Apple[, c(1, 2, 3, 4, 5, 7, 6)]

#Renaming our prices dataset columns names

names(prices)[names(prices)== "symbol"] <- "Symbol"

names(prices)[names(prices)== "date"] <- "Date"
```

Figure 15: It displays some data manipulation codes from RStudio.

4.3 Exploratory Data analysis.

EDA is an approach to analyze our datasets to summarise their characteristics, it frequently used data visualisations and statistical techniques. Whereas in EDA mostly graphical techniques are used scatter plot, line graph, histogram, count map reduces, etc.

Have you heard of the phrase "*garbage in, garbage out*". With EDA it is more possible to "garbage in, perform EDA possible garbage out". While performing EDA in our datasets provides us a better understanding of our variables *("you do not know what you do not know, and if you do not know what you do not know then how you are supposed to know whether your insights make sense or not? "*), clean up our datasets, analyze relationships between them. For a better understanding of my dataset's variable, first of all, I displayed names of all columns and then changed their names if necessary, in RStudio.

```
#Checking names of column

names(Apple)

names(amazon)

#Renaming some apple dataset column names

names(Apple)[names(Apple)=="Date"] <- "Dates (2010-2016)"
```

Figure 16: Checking attributes from data and changing their names.

A graphical representation of our amazon closing stock prices, from 2010 to 2016. Which displays the market will close at different stock prices with a slight difference.

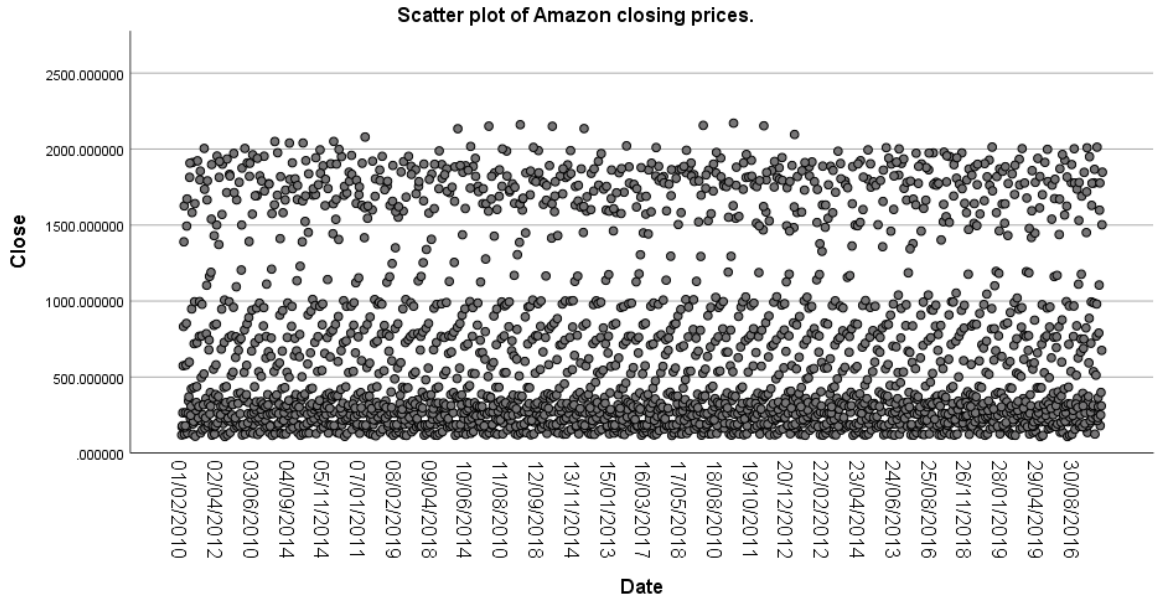


Figure 17: Scatter plot of amazon closing prices.

Descriptive statistics is performed to explore our datasets for better understanding, whereas it is performed in both IBM SPSS (**section 2.4**) and RStudio. Interquartile Statistics is performed which is a measure of variability, based on dividing the dataset into quartile. Quantile provides the range of probability distributions into continuous intervals with equal probability. Variance is the expectations of the square deviations of the random variable from its means. Standard deviations are the set amount of variation of a set of a value low standard deviation shows that the value tends to be close to the mean of the set. The range is the value that provides the difference between the largest and smallest values. The weighted mean is similar to the arithmetic mean, except each data point contributed equally to the final average. The mode in statistics provides us the most existing values in our datasets. Median in statistics which provides us the center value in a dataset. Mean in descriptive statistics which provides us the most average values of datasets.

```
mean_Open <- mean(prices$Open, na.rm = TRUE)
median_Open <- median(prices$Open, na.rm = TRUE)
mlv(prices$Symbol, method = "mfv")
wg_Mean_Open <- weighted.mean(prices$Open, na.rm = TRUE)
range_open_prices <- range(prices$Open, na.rm = TRUE)
sd_open_prices <- sd(prices$Open, na.rm = TRUE)
variance_Open <- var(prices$Open, na.rm = TRUE)
quantile_Open <- quantile(prices$Open, na.rm = TRUE)
iqr_Open <- IQR(prices$Open, na.rm = TRUE)
```

Figure 18: R code to represent some statistical method carried out.

Dual y-axis graph represents amazon's high and low value of stock prices, by taking a look at the below graph we can say that there is not much difference between market low and high price.

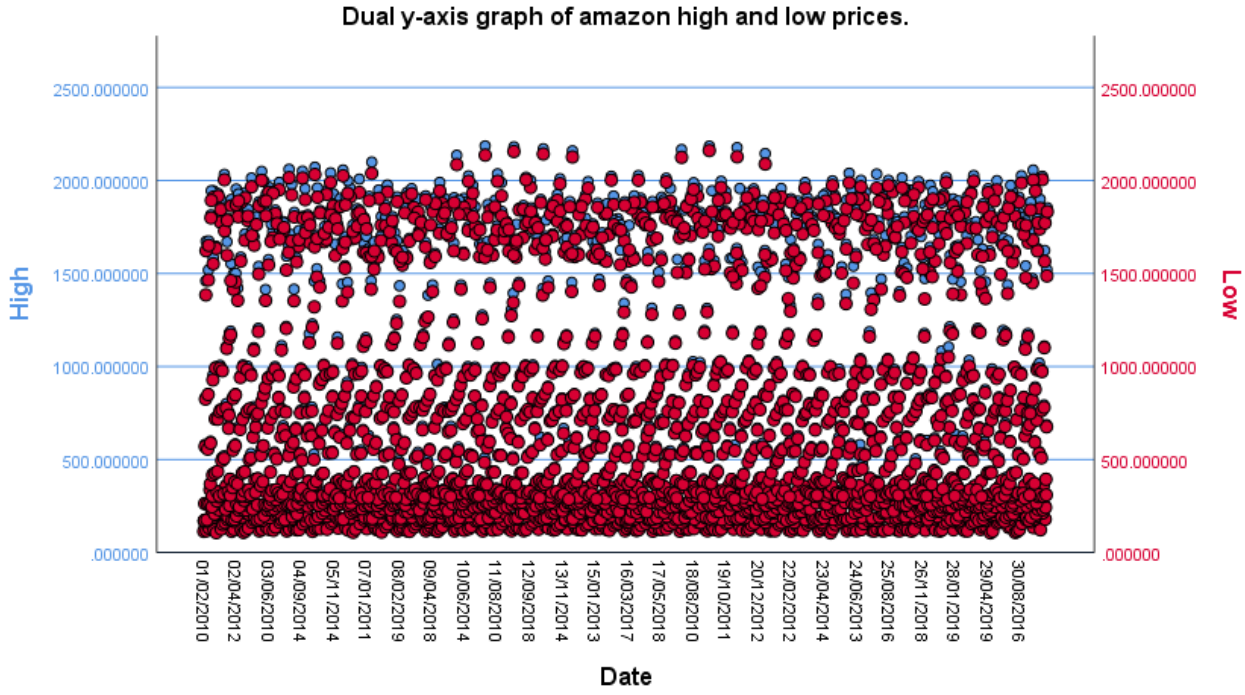


Figure 18: represents the difference between amazon's low and high stocks.

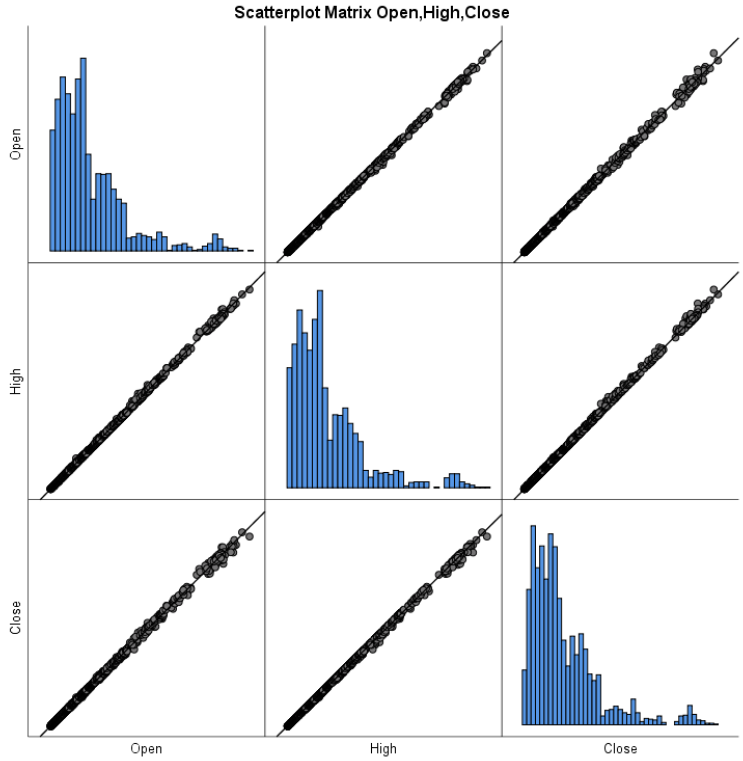


Figure 19: Scatter matrix open, high, close for Apple dataset.

4.4 Principal component analysis (PCA)

PCA is mostly a technique that transforms nation of numbers, it performs that by using correlation hence we can take from the set of large dimension and reduce that into a very smaller subject which can be known as the principal, whereas it can be used to uncover facts around data and realize its trends. I performed PCA analysis in RStudio on our Apple and Amazon datasets.

Summary of PCA of apple dataset.

	PC1	PC2	PC3	PC4
Standard Deviation.	1213.3280	10.12023	8.79200	3.29055
Proportion of variation	0.9999	0.00007	0.00005	0.00001
Cumulative Proportion.	0.9999	0.99994	9.99999	1.00000

Figure 20: principal component analysis summary of apple dataset.

Summary of PCA of Amazon dataset.

	PC1	PC2	PC3	PC4
Standard Deviation	45.8583	0.47217	0.39628	0.17651
Proportion of variation	0.9998	0.00011	0.00007	0.00001
Cumulative Proportion	0.9998	0.99991	0.99999	1.00000

Figure 21: principal component analysis of amazon dataset.

Whereas from above summarise of both Apple and Amazon datasets, we have four different types of component based on our input data which have four features so its calculated four features and depends on the number of features that we pass. Let's try to understand, what does this mean, lets start with a cumulative proportion of apple datasets it is telling us that our component one is a collection of close to 99 percent of the variation in our data, similar to our amazon dataset it is also 99 percent and close to variation of our data. It is a downward order of how much component is reflection variation in whole data.

4.5 KMeans and Hierarchical clustering.

These techniques are performed with R programming language, k-means clustering is performed in RStudio for apple stock dataset, whereas k-means clustering is a method of vector quantization, that aims are to partitions n observations into k cluster in which each observation belongs to the cluster with the nearest mean. Hierarchical clustering technique performed in RStudio for apple dataset, in data mining and statistics hierarchical clustering or HCA is a cluster analysis which is used to build a hierarchy of clusters. This type of analysis is mostly produced in the dendrogram. The advantage of HCA that it is easy to understand and implement. The dendrogram output of the algorithm can be used to understand the big picture as well as the groups in our data.

```
#Let's See our dataset
head(Apple)
#Standarize my data
Apple_cluster_Analysis<- scale(Apple[-1])
#Running KMeans in our standarize data
ourGroups <- kmeans(Apple_cluster_Analysis, 3)
#Visualizing our cluster Analysis
clusplot(Apple_cluster_Analysis, ourGroups$cluster)
#Summarize our data
ourGroups$size#It gives us size of our three groups that we just created above
ourGroups$betweenss
```

Figure 22: displays r code for Kmeans clustering.

```
#hierarchical clustering
idx <- sample(1:dim(Apple)[2], 40, replace = TRUE)
appleSample <- Apple[idx,]
appleSample$Open <- NULL

hc <- hclust(dist(appleSample), method="ave")
plot(hc, hang = -1, labels = Apple$Open[idx])
```

Figure 23: displays r code for hierarchical clustering.

4.6 Artificial recurrent neural networks (Long short-term memory).

AI recurrent neural network is performed in PyCharm for predicting stock prices of apple and amazon. Whereas a long short-term memory network is used which is a type of recurrent neural network and capable of learning order dependence in sequence prediction problems. How was LSTM used in PyCharm? First of all, I imported all my packages like Keras, TensorFlow, and NumPy, etc. and imported my datasets in PyCharm. Next, I build a new data frame for the prediction column (e.g., it can be closing, opening, high or low prices). After that, I convert my data frame to NumPy array and take several rows for training our model. Then, I performed a scale training dataset and convert it to the x and y train array, and NumPy array. Later, I convert the x train dataset to NumPy array and applied the LSTM model to compile the model and train model. Once my model was tested from 2D to 3D then I predict my model, find the root mean square value and visualise my graph. (Below flow chart represents steps by steps implementation process)

Why did I decide to use LSTM? It is an AI RNN network used in deep learning because our datasets are time-series datasets and it is nicely suitable for classifying, processing and make a prediction on the stock price. For the evaluation of our model, I performed the root mean square error value, RMSE is a nice measure to check how accurate is the model predictor and it is the standard deviation of residuals, as well as a lower value of our RMSE, indicate a better fit model accuracy. As our RMSE value for predicting closing stock price is 2.965824339898993, signify that our model is accurate. Whereas RMSE value for amazon closing stock price is 0.5208368889213619, it is a very lower value and represents a better fit accuracy of our model.

```
#Here I am creating new array containing scaled values from index 2134
test_data= scaled_data[training_data_len - 60: , :]
#Creating dataset x test and y test
x_test= []
y_test=dataset[training_data_len: , :]
for i in range(60, len(test_data)):
    x_test.append(test_data[i-60: i, 0])
#Converting the test data into numpy array
x_test=np.array(x_test)
#Reshaping our x test data from 2D to 3D
x_test=np.reshape(x_test, (x_test.shape[0], x_test.shape[1], 1))
#Models predicted price values
predictions=model.predict(x_test)
predictions=scaler.inverse_transform(predictions)
#Finding root mean square value
rmse=np.sqrt(np.mean(predictions-y_test)**2)
print(rmse)
```

Figure 24: Represents python for getting our RMSE value.

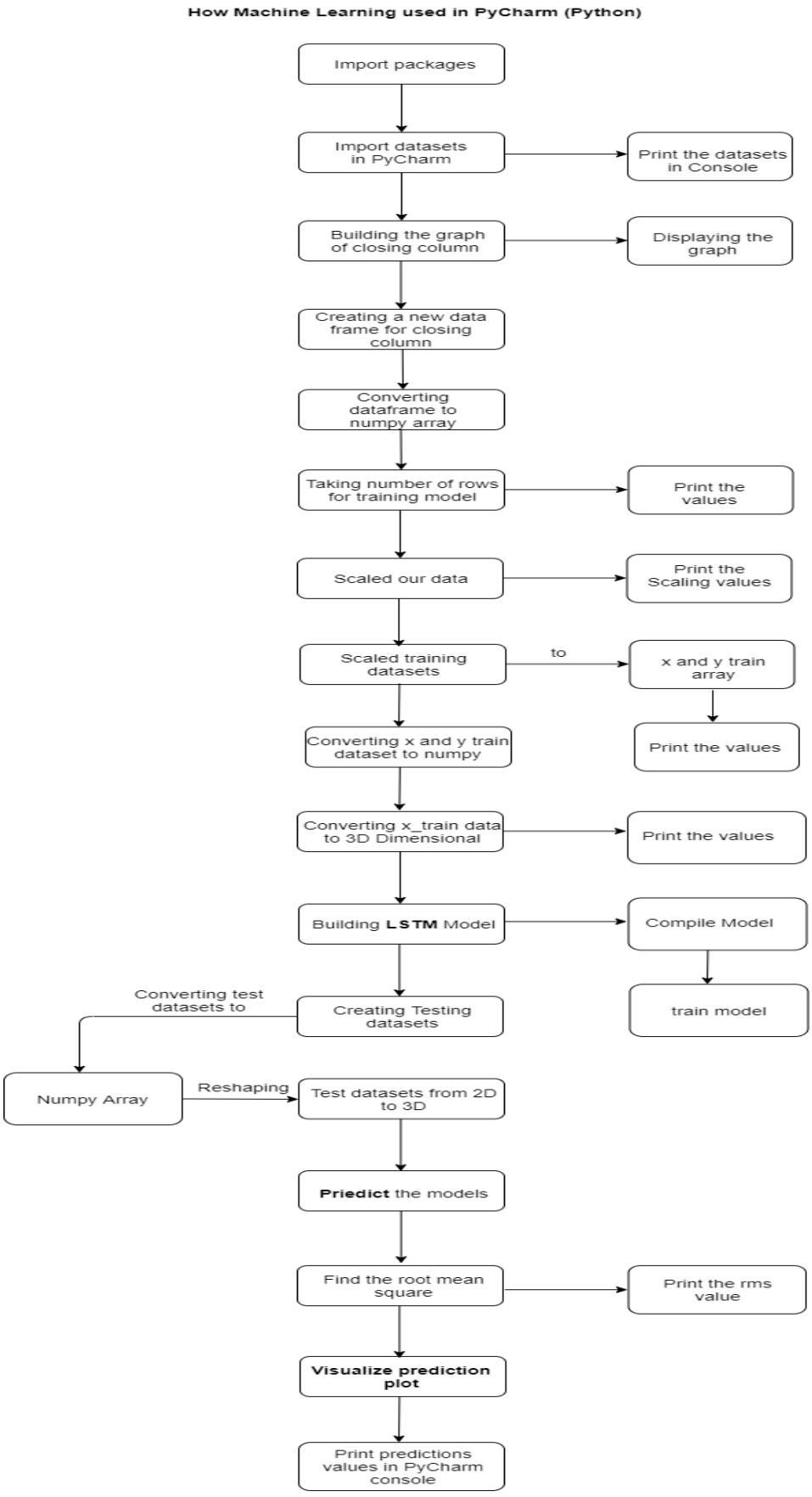


Figure 25: Flow chart of Long short-term memory.

4.7 ARIMA forecasting.

ARIMA is a model which is used in statistics and econometrics to measure events that happened over a while. This model can be used to understand past data and predict future data in series. ARIMA is applied in both datasets Apple Inc and Amazon to predict the future trends of stocks. During implementation, first I imported all my packages e.g., Quantmod, tseries, timeseries, forecast, and xts. Next, I import my data in RStudio of the last five years and plot a graph of a column which is going to predict. Then, I graph for ACF and PACF to check the difference between series and print adf value for amazon closing stock which is 0.01. Further, I have some 4 fit values, based on these four fit values I performed graphs, and test their accuracy. Once I have all four ARIMA models (The first one is auto ARIMA, the second one is custom ARIMA, the third one is guessing on different values based on our auto ARIMA, the fourth one is standard de facto default) then I plot them all together and check their accuracy.

```
fitA=auto.arima(AMZN_Close_Price, seasonal = FALSE)
tsdisplay(residuals(fitA), lag.max=4, main='(3,1,4)Residuals')
auto.arima(AMZN_Close_Price, seasonal = FALSE)#AIC/BIC47259.56/47286.13
fitB= arima(AMZN_Close_Price, order = c(1,2,4))
tsdisplay(residuals(fitB), lag.max=4, main='(1,2,4)Residuals')
fitC=arima(AMZN_Close_Price, order = c(5,1,4))
tsdisplay(residuals(fitC), lag.max=4, main='(5,1,4)Residuals')
fitD=arima(AMZN_Close_Price, order = c(1,1,1))
tsdisplay(residuals(fitD), lag.max=4, main='(1,1,1)Residuals')
par(mfrow= c(2,2))
term <- 100
fcast1 <- forecast(fitA, h=term)
plot(fcast1)
fcast2 <- forecast(fitB, h=term)
plot(fcast2)
fcast3 <- forecast(fitC, h=term)
plot(fcast3)
fcast4 <- forecast(fitD, h=term)
plot(fcast4)
#Mape accuracy
accuracy(fcast1)
accuracy(fcast2)
accuracy(fcast3)
accuracy(fcast4)
```

Figure 26: Displays codes of all my four fit ARIMA model, plot, and their accuracy.

4.8 Some important packages used during the implementation process.

Keras: for developing and evaluating my LSTM model.

TensorFlow: for training inference of deep neural networks.

NumPy: for importing large and multi-dimensional arrays.

Cluster: for grouping our set of an object to perform Kmeans clustering.

Forecast: tools for displaying and analyzing time-series data.

Quantmod: use of development, testing and training model.

Readxl: for reading our excel file.

Dplyr: for manipulation of our datasets.

5.0 Results

Principal Component Analysis for Apple dataset.

Standard deviations of all four inserted PCA.

	PC1	PC2	PC3	PC4
Standard Deviation	45.8582836	0.4721697	0.3962777	0.1765095

Figure 27: Display standard deviation of PCA

Rotation of our four Principal component analysis.

Rotation	PC1	PC2	PC3	PC4
Low	0.4928549	-0.3136895	0.6794702	0.4438617
High	0.5068845	0.2855968	-0.5967898	0.5525799
Open	0.4999929	0.6501982	0.2677424	-0.5055327
Close	0.5001692	-0.6302981	-0.3323796	-0.4920152

Figure 28: Display rotation of PCA.

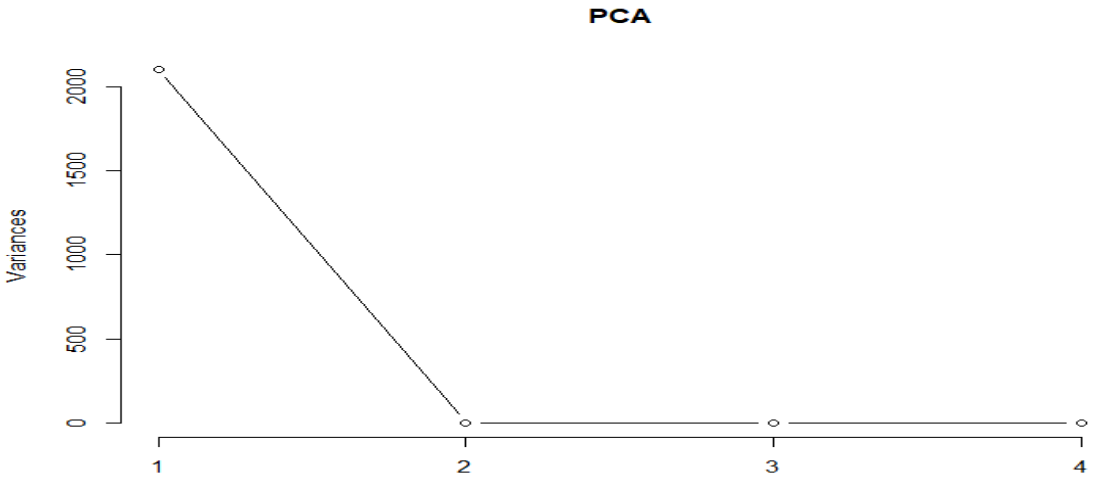


Figure 29: Display line graph for apple PCA analysis.

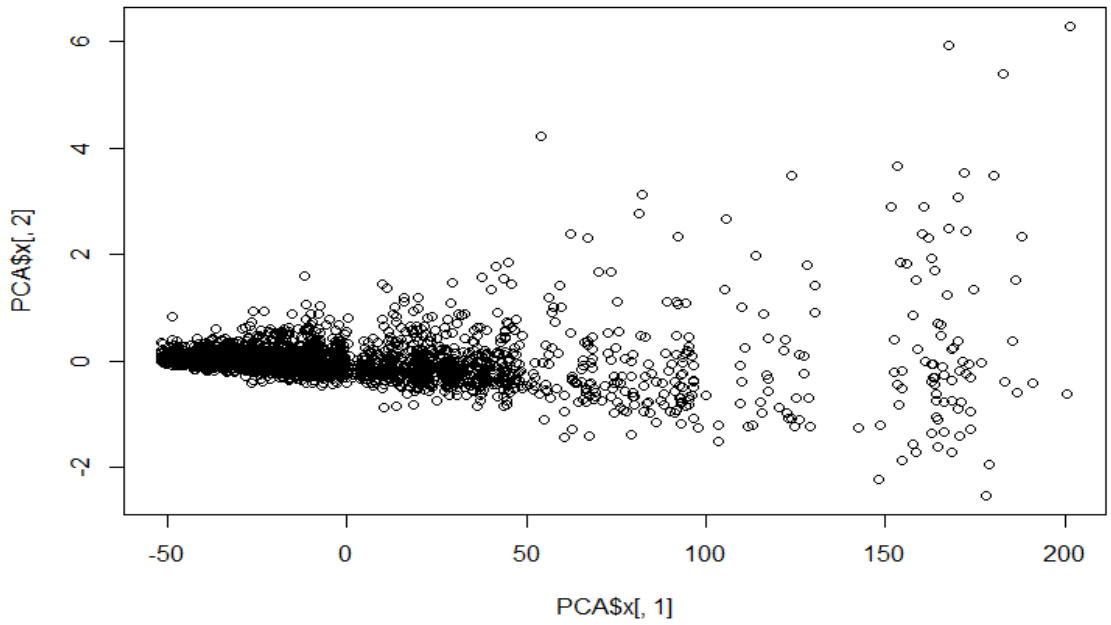


Figure 30: Displays a scatter plot between apple PCA1 and PCA2.

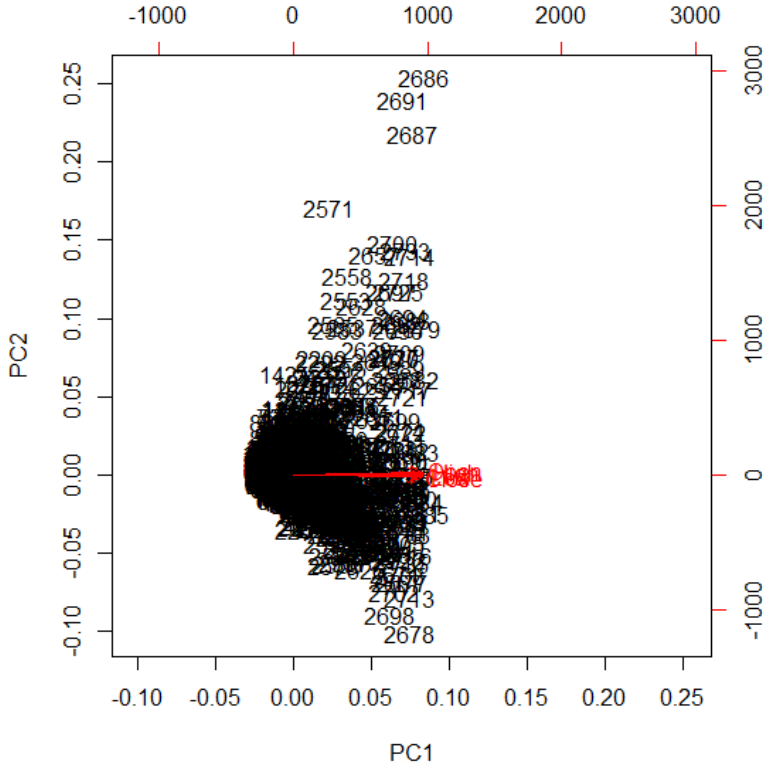


Figure 31: Bipolot a better graphical representation of apple PCA.

A scree plot clearly defines that our component PC has the highest variation than others.

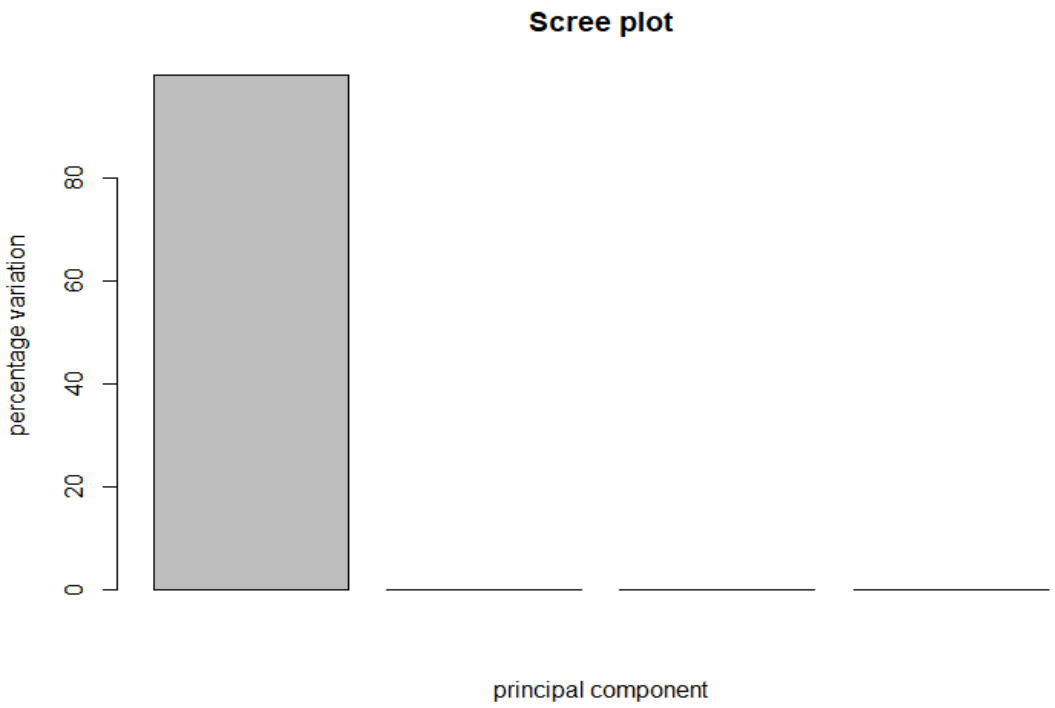


Figure 32: Scree plot for our PC1, PC2, PC3, and PC4.

