Project: Stock Price Forecast using Time Series Analysis

1. Introduction

The general research associated with the stock or share market is highly focusing on neither buy nor sell but it fails to address the dimensionality and expectancy of a new investor. The general sentiment in society against the stock market is that it is highly risky for investment or unfit for trade, so most people are not even interested. The seasonal volatility and steady movement of any index would allow both current and naïve investors to understand the stock/share market and decide to invest.

The time series analysis would be the best method for predicting the trend, or even the future, to solve these types of problems. The trend map should provide the investor with the necessary guidance. Transforming the time series using ARIMA is a better algorithmic approach than directly forecasting, as it gives more accurate and reliable results.

Autoregressive Integrated Moving Average (ARIMA) Model converts non-stationary data to stationary data before working on it. It is one of the most popular models to predict linear time series data. ARIMA model has been widely used in finance and economics, as it is reliable, efficient, and has a good potential for short-term market share prediction.

For this project Weekly closing stock prices are used to build ARIMA model and forecast ability of model is validated. The adequacy of the model is checked using various diagnostic and residual plots. To study different types of model, stock prices of 3 different companies are taken as dataset.

2. Data Collection

The dataset consists of stock market data of Apple Inc., Tesla Inc., and Microsoft Corporation which is collected from Yahoo Finance. For this analysis, we are isolate weekly close prices as the benchmark for future predictions, as those values are assumed to best reflect the changes in real values of the stock during the time frame.

2.1 Company Profiles

2.1.1 Apple Inc. (AAPL)

Apple Inc. was founded in 1977 and is headquartered in Cupertino, California. Apple Inc. designs, manufactures, and markets smartphones, personal computers, tablets, wearables, and

accessories worldwide. It also sells various related services. The company offers iPhone, a line of smartphones; Mac, a line of personal computers; iPad, a line of multi-purpose tablets; and wearables, home, and accessories comprising AirPods, Apple TV, Apple Watch, Beats products, HomePod, iPod touch, and other Apple-branded and third-party accessories. It also provides digital content stores and streaming services.

2.2.2 Tesla, Inc. (TSLA)

Tesla, Inc. was founded in 2003 and is headquartered in Palo Alto, California. Tesla, Inc. designs, develops, manufactures, leases, and sells electric vehicles, and energy generation and storage systems in the United States, China, Netherlands, Norway, and internationally. The company operates in two segments, Automotive and Energy Generation and Storage. The Automotive segment offers sedans and sport utility vehicles. It also provides electric powertrain components and systems; and services for electric vehicles through its company-owned service locations, and Tesla mobile service technicians. The Energy Generation and Storage segment offers energy storage products, such as rechargeable lithium-ion battery systems for use in homes, industrial, commercial facilities, and utility grids; and designs, manufactures, installs, maintains, leases, and sells solar energy generation and energy storage products to residential and commercial customers. It also provides vehicle insurance services, as well as renewable energy.

2.1.3 Microsoft Corporation (MSFT)

Microsoft Corporation was founded in 1975 and is headquartered in Redmond, Washington. Microsoft Corporation develops, licenses, and supports software, services, devices, and solutions worldwide. Its Productivity and Business Processes segment offers Office, Exchange, SharePoint, Microsoft Teams, Office 365 Security and Compliance, and Skype for Business, as well as related Client Access Licenses (CAL); and Skype, Outlook.com, and OneDrive. It also provides LinkedIn that includes Talent and marketing solutions, and subscriptions; and Dynamics 365, a set of cloud-based and on-premises business solutions for small and medium businesses, large organizations, and divisions of enterprises. Its Intelligent Cloud segment licenses SQL and Windows Servers, Visual Studio, System Center, and related CALs; GitHub that provides a collaboration platform and code hosting service for developers; and Azure, a cloud platform. It also provides support services and Microsoft consulting services to assist customers in developing, deploying, and managing Microsoft server and desktop solutions, and training and certification to developers and IT professionals on various Microsoft products.

2.2. Data overview

Weekly closing stock price data is collected in such a way that training set will have at least 5 year of data and testing set will have enough data in order to check predictive ability of model.

Weekly Closing Stock price data collected,

From date: 01/01/2014

Till date: 28/04/2020.

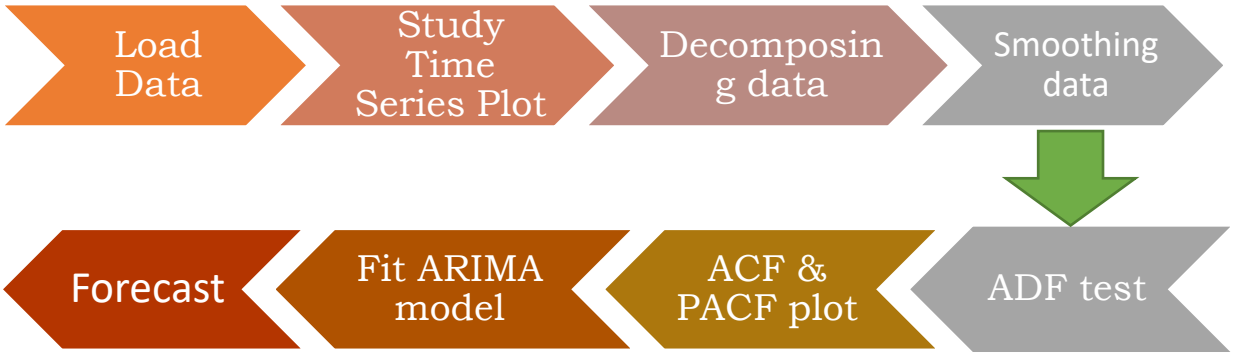
Data till end of 2019 is used for training the ARIMA model and remaining data is used for testing the forecast ability of model.

3. Data Analysis & Result discussion

Stock price Data for each company is analyzed separately to build ARIMA model and with help of model, future trend in stock prices is identified. The analysis is carried out using R studio.

3.1 Procedure

Below is the steps followed for time series analysis,



3.2. Stock price analysis

AAPL:

Step 1: Load Data

The data can be loaded in R studio in several different ways. There are some packages available in R that allows downloading weekly stock data from yahoo finance. The data is downloaded from January 1, 2014 to April 28, 2020. Only date and closing stock price data is kept and all other data is removed.

Step 2: Time series Plot

Time series plot gives picture about how stock price has changed over study period. It is also helpful in identifying any trend, seasonality in the data.

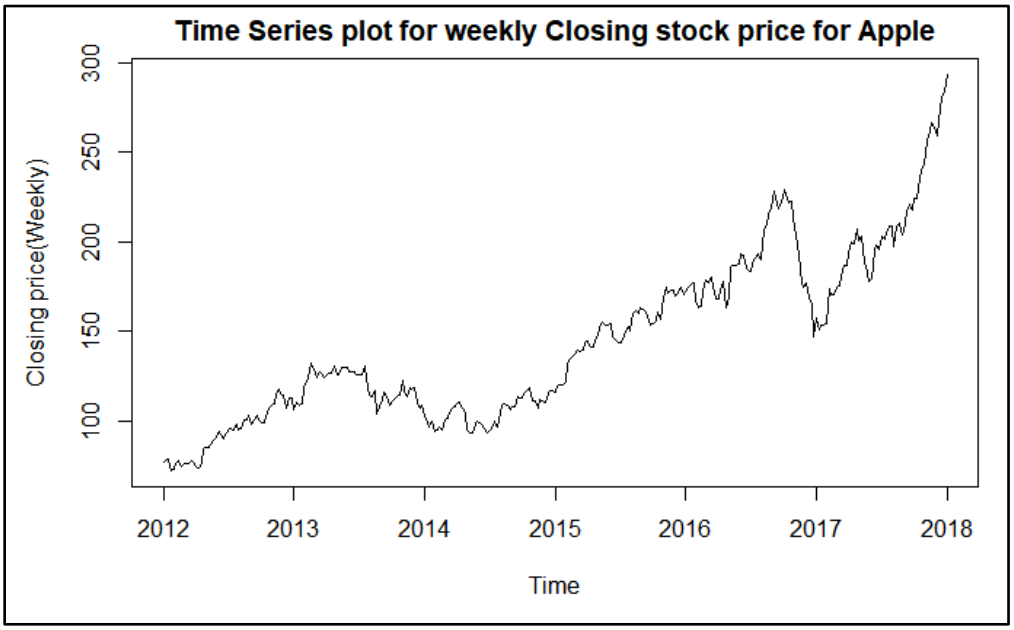


Fig. 1 – Time series Plot for Apple stock

Looking at time series plot, we can see that stock price has increased over time. There seems to be constant volatility in stock prices, still no cyclicity is observed. The stock prices have spiked recently.

Step 3: Decomposing data

To further analyze trends in the data, we can decompose the data into seasonal, and trend components. Decomposition is the foundation for building the ARIMA model by providing insight towards the changes in stock fluctuations.

- The observed data section shows time series plot same as in Step 2.
- The trend component describes the overall pattern of the series over the entire range of time, considering increases and decreases in prices together. From the plot below, the trend is overall increasing.
- The seasonal component describes the fluctuations in stock price based on the calendar year. The pick in stock prices occurs every year in Q1 (October, November, December) with immediate trough in Q2 (January, February, March). Overall, quarterly fluctuations are observed in stock prices that indicates presence of cyclicity.

- The random (residual) error, or noise section describes the trends that cannot be explained by trend or seasonal components. Statistically, these errors are the difference between the observed price and the estimated price. Random error is particularly important for this project because a statistical model can only be fit if the residuals are independent and independently distributed. The random plot shows fluctuations in data indicating white noise.

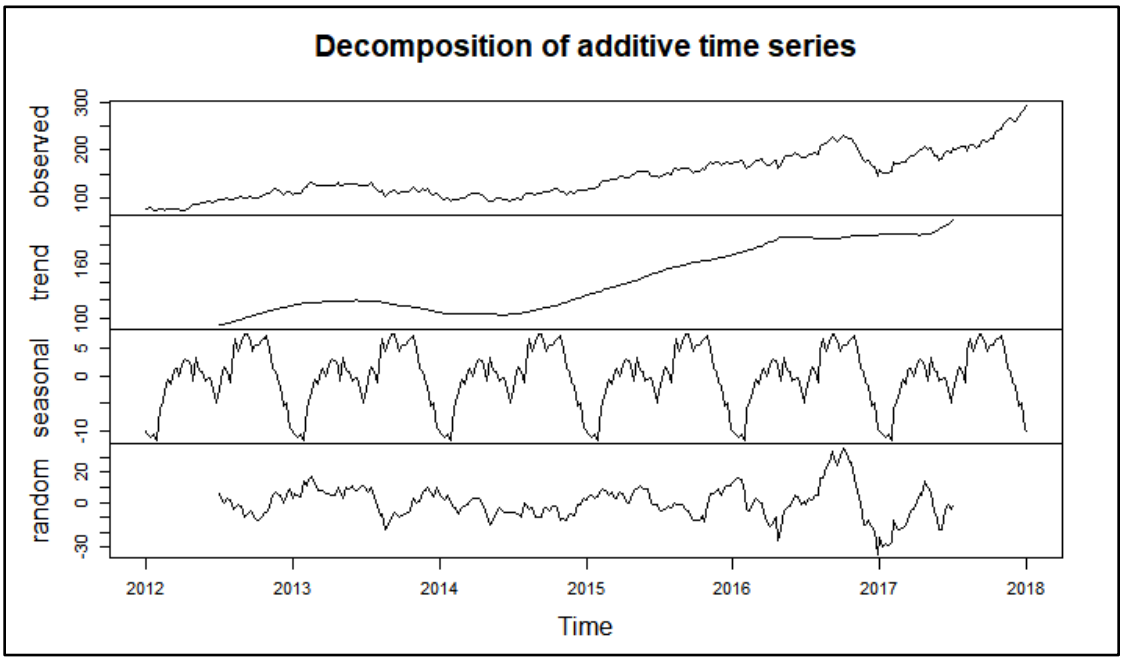


Fig. 2 – Decomposing data into components

Step 4: Smoothing data

As seen from decomposition plots data has variance and in order to stabilize the data smoothing techniques are implied.

In this project three types of smoothing techniques are used to stabilize the data.

1. Logarithmic returns

The main reason to use logarithmic returns is that the fluctuations observed in stock prices can be better compared over time and help in describing trends. It returns smoothed curve with reduced variation hence more accurate forecast model can be build.

The Logarithmic plot below is smoother than time series plot from step 2. The curve exactly shows trend in dataset by reducing fluctuations.