```
pca.var=PCA$sdev^2
pca.var.per<-round(pca.var/sum(pca.var)*100,2)
barplot(pca.var.per,main="Scree plot",xlab="principal component",ylab="percentage variation")
```

Figure 33: Display code, how scree plot was built in RStudio.

Principal Component Analysis for amazon dataset.

Standard deviation for amazon inserted all four PCA.

	PC1	PC2	PC3	PC4
Standard Deviation	1213.328025	10.120232	8.792002	3.290552

Figure 34: Display standard deviation of amazon PCA.

Rotation for amazon for our four principal component analysis.

Rotation	PC1	PC2	PC3	PC4
Low	0.4947542	-0.3019312	0.6733979	0.4589020
High	0.5049221	0.2718949	-0.6166203	0.5393572
Open	0.5002514	0.6579410	0.2656571	-0.4962747
Close	0.5000205	-0.6340539	-0.3094203	-0.5022094

Figure 35: Displays rotation of amazon for four PCA.

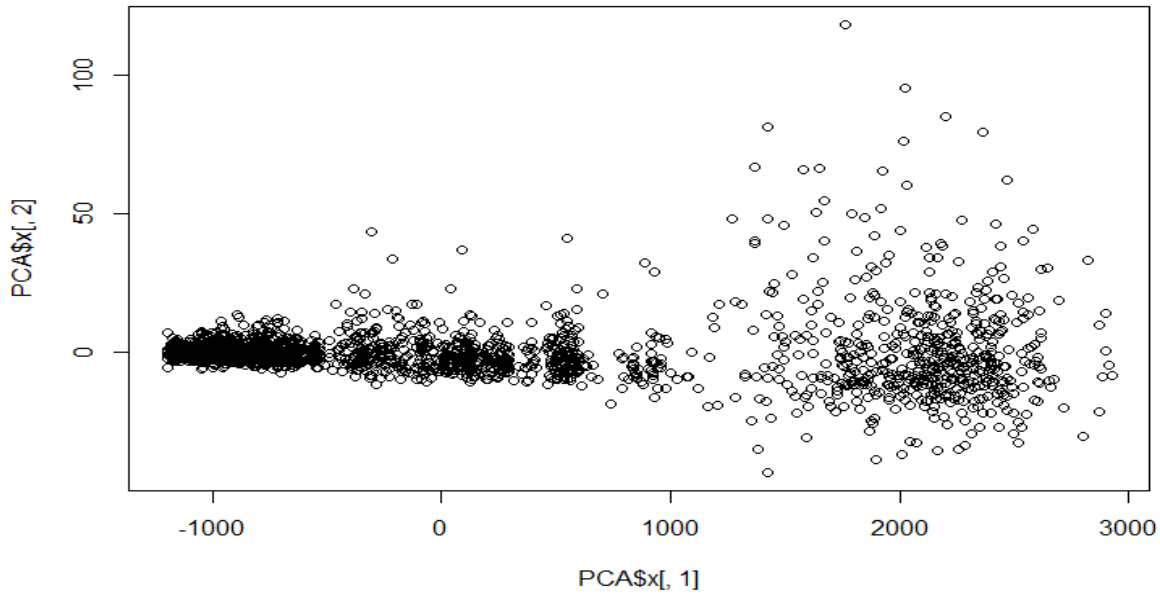


Figure 36: A scatterplot difference between amazon PC1 and PC2.

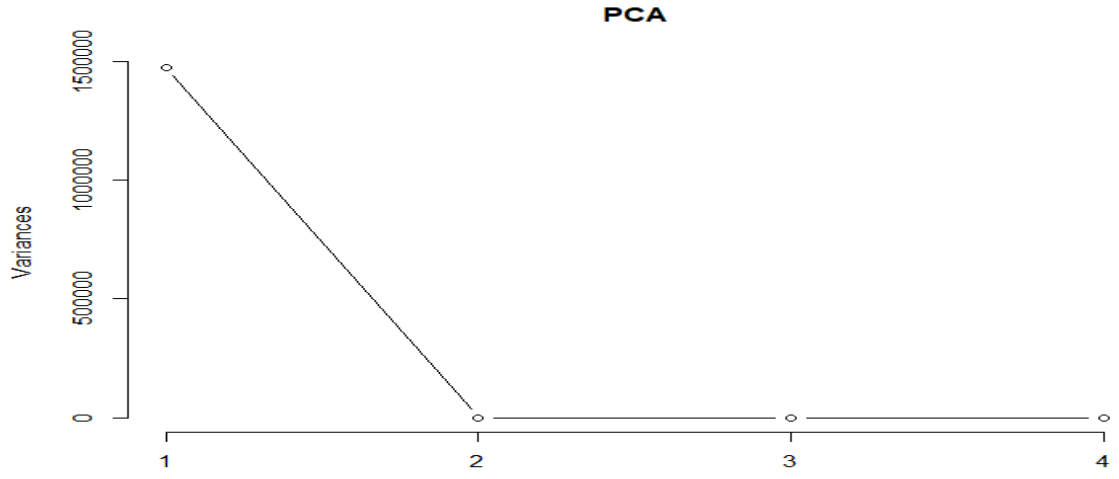


Figure 37: A line graph representation for our amazon PCA.

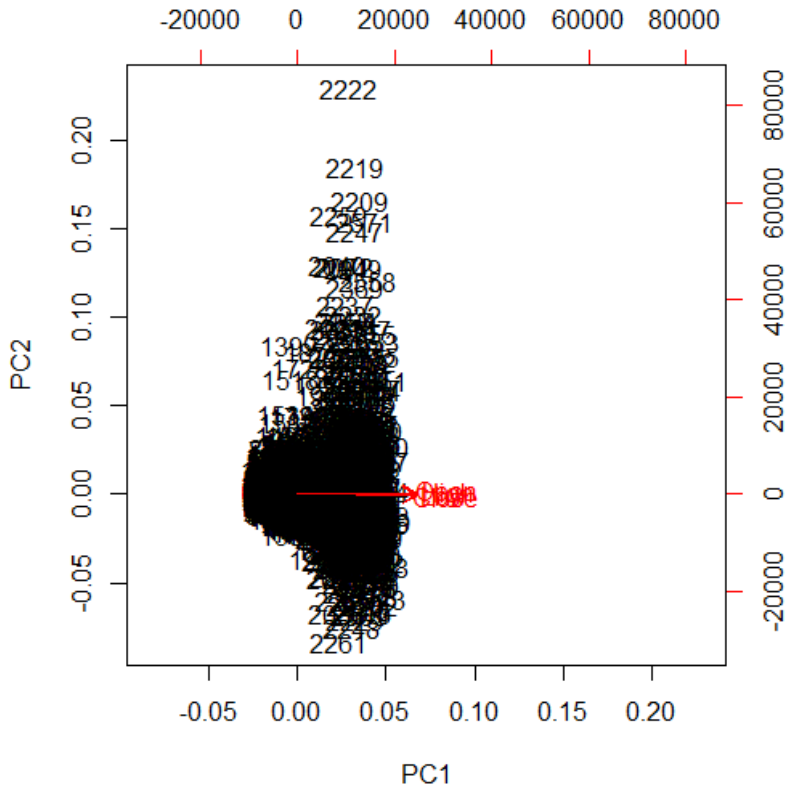


Figure 38: Graphical representation of our amazon in the shape of Biplot.

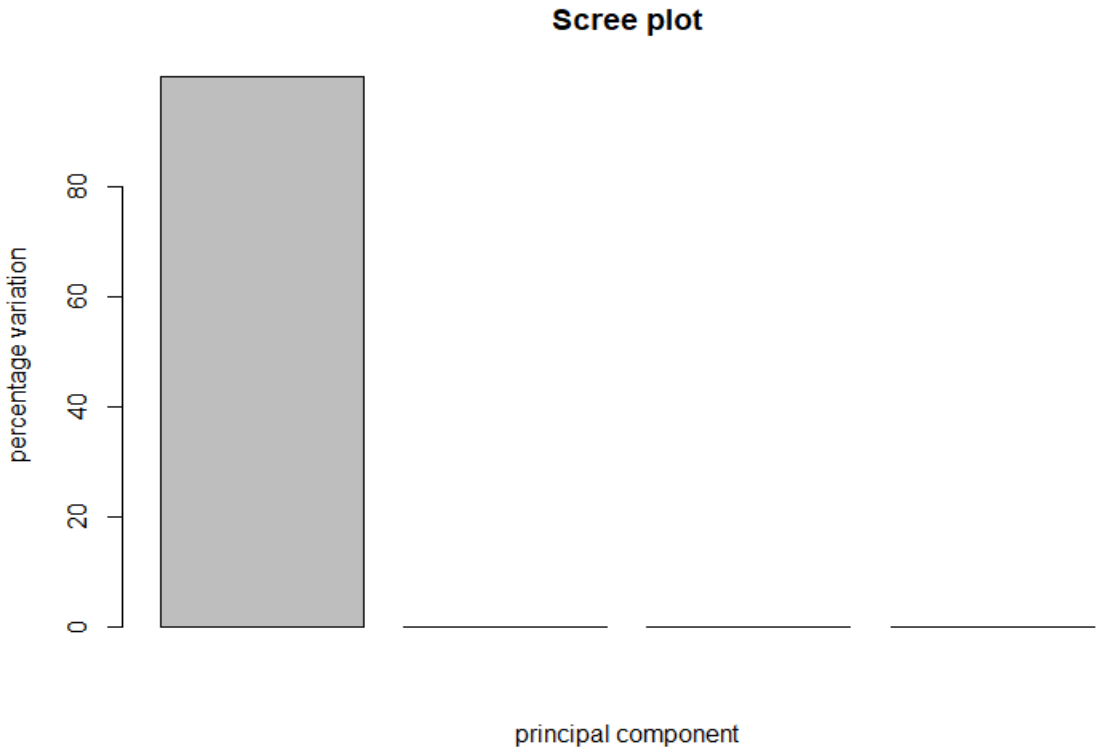


Figure 39: A scree plot for amazon principal component analysis.

Regression analysis between apple's lowest stock prices and highest stock prices displays there is not much difference in opening and closing stock market price.

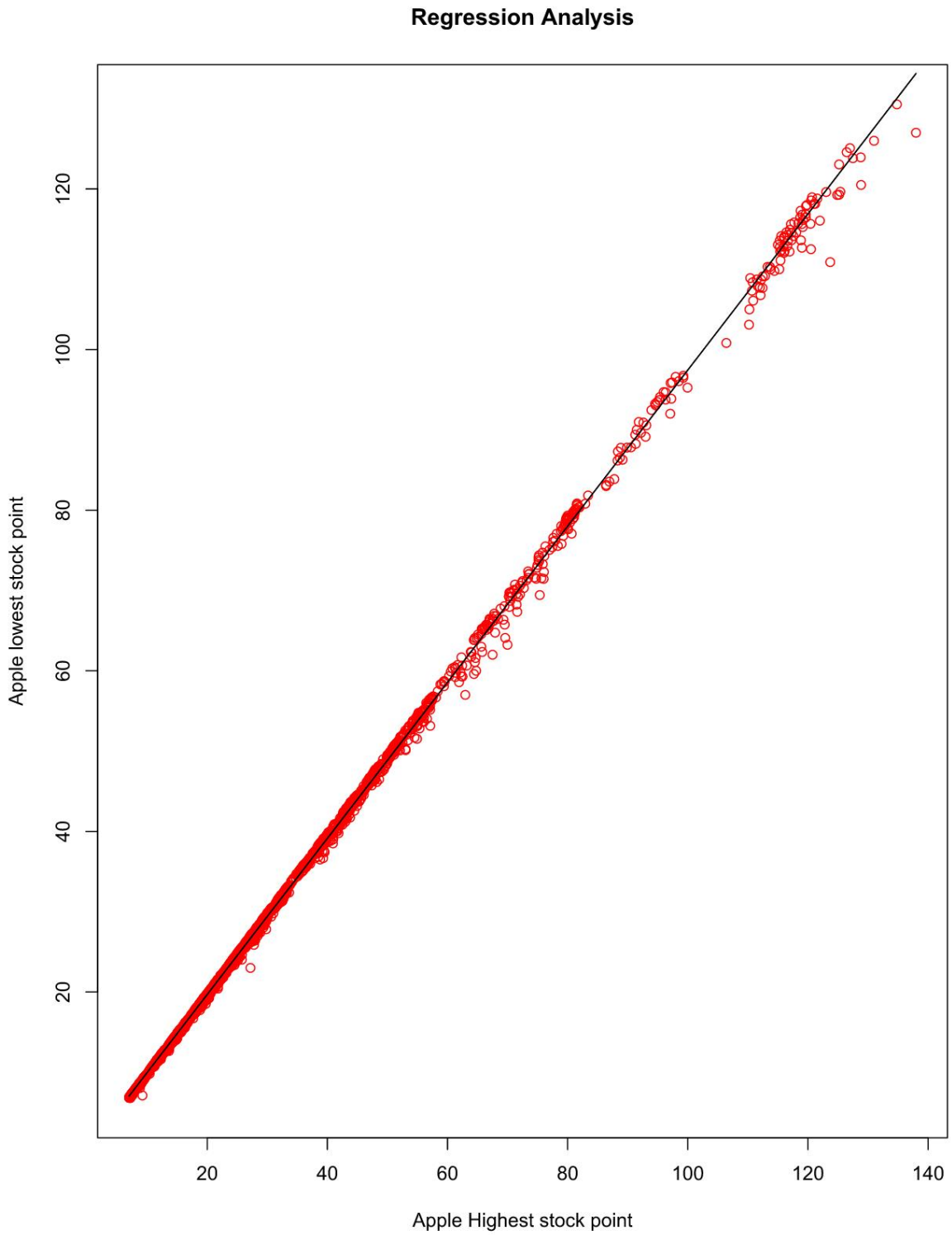


Figure 40: regression analysis between apple's high and low price.

KMeans and Hierarchical clustering output.

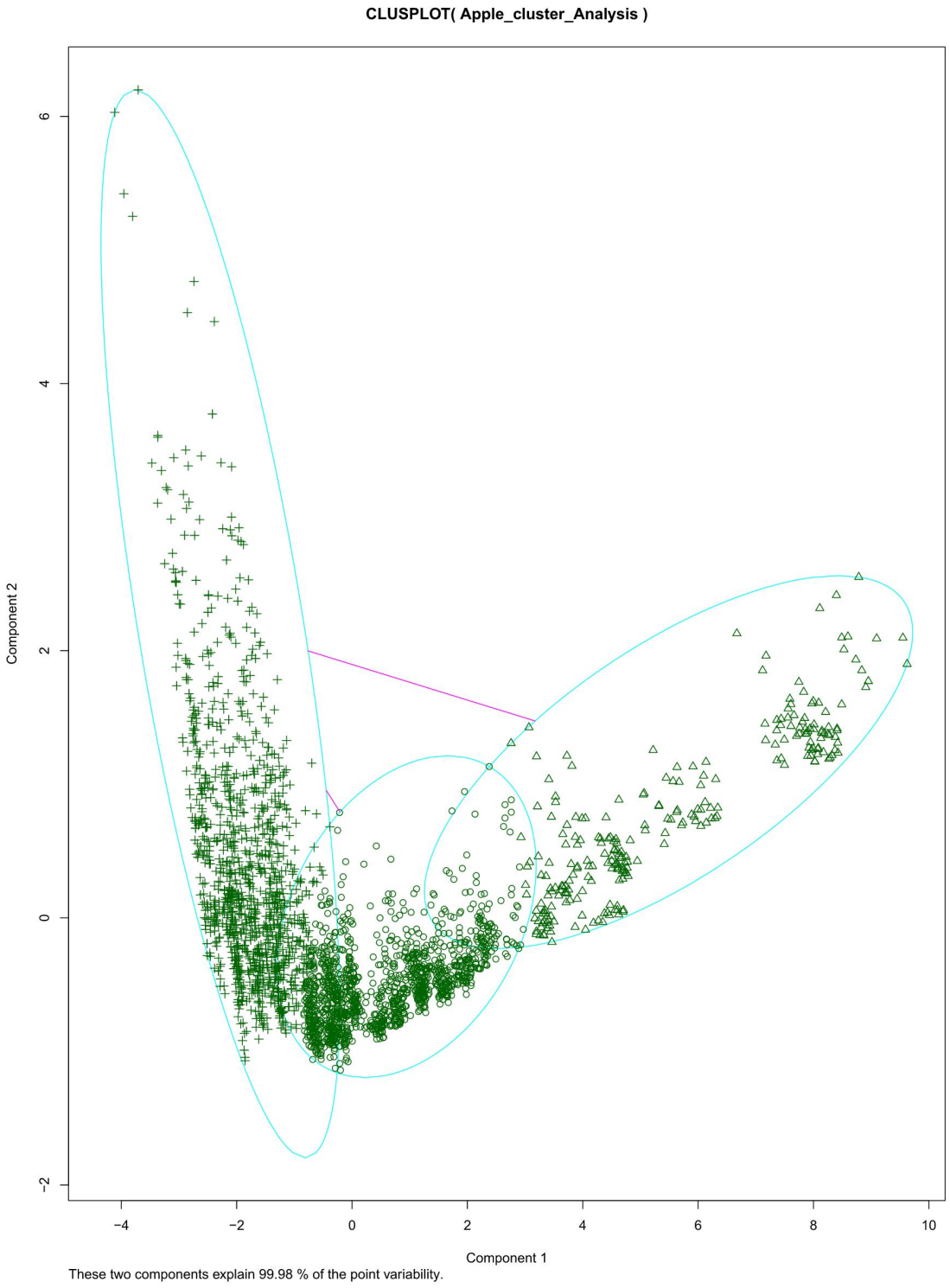


Figure 41: represents Kmeans clustering analysis on the dataset.

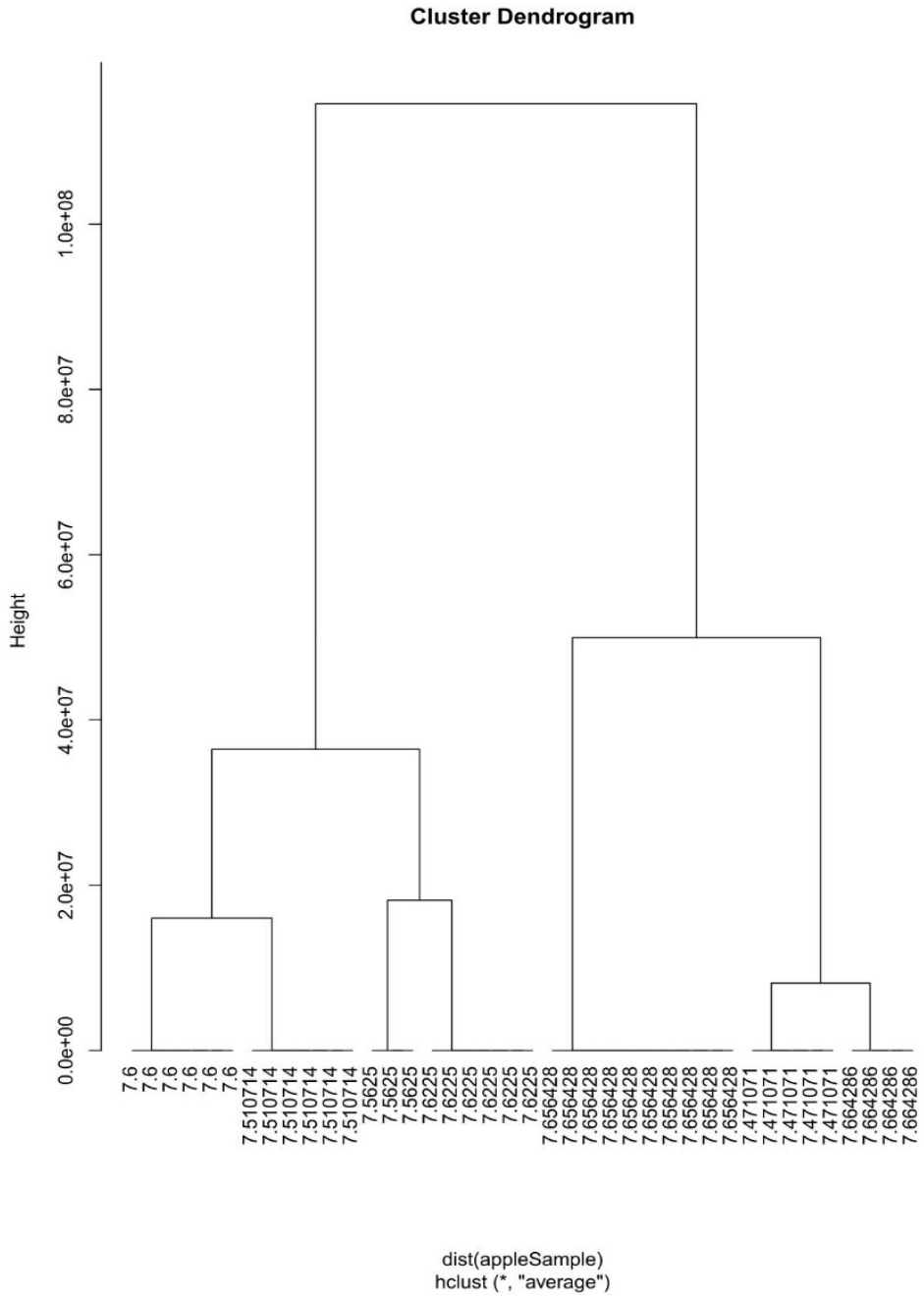


Figure 42: Represent hierarchical clustering analysis in cluster dendrogram of first 40 values open column from Apple dataset.

ARIMA Forecasting.

ARIMA forecasting is used for both Apple Inc and Amazon datasets to predict the next 100 days' stock price trends. All implementation is done in RStudio by using the R programming language.

ARIMA forecasting for amazon high stock prices.

Figure 43: Line graph on last 5 years amazon high stock prices with rising and fall with series of time from 2015-01-02 to 2021-05-03.

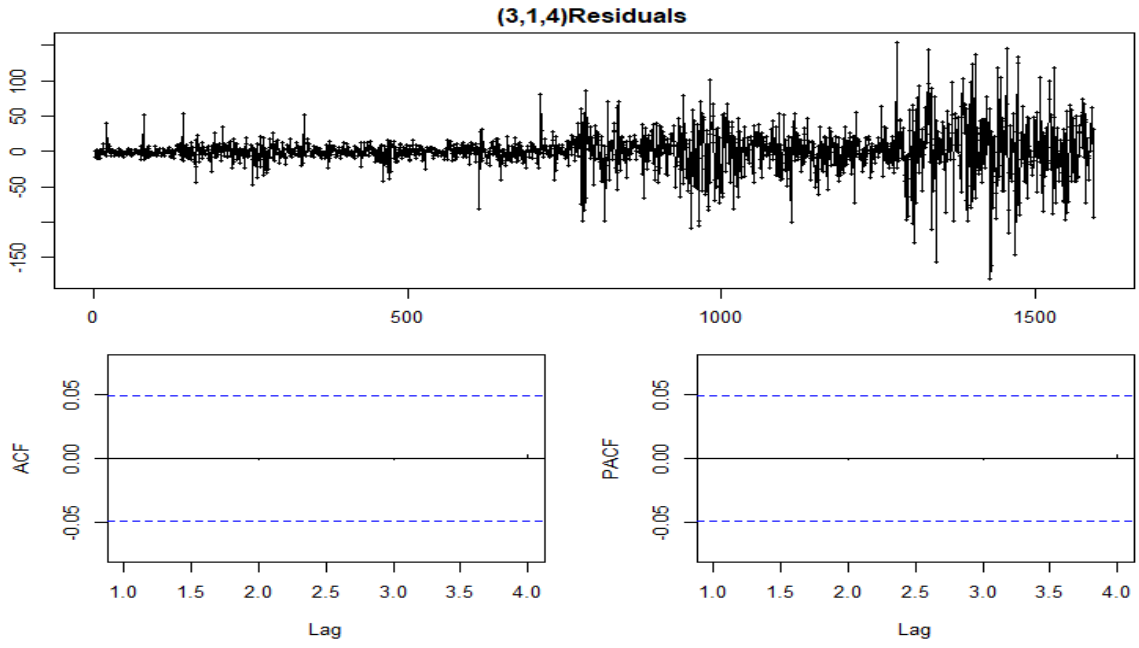


Figure 44: It is a model residual an auto ARIMA of Amazon high stock prices with AIC and BIC values (15426.84/15453.71).

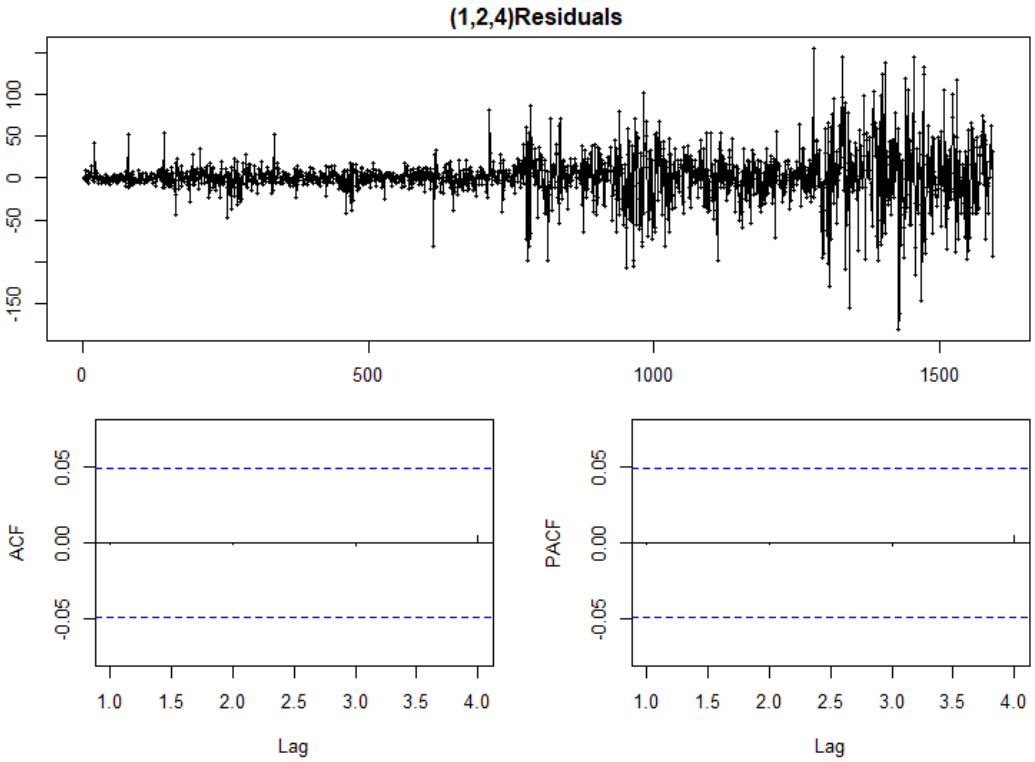


Figure 45: It represents model residual a custom ARIMA (1,2,4) of amazon high stock prices.

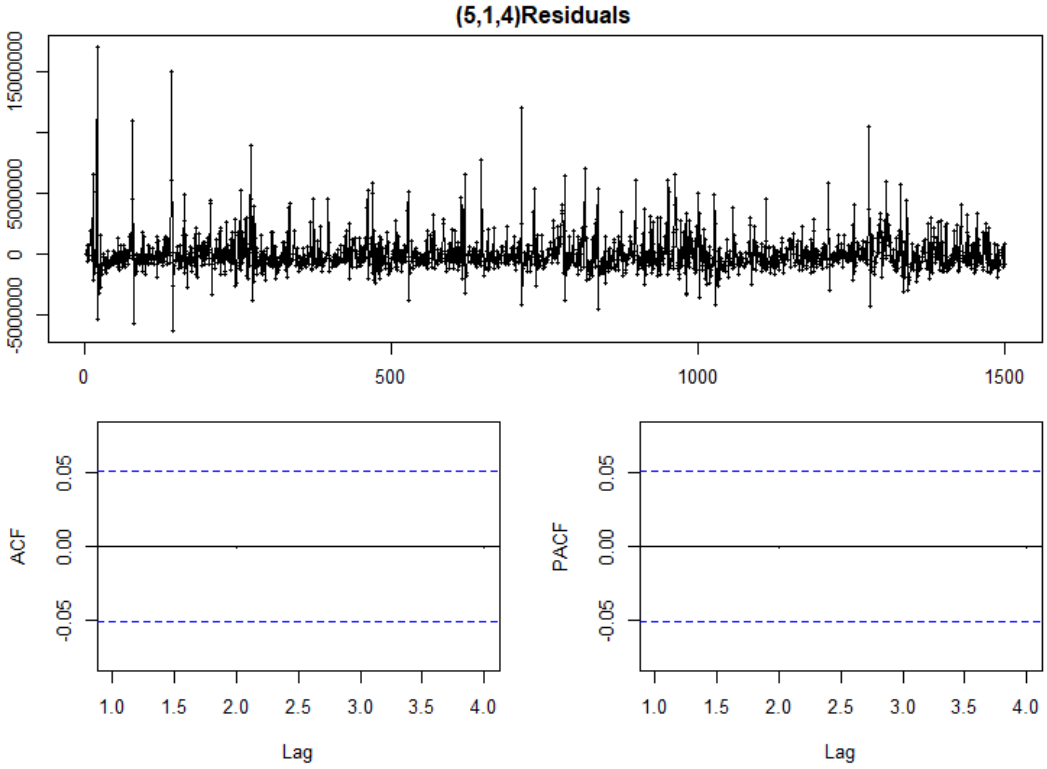


Figure 46: It represents guessing on different values based on our auto ARIMA.

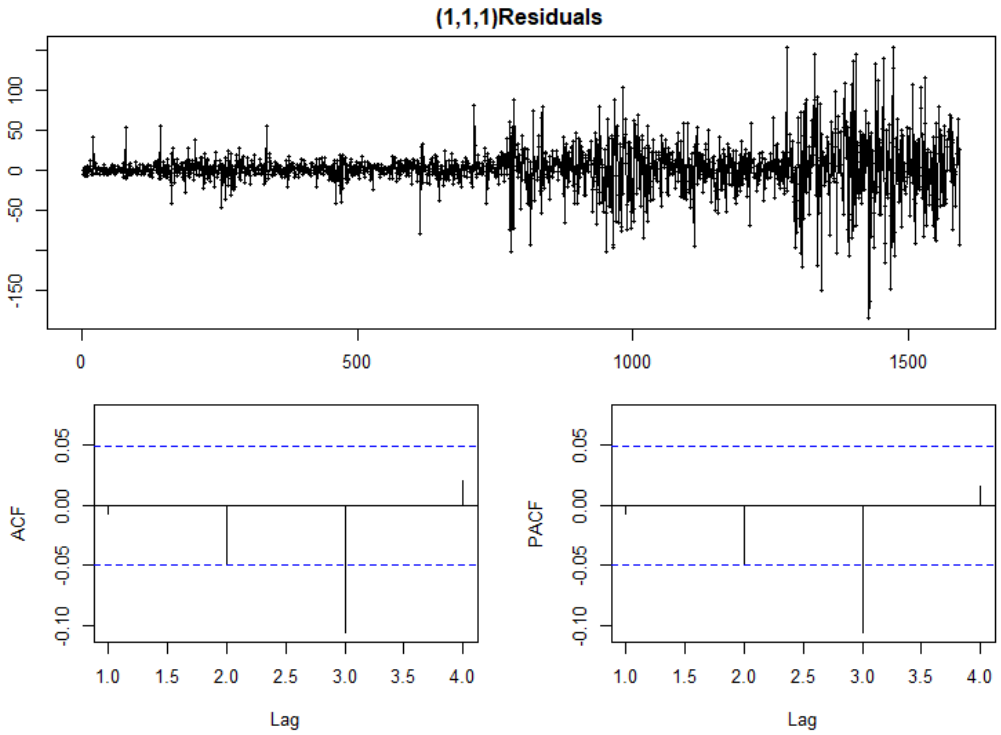


Figure 47: Standard de facto default ARIMA on amazon high stock prices.

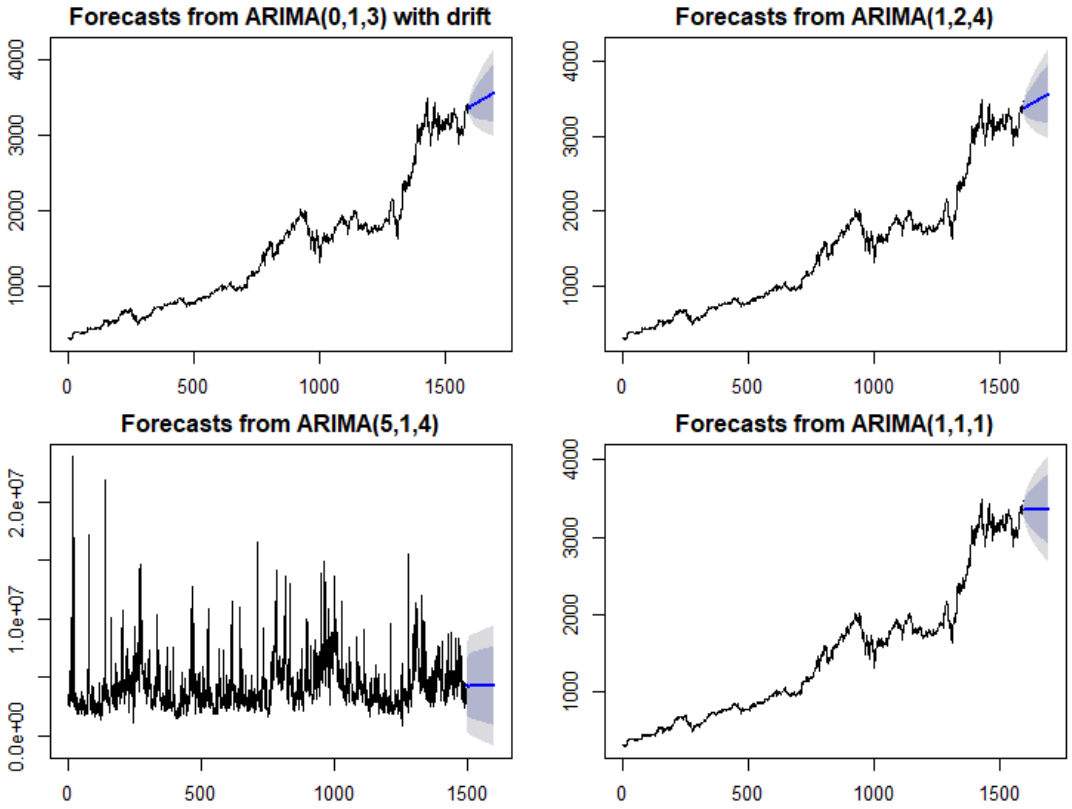


Figure 48: It shows us all four ARIMA models together with the prediction of the next hundred days, So we can say that based on our auto and custom ARIMA there is an increase in amazon high stock prices.

ARIMA forecasting for amazon close stock prices.

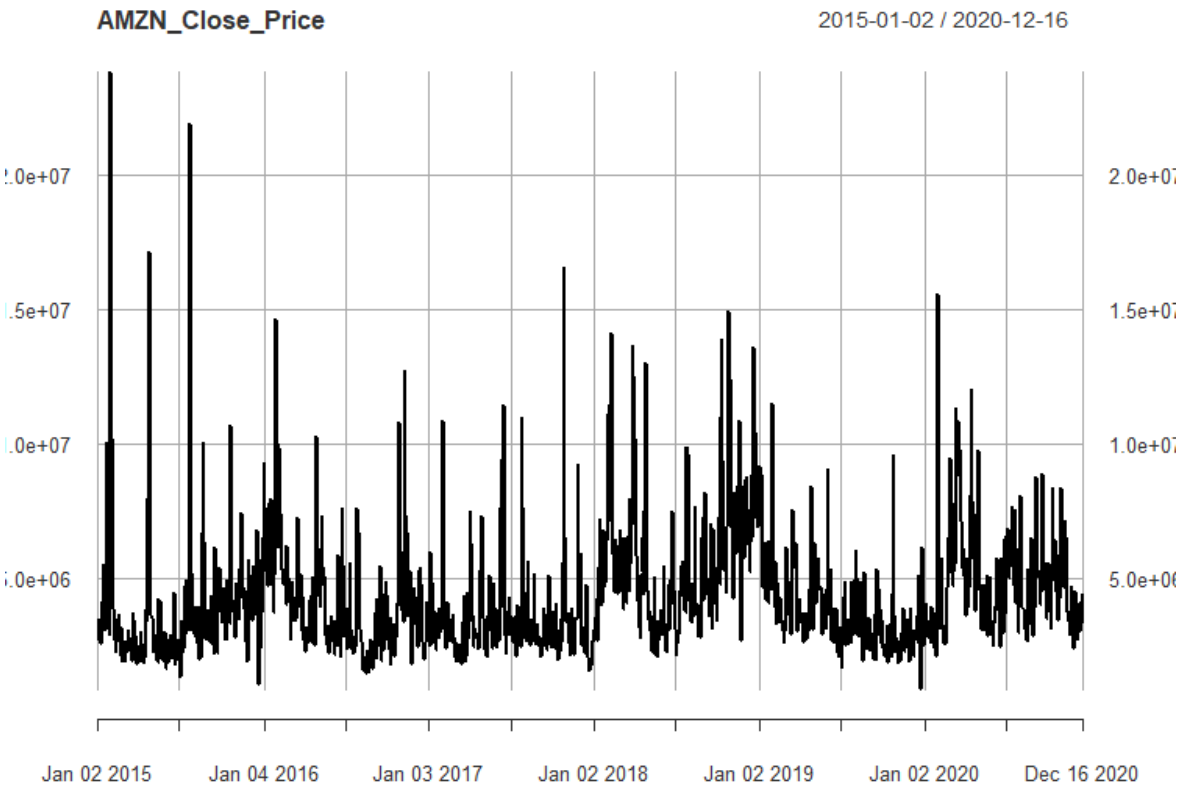


Figure 49: It represents amazon closing stock prices, from 2015-01-02 to 2020-12-16.

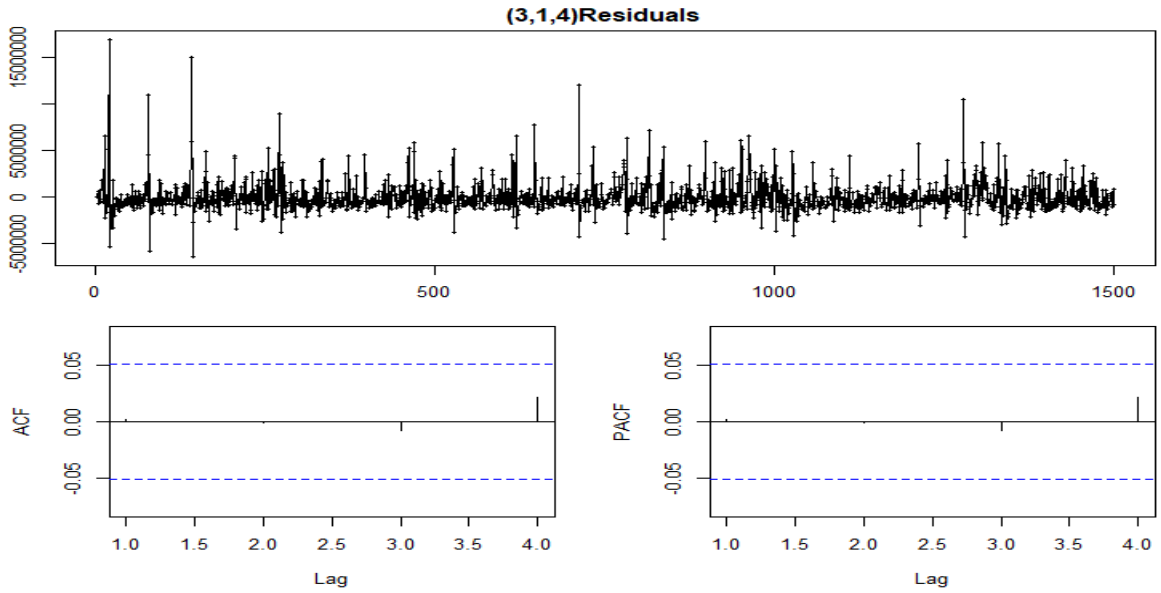


Figure 50: Model residual auto ARIMA of amazon closing price with AIC 47259.56and BIC 47286.13.

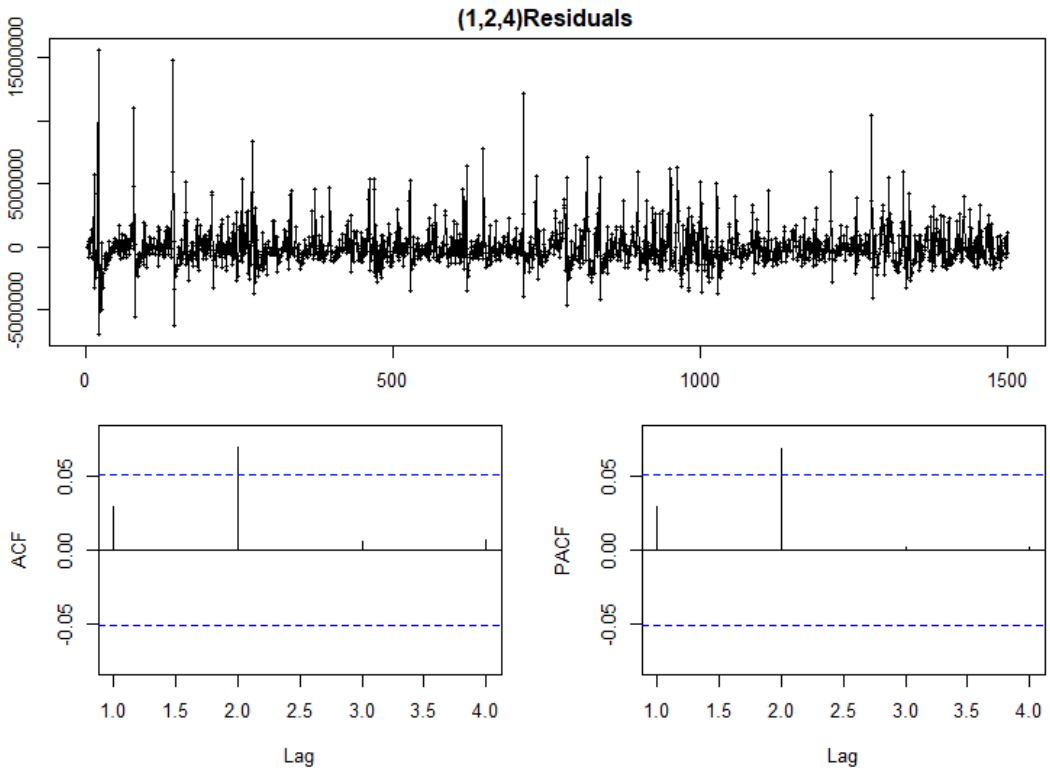


Figure 51: Model residual (1,2,4) custom ARIMA for amazon closing stock price.

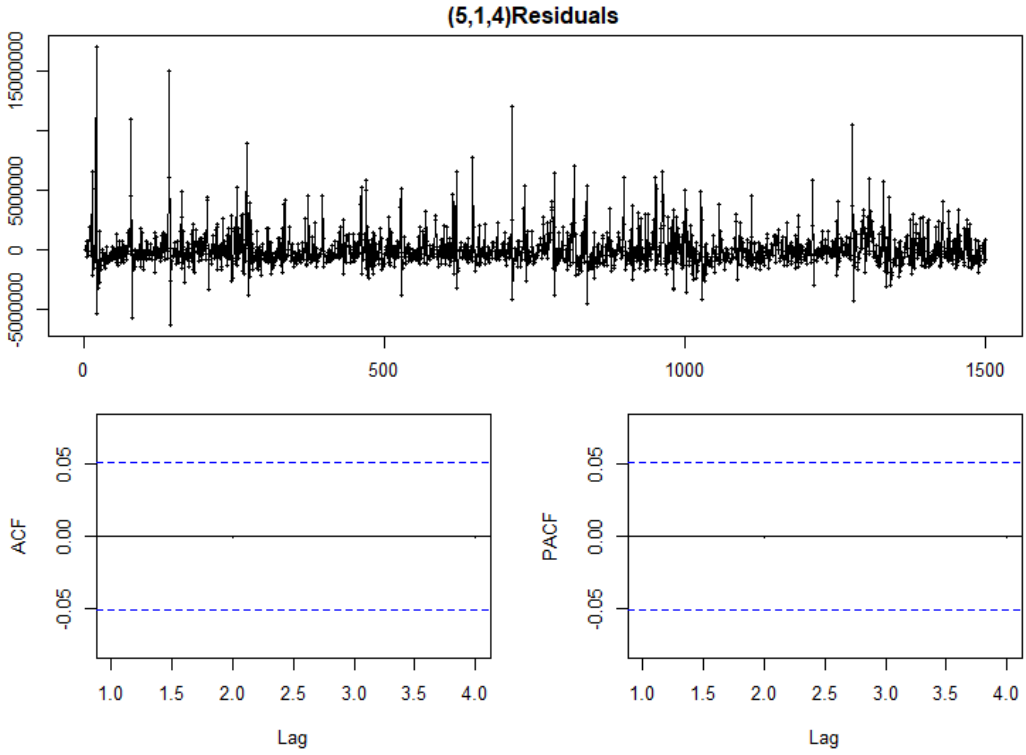


Figure 52: Model residual (5,1,4) guessing on different values based on auto ARIMA.

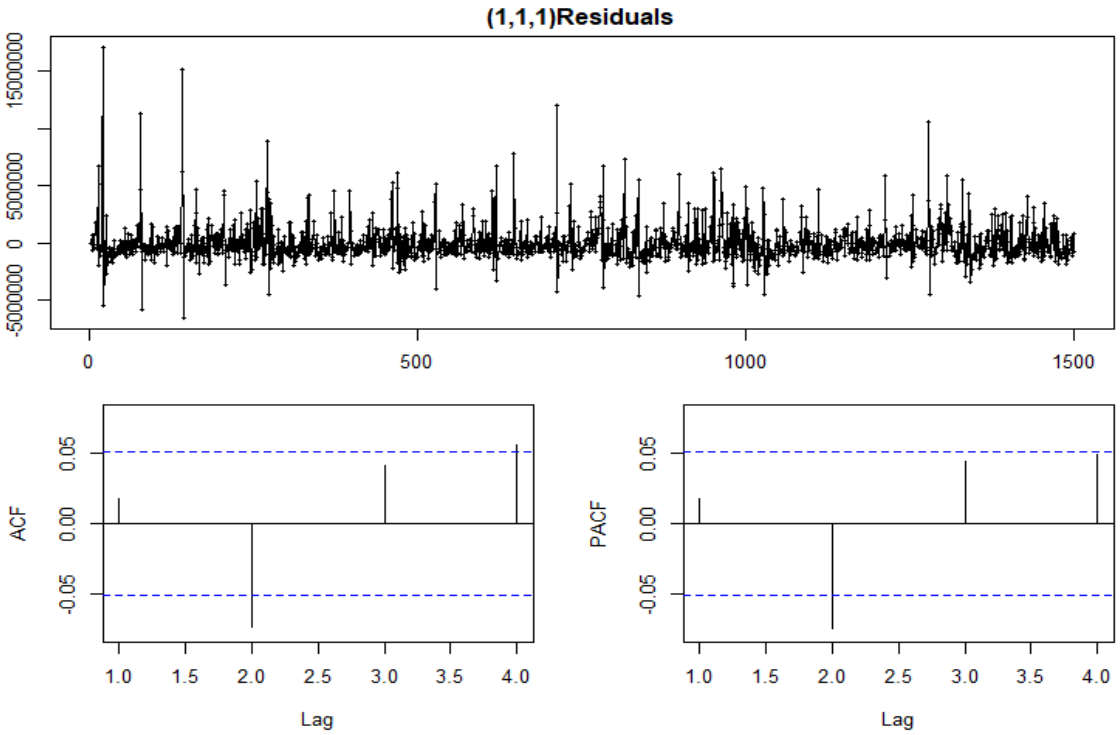


Figure 53: Model residual (1,1,1) standard defacto default ARIMA.

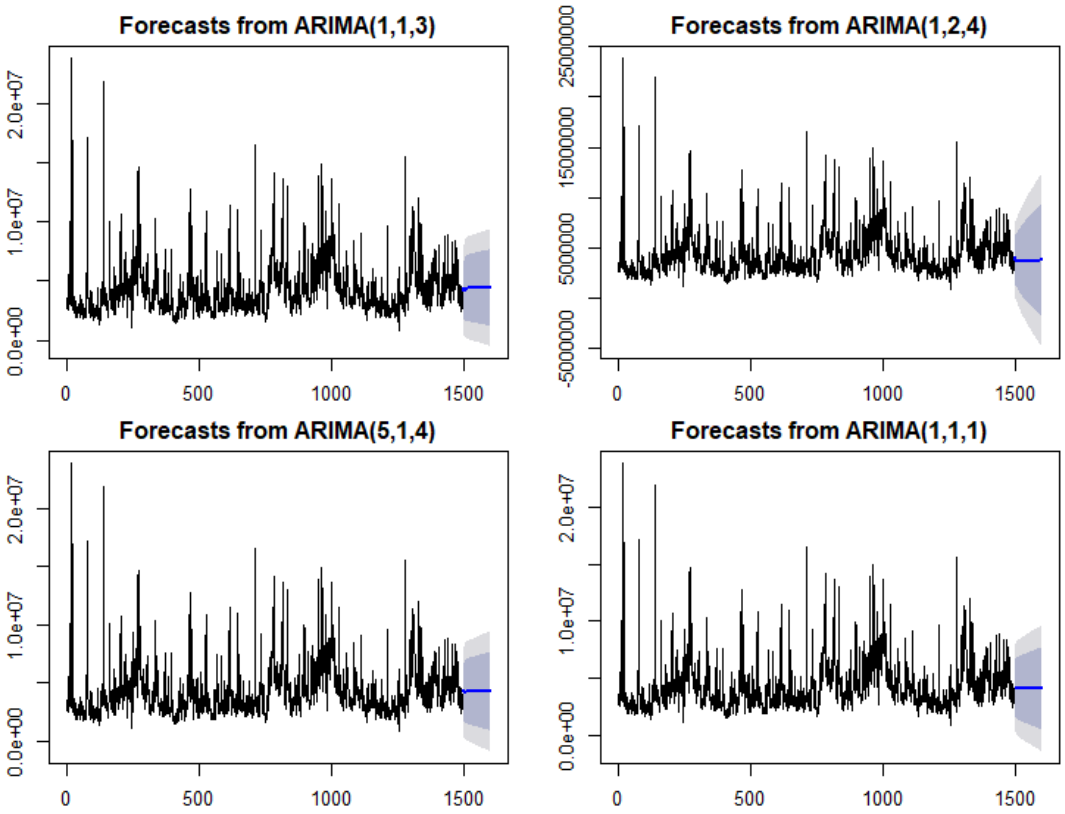


Figure 54: It shows us all four ARIMA models together with a prediction of the next hundred days, So we can say that based on all ARIMA model prices of our closing stock prices of amazon will slightly remain constant.

ARIMA forecasting for apple high stock prices.

Figure 55: It represents apple's high stock prices from 2015-01-02 to 2021-05-03.

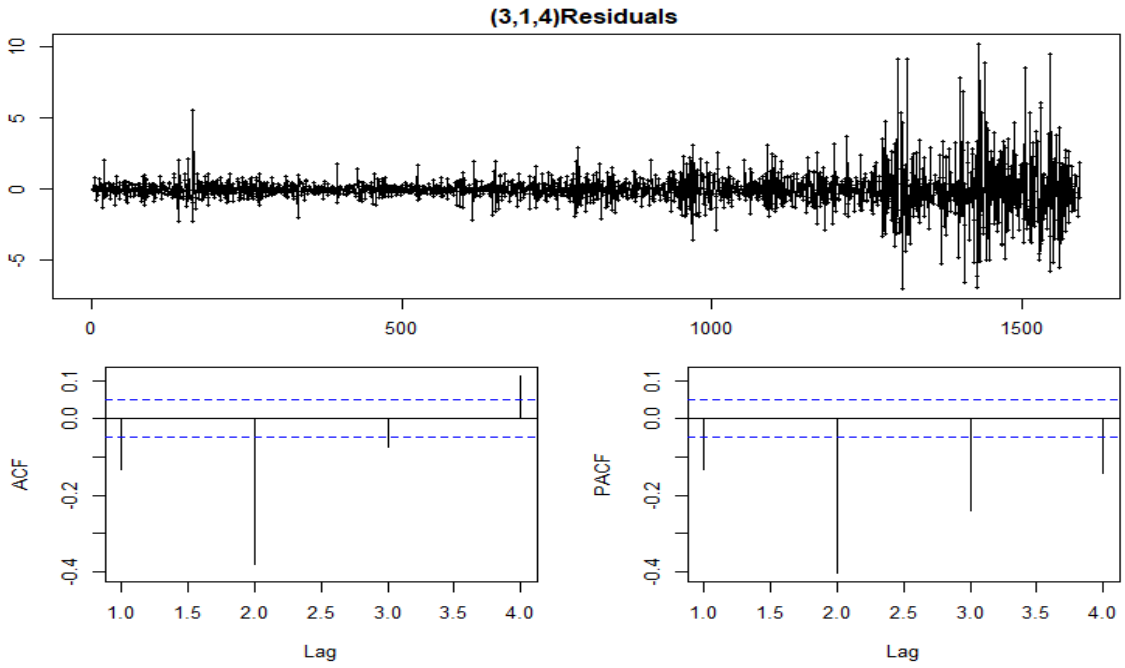


Figure 56: It represents model residual (3,1,4) auto ARIMA of apple high stock prices with AIC 15426.84, BIC 15453.71.

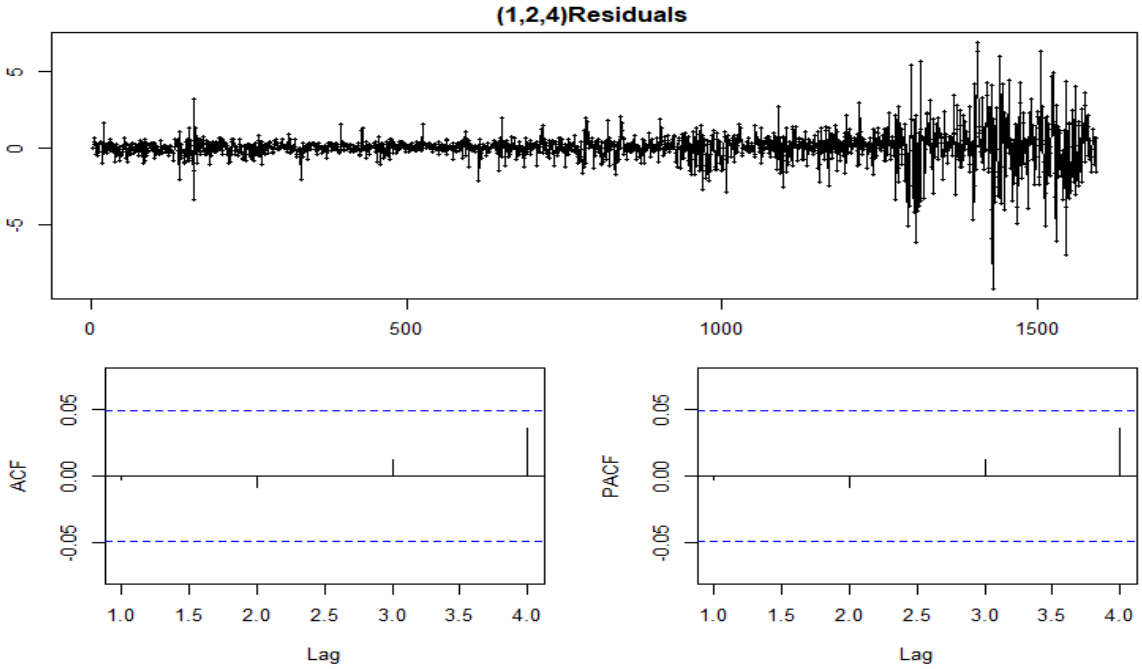


Figure 57: Custom model residual (1,2,4) for apple high stock prices.

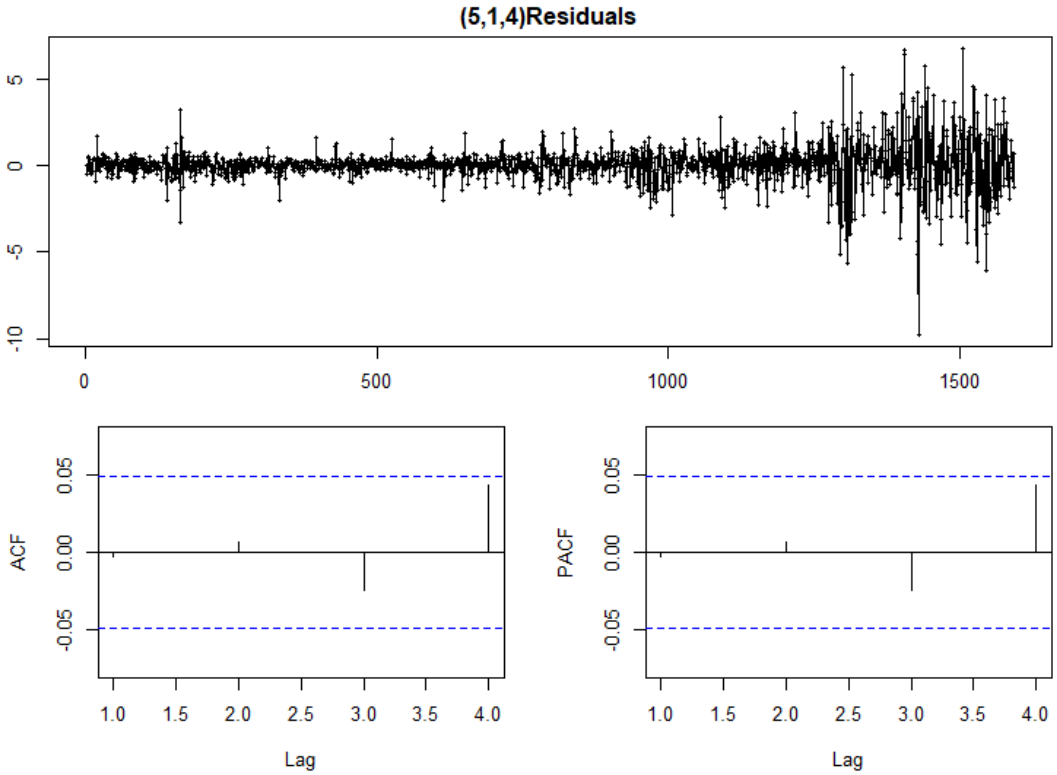


Figure 58: Guessing on different values on auto Arima model residual for apple high stock prices.

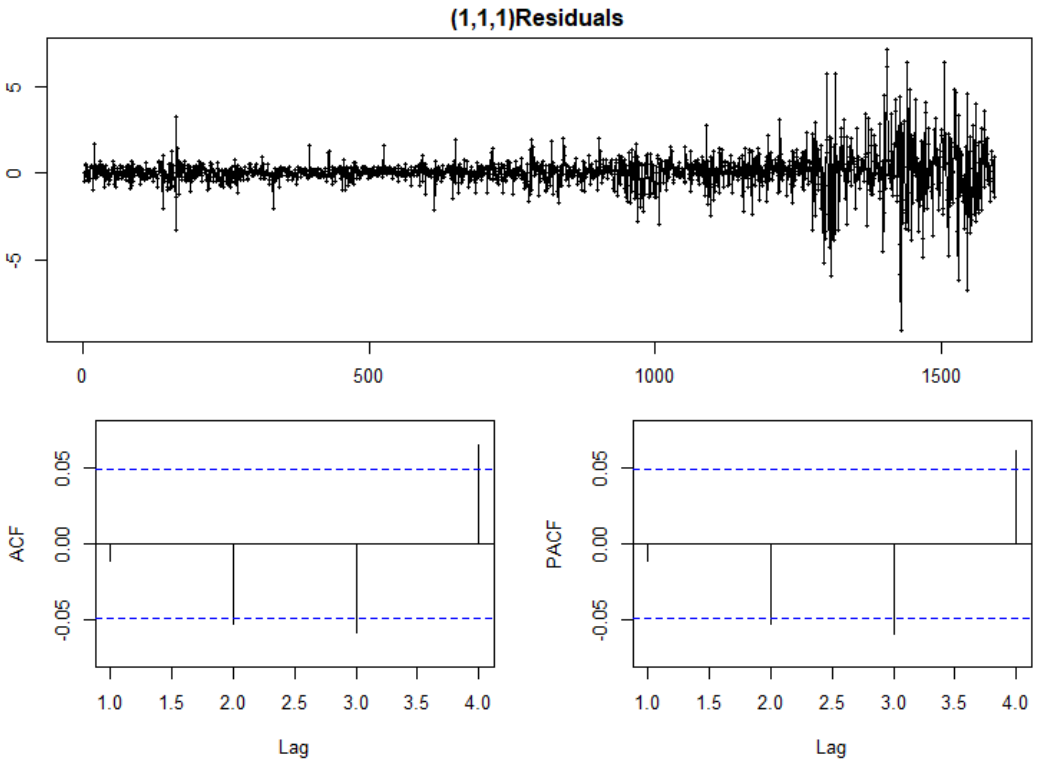


Figure 59: Standard de facto model residual for apple high stock prices.

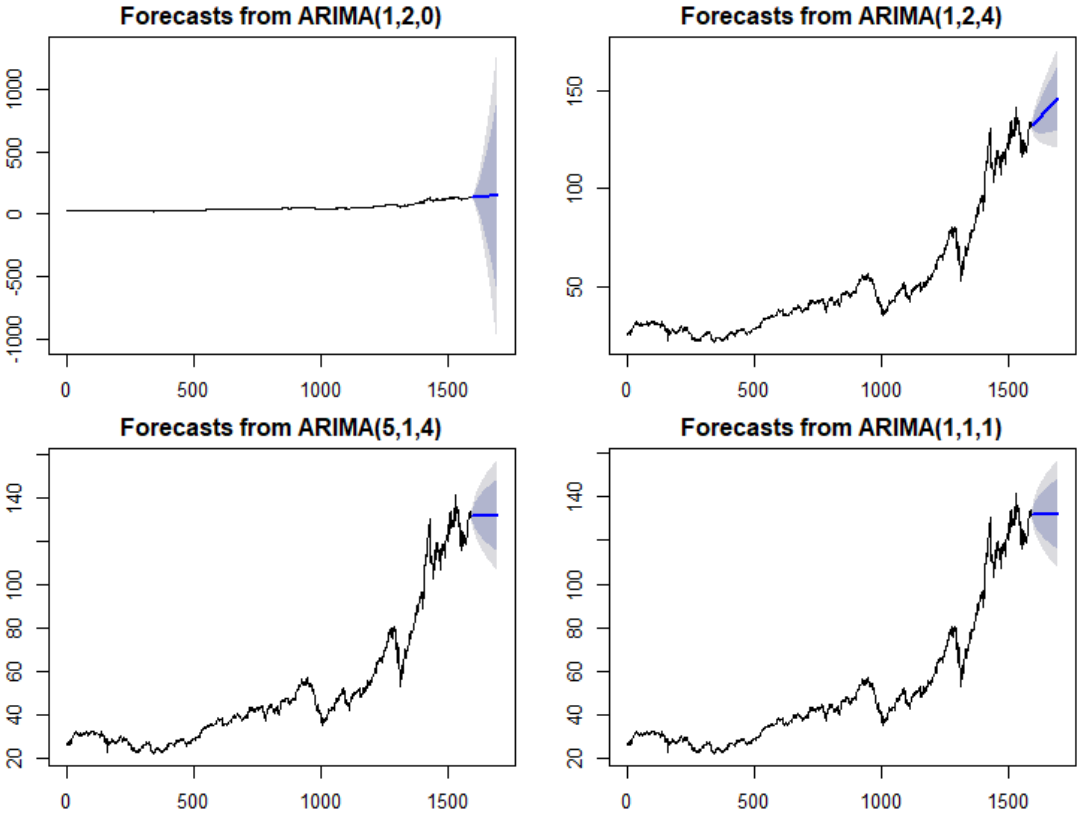


Figure 60: From the above picture analysis, based on our custom arima we can say that prices of an apple high stock prices will increase for the next 100 days.

ARIMA forecasting for apple close stock prices.

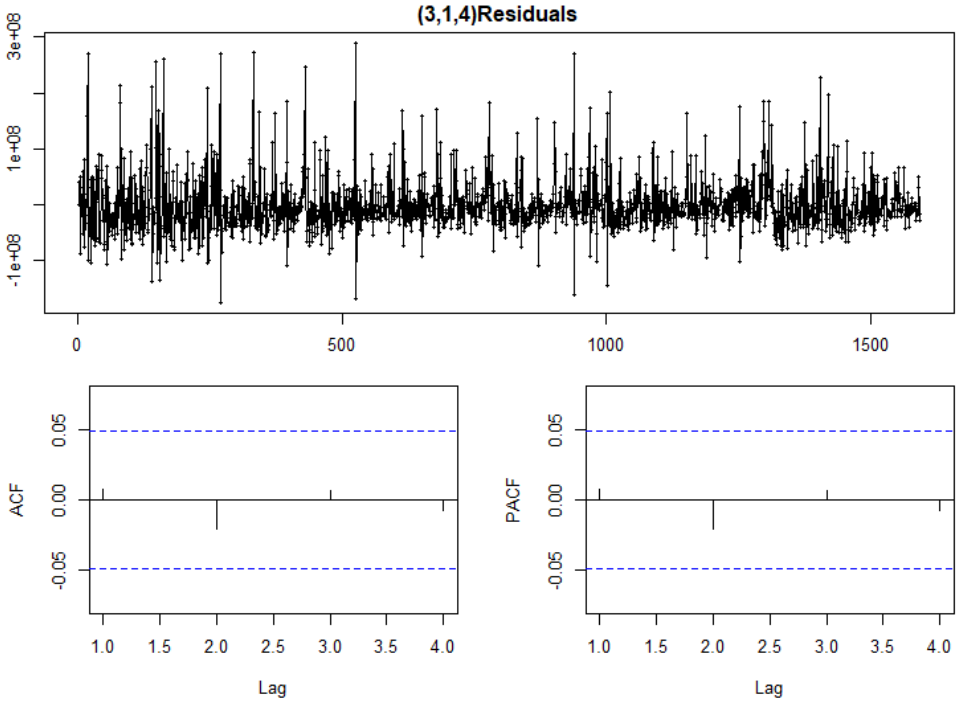


Figure 61: It represents auto arima model residual (3,1,4) with AIC 60989.93 and, BIC 61006.05.

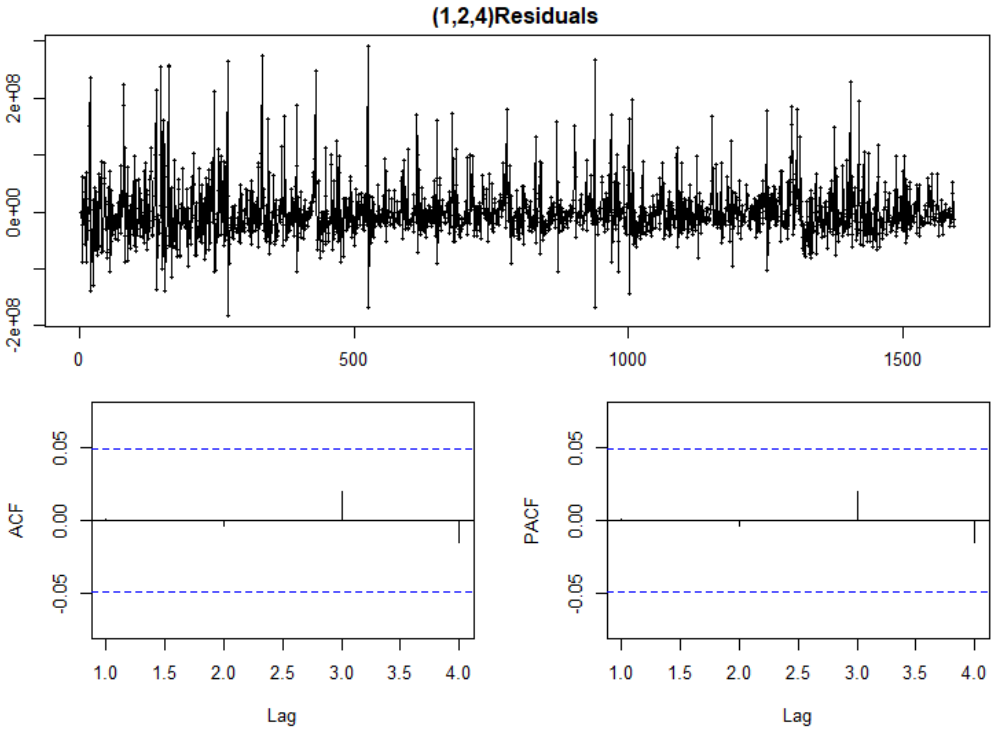


Figure 62: It represents a custom arima model for apple closing stock prices.

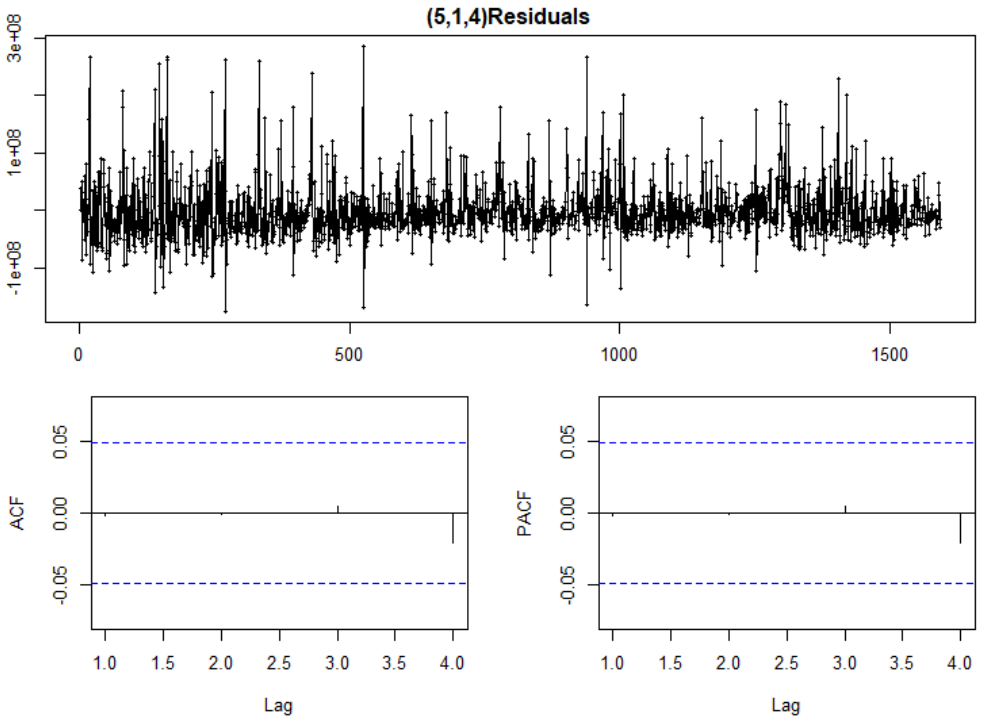


Figure 63: Model residual (5,1,4), guessing on different values on auto arima model for apple closing stock prices.

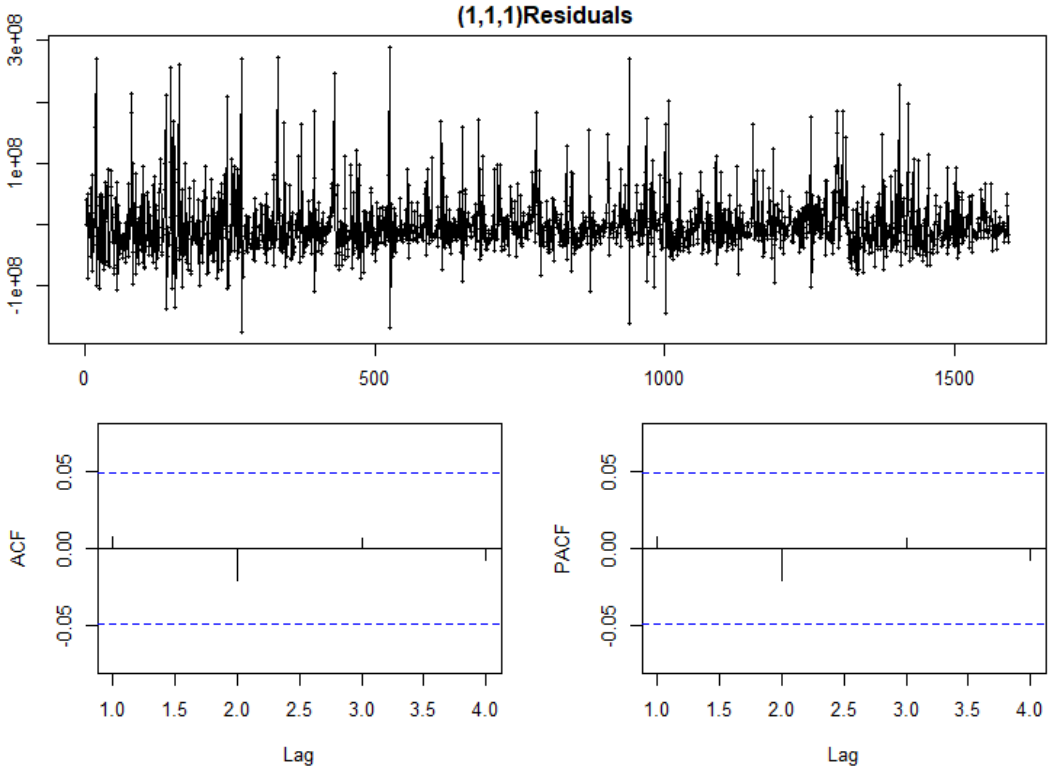


Figure 64: Standard de facto model residual (1,1,1) for apple closing stock prices.

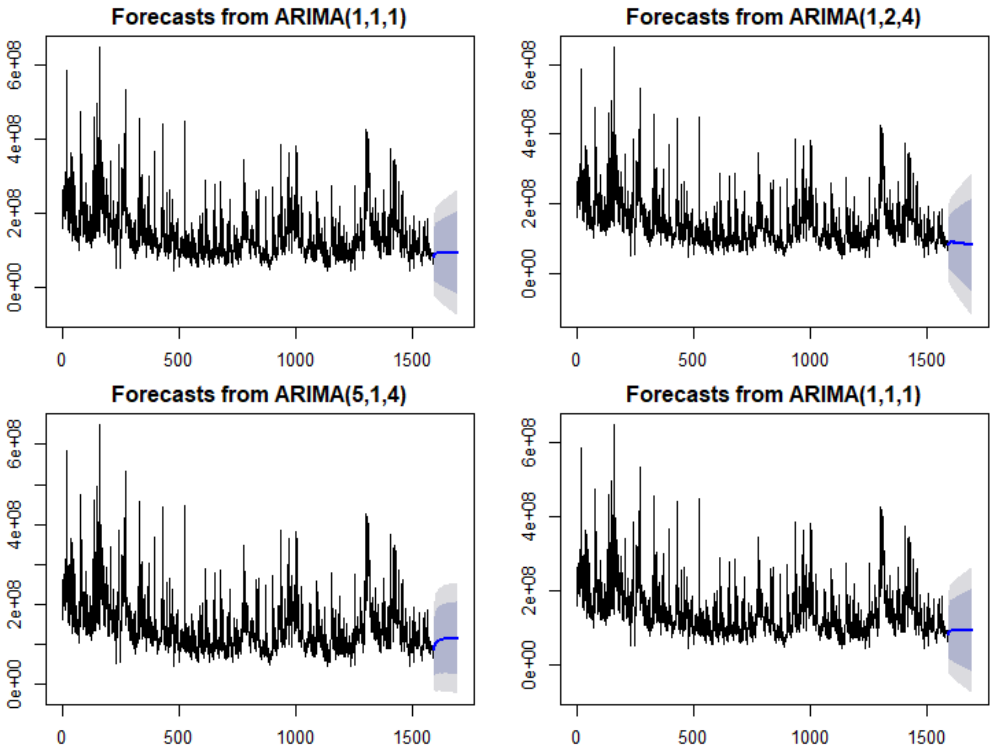


Figure 65: It represents all arima models together in one plot with predictions. Based on our above prediction we can say that prices of apple closing stock for the next 100 days are meant to remain the same with a slight change.

AI recurrent neural network/ LSTM

Predicted model of Apple closing stock values in the line graph.

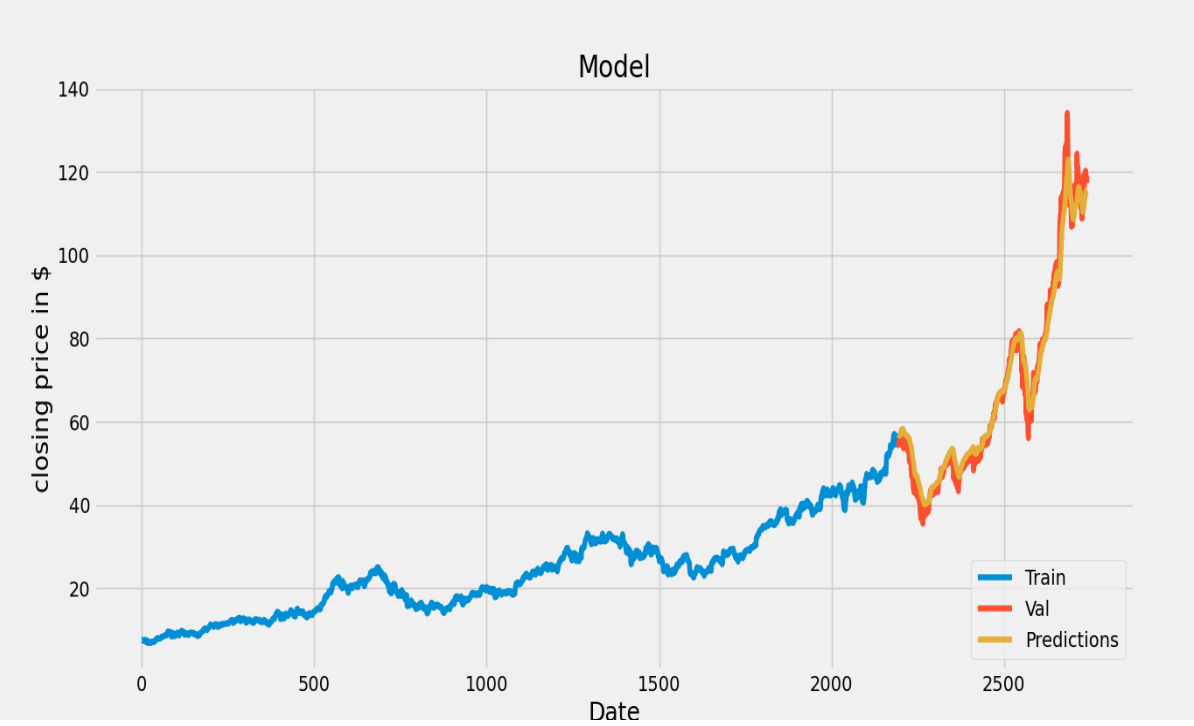


Figure 66: represent prediction of apple closing stock price in a line graph.

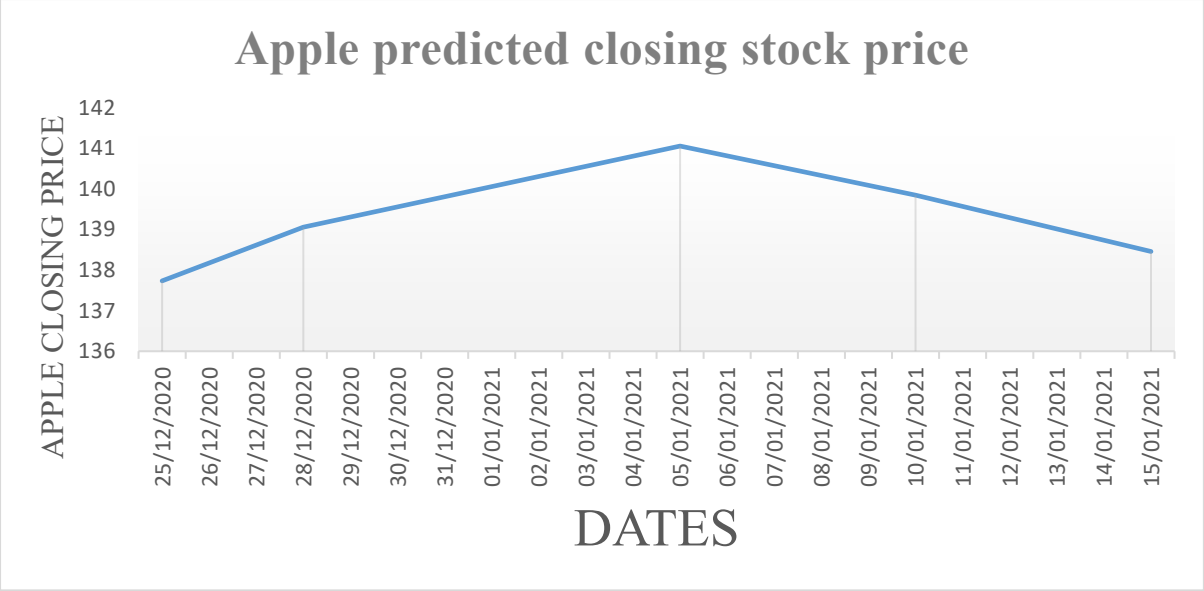


Figure 67: Apple predicted stocks value.

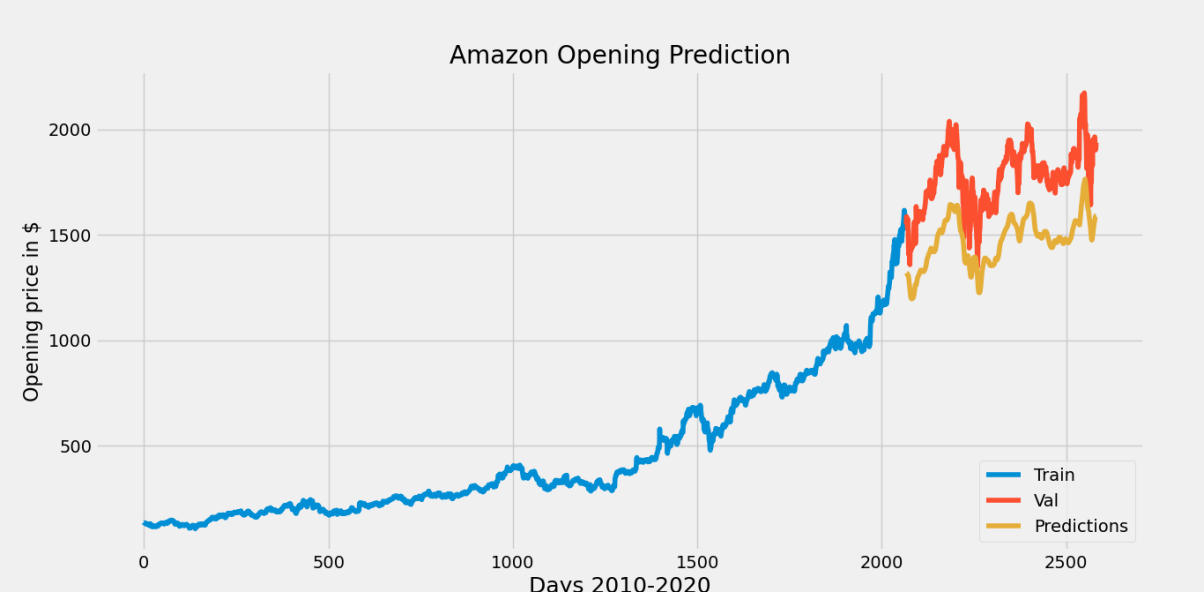


Figure 68: It represents the predicted model of amazon opening stock price.

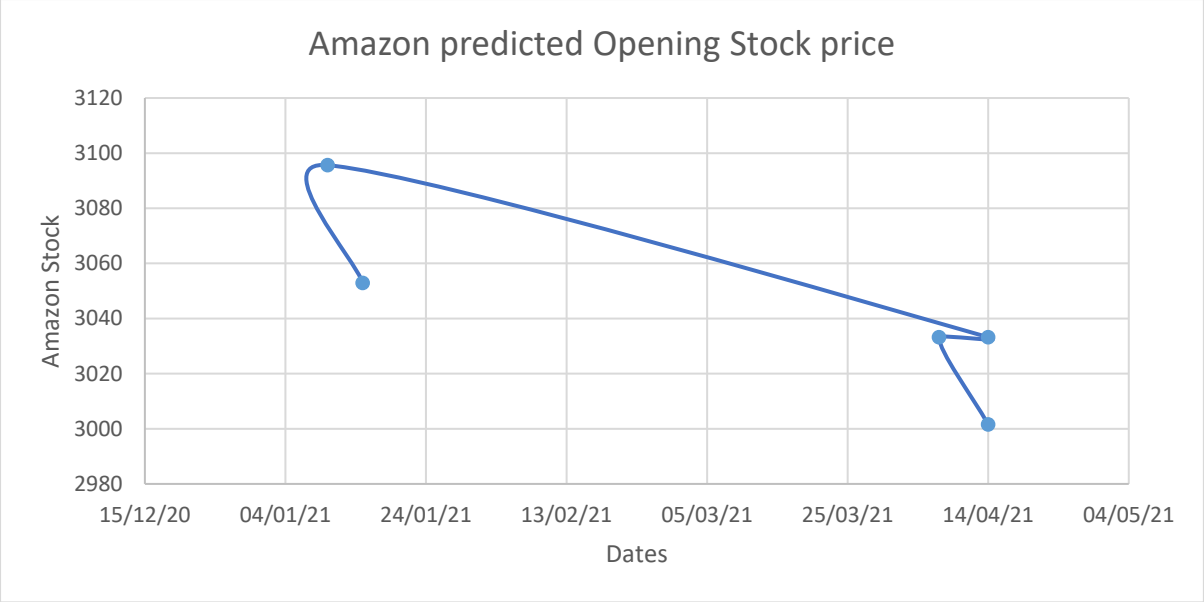


Figure 69: It represents predicted amazon open stock prices.

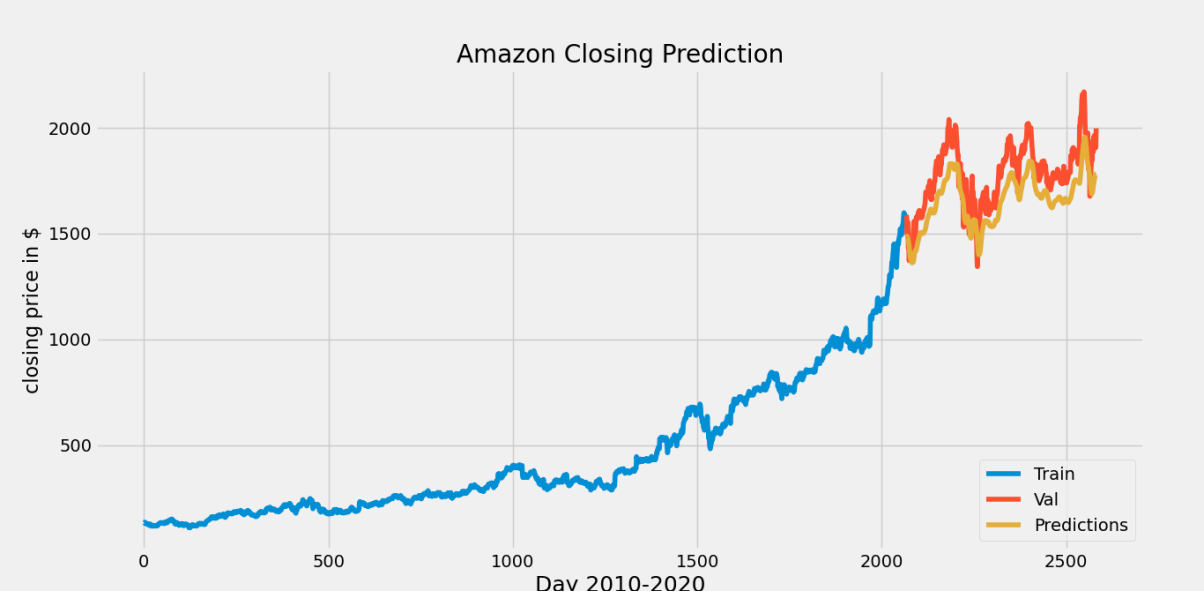


Figure 70: It represents the prediction of amazon closing stock price.

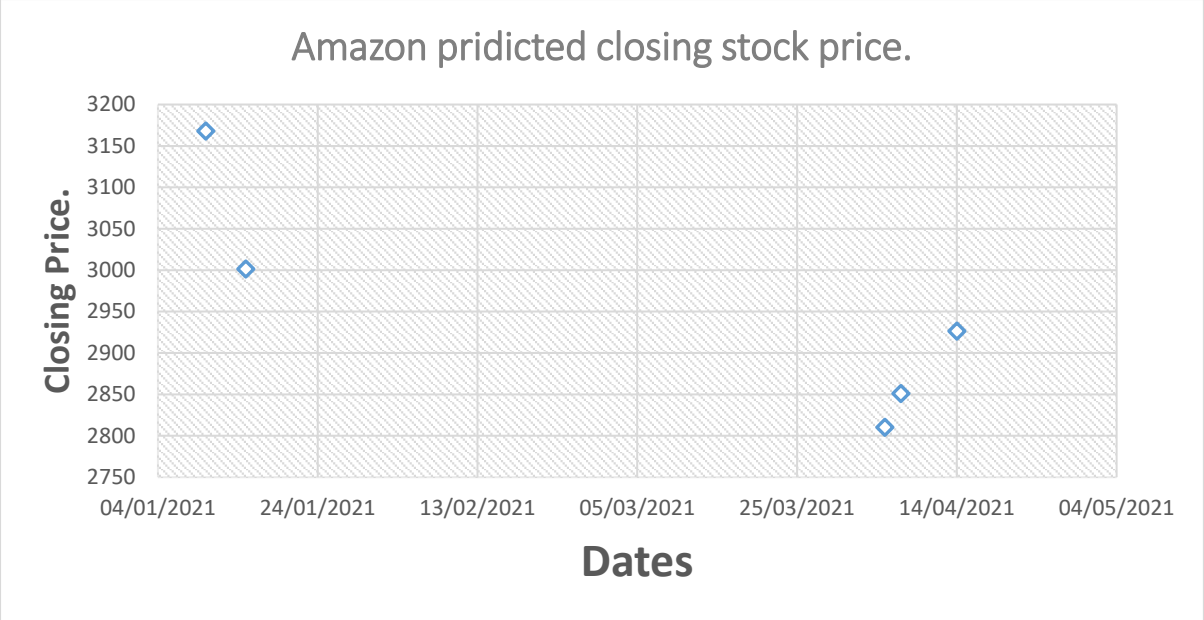


Figure 71: Predicted closing stock price of amazon of some selected dates.

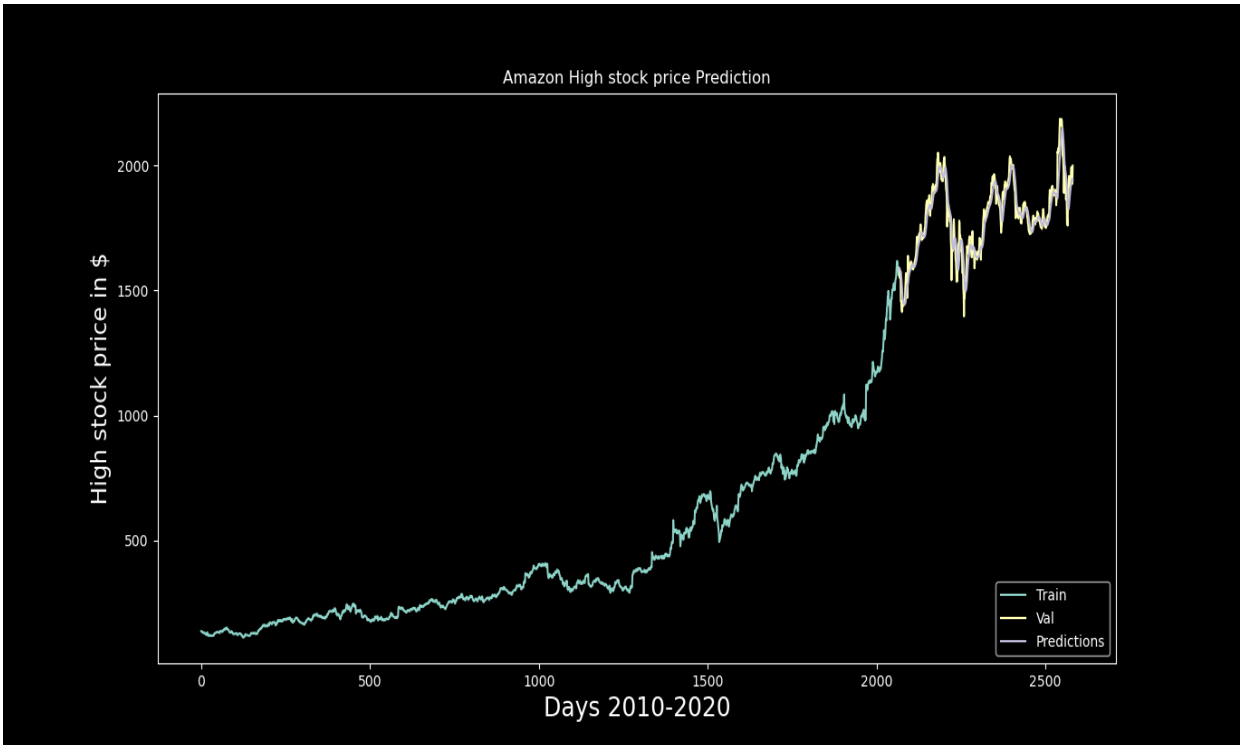


Figure 72: It represents the graphical representation of amazon's high stock price prediction.

Figure 73: Amazon predicted high stock prices, where we can see that on 05/04/21 price of amazon's high stock will be 2945.94.

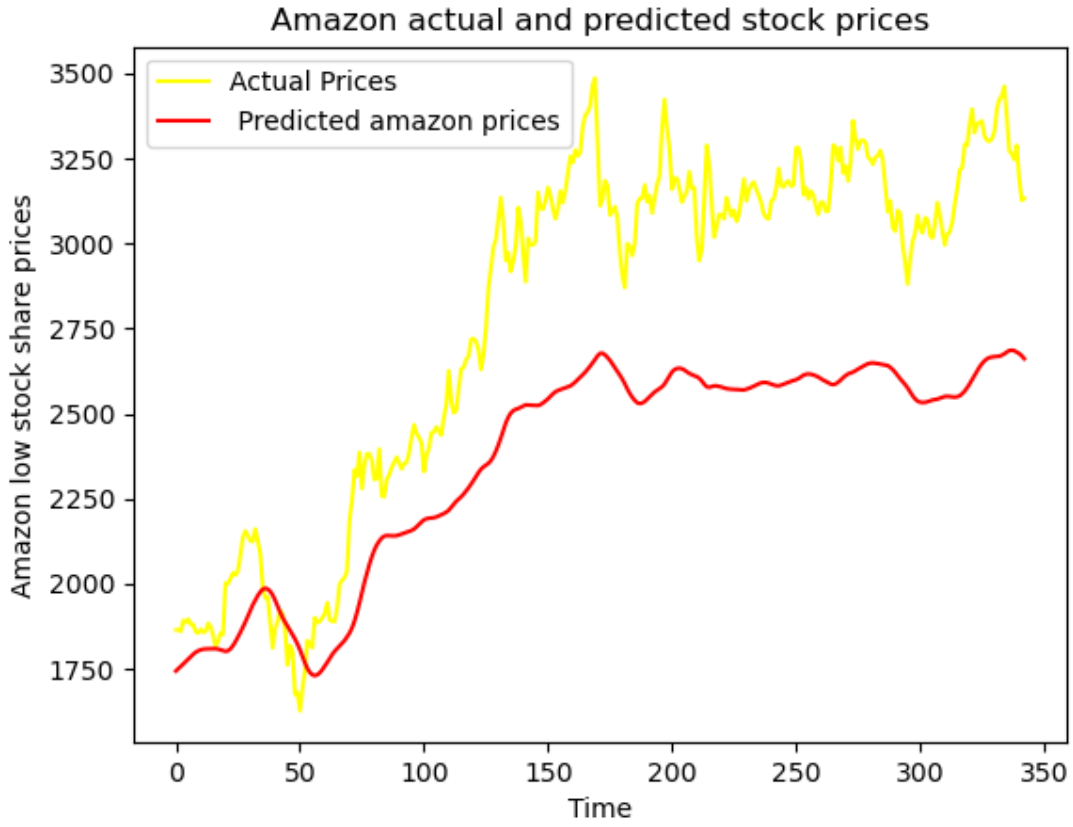


Figure 74: It represents amazon's actual and predicted stock Low price of amazon in the year 2020.

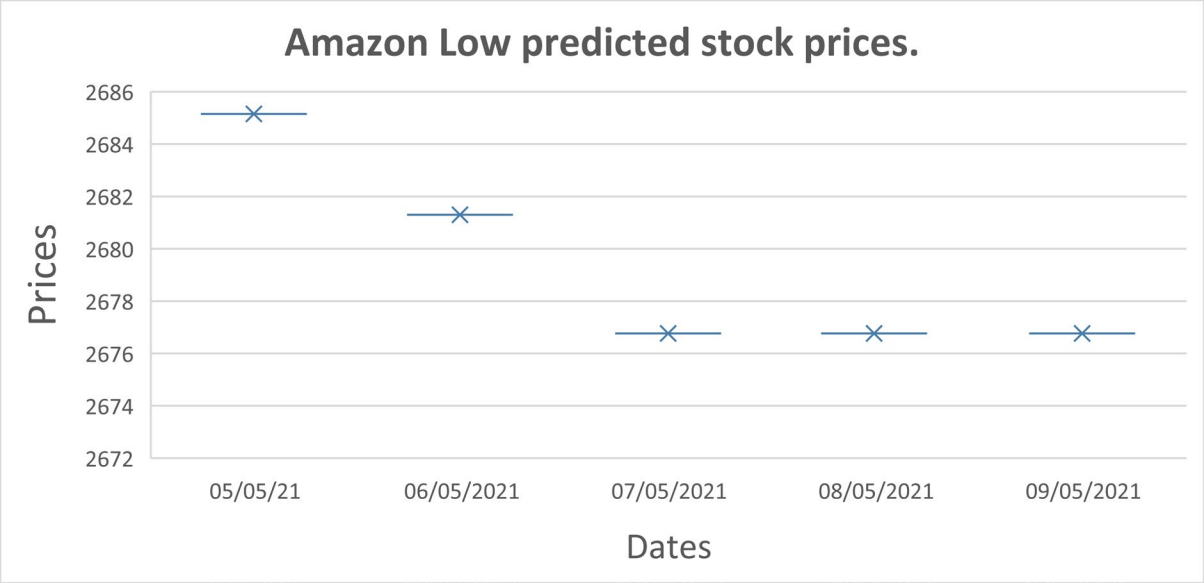


Figure 75: Amazon's low predicted stock prices, with different dates as shown in the picture.

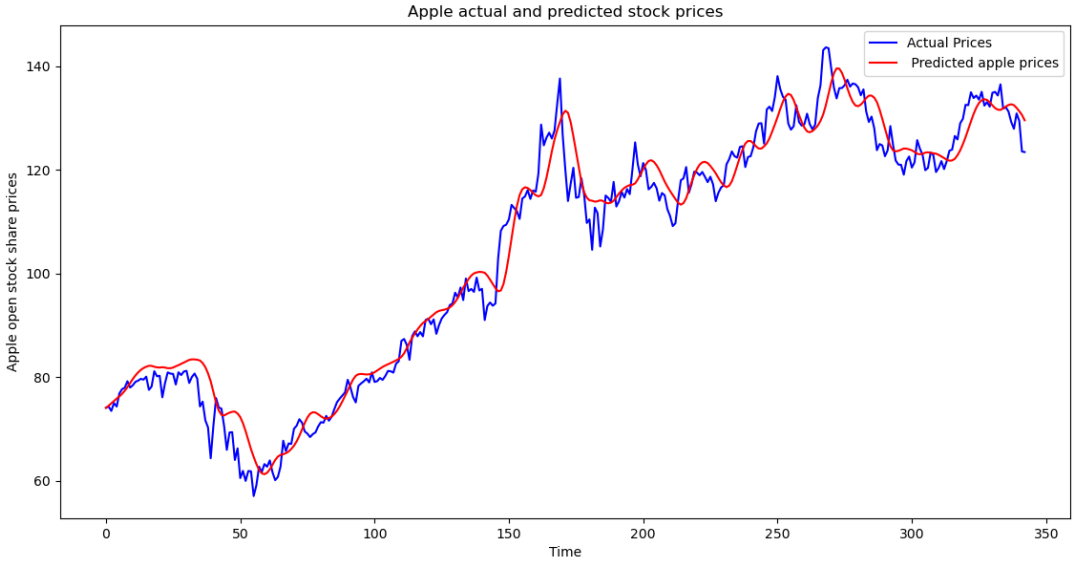


Figure 76: Apple predicted and actual stock prices of the year 2020.

6.0 Conclusions

Therefore, from the above analysis, we can see our model predicts the closing, opening, high and low stock prices of Apple Inc and Amazon, which is slightly different from real stock prices and actual prices. LSTM proves us a prediction of at least one or more than one day's stock price (e.g., prediction of amazon opening stock price on 2021-04-05 is 3001.5588). Whereas ARIMA forecasting provides us increase or decrease of stock prices of next 100 days. Is the stock price rising or decreasing? Well based on the above analysis we can clearly say that there is an increase in Amazon stock price for the next 100 days based on ARIMA forecasting and stock prices of apple closing are meant to remain in the same series with a slight change of trends based on ARIMA forecasting for next 100 days.

We will gain some valuable insights, Prediction gets you into the habit of looking at past and real-time data to predict future demand. We will learn from past mistakes we do not need to start from scratch if the prediction was nowhere close to where we expected it to be, and it will hint from again starting point. Predictions can never be 100% accurate even that the predictions model we have is slightly different from the original value, but it will give us an accurate 90 percent chance of where the closing prices will go to.

7.0 Testing.

7.1 Normality Test.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
High	2742	100.0%	0	0.0%	2742	100.0%
Low	2742	100.0%	0	0.0%	2742	100.0%

Figure 77: Case processing summary for normality test.

Normality test has been caried out in IBM SPSS, it is used to determine if our datasets are well modelled by normal distribution and compute how likely it is for random variable underlying datasets to be normal distributed. It is used to determine whether our sample data has been drawn from a normal distributed population. Normality test has been performed for our apple dataset for high and low stock prices. Whereas from our case summary processing we can see that we have total 2742 values which indicate that we did not miss any value from our dataset. To perform normality test first of all descriptive statistics is performed for our apple high and low stock prices (mean, median, mode, and range) as seen in below figure but the one that I am interested is skewness for high column is 1.820, which indicates that it is not normal distributed.