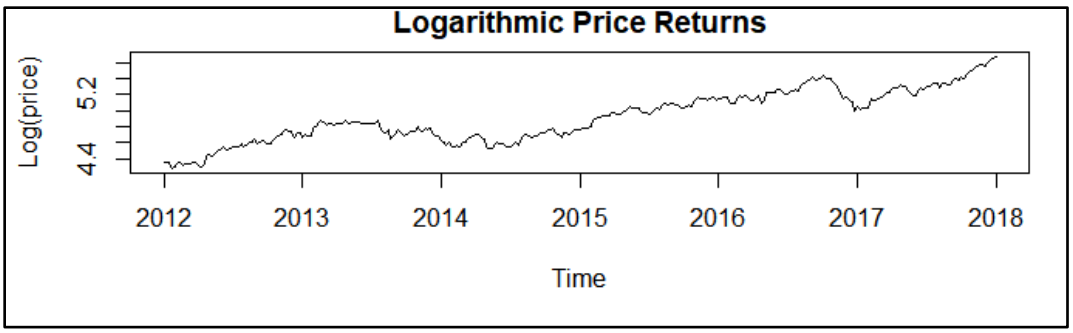


Fig. 3 – Logarithmic price returns plot

2. Square root values

Square root values instead of raw prices are used to scale the volatility between points to manage the time horizon of the stock. This is especially important because the longer a position is held, the greater a potential loss can be found.

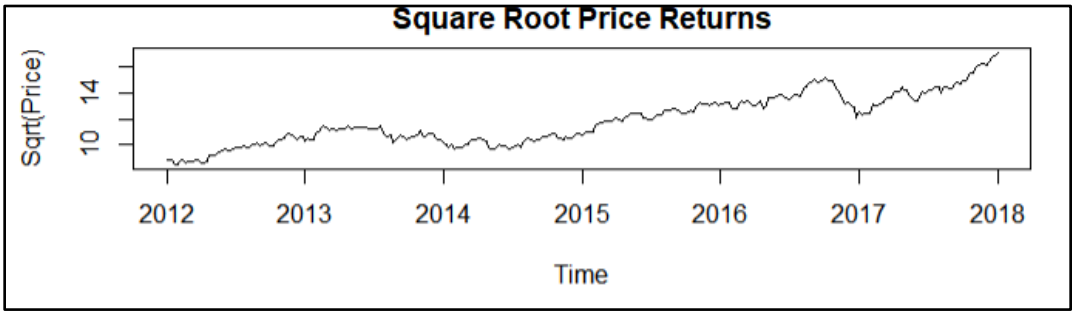


Fig. 4 – Square price returns plot

3. Differencing

Differencing is used to make data stationary. This is an important step in ARIMA model, as making data stationary allows analyst to make assumption that prediction trend will be same in the future as it was in the past.

The stock price data for APPLE requires first order differencing.

Both differenced plots below show that data have become stationary after first differencing.

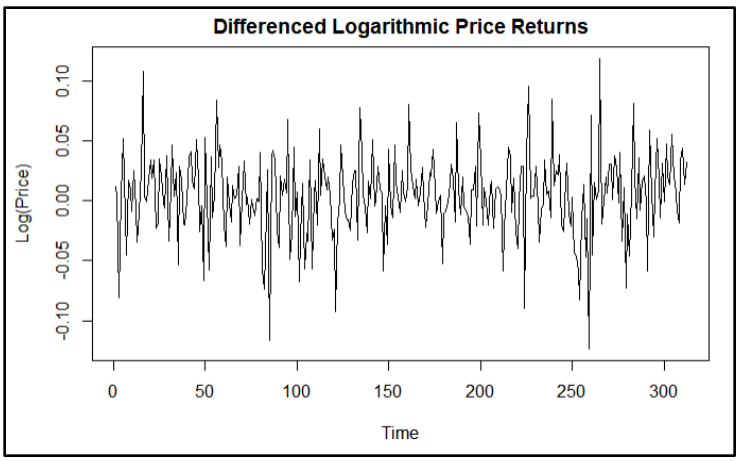


Fig. 5 – Differenced Logarithmic price returns plot

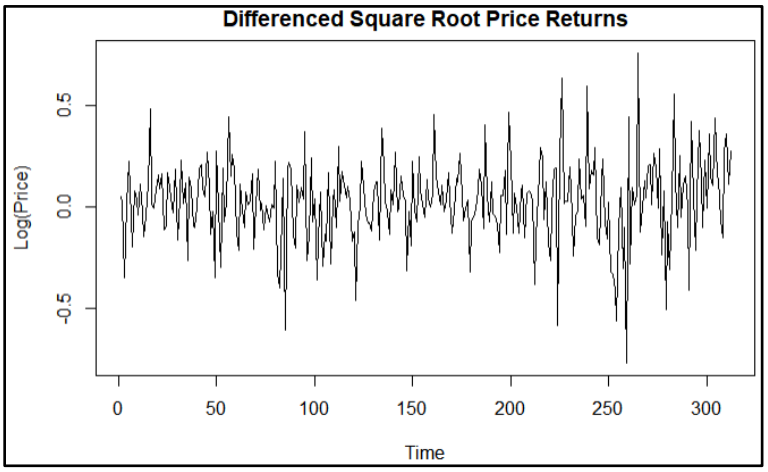


Fig. 6 – Differenced Square price returns plot

Step 5: ADF Test

The Augmented Dickey Fuller Test is used to determine if data is stationary. It calculates p-value that determines which hypothesis to choose.

Hypothesis are as follows,

Null Hypothesis (H0) (p-value>0.05): The time series data is non-stationary.

Alternate hypothesis (Ha) (p-value<0.05: The time series in stationary.

p-value for first two statistics is above 0.05, which indicates that null hypothesis can not be rejected. So, logarithmic return and square root return data is non-stationary.

Whereas, p-value for above two return prices after differencing is below 0.05, hence null hypothesis can be rejected. This indicates that first order differencing has made data stationary.

```
        Augmented Dickey-Fuller Test

data:  logprice[, 2]
Dickey-Fuller = -2.0311, Lag order = 6, p-value = 0.5633
alternative hypothesis: stationary


        Augmented Dickey-Fuller Test

data:  sqrtprice[, 2]
Dickey-Fuller = -1.6319, Lag order = 6, p-value = 0.7316
alternative hypothesis: stationary

p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  dlogprice
Dickey-Fuller = -6.5829, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  dsqrtprice
Dickey-Fuller = -6.3521, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Step 6: ACF & PACF plot

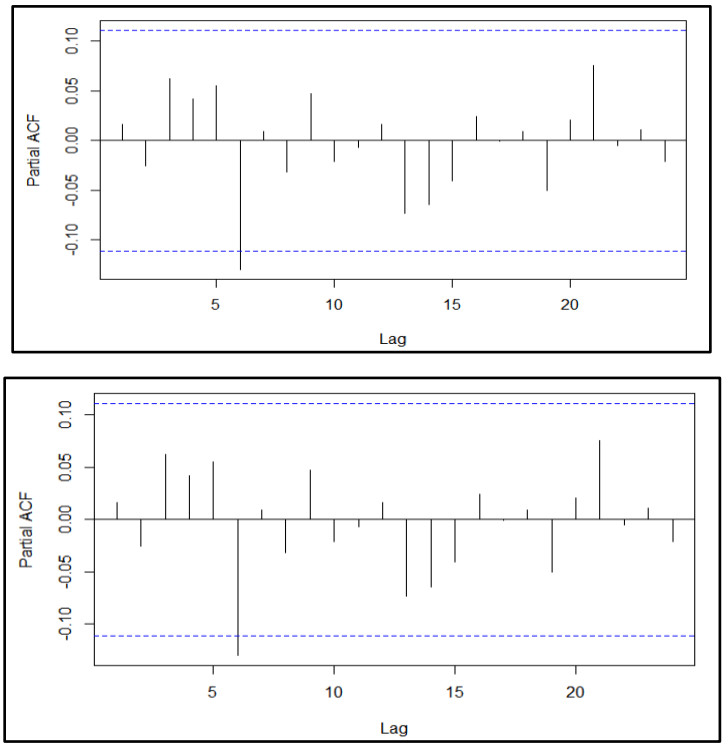


Fig. 7 – ACF & PACF plot

ARIMA models integrate two types of correlograms,

- The AutoCorrelation Function (ACF) displays the correlation between series and lags for the Moving Average (q) of the ARIMA model. Here, ACF plot shows no lag is cutting the

blue line that means the q-notation is 0. Although one lag at 7 is cutting the blue line, it can be ignored.

- The Partial AutoCorrelation Function (PACF) displays the correlation between returns and lags for the Auto-Regression (p) of the ARIMA model. Here, PACF plot have no lag cutting beyond blue line that means the p-notation is also 0.

Based on ACF & PACF plot analysis fitting ARIMA (0,0,0) model would yield best results.

Step 7: Fit ARIMA model

The model to fit onto the logarithmic price returns is ARIMA (0,0,0). This particular ARIMA model represents white noise, which means no model will be able to fit the square root values for this stock.

```
Call:
arima(x = dlogprice, order = c(0, 0, 0))