Descriptives

			Statistic	Std. Error
High	Mean		33.20358927	.443942464
	95% Confidence Interval for Mean	Lower Bound	32.33309365	
		Upper Bound	34.07408490	
	5% Trimmed Mean		30.53587566	
	Median		27.03624900	
	Variance		540.407	
	Std. Deviation		23.24665195	
	Minimum		7.000000	
	Maximum		137.979996	
	Range		130.979996	
	Interquartile Range		25.446696	
	Skewness		1.820	.047
	Kurtosis		3.761	.093
Low	Mean		32.52427062	.431664087
	95% Confidence Interval for Mean	Lower Bound	31.67785079	
		Upper Bound	33.37069044	
	5% Trimmed Mean		29.95946877	
	Median		26.56125100	
	Variance		510.928	
	Std. Deviation		22.60370568	
	Minimum		6.794643	
	Maximum		130.529999	
	Range		123.735356	
	Interquartile Range		25.153393	
	Skewness		1.785	.047
	Kurtosis		3.595	.093

Figure 78: Descriptive statistics for our normality test.

Below figure 71 tells us that both Kolmogorov-Smirnov and Shapiro-Wilk. In our high stock prices, our statistics value is 0.154 and p value is 0.000 is lower than our alpha value which tells us that our sample test is not normal distributed (High). Also, in our low stock price statistics value is 0.151 and p value is 0.000 which is also lower than our alpha value therefore this sample test is not sample distributed.

Hence, our p value for both high and low stock is less than our chosen alpha value which is 0.05 and we reject our null hypothesis, and it is an evidence that data tested is not normal distributed.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
High	.154	2742	.000	.825	2742	.000
Low	.151	2742	.000	.829	2742	.000

a. Lilliefors Significance Correction

Figure 79: Test of Normality.

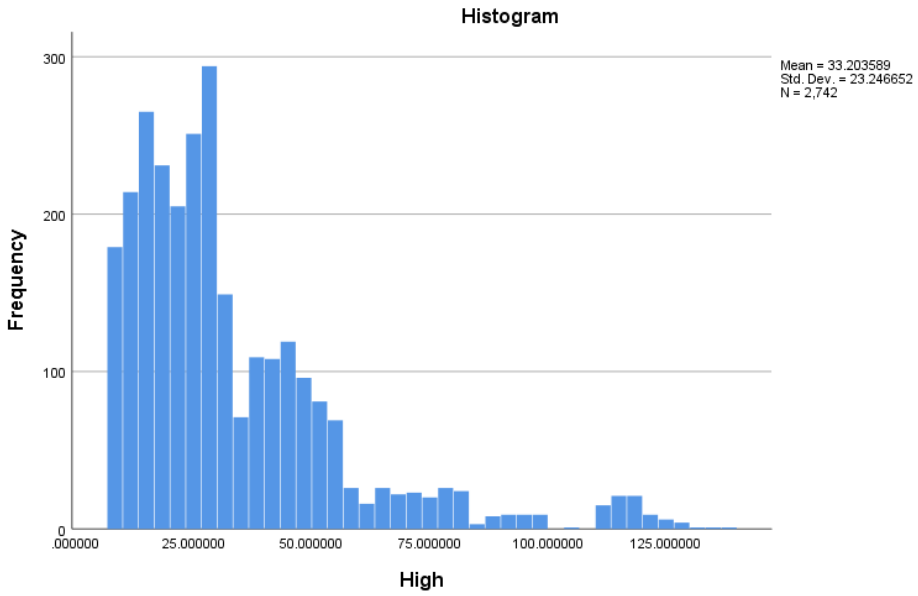


Figure 80: Histogram for high stocks prices.

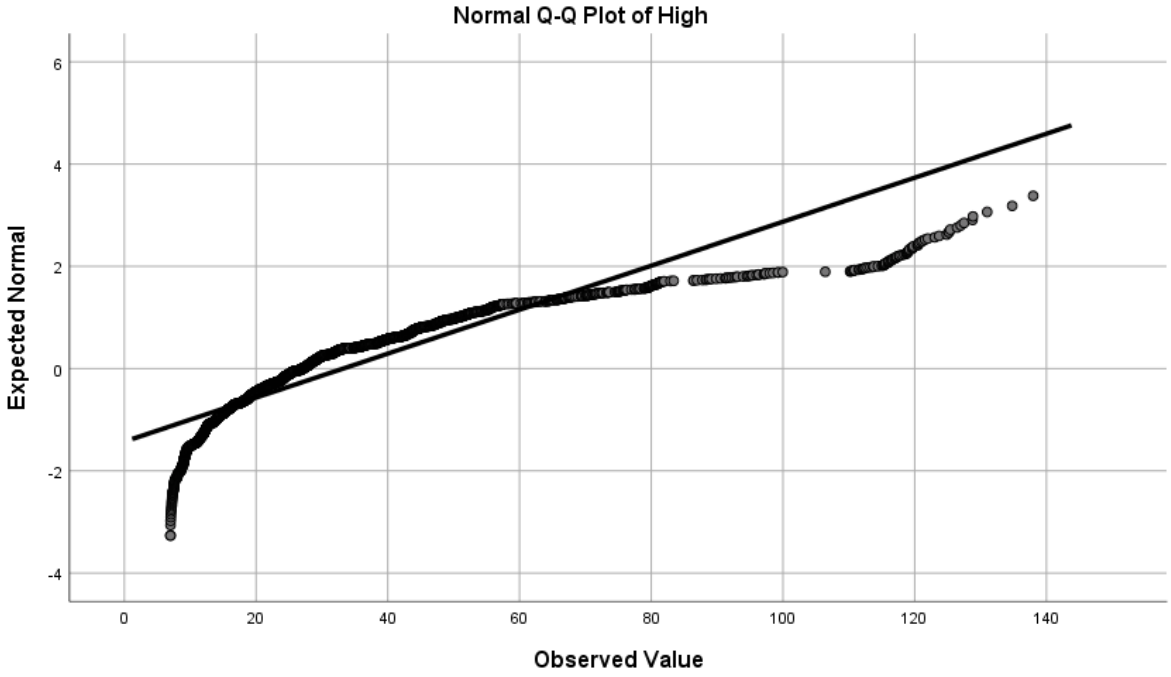


Figure 81: Normal Q-Q plot of High stock prices.

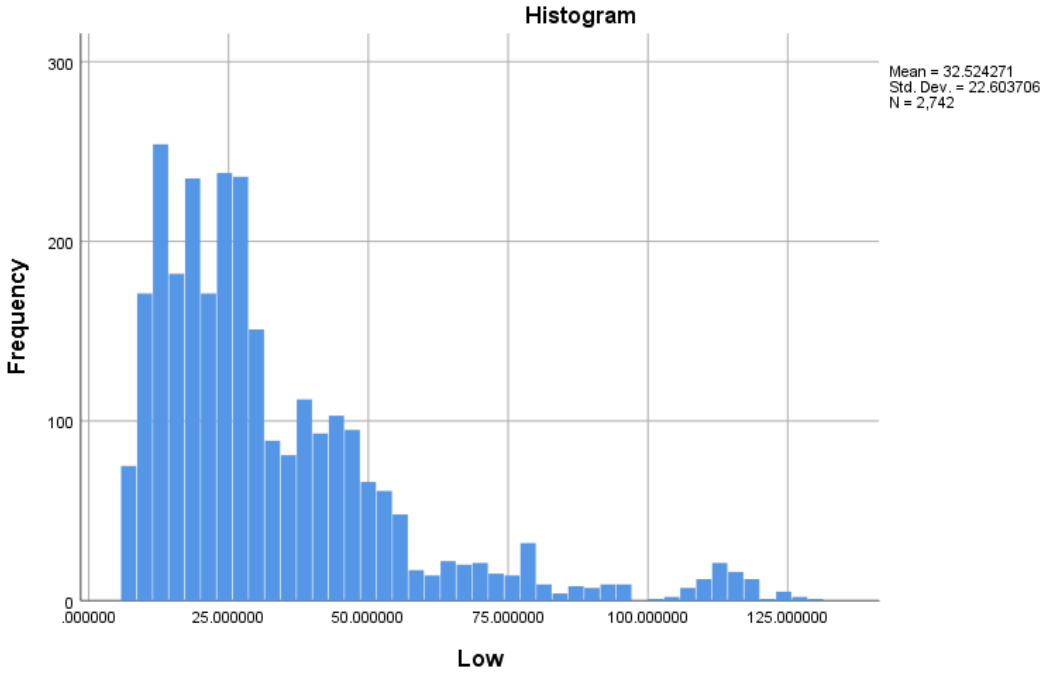


Figure 82: Histogram for low stock prices.

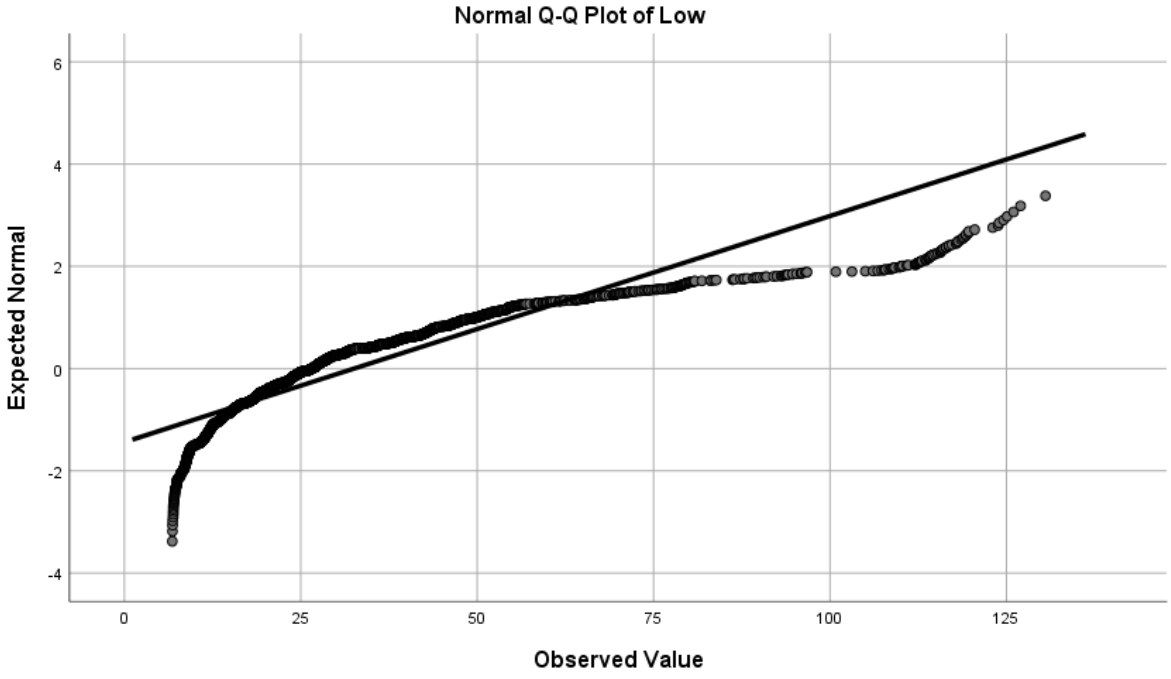


Figure 83: Normal Q-Q pot for low stocks.

7.2 Mann-Whitney test between apple closing and opening price dataset.

Mann-Whitney test is performed between apple opening and closing stock prices and compare them, it is a nonparametric test for null hypothesis, random selected values and for two population.

Test Statistics[a,b]

		Number of Runs	Z	Asymp. Sig. (1-tailed)
Close	Minimum Possible	12[c]	-21.491	.000
	Maximum Possible	12[c]	-21.491	.000
Open	Exact Number of Runs	14[d]	-21.313	.000

a. Wald-Wolfowitz Test

b. Grouping Variable: Group

c. There are 1 inter-group ties involving 2 cases.

d. No inter-group ties encountered.

Figure 84: It is test statistics for our closing and opening, whereas our p value is less than 0.05 so we will reject our null hypothesis.

Test Statistics[a]

	Close	Open
Mann-Whitney U	56.500	73.000
Wilcoxon W	31934.500	31951.000
Z	-19.388	-19.378
Asymp. Sig. (2-tailed)	.000	.000

a. Grouping Variable: Group

Figure 85: It provide us our Mann-Whitney for apple closing which is 56.500 and for opening 73.00 of total number of values 2742. Whereas our p value is 0.000 so which is less than our alpha value 0.05, so we will reject our null hypothesis.

7.3 Independent Samples Kruskal-Wallis test.

Kruskal-Wallis test is performed for amazon opening and closing stock price, it is non-parametric method for testing whether samples are originating from same sample distribution. From our hypothesis test summary, our significance or p value is 0.000 which is lower than our selected alpha value and null hypothesis will be rejected.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Open is the same across categories of Group.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.
2	The distribution of Close is the same across categories of Group.	Independent-Samples Kruskal-Wallis Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .050.

Figure 86: Hypothesis test summary of independent sample Kruskal-Wallis test.

Kruskal-Wallis test for amazon open stock prices (Figure 77). Total number of values is 2582, test statistics is 2503.082 with degree of freedom 10. Our p value is 0.000 which is lower than our selected alpha value 0.05, it means we will reject our null hypothesis.

Independent-Samples Kruskal-Wallis Test Summary

Total N	2582
Test Statistic	2503.082[a]
Degree Of Freedom	10
Asymptotic Sig.(2-sided test)	.000

a. The test statistic is adjusted for ties.

Figure 87: Kruskal-Wallis test summary.

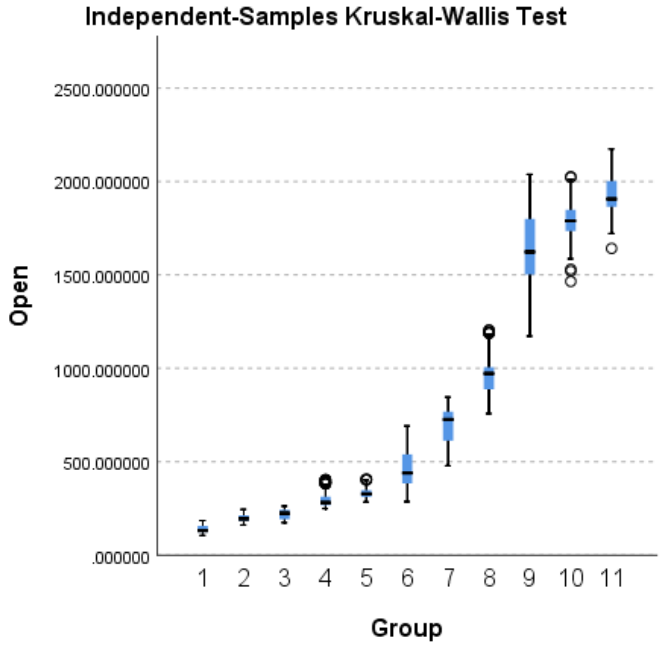


Figure 88: Boxplot for representation of Kruskal-Wallis test (Group represents years 2010-2020).

Kruskal Wallis test for amazon closing stock price from 2010 to 2020. From figure 79 we have total number of values 2582, test statistic 2503.958 with degree of freedom 10. Our significance or p value is also 0.000 which is lower than selected alpha value and we will reject our null hypothesis.

Independent-Samples Kruskal-Wallis Test Summary

Total N	2582
Test Statistic	2503.958[a]
Degree Of Freedom	10
Asymptotic Sig.(2-sided test)	.000

a. The test statistic is adjusted for ties.

Figure 89: Kruskal Wallis test summary.

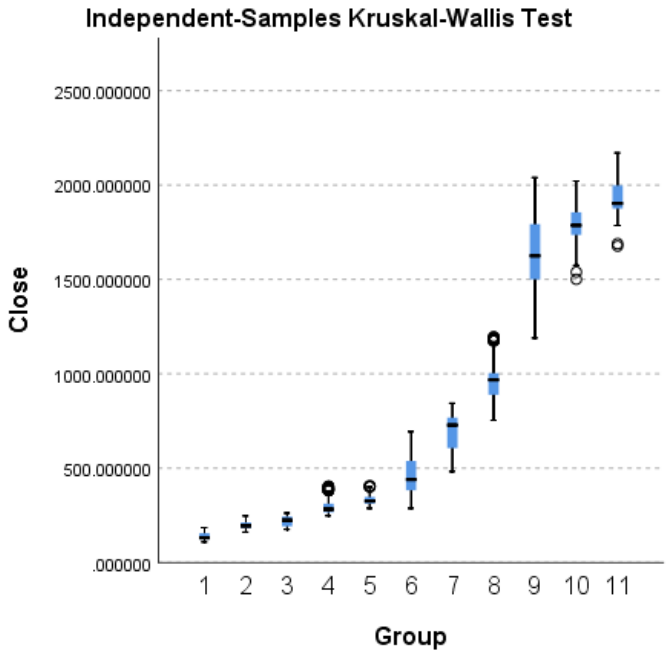


Figure 90: Boxplot for representation of amazon closing Kruskal-Wallis test (Group represents years 2010-2020).

7.4 Root Mean Squared Error (RMSE).

Root mean square error value is performed, RMSE is nice measure to check how accurate is model predictor and it is standard deviation of residuals as well as a lower value of our RMSE indicate a better fit model accuracy. As our RMSE value for predicting closing stock price is 2.965824339898993, signify that our model is accurate. Whereas RMSE value for amazon closing stock price is 0.5208368889213619, it is very lower value and represent a better fit accuracy of our model.

7.5 MAPE Accuracy for ARIMA forecasting.

To check, accuracy of our amazon closing stock prices, we find MAPE value of all our four different ARIMA forecast model. Whereas accuracy can be finding out by subtracting our MAPE value from 100 (where one hundred is our next future days). Accuracy of ARIMA forecasting (fcast1) is 75.15637%, accuracy of our custom ARIMA forecasting (facst2) is 74.60634 %, accuracy of guessing on random values of auto ARIMA (fcast3) 75.19114%, and accuracy of our last standard de facto ARIMA forecasting is (fcast4) 75.21762%.

Amazon closing ARIMA MAPE accuracy, subtract from next 100 days.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Accuracy (fcast1)	5745.419	1672975	1079181	-9.13778	24.84363	0.927691 6	0.001659516
Accuracy (fcast2)	-13228.07	1692137	1109422	-8.829701	25.39366	0.953688 3	0.0290624
Accuracy (fcast3)	4118.381	1672449	1078361	-9.082479	24.80886	0.926987	-0.000153155 5
Accuracy (facst4)	2969.758	1682001	1079244	-9.08617	24.78238	0.927746 2	0.01789251

Figure 91: ARIMA MAPE accuracy for amazon closing.

MAPE accuracy for amazon high prices, whereas our auto ARIMA (fcast1) has 98.726568% accuracy, custom ARIMA (fcast2) has accuracy 98.730256%, guessing on random values auto ARIMA (fcast3) has 75.19114%, and accuracy of standard de facto ARIMA (fcast4) is 98.729038%.

Amazon high prices, ARIMA forecasting MAPE accuracy, subtract from next 100 days.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Accuracy (fcast1)	-0.001610998	30.55905	19.13306	-0.06482996	1.273432	0.9949444	0.000150058
Accuracy (fcast2)	0.6419429	30.55935	19.14057	0.01897482	1.269744	0.9953351	-0.000988039 4
Accuracy (fcast3)	4118.381	1672449	1078361	-9.082479	24.80886	0.926987	-0.000153155 5
Accuracy (facst4)	1.691795	30.82096	19.2041	0.117981	1.270962	0.9986383	-0.00740774

Figure 92: ARIMA MAPE accuracy for amazon high.

Apple closing ARIMA MAPE accuracy, subtract from next 100 days.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Accuracy (fcast1)	-673538.5	4970489 6	3411468 6	-8.42357 1	23.7339 2	0.906617 1	0.00767708 8
Accuracy (fcast2)	244521. 2	4973049 1	3390736 4	-7.34667 5	23.3287 5	0.901107 4	0.00141162 8
Accuracy (fcast3)	-149089 8	4945123 1	3434423 4	-9.31621 5	24.1691 2	0.912717 5	-0.00225507 4
Accuracy (facst4)	-673538. 5	4970489 6	3411468 6	-8.42357 1	23.7339 2	0.906617 1	0.00767708 8

Figure 93: ARIMA MAPE accuracy for apple closing.

Similar to apple dataset, MAPE accuracy is performed to check how accurate is our ARIMA forecasting, from above picture we have our MAPE values and accuracy can be find out by subtracting MAPE value by 100. Accuracy for our fcast1 of auto ARIMA is 76.26608%, custom ARIMA have accuracy 76.67125, fcast3 (guessing on different values on auto ARIMA) has 75.83088%, and standard de facto ARIMA has 76.26608.

Apple high prices, ARIMA forecasting MAPE accuracy, subtract from next 100 days.

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Accuracy (fcast1)	0.00053415 06	1.4143 34	0.82788 69	0.0032088 03	1.4819 47	1.23923 1	-0.1313386
Accuracy (fcast2)	0.02414644	1.1334 08	0.65814 53	0.0188893 3	1.1864 79	0.98515 17	-0.0026073 08
Accuracy (fcast3)	0.05655581	1.1202 47	0.65955 87	0.0732692 1	1.1948 7	0.98726 74	-0.0031431 51
Accuracy (facst4)	0.05988287	1.1393 54	0.66205 11	0.0770792 5	1.1893 72	0.99099 81	-0.0109197

Figure 94: ARIMA MAPE accuracy for apple closing.

Auto ARIMA Accuracy is 98.52%.

Custom ARIMA accuracy is 98.81%.