Coefficients:
      intercept
         0.0043
s.e.     0.0019

sigma^2 estimated as 0.001146:  log likelihood = 613.68,  aic = -1223.36

Training set error measures:
                      ME       RMSE        MAE  MPE MAPE      MASE       ACF1
Training set 2.055152e-18 0.03384859 0.02494771 -Inf  Inf 0.6826687 0.01653934
```

Model adequacy can be checked using dignostic plots.

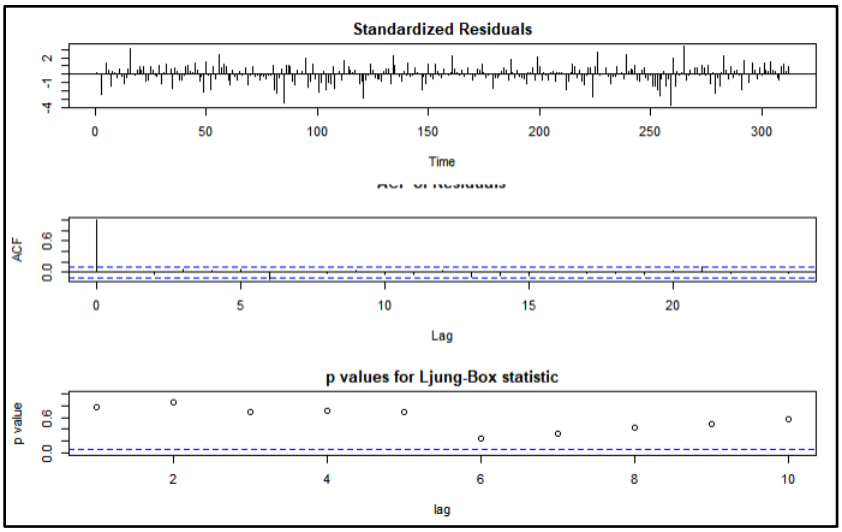


Fig. 8 – Diagnostic plot

Dignostic plots looks fine and represents adquacy of model. As ACF plot of residuals have no large lag, p-values are above 0.05. The histogram of residuals indicates that residuals are normally distributed. Normal distribution of residuals shows that model is adequate.

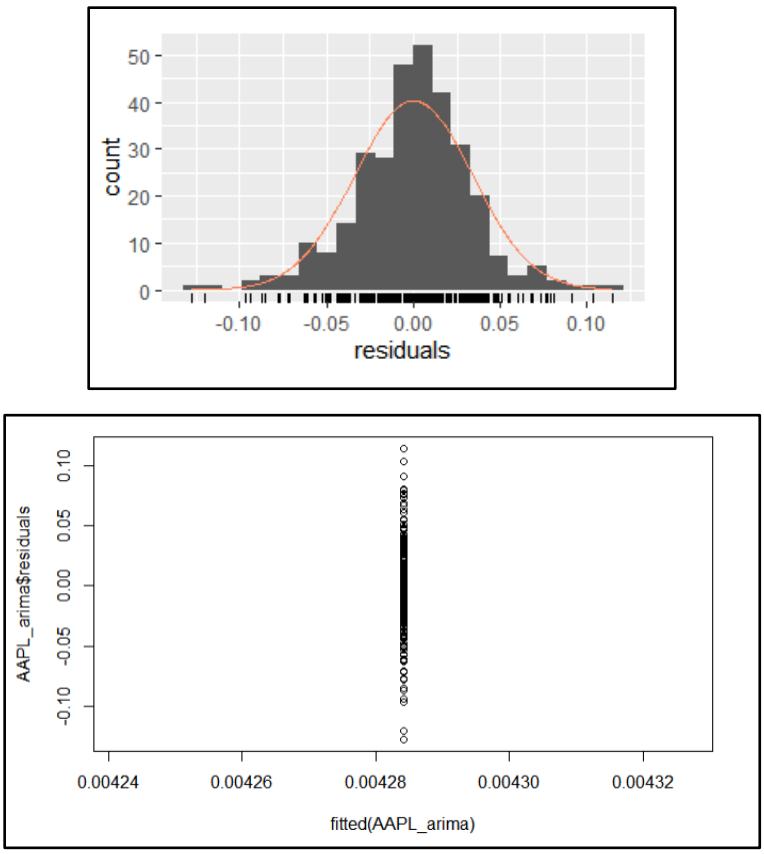


Fig. 9 – Residual plot

Step 8: Forecast

Note: For comparison, Actual values are differenced log of closing prices.

Week	fit	Lo95	Hi95	Actual	Within Range
313	0.004284	-0.06206	0.070626	0.046779	Y
314	0.004284	-0.06206	0.070626	0.012364	Y
315	0.004284	-0.06206	0.070626	0.003532	Y
316	0.004284	-0.06206	0.070626	0.003645	Y
317	0.004284	-0.06206	0.070626	0.002381	Y
318	0.004284	-0.06206	0.070626	-0.00191	Y
319	**0.004284**	**-0.06206**	**0.070626**	**-0.10195**	**N**
320	0.004284	-0.06206	0.070626	0.004295	Y
321	0.004284	-0.06206	0.070626	-0.01385	Y
322	**0.004284**	**-0.06206**	**0.070626**	**-0.12085**	**N**
323	**0.004284**	**-0.06206**	**0.070626**	**-0.12085**	**N**
324	0.004284	-0.06206	0.070626	-0.02393	Y
325	0.004284	-0.06206	0.070626	0.029573	Y
326	0.004284	-0.06206	0.070626	0.020012	Y
327	**0.004284**	**-0.06206**	**0.070626**	**0.101169**	**N**
328	**0.004284**	**-0.06206**	**0.070626**	**-0.06729**	**N**
329	0.004284	-0.06206	0.070626	0.053681	Y

12 out of 17 observations are within 95% PI. Hence, we can conclude that model does well in predicting future values.

Summary:

The Apple stock prices show upward trend and have seasonal components. The logarithmic of prices is used to build ARIMA model. The first order differencing is used to make data stationary, as ADF test indicated non-stationarity. Based on ACF and PACF model analysis ARIMA (0,0,0) model seem best for model fitting. The negligible sigma value and dignostic & residual plots prooved that ARIMA (0,0,0) model is adequate. Finally, 70,5% forecasted values were in 95% prediction interval, which indicates that ARIMA(0,0,0) was best fit.

MSFT

Step 1: Load Data

The data is gathered similarly as done before and for same time period.

Step 2: Time series Plot

Time series plot gives picture about how stock price has changed over study period. It is also helpful in identifying any trend, seasonality in the data.

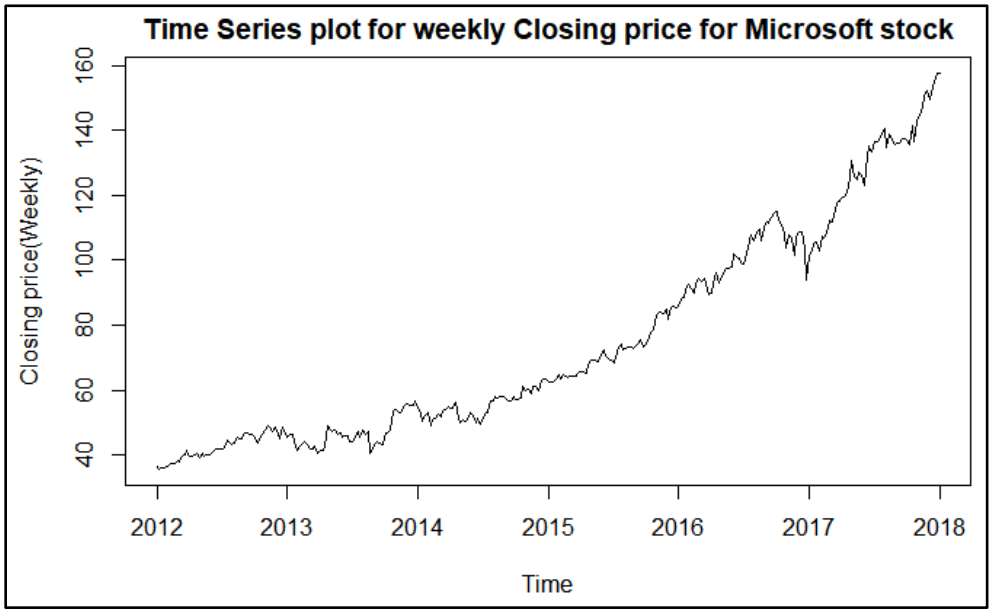


Fig. 10 – Time series Plot for Microsoft stock

Looking at time series plot, we can see that stock price has increased over time. The stock prices have spiked in last 2 years.

Step 3: Decomposing data

To further analyze trends in the data, we can decompose the data into seasonal, and trend components. Decomposition is the foundation for building the ARIMA model by providing insight towards the changes in stock fluctuations.

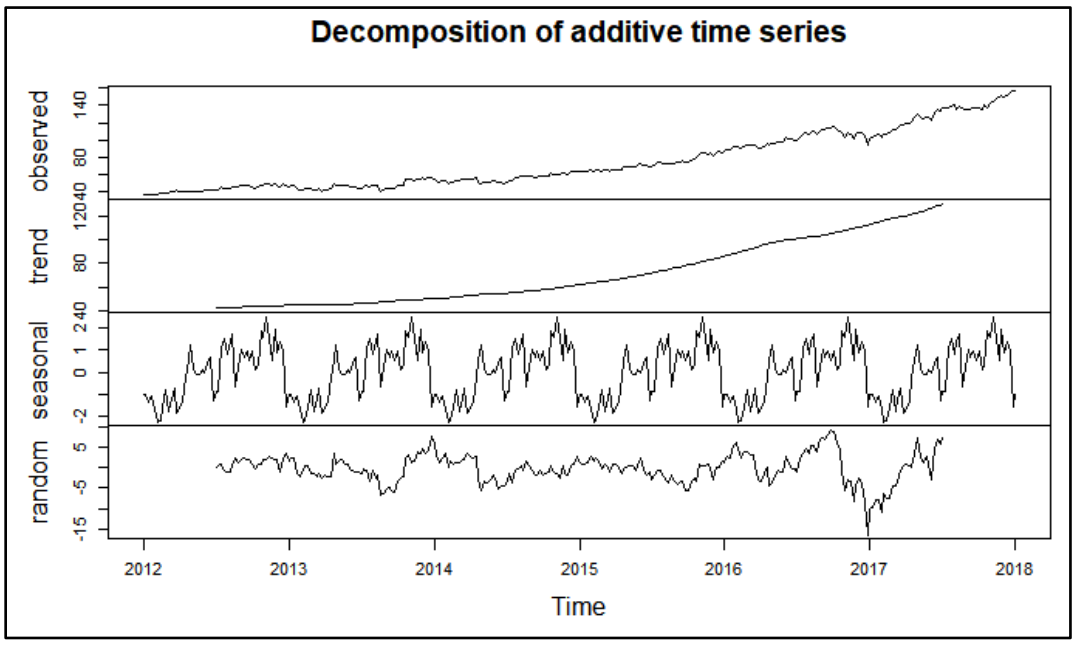
- The observed data section shows time series plot same as in Step 2.
- The trend component describes the overall pattern of the series over the entire range of time, considering increases and decreases in prices together. From the plot below, the trend is overall increasing.
- The seasonal component describes the fluctuations in stock price based on the calendar year. The stock prices have experienced many fluctuations over study period. There seems to be cyclic order in fluctuations.
- The random (residual) error, or noise section describes the trends that cannot be explained by trend or seasonal components. Statistically, these errors are the difference between the observed price and estimated price. The random noise plot shows that there is more fluctuations since 2017, which indicates that there are more statistical error since then.
- 

Fig. 11 – Decomposing data into components

Step 4: Smoothing data

As seen from decomposition plots data has variance and in order to stabilize the data smoothing techniques are implied.

In this project three types of smoothing techniques are used to stabilize the data.

1. Logarithmic returns

The main reason to use logarithmic returns is that the fluctuations observed in stock prices can be better compared over time and help in describing trends. It returns smoothed curve with reduced variation hence more accurate forecast model can be build.

The Logarithmic plot below is smoother than time series plot from step 2. The curve exactly shows trend in dataset by reducing fluctuations.

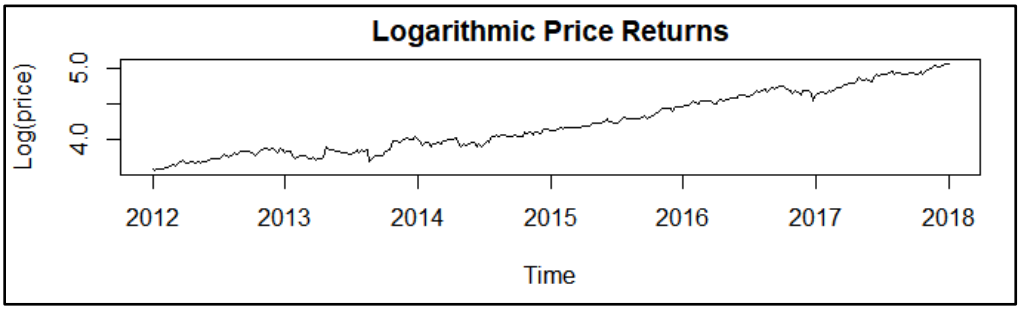


Fig. 13 – Logarithmic price returns plot

2. Square root values

Square root values instead of raw prices are used to scale the volatility between points to manage the time horizon of the stock. This is especially important because the longer a position is held, the greater a potential loss can be found. Again data seems smother than that of in step 2.

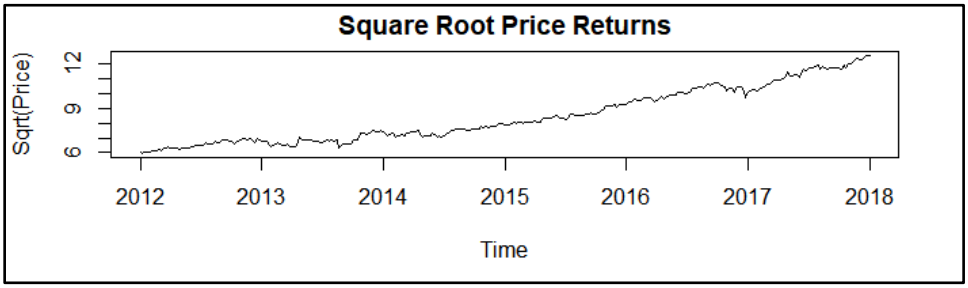


Fig. 14 – Square price returns plot

3. Differencing

Differencing is used to make data stationary. This is an important step in ARIMA model, as making data stationary allows to make assumption that prediction trend will be same in the future as it was in the past.

The stock price data for Microsoft requires first order differencing.

Both differenced plots below show that data have become stationary after first differencing.

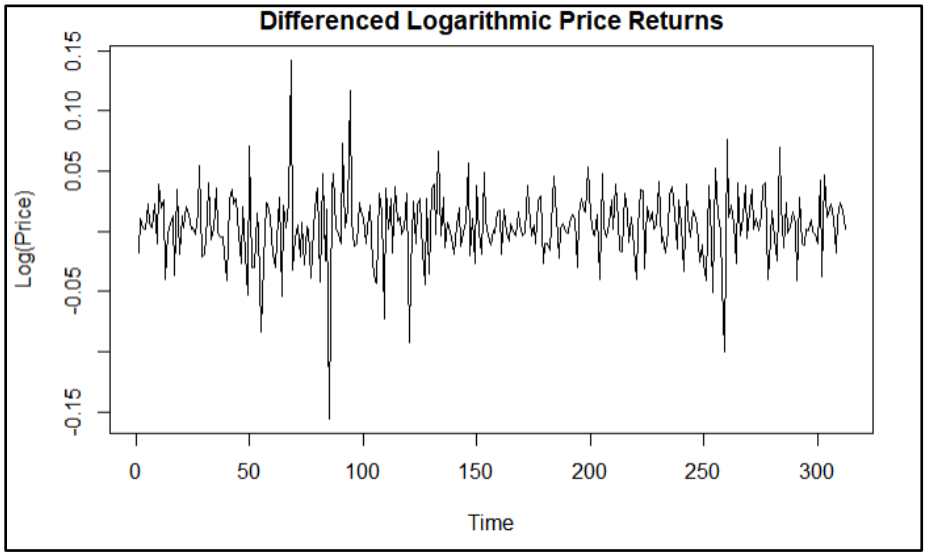


Fig. 15 – Differenced Logarithmic price returns plot

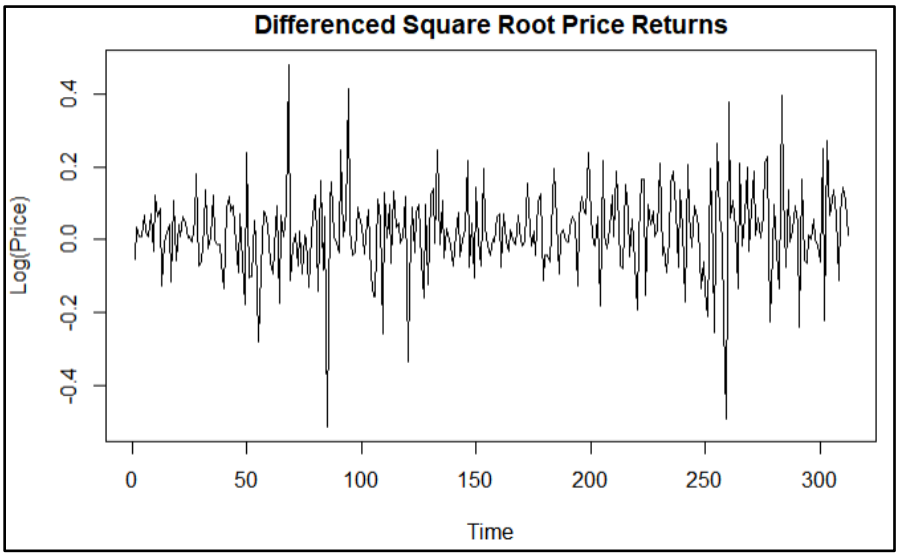


Fig. 16 – Differenced Square price returns plot

Step 5: ADF Test

The Augmented Dickey Fuller Test is used to determine if data is stationary. It calculates p-value that determines which hypothesis to choose.

Hypothesis are as follows,

Null Hypothesis (H0) (p-value>0.05): The time series data is non-stationary.

Alternate hypothesis (Ha) (p-value<0.05: The time series in stationary.

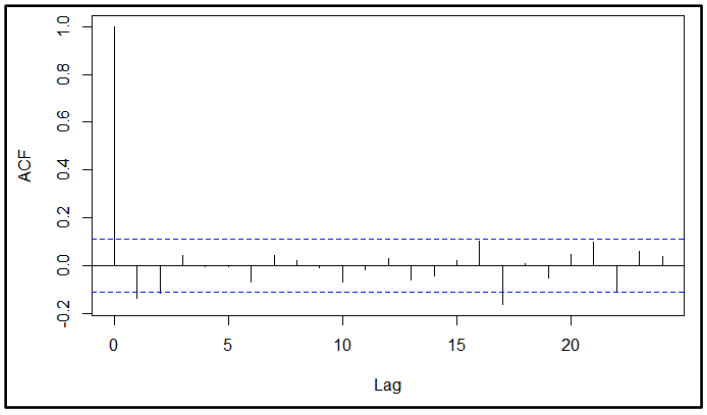
p-value for first two statistics is above 0.05, which indicates that null hypothesis cannot be rejected.

So, logarithmic return and square root return data is non-stationary.

Whereas, p-value for above two return prices after differencing is below 0.05, hence null hypothesis can be rejected. This indicates that first order differencing has made data stationary.