Guessing on different values on auto ARIMA accuracy 98.81%.

Standard de facto ARIMA accuracy 98.81%.

8.0 Further Development or Research

In future, I am planning to use artificial neural network, at the moment I used recurrent neural network to predict stocks. Reason for using different AI network is to find out is there any difference in accuracy of model. Then I want to apply random forest on different machine for predicting stock prices of our datasets.

9.0References

En.wikipedia.org. 2020. Stock Market Prediction. [online] Available at: https://en.wikipedia.org/wiki/Stock_market_prediction#Fundamental_analysis

[Accessed 20 December 2020].

En.wikipedia.org. 2020. Cluster Analysis. [online] Available at: https://en.wikipedia.org/wiki/Cluster_analysis#:~:text=Cluster%20analysis%20or%20clustering%20is,in%20other%20groups%20(clusters).

[Accessed 20 December 2020].

Small Business - Chron.com. 2020. How The Stock Market Was Started & By Whom. [online] Available at: https://smallbusiness.chron.com/stock-market-started-whom-14745.html

[Accessed 20 December 2020].

Built-In. 2020. How AI Trading Technology Is Making Stock Market Investors Smarter. [online] Available at: https://builtin.com/artificial-intelligence/ai-trading-stock-market-tech

[Accessed 20 December 2020].

DBD, U., 2020. KDD Process/Overview. [online] Www2.cs.uregina.ca. Available at: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html#:~:text=The%20term%20Knowledge%20Discovery%20in,of%20particular%20data%20mining%20methods.&text=The%20unifying%20goal%20of%20the,the%20context%20of%20large%20databases.

[Accessed 20 December 2020].

Saedsayad.com. 2020. K-Means. [online] Available at: https://www.saedsayad.com/clustering_kmeans.htm#:~:text=The%20objective%20of%20K%2DMeans,to%20the%20Euclidean%20distance%20function.

[Accessed 20 December 2020].

Lynda.com - from LinkedIn. 2020. Why Use Rstudio?. [online] Available at: https://www.lynda.com/RStudio-tutorials/Why-use-RStudio/452087/490021-4.html#:~:text=RStudio%20is%20a%20powerful%20IDE,and%20version%20control%20management%20tools.

[Accessed 20 December 2020].

Researchconnections.org. 2020. Descriptive Statistics. [online] Available at: https://www.researchconnections.org/childcare/datamethods/descriptivestats.jsp#:~:text=Descriptive%20statistics%20can%20be%20useful,Graphical%2FPictorial%20Methods

[Accessed 20 December 2020].

En.wikipedia.org. 2020. Standard Deviation. [online] Available at: https://en.wikipedia.org/wiki/Standard_deviation

[Accessed 20 December 2020].

En.wikipedia.org. 2020. Hierarchical Clustering. [online] Available at: https://en.wikipedia.org/wiki/Hierarchical_clustering#:~:text=In%20data%20mining%20and%20statistics,build%20a%20hierarchy%20of%20clusters.

[Accessed 20 December 2020].

Hodgson, E., 2020. Is Hierarchical Clustering Worth Pursuing?. [online] Dotactiv.com. Available at: https://www.dotactiv.com/blog/is-hierarchical-clustering-worth-pursuing

[Accessed 20 December 2020].

Available at: https://blog.johngalt.com/3-advantages-disadvantages-of-forecasting

[Accessed 20 December 2020].

Wicklin, R. and Wicklin, R., 2021. Should you use principal component regression?. [online] The DO Loop. Available at: https://blogs.sas.com/content/iml/2017/10/25/principal-component-regression-drawbacks.html

[Accessed 1 May 2021].

Youtube.com. 2021. Before you continue to YouTube. [online] Available at: https://www.youtube.com/watch?v=NLrb41ls4qo

[Accessed 1 May 2021].

Medium. 2021. An Extensive Step by Step Guide to Exploratory Data Analysis. [online] Available at: https://towardsdatascience.com/an-extensive-guide-to-exploratory-data-analysis-ddd99a03199e

[Accessed 14 May 2021].

10.0 Appendices

This section should contain information that is supplementary to the main body of the report.

10.1. Project Plan

	1 Initiation	**23 days?**	**Mon 28/09/20**	**Wed 28/10/20**
	1.1 Research on project idea	**9 days?**	**Wed 30/09/20**	**Mon 12/10/20**
	1.1.1 Finding a suitable idea fimillar to specillization	**4 days?**	**Wed 30/09/20**	**Sat 03/10/20**
	1.1.1.1 Research on internet	4 days	Wed 30/09/20	Sat 03/10/20
	1.1.1.2 Discussing project with last year students	4 days	Wed 30/09/20	Sat 03/10/20
	1.1.1.3 Sharing idea with lecturer and getting feedback	4 days	Wed 30/09/20	Sat 03/10/20
	1.1.1.4 Discussing with friends	4 days	Wed 30/09/20	Sat 03/10/20
	1.1.1.5 Explore emerging platform	**6 days?**	**Sun 11/10/20**	**Fri 16/10/20**
	1.1.1.5.1 Research on google analytics platform	**6 days?**	**Fri 30/10/20**	**Fri 06/11/20**
	1.1.1.5.1.1 Traffic Channels	6 days?	Fri 30/10/20	Fri 06/11/20
	1.1.1.5.1.2 Blog post engagement	6 days?	Fri 30/10/20	Fri 06/11/20
	1.1.1.5.1.3 performance by platform	6 days?	Fri 30/10/20	Fri 06/11/20
	1.1.1.5.1.4 Google big data analytics	6 days?	Fri 30/10/20	Fri 06/11/20
	1.1.1.6 Research on github for new ideas	6 days?	Fri 30/10/20	Fri 06/11/20

	1.2 Project risks and issues	**4 days?**	**Wed 07/10/20**	**Mon 12/10/20**
	1.2.1 Data quality	4 days?	Wed 07/10/20	Mon 12/10/20
	1.2.2 Server or sites down	4 days?	Wed 07/10/20	Mon 12/10/20
	1.2.3 Cleaning big data	4 days?	Wed 07/10/20	Mon 12/10/20
	1.3 Doing fesability study on project	**6 days?**	**Tue 13/10/20**	**Tue 20/10/20**
	1.3.1 Market Research	6 days?	Tue 13/10/20	Tue 20/10/20
	1.3.2 Financial Research	6 days?	Tue 13/10/20	Tue 20/10/20
	1.3.3 What is stock predictions			
	1.3.4 Capability of project	6 days?	Tue 13/10/20	Tue 20/10/20
	1.3.4 Studying online tutorial	**9 days?**	**Thu 22/10/20**	**Tue 03/11/20**
	1.3.4.1 How to use Tableau for big data analysis	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.2 Advanced R programming tutorial from linkedIn learning	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.3 IBM SPSS tutorial	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.4 Advance statistical tutorial	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.5 Data Mining	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.6 Advance visualization tutorial	9 days?	Thu 22/10/20	Tue 03/11/20

1.3.4.6 Machine learning Tutorials	**5 days?**	**Thu 05/11/20**	**Wed 11/11/20**
1.3.4.6.1 Machine Learning LSTM	5 days?	Thu 05/11/20	Wed 11/11/20
1.3.4.6.3 Clustering Algorithms	5 days?	Thu 05/11/20	Wed 11/11/20
1.3.4.7 MapReduce from linkedIn learning	5 days?	Thu 05/11/20	Wed 11/11/20
1.4 Finding Stakeholder of project	**2 days?**	**Fri 13/11/20**	**Mon 16/11/20**
1.4.1 Appointed the project	**2 days?**	**Fri 13/11/20**	**Mon 16/11/20**
1.4.1.1 Umer Iqbal	2 days?	Fri 13/11/20	Mon 16/11/20
1.4.2 Appointed supervisor by college	**2 days?**	**Fri 13/11/20**	**Mon 16/11/20**
1.4.2.1 Keith Maycock	2 days?	Fri 13/11/20	Mon 16/11/20
1.4.3 Setting meeting with supervisor for feedback on project	**2 days?**	**Wed 18/11/20**	**Thu 19/11/20**
1.4.3.1 Meeting 1	2 days?	Wed 18/11/20	Thu 19/11/20
1.4.3.2 Meeting 2	2 days?	Wed 18/11/20	Thu 19/11/20
1.4.3.3 Meeting 3	2 days?	Wed 18/11/20	Thu 19/11/20
1.5 Project pitch video	**4 days?**	**Mon 12/10/20**	**Thu 15/10/20**
1.5.1 Making notes for video	4 days?	Mon 12/10/20	Thu 15/10/20
1.5.2 Recording video on Laptop	4 days?	Mon 12/10/20	Thu 15/10/20
1.5.3 Editing video on adobe workshop	4 days?	Mon 12/10/20	Thu 15/10/20
1.5.4 Uploading project pitch video on Moodle	4 days?	Mon 12/10/20	Thu 15/10/20
1.5.5 Email supervisor for feedback	4 days?	Mon 12/10/20	Thu 15/10/20
1.6 Review project with lecturer and supervisor	4 days?	Mon 12/10/20	Thu 15/10/20
2 Planning	**18 days?**	**Thu 19/11/20**	**Mon 14/12/20**
2.1 Discussing project requirement with supervisor	**3 days?**	**Thu 19/11/20**	**Mon 23/11/20**
2.1.1 Setting up meeting	**3 days?**	**Thu 19/11/20**	**Mon 23/11/20**
2.1.1.1 Meeting first	3 days?	Thu 19/11/20	Mon 23/11/20
2.1.1.2 Second Meeting	3 days?	Thu 19/11/20	Mon 23/11/20
2.1.1.3 Third Meeing	3 days?	Thu 19/11/20	Mon 23/11/20
2.1.1.4 Providing feedback of meetings	3 days?	Thu 19/11/20	Mon 23/11/20
2.1.2 Project Market search	**4 days?**	**Tue 24/11/20**	**Fri 27/11/20**
2.1.2.1 Quantitative Market research	**4 days?**	**Tue 24/11/20**	**Fri 27/11/20**
2.1.2.1.1 Multiple Regression	4 days?	Tue 24/11/20	Fri 27/11/20
2.1.2.1.2 Discriminant Analysis	4 days?	Tue 24/11/20	Fri 27/11/20
2.1.2.1.4 Cluster Analysis	4 days?	Tue 24/11/20	Fri 27/11/20
2.2 Analysing previous project	**2 days?**	**Sat 28/11/20**	**Mon 30/11/20**
2.2.1 Stock market predictions	**2 days?**	**Sat 28/11/20**	**Mon 30/11/20**
2.2.1.1 technologies used	2 days?	Sat 28/11/20	Mon 30/11/20
2.2.1.2 Software used	2 days?	Sat 28/11/20	Mon 30/11/20
2.2.1.3 Big data analysis	2 days?	Sat 28/11/20	Mon 30/11/20
2.2.2 Making notes how to make it different from other projects	2 days?	Sat 28/11/20	Mon 30/11/20
2.3 Gathering any further information before next meeting with supervisor	2 days?	Sat 28/11/20	Mon 30/11/20
2.4 Gathering Big Data for project	**8 days?**	**Tue 01/12/20**	**Thu 10/12/20**
2.4.1 Downloading New York stock exchange data from keggle.com	8 days?	Tue 01/12/20	Thu 10/12/20
2.4.2 Downloading historical stock data of apple and amazon from Yahoo finance	8 days?	Tue 01/12/20	Thu 10/12/20
2.5 Collecting data from websites	**3 days?**	**Fri 11/12/20**	**Tue 15/12/20**
2.5.1 Finding details for stock predictions from internet	3 days?	Fri 11/12/20	Tue 15/12/20

	2.5.2 Web Scrapping	**3 days?**	**Fri 11/12/20**	**Tue 15/12/20**
	2.5.2.1 Using R programming language for collecting data from internet	3 days?	Fri 11/12/20	Tue 15/12/20