```
        Augmented Dickey-Fuller Test

data:  logprice_m[, 2]
Dickey-Fuller = -1.9695, Lag order = 6, p-value = 0.5893
alternative hypothesis: stationary


        Augmented Dickey-Fuller Test

data:  sqrtprice_m[, 2]
Dickey-Fuller = -1.0397, Lag order = 6, p-value = 0.9311
alternative hypothesis: stationary

p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  dlogprice_m
Dickey-Fuller = -7.2897, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

p-value smaller than printed p-value
        Augmented Dickey-Fuller Test

data:  dsqrtprice_m
Dickey-Fuller = -7.0411, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Step 6: ACF & PACF plot



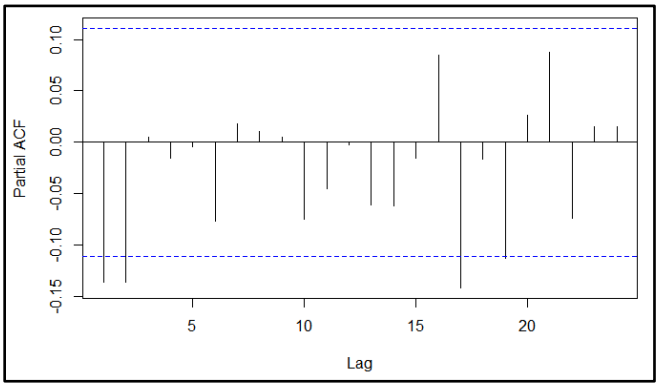


Fig. 17 – ACF & PACF plot

ARIMA models integrate two types of correlograms,

- The AutoCorrelation Function (ACF) displays the correlation between series and lags for the Moving Average (q) of the ARIMA model. Here, ACF plot shows that cutoff for strong correlation is after lag 2, hence q-notation is 2.
- The Partial AutoCorrelation Function (PACF) displays the correlation between returns and lags for the Auto-Regression (p) of the ARIMA model. Here, PACF plot shows that cutoff for strong correlation is after lag2, hence p-notation is 2.

Based on ACF & PACF plot analysis fitting ARIMA (2,0,2) model would yield best results.

Step 7: Fit ARIMA model

The model to fit onto the logarithmic price returns is ARIMA (2,0,2).

```
Call:
arima(x = dlogprice_m, order = c(2, 0, 2))

Coefficients:
          ar1     ar2     ma1      ma2  intercept
      -0.4608  0.1102  0.3085  -0.2970     0.0047
s.e.   0.3120  0.2827  0.3009   0.2739     0.0012

sigma^2 estimated as 0.0008204:  log likelihood = 665.76,  aic = -1319.52

Training set error measures:
                        ME       RMSE        MAE MPE MAPE      MASE          ACF1
Training set -1.682643e-05 0.02864179 0.02008919 NaN  Inf 0.6288257 -0.0003839116
```

Model adequacy can be checked using dignostic plots.