	2.5.2.2 collect first data and store in .csv file	3 days?	Fri 11/12/20	Tue 15/12/20
	2.6 Discussing project name with friends and supervisor	1 day	Wed 16/12/20	Wed 16/12/20
	2.7 Project Logo	**1 day?**	**Thu 17/12/20**	**Thu 17/12/20**
	2.7.1 Designing and editing project logo on adobe	1 day?	Thu 17/12/20	Thu 17/12/20
	2.8 Installing tools in computer	**1 day?**	**Fri 18/12/20**	**Fri 18/12/20**
	2.8.1 Rstudio	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.2 Tableau	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.3 IBM SPSS	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.4 MapReduce	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.5 Hadoop	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.6 Pandas Software	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.7 PyCharm			
	2.9 Submiting project proposal	**3 days?**	**Mon 02/11/20**	**Wed 04/11/20**
	2.9.1 Filling project proposal template	3 days?	Mon 02/11/20	Wed 04/11/20
	2.9.2 Uploading project proposal on Moodle	3 days?	Mon 02/11/20	Wed 04/11/20
	3 Execution	**90 days?**	**Tue 10/11/20**	**Mon 15/03/21**
	3.1 Importing data sets in Rstudio	**4 days?**	**Wed 11/11/20**	**Mon 16/11/20**
	3.1.1 Setting up GitHub and uploading codes regularly	4 days?	Wed 11/11/20	Mon 16/11/20
	3.2 Clean Datasets in Rstudio	**23 days?**	**Tue 17/11/20**	**Thu 17/12/20**
	3.2.1 Delete unnecessary column	23 days?	Tue 17/11/20	Thu 17/12/20
	3.2.2 Deleting unnecessary rows	23 days?	Tue 17/11/20	Thu 17/12/20
	3.2.3 Changing the name of column if possible	23 days?	Tue 17/11/20	Thu 17/12/20
	3.2.4 Re-ordering positions of columns			
	3.2.4 Deals with missing values	**23 days?**	**Tue 17/11/20**	**Thu 17/12/20**
	3.2.4.1 Filling the missing values of rows and columns	23 days?	Tue 17/11/20	Thu 17/12/20
	3.2.4.2 Entering specific values or null related to project	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3 Combining all datasets	**23 days?**	**Tue 17/11/20**	**Thu 17/12/20**
	3.3.1 Building visualizations	**23 days?**	**Tue 17/11/20**	**Thu 17/12/20**
	3.3.1.1 Building scatter plot	23 days?	Tue 17/11/20	Thu 17/12/20

	3.3.1.2 Building Histogram plot	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3.1.3 Building Bar chart	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3.1.4 Building Candle Stick chart	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3.1.5 Building ggplot2	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3.1.6 Building line graphs	23 days?	Tue 17/11/20	Thu 17/12/20
	3.4 Advanced Statistical Analysis	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.4.1 Mean, Median and Mode of all numerical values	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.2 Standard deviation, skewiss and kurtosis of all numerical datasets	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.3 Probability and combinations	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.5 Linear Regression	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.4.5.1 Simple Linear regression	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.5.2 Regression Analysis	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.6 Discriminant Analysis	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.4.6.1 Linear discriminant Analysis	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.6.2 Quadratic discriminant Analysis	10 days?	Mon 04/01/21	Fri 15/01/21
	3.5 Analysing data in Tableau	10 days?	Mon 04/01/21	Fri 15/01/21
	3.6 Applying MapReduce or Machine Learning RNN (LSTM)	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.6.1 MapReduce for big data	10 days?	Mon 04/01/21	Fri 15/01/21
	3.6.2 Applying Machine learning to predict apple stock	10 days?	Mon 04/01/21	Fri 15/01/21
	3.7 Meeting with supervisor before Mid-presentation	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.7.1 First Meeting and making changing	10 days?	Mon 04/01/21	Fri 15/01/21
	3.7.2 Second meeting and making changing	10 days?	Mon 04/01/21	Fri 15/01/21
	3.7.3 Third Meetig and making changing	10 days?	Mon 04/01/21	Fri 15/01/21
	3.8 High Level Analysis	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.8.1 Applying Methdologies	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.8.1.1 Knowledge Discovery Analysis	10 days?	Mon 04/01/21	Fri 15/01/21
	3.8.1.2 KMEans Clustering and Hererichal clutering	10 days?	Mon 04/01/21	Fri 15/01/21
	3.9 Uploading Mid-point Implementation on Moodle	1 day	Tue 22/12/20	Tue 22/12/20

	3.10 Uploading Mid-point documentation on Moodle	1 day	Tue 22/12/20	Tue 22/12/20
	3.11 Uploading Mid-Point preentation	1 day	Tue 22/12/20	Tue 22/12/20
	3.12 Applying Machine Learning	**25 days?**	**Sat 16/01/21**	**Thu 18/02/21**
	3.12.1 Long short term model in amazon dataset	25 days?	Sat 16/01/21	Thu 18/02/21
	3.12.2 LSTM in New York stock exchange datasets	25 days?	Sat 16/01/21	Thu 18/02/21
	3.13 Applying Web and Data Mining	**25 days?**	**Fri 19/02/21**	**Thu 25/03/21**
	3.13.1 Defining the problems	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.2 Identifying required data	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.3 Preparing Data/Clean Data	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.4 Model the Data	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.5 Train and test data	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.6 Verify and Deploy	25 days?	Fri 19/02/21	Thu 25/03/21
	3.14 Advanced Data Analysis	25 days?	Fri 19/02/21	Thu 25/03/21
	4 Monitoring and Controlling	**10 days?**	**Fri 26/03/21**	**Thu 08/04/21**
	4.1 Observing the behaviour of project	**10 days?**	**Fri 26/03/21**	**Thu 08/04/21**
	4.1.1 Testing Method	10 days?	Fri 26/03/21	Thu 08/04/21
	4.1.2 Case Study Method	10 days?	Fri 26/03/21	Thu 08/04/21
	4.1.3 Cross Sectional Methods	10 days?	Fri 26/03/21	Thu 08/04/21
	4.2 Detect and diagnose performance problems	10 days?	Fri 26/03/21	Thu 08/04/21
	4.3 Updating project performance to supervisor	10 days?	Fri 26/03/21	Thu 08/04/21
	4.4 perform Changing	**4 days?**	**Fri 09/04/21**	**Wed 14/04/21**
	4.4.1 Meeting with supervisor	4 days?	Fri 09/04/21	Wed 14/04/21
	4.4.2 Making changes if needed	4 days?	Fri 09/04/21	Wed 14/04/21
	4.4.3 Performing all possible changes	4 days?	Fri 09/04/21	Wed 14/04/21
	5 Project Closure	**2 days?**	**Thu 15/04/21**	**Fri 16/04/21**
	5.1 Conduct review for all possible changes	2 days?	Thu 15/04/21	Fri 16/04/21
	5.2 Review all the codes	2 days?	Thu 15/04/21	Fri 16/04/21
	5.3 Review all visualizations	2 days?	Thu 15/04/21	Fri 16/04/21
	5.4 Review final implementation	**1 day?**	**Sun 09/05/21**	**Sun 09/05/21**
	5.4.1 Uploading final implementation on Mood	1 day?	Sun 09/05/21	Sun 09/05/21
	5.5 Reviewing final documentation	**1 day?**	**Sun 09/05/21**	**Sun 09/05/21**
	5.5.1 Uploading final document on Moodle	1 day?	Sun 09/05/21	Sun 09/05/21
	5.6 Preparing Final presentation	**1 day?**	**Fri 14/05/21**	**Fri 14/05/21**
	5.6.1 Uploading final presentation on Moodle	1 day?	Fri 14/05/21	Fri 14/05/21
	5.7 Gain formal acceptance form supervisor	1 day?	Fri 14/05/21	Fri 14/05/21
	5.8 Documentation	**3 days?**	**Thu 20/05/21**	**Mon 24/05/21**
	5.8.1 Collecting all documents for future	3 days?	Thu 20/05/21	Mon 24/05/21
	5.8.2 Deleting all unecessary records	3 days?	Thu 20/05/21	Mon 24/05/21
	5.9 Lesson learned for future projects	3 days?	Thu 20/05/21	Mon 24/05/21

10.2. Reflective Journals

Reflective Journal

Student Name: Umer Iqbal **Student Number:** x17111854

Course: BSHCDA **Month:** October

Finalized Project Idea:

In every project, the first thing that need to be done is finalizing the project idea. In the beginning of this month I started to work on my project idea which will belong to my specialization data analytics. After few days research I came up with an idea to Analyse and predict Credit fraud detection. A lot of fraud is happening in big datasets like transection, credit card data and most of the companies are working on similar model.

Find data source:

After deciding my project idea, I researched online to find credit card datasets. Here are some sites from where I downloaded some datasets:

https://www.kaggle.com/datasets

https://aws.amazon.com/datasets/

https://data.gov.ie/data

https://toolbox.google.com/datasetsearch

After downloading some datasets from these sites, I started to analyse what type of data is stored in these files, how big is datasets, is it suitable for my projects, can I combine these datasets and make some insights like visualizations, how to perform statistics, how to use programming for big data and Machine Learning.

What technologies I decided to use:

Here are some technologies I installed in my PC for projects:

Tableau:

I downloaded and install tableau in my PC. Because this software can help me to see and understand my datasets. I can connect to any type of database; I just need to drag and drop and create my own visualizations. I can perform many statistical tools, calculations, and computation with just drag and drop.

RStudio:

RStudio is an IDE for R programming language to perform statistical computing and graphs. Every organization are moving to R and python and other open source language to make sense of data.

MapReduce:

It is programming model or pattern within the Hadoop framework which is used to access big data stored in the Hadoop File System.

SPSS:

Some of the data sets I found was in in the form of SPSS file. So, to run SPSS file it is important for me to install SPSS in my computer to run and analyse these files. The IBM SPSS software will provide us advanced statistical analysis and a massive machine learning algorithm.

Clustering Algorithm:

I did some research how I can use clustering algorithm and watch some online tutorials. It involves automatically discovering natural grouping in data.

Statistical Analysis:

After collecting my datasets, I did some statistical analysis in IBM SPSS and Microsoft Excel. I find average value of my numerical values in datasets e.g. how many transections are happened on daily, weekly, or monthly basis. Then, I find mean, median, mode standard deviation variance, range etc.

Then, I did some visualizations, build some graphs like line graphs scatter graphs, histogram graph, Bar chart etc.

Further What I decided to do:

In coming weeks, I decided to import my datasets in RStudio, clean my datasets, delete some columns if they are not useful for my project, deal with missing values, combine my datasets and build some visualizations by using R programming language. Then, I want to collect credit card fraud data from internet because ninety percent of data is available on internet by using web scrapping.

Reflective Journal

Student Name: Umer Iqbal

Student Number: x17111854

Course: BSHCDA

Month: November

RStudio:

I imported my new dataset (of my new project idea analysing and predicting stocks prices) in RStudio for further processing. I cleaned my datasets with R programming language delete values, added new columns, re-arranged them, changing the name of my columns, checking the empty rows, fill the empty rows etc.

Once my datasets were cleaned, I performed some distributed statistics analysis for both New York Stock exchange and apple dataset:

- Mean
- Median
- Mode
- Quartiles

- Range
- Weighted Mean
- Standard Deviation
- Variance
- Interquartile Range

Visualization:

Once I finished my statistics, I did some visualization like:

- Candle Stick chart for stock prices.
- Bar chart.
- Scatter Plot.
- Line plot.
- Weighted plot.
- Regression Analysis.

- Cluster Analysis.

PyCharm (Python)

I imported my datasets in PyCharm IDE and using Machine Learning tutorials for predicting stock prices.

Reflective Journal

Student Name: Umer Iqbal **Student Number:** x17111854

Course: BSHCDA **Month:** December

In this month, most of the time for Analysing and predicting stock price project was spent on mid-point documentation. First of all, I finished my Machine learning implementations, in this month by changing my python 3.8 to 3.9 because I was getting a lot of errors in my PyCharm IDE because some machine learning and deep learning packages does not support in python 3.8 like Keras or TensorFlow etc.

After, predicting apple stock price then I decided to predict the price of just one day of Apple 25th of December and compare it to the actual closing value and see how much difference we have in our predicted model and in actual closing price.

After that, I focus on my documentation explain all parts in the report one by one, the introduction of project what is goal and background of it. What methodology I used like KDD explain it briefly in the project. Writing the analysis part in the report and explain them.

Further, I am planning to more Machine Learning prediction for my datasets in PyCharm for amazon datasets and apply advance Data Mining in my project.

Reflective Journal

Student Name: Umer Iqbal **Student Number:** x17111854

Course: BSHCDA **Month:** January

After submitting my mid-point implementation, I researched mostly on data mining for analysing and predicting stock prices using R programming language in RStudio. Further, I decided to use time series analysis in this project which will provide me more insights and statistical analysis.

I imported my datasets to Tableau and performed some pre-processing like finding missing values, deleting unnecessary columns and rows, changing names of attributes if possible. Once my datasets are cleaned in Tableau, I performed some visualizations for some columns like high and low of New York stock exchange. I performed some bar charts, candle sticks charts etc.

After that I performed some machine learning predictions for more apple stocks like opening and high columns in PyCharm and store the predicted values. Further, I decided to find out predicted values of just one day of stock price like 1st February or 3rd February and compare it with original values and find difference between them.

Reflective Journal

Student Name: Umer Iqbal **Student Number:** x17111854

Course: BSHCDA **Month:** March

The month I spent most of my time working on data mining techniques like random forest, time series analysis, clustering, and decision tree.

Start working on final project report.

Applying pre-processing method to amazon data sets.

Applying LSTM techniques to amazon datasets for predictions.

Applying exploratory data analysis on my both apple and amazon data sets.

Built candle stick chart for both of my datasets.

Searching online for advanced visualisations how people displayed their stocks data in graphical methods.

Starting data mining techniques in R studio using different algorithm and testing them to see if it is working properly.

So far, my main goal is to do more research on data mining and predictions and applied it on project so I can explain all the outcomes in my project report at the end.

Reflective Journal

Student Name: Umer Iqbal **Student Number:** x17111854

Course: BSHCDA **Month:** May

The month I spent most of my time working on data mining techniques like random forest, time series analysis, clustering, and decision tree.

Start working on final project report.

Applying LSTM techniques to amazon datasets for predictions.

Applied ARIMA forecasting on last 5-year data and predict next 100 days.

Applying normality test for both Apple and Amazon dataset.

Applied Kruskal Wallis test for both Apple and Amazon dataset.

Applied Mann-Whitney test for both Apple and Amazon dataset.

Applied Wilcoxon Signed rank test in SPSS for both Apple and Amazon test.

The aim of this month for me was to finish my final report, finalise my output, apply some testing, and evaluate my analysis.

10.3. Project Proposal

National College of Ireland

Project Proposal

Analysing and predict stock prices

Date: 07-11-2020

Bachelor of Science HONS in Computing (BSHCDA)

Data Analytics

Academic Year i.e. 2020/2021

Umer Iqbal

X17111854

X17111854@student.ncirl.ie

Contents

Objectives

KDD is a multi-step process that involves data preparation, pattern searching and knowledge evaluation. KDD is knowledge discovery in database/data mining. Here are the objectives of my project:

Selecting and understanding Data from different internet sources.

Understanding the Data process starts at an early stage of Data collection and I want to get familiar with my datasets by performing some activities (like reading attributes, size of data, checking empty rows, structured or unstructured datasets, etc,) to recognize some data quality problems, I need to identify the first visualization in datasets.

Some online sites from where I can collect different companies/exchange historical stock data for statistical analysis, visualization, Machine Learning and Data Mining, etc.

a) https://www.kaggle.com/datasets
b) https://finance.yahoo.com/

Pre-processing/Clean Datasets which is collected from online resources.

It is important in every data analytics project that data we collected from online resources or websites need to be clean. Importing datasets into RStudio for the cleaning process the first thing to do is removing the irrelevant piece of data, deleting all duplicate values, changing attributes, adding new columns filling missing values, etc.

Machine Learning

Machine learning algorithms like recurrent neural networks can be used with training sets to predict predictions of opening, closing, high, low, etc. of stock prices.

Transformation processing.

It is a process of changing the format, structure, or values of data. Storing a big volume of data in data migration, data warehousing, or data wrangling for data transformation. Whereas data will be transformed to amazon web services to do run time analysis.

Data Mining

Data mining techniques can be used in the New York stock exchange, apple, amazon datasets. Different types of techniques will use for example in RStudio KMeans clustering, Hierarchical clustering, etc.

Interpretation/Evaluation

In interpretation and evaluation steps all outcomes will be performed in the form of graphs e.g., boxplots, Q-Q plots, scatter plots, Linear regression, etc. In this step we can decide should we buy stocks or not. Is prices of different opening and closing are going up or down.

Background

Stock market predictions are the act of trying to determine the future value of a company of stock or other financial traded on an exchange. In the past data, scientists provided a lot of profit to financial companies by predicting the future stock value.

Organizations and Corporations in the financial service sector are using data analytics tools to get into the stock market trends to make some wonderful decisions for their organizations. The stock market is very dynamic that thousands of transactions are happening every second around the world. Big Data enables the investor to analyze the data with some complex mathematical formulas and algorithms. In algorithm trading, computers analyze data quickly and provide rapid responses to investors. Whereas both humans and computers are complementing with each other, humans can create content and analyze data, but machines can process stock data in a fraction of a second.

There is a different type of analysis people used in the past to predict stocks.

Fundamental Analysis.

It is concerned with the companies that underlie the stock itself. They calculate the companies past performance and credibility of its accounts. Many performance ratios are created that aid the fundamental analysts with assessing the validity of stock e.g., P/E ratio.

Warren Buffett is one of the famous fundamental analysts, he used the overall market capitalizations to GDP ratio to indicate the relative ratio of the stock market.

Fundamental analysis in the stock market is to try to achieve the true value of stock and compare it with traded value. This type of analysis is thought of more as a long-term strategy.

Technical Analysis.

Data Scientists use their technical techniques to find the future values of stock based on their past prices. In past, analysts prefer to use short-term strategies rather than long-term strategies. The Candlestick chart which is developed by the Japanese rice merchants is one of the charts which is widely using in these days by all technical analysts.

There are some basics assumptions used in this analysis, the first being that everything significant about the company is already priced into a stock, other being that prices move in trends and lastly that history tends to repeat itself this is because of market psychology.

Technical Approach

Once I downloaded my historical stock prices datasets from the online resources then, I will import them to RStudio for cleaning datasets, I must delete all unnecessary columns and delete all unnecessary rows and change the names of columns if possible, deals with missing values (fill all the missing values if possible). After combining all my datasets, I will perform some visualizations to see some insights like building scatter plots, Histogram plots, Bar plots, ggplot2 and line graphs, etc.

Descriptive Statistics Analysis.

After that, I want to do some statistical analysis.

- Find the mean, median and mode of my dataset's columns.
- Standard deviations, skewness, and kurtosis.
- Quartiles, Range
- Weighted mean, variance, Interquartile Range

After statistical, I would like to some Maps Reduce for big data analysis by using the python language and machine learning for some prediction in datasets e.g., find the predictions of past dates compare it with actual values and see how much the difference is. Further, I want to do some data mining for my project, with the data mining process I can find anomalies, patterns within large data.

Special Resources Required

LinkedIn Learning

LinkedIn Learning is an online video course that is run by LinkedIn learning experts in the software. For my special resources, I downloaded some courses from LinkedIn learning sites that can help in my project like the Machine learning course, RStudio course, data mining course, etc.

Project Plan

	1 Initiation	**23 days?**	**Mon 28/09/20**	**Wed 28/10/20**
	1.1 Research on project idea	**9 days?**	**Wed 30/09/20**	**Mon 12/10/20**
	1.1.1 Finding a suitable idea fimillar to specillization	**4 days?**	**Wed 30/09/20**	**Sat 03/10/20**
	1.1.1.1 Research on internet	4 days	Wed 30/09/20	Sat 03/10/20
	1.1.1.2 Discussing project with last year students	4 days	Wed 30/09/20	Sat 03/10/20
	1.1.1.3 Sharing idea with lecturer and getting feedback	4 days	Wed 30/09/20	Sat 03/10/20
	1.1.1.4 Discussing with friends	4 days	Wed 30/09/20	Sat 03/10/20
	1.1.1.5 Explore emerging platform	**6 days?**	**Sun 11/10/20**	**Fri 16/10/20**
	1.1.1.5.1 Research on google analytics platform	**6 days?**	**Fri 30/10/20**	**Fri 06/11/20**
	1.1.1.5.1.1 Traffic Channels	6 days?	Fri 30/10/20	Fri 06/11/20
	1.1.1.5.1.2 Blog post engagement	6 days?	Fri 30/10/20	Fri 06/11/20
	1.1.1.5.1.3 performance by platform	6 days?	Fri 30/10/20	Fri 06/11/20
	1.1.1.5.1.4 Google big data analytics	6 days?	Fri 30/10/20	Fri 06/11/20
	1.1.1.6 Research on github for new ideas	6 days?	Fri 30/10/20	Fri 06/11/20

	1.2 Project risks and issues	**4 days?**	**Wed 07/10/20**	**Mon 12/10/20**
	1.2.1 Data quality	4 days?	Wed 07/10/20	Mon 12/10/20
	1.2.2 Server or sites down	4 days?	Wed 07/10/20	Mon 12/10/20
	1.2.3 Cleaning big data	4 days?	Wed 07/10/20	Mon 12/10/20
	1.3 Doing fesability study on project	**6 days?**	**Tue 13/10/20**	**Tue 20/10/20**
	1.3.1 Market Research	6 days?	Tue 13/10/20	Tue 20/10/20
	1.3.2 Financial Research	6 days?	Tue 13/10/20	Tue 20/10/20
	1.3.3 What is stock predictions			
	1.3.4 Capability of project	6 days?	Tue 13/10/20	Tue 20/10/20
	1.3.4 Studying online tutorial	**9 days?**	**Thu 22/10/20**	**Tue 03/11/20**
	1.3.4.1 How to use Tableau for big data analysis	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.2 Advanced R programming tutorial from linkedIn learning	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.3 IBM SPSS tutorial	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.4 Advance statistical tutorial	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.5 Data Mining	9 days?	Thu 22/10/20	Tue 03/11/20
	1.3.4.6 Advance visualization tutorial	9 days?	Thu 22/10/20	Tue 03/11/20

	1.3.4.6 Machine learning Tutorials	**5 days?**	**Thu 05/11/20**	**Wed 11/11/20**
	1.3.4.6.1 Machine Learning LSTM	5 days?	Thu 05/11/20	Wed 11/11/20
	1.3.4.6.3 Clustering Algorithms	5 days?	Thu 05/11/20	Wed 11/11/20
	1.3.4.7 MapReduce from linkedIn learning	5 days?	Thu 05/11/20	Wed 11/11/20
	1.4 Finding Stakeholder of project	**2 days?**	**Fri 13/11/20**	**Mon 16/11/20**
	1.4.1 Appointed the project	**2 days?**	**Fri 13/11/20**	**Mon 16/11/20**
	1.4.1.1 Umer Iqbal	2 days?	Fri 13/11/20	Mon 16/11/20
	1.4.2 Appointed supervisor by college	**2 days?**	**Fri 13/11/20**	**Mon 16/11/20**
	1.4.2.1 Keith Maycock	2 days?	Fri 13/11/20	Mon 16/11/20
	1.4.3 Setting meeting with supervisor for feedback on project	**2 days?**	**Wed 18/11/20**	**Thu 19/11/20**
	1.4.3.1 Meeting 1	2 days?	Wed 18/11/20	Thu 19/11/20
	1.4.3.2 Meeting 2	2 days?	Wed 18/11/20	Thu 19/11/20
	1.4.3.3 Meeting 3	2 days?	Wed 18/11/20	Thu 19/11/20
	1.5 Project pitch video	**4 days?**	**Mon 12/10/20**	**Thu 15/10/20**
	1.5.1 Making notes for video	4 days?	Mon 12/10/20	Thu 15/10/20
	1.5.2 Recording video on Laptop	4 days?	Mon 12/10/20	Thu 15/10/20

	1.5.3 Editing video on adobe workshop	4 days?	Mon 12/10/20	Thu 15/10/20
	1.5.4 Uploading project pitch video on Moodle	4 days?	Mon 12/10/20	Thu 15/10/20
	1.5.5 Email supervisor for feedback	4 days?	Mon 12/10/20	Thu 15/10/20
	1.6 Review project with lecturer and supervisor	4 days?	Mon 12/10/20	Thu 15/10/20
	2 Planning	**18 days?**	**Thu 19/11/20**	**Mon 14/12/20**
	2.1 Discussing project requirement with supervisor	**3 days?**	**Thu 19/11/20**	**Mon 23/11/20**
	2.1.1 Setting up meeting	**3 days?**	**Thu 19/11/20**	**Mon 23/11/20**
	2.1.1.1 Meeting first	3 days?	Thu 19/11/20	Mon 23/11/20
	2.1.1.2 Second Meeting	3 days?	Thu 19/11/20	Mon 23/11/20
	2.1.1.3 Third Meeing	3 days?	Thu 19/11/20	Mon 23/11/20
	2.1.1.4 Providing feedback of meetings	3 days?	Thu 19/11/20	Mon 23/11/20
	2.1.2 Project Market search	**4 days?**	**Tue 24/11/20**	**Fri 27/11/20**
	2.1.2.1 Quantitative Market research	**4 days?**	**Tue 24/11/20**	**Fri 27/11/20**
	2.1.2.1.1 Multiple Regression	4 days?	Tue 24/11/20	Fri 27/11/20
	2.1.2.1.2 Discriminant Analysis	4 days?	Tue 24/11/20	Fri 27/11/20
	2.1.2.1.4 Cluster Analysis	4 days?	Tue 24/11/20	Fri 27/11/20

	2.2 Analysing previous project	**2 days?**	**Sat 28/11/20**	**Mon 30/11/20**
	2.2.1 Stock market predictions	**2 days?**	**Sat 28/11/20**	**Mon 30/11/20**
	2.2.1.1 technologies used	2 days?	Sat 28/11/20	Mon 30/11/20
	2.2.1.2 Software used	2 days?	Sat 28/11/20	Mon 30/11/20
	2.2.1.3 Big data analysis	2 days?	Sat 28/11/20	Mon 30/11/20
	2.2.2 Making notes how to make it different from other projects	2 days?	Sat 28/11/20	Mon 30/11/20
	2.3 Gathering any further information before next meeting with supervisor	2 days?	Sat 28/11/20	Mon 30/11/20
	2.4 Gathering Big Data for project	**8 days?**	**Tue 01/12/20**	**Thu 10/12/20**
	2.4.1 Downloading New York stock exchange data from keggle.com	8 days?	Tue 01/12/20	Thu 10/12/20
	2.4.2 Downloading historical stock data of apple and amazon from Yahoo finance	8 days?	Tue 01/12/20	Thu 10/12/20
	2.5 Collecting data from websites	**3 days?**	**Fri 11/12/20**	**Tue 15/12/20**
	2.5.1 Finding details for stock predictions from internet	3 days?	Fri 11/12/20	Tue 15/12/20
	2.5.2 Web Scrapping	**3 days?**	**Fri 11/12/20**	**Tue 15/12/20**
	2.5.2.1 Using R programming language for collecting data from internet	3 days?	Fri 11/12/20	Tue 15/12/20
	2.5.2.2 collect first data and store in .csv file	3 days?	Fri 11/12/20	Tue 15/12/20
	2.6 Discussing project name with friends and supervisor	1 day	Wed 16/12/20	Wed 16/12/20
	2.7 Project Logo	**1 day?**	**Thu 17/12/20**	**Thu 17/12/20**
	2.7.1 Designing and editing project logo on adobe	1 day?	Thu 17/12/20	Thu 17/12/20
	2.8 Installing tools in computer	**1 day?**	**Fri 18/12/20**	**Fri 18/12/20**