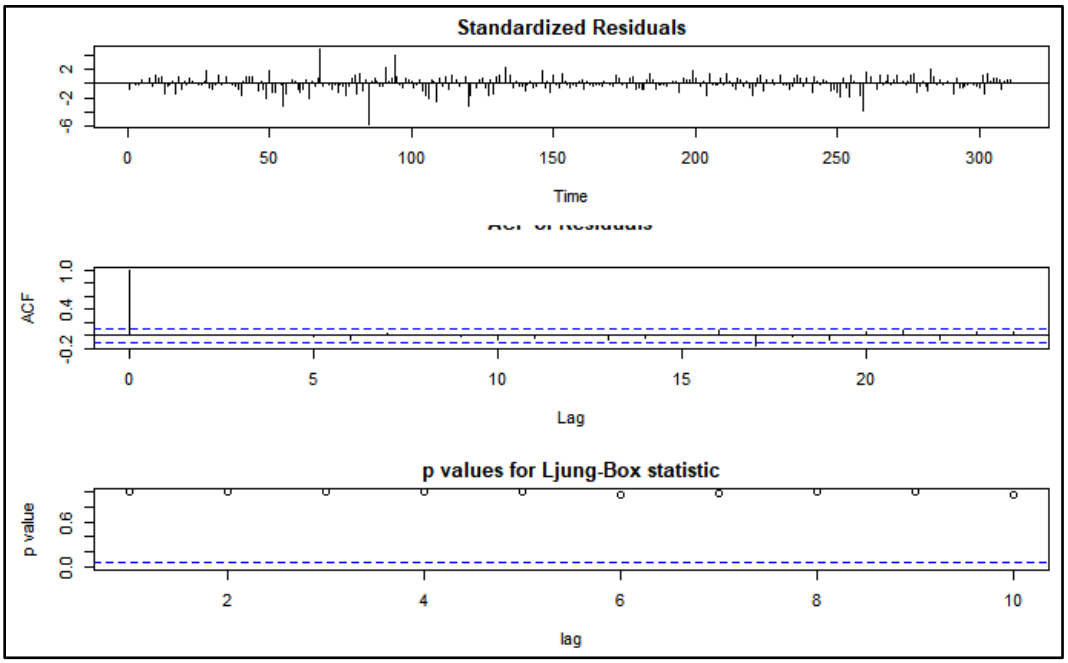


Fig. 18 – Diagnostic plot

Dignostic plots looks fine and represents adquacy of model. As ACF plot of residuals have no large lag, p-values are above 0.05. The histogram of residuals indicates that residuals are normally distributed. Normal distribution of residuals shows that model is adequate.

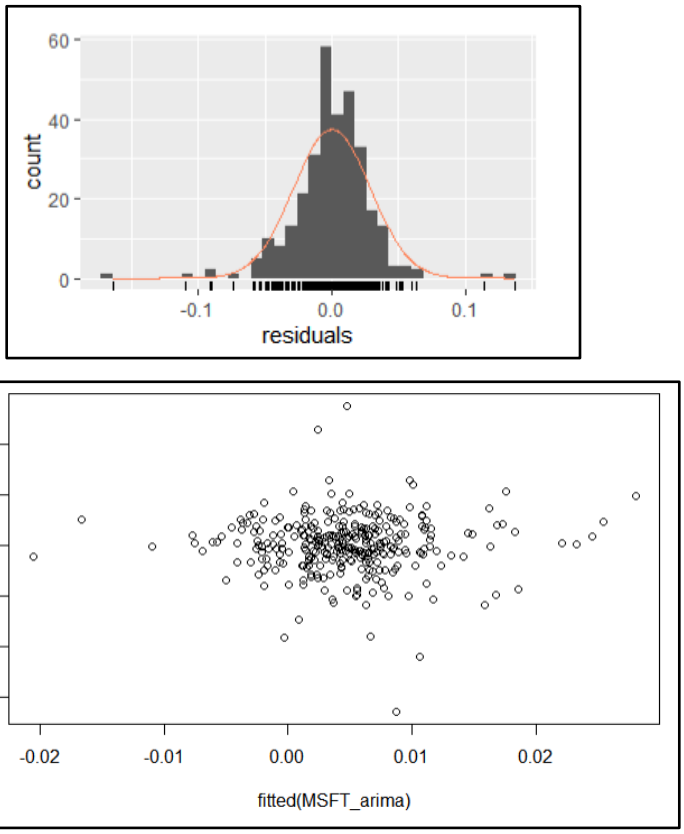


Fig. 19 – Residual plot

Step 8: Forecast

Week	fit	Lo95	Hi95	Actual	Within Range
313	0.002389	-0.05375	0.058526	-0.00076	Y
314	0.005305	-0.05148	0.062089	0.028465	Y
315	0.004175	-0.05298	0.061336	0.026597	Y
316	0.005017	-0.05218	0.062215	-0.00627	Y
317	**0.004505**	**-0.05272**	**0.061728**	**0.084894**	**N**
318	0.004834	-0.0524	0.062065	0.023701	Y
319	0.004626	-0.05261	0.061861	0.015014	Y
320	**0.004758**	**-0.05248**	**0.061994**	**-0.10796**	**N**
321	0.004674	-0.05256	0.061911	-0.02141	Y
322	0.004727	-0.05251	0.061965	-0.02206	Y
323	**0.004694**	**-0.05254**	**0.061931**	**-0.0934**	**N**
324	0.004715	-0.05252	0.061952	0.012004	Y
325	0.004701	-0.05254	0.061939	0.061251	Y
326	0.00471	-0.05253	0.061947	0.035994	Y
327	0.004705	-0.05253	0.061942	0.060578	Y
328	0.004708	-0.05253	0.061945	-0.03444	Y
329	0.004706	-0.05253	0.061943	0.036451	Y

14 out of 17 observations are within 95% PI. Hence, we can conclude that model does well in predicting future values.

Summary:

The Microsoft stock prices show upward trend and have seasonal components. The logarithmic of prices is used to build ARIMA model. The first order differencing is used to make data stationary, as ADF test indicated non-stationarity. Based on ACF and PACF model analysis ARIMA (2,0,2) model seem best for model fitting. The negligible sigma value and dignostic & residual plots prooved that ARIMA (2,0,2) model is adequate. Finally