	2.8.1 Rstudio	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.2 Tableau	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.3 IBM SPSS	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.4 MapReduce	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.5 Hadoop	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.6 Pandas Software	1 day?	Fri 18/12/20	Fri 18/12/20
	2.8.7 PyCharm			
	2.9 Submiting project proposal	**3 days?**	**Mon 02/11/20**	**Wed 04/11/20**
	2.9.1 Filling project proposal template	3 days?	Mon 02/11/20	Wed 04/11/20
	2.9.2 Uploading project proposal on Moodle	3 days?	Mon 02/11/20	Wed 04/11/20
	3 Execution	**90 days?**	**Tue 10/11/20**	**Mon 15/03/21**
	3.1 Importing data sets in Rstudio	**4 days?**	**Wed 11/11/20**	**Mon 16/11/20**
	3.1.1 Setting up GitHub and uploading codes regularly	4 days?	Wed 11/11/20	Mon 16/11/20
	3.2 Clean Datasets in Rstudio	**23 days?**	**Tue 17/11/20**	**Thu 17/12/20**
	3.2.1 Delete unnecessary column	23 days?	Tue 17/11/20	Thu 17/12/20
	3.2.2 Deleting unnecessary rows	23 days?	Tue 17/11/20	Thu 17/12/20
	3.2.3 Changing the name of column if possible	23 days?	Tue 17/11/20	Thu 17/12/20
	3.2.4 Re-ordering positions of columns			
	3.2.4 Deals with missing values	**23 days?**	**Tue 17/11/20**	**Thu 17/12/20**
	3.2.4.1 Filling the missing values of rows and columns	23 days?	Tue 17/11/20	Thu 17/12/20
	3.2.4.2 Entering specific values or null related to project	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3 Combining all datasets	**23 days?**	**Tue 17/11/20**	**Thu 17/12/20**
	3.3.1 Building visualizations	**23 days?**	**Tue 17/11/20**	**Thu 17/12/20**
	3.3.1.1 Building scatter plot	23 days?	Tue 17/11/20	Thu 17/12/20

	3.3.1.2 Building Histogram plot	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3.1.3 Building Bar chart	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3.1.4 Building Candle Stick chart	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3.1.5 Building ggplot2	23 days?	Tue 17/11/20	Thu 17/12/20
	3.3.1.6 Building line graphs	23 days?	Tue 17/11/20	Thu 17/12/20
	3.4 Advanced Statistical Analysis	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.4.1 Mean, Median and Mode of all numerical values	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.2 Standard deviation, skewiss and kurtosis of all numerical datasets	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.3 Probability and combinations	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.5 Linear Regression	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.4.5.1 Simple Linear regression	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.5.2 Regression Analysis	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.6 Discriminant Analysis	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.4.6.1 Linear discriminant Analysis	10 days?	Mon 04/01/21	Fri 15/01/21
	3.4.6.2 Quadratic discriminant Analysis	10 days?	Mon 04/01/21	Fri 15/01/21
	3.5 Analysing data in Tableau	10 days?	Mon 04/01/21	Fri 15/01/21
	3.6 Applying MapReduce or Machine Learning RNN (LSTM)	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.6.1 MapReduce for big data	10 days?	Mon 04/01/21	Fri 15/01/21
	3.6.2 Applying Machine learning to predict apple stock	10 days?	Mon 04/01/21	Fri 15/01/21
	3.7 Meeting with supervisor before Mid-presentation	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.7.1 First Meeting and making changing	10 days?	Mon 04/01/21	Fri 15/01/21
	3.7.2 Second meeting and making changing	10 days?	Mon 04/01/21	Fri 15/01/21
	3.7.3 Third Meetig and making changing	10 days?	Mon 04/01/21	Fri 15/01/21
	3.8 High Level Analysis	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.8.1 Applying Methdologies	**10 days?**	**Mon 04/01/21**	**Fri 15/01/21**
	3.8.1.1 Knowledge Discovery Analysis	10 days?	Mon 04/01/21	Fri 15/01/21
	3.8.1.2 KMEans Clustering and Hererichal clutering	10 days?	Mon 04/01/21	Fri 15/01/21
	3.9 Uploading Mid-point Implementation on Moodle	1 day	Tue 22/12/20	Tue 22/12/20

	3.10 Uploading Mid-point documentation on Moodle	1 day	Tue 22/12/20	Tue 22/12/20
	3.11 Uploading Mid-Point preentation	1 day	Tue 22/12/20	Tue 22/12/20
	3.12 Applying Machine Learning	**25 days?**	**Sat 16/01/21**	**Thu 18/02/21**
	3.12.1 Long short term model in amazon dataset	25 days?	Sat 16/01/21	Thu 18/02/21
	3.12.2 LSTM in New York stock exchange datasets	25 days?	Sat 16/01/21	Thu 18/02/21
	3.13 Applying Web and Data Mining	**25 days?**	**Fri 19/02/21**	**Thu 25/03/21**
	3.13.1 Defining the problems	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.2 Identifying required data	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.3 Preparing Data/Clean Data	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.4 Model the Data	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.5 Train and test data	25 days?	Fri 19/02/21	Thu 25/03/21
	3.13.6 Verify and Deploy	25 days?	Fri 19/02/21	Thu 25/03/21
	3.14 Advanced Data Analysis	25 days?	Fri 19/02/21	Thu 25/03/21
	4 Monitoring and Controlling	**10 days?**	**Fri 26/03/21**	**Thu 08/04/21**
	4.1 Observing the behaviour of project	**10 days?**	**Fri 26/03/21**	**Thu 08/04/21**
	4.1.1 Testing Method	10 days?	Fri 26/03/21	Thu 08/04/21
	4.1.2 Case Study Method	10 days?	Fri 26/03/21	Thu 08/04/21
	4.1.3 Cross Sectional Methods	10 days?	Fri 26/03/21	Thu 08/04/21
	4.2 Detect and diagnose performance problems	10 days?	Fri 26/03/21	Thu 08/04/21
	4.3 Updating project performance to supervisor	10 days?	Fri 26/03/21	Thu 08/04/21
	4.4 perform Changing	**4 days?**	**Fri 09/04/21**	**Wed 14/04/21**
	4.4.1 Meeting with supervisor	4 days?	Fri 09/04/21	Wed 14/04/21
	4.4.2 Making changes if needed	4 days?	Fri 09/04/21	Wed 14/04/21
	4.4.3 Performing all possible changes	4 days?	Fri 09/04/21	Wed 14/04/21
	5 Project Closure	**2 days?**	**Thu 15/04/21**	**Fri 16/04/21**
	5.1 Conduct review for all possible changes	2 days?	Thu 15/04/21	Fri 16/04/21
	5.2 Review all the codes	2 days?	Thu 15/04/21	Fri 16/04/21
	5.3 Review all visualizations	2 days?	Thu 15/04/21	Fri 16/04/21
	5.4 Review final implementation	**1 day?**	**Sun 09/05/21**	**Sun 09/05/21**
	5.4.1 Uploading final implementation on Mood	1 day?	Sun 09/05/21	Sun 09/05/21
	5.5 Reviewing final documentation	**1 day?**	**Sun 09/05/21**	**Sun 09/05/21**
	5.5.1 Uploading final document on Moodle	1 day?	Sun 09/05/21	Sun 09/05/21
	5.6 Preparing Final presentation	**1 day?**	**Fri 14/05/21**	**Fri 14/05/21**
	5.6.1 Uploading final presentation on Moodle	1 day?	Fri 14/05/21	Fri 14/05/21
	5.7 Gain formal acceptance form supervisor	1 day?	Fri 14/05/21	Fri 14/05/21
	5.8 Documentation	**3 days?**	**Thu 20/05/21**	**Mon 24/05/21**
	5.8.1 Collecting all documents for future	3 days?	Thu 20/05/21	Mon 24/05/21
	5.8.2 Deleting all unecessary records	3 days?	Thu 20/05/21	Mon 24/05/21
	5.9 Lesson learned for future projects	3 days?	Thu 20/05/21	Mon 24/05/21

Technical Details

Tableau:

This software can help me to see and understand my datasets. I can connect to any type of database; I just need to drag and drop and create my visualizations. I can perform many statistical tools, calculations, and computation with just drag and drop.

RStudio:

RStudio is an IDE for the R programming language to perform statistical computing and graphs. Every organization is moving to R and python and other open-source languages to make sense of data.

MapReduce:

It is a programming model or pattern within the Hadoop framework which is used to access big data stored in the Hadoop File System.

SPSS:

The IBM SPSS software will provide us advanced statistical analysis and a massive machine learning algorithm.

Clustering Algorithm:

It is a task of grouping a set of objects, in a way that objects in the same group are more similar to each other than to those in other groups.

Web Scrapping:

It is a technique where I can use to collect a large amount of data from online resources.

Data Mining:

It is the process in which we can find some anomalies in my datasets and can use them to find a pattern in my credit card fraud datasets.

Evaluation

Organizing or cleaning my datasets and make sure it is ready to analyze. Quantitative analyses that count things like tallying response counting program activities and calculating some changes in results.

After performing all my technical approach, cleaning my datasets, performing visualizations, Machine learning algorithm, etc. where end-user can see an output in the form graph for all data analysis like when prices of stocks will go up or down, how much profit they can make based on predictions model if they want to buy stocks.

Reference

upGrad blog. 2020. Data Cleaning Techniques: Learn Simple & Effective Ways To Clean Data. [online] Available at:

https://www.upgrad.com/blog/data-cleaning-techniques/

[Accessed 7 November 2020].

Available at: https://link.springer.com/chapter/10.1007/978-3-030-11928-7_17#:~:text=Data%20mining%20(DM)%20plays%20an,the%20different%20methods%20of%20DM.

[Accessed 7 November 2020].

Available at: https://www.cottagehealth.org/population-health/learning-lab/toolkit/analyze-interpret-evaluation-data/

Accessed 7 November 2020].

Available at: https://en.wikipedia.org/wiki/Stock_market_prediction

Accessed 7 November 2020].

Available at: https://www.cioinsiderindia.com/tech-buzz/benefits-of-data-analysis-in-stock-market-prediction-tbid-1453.html

Accessed 7 November 2020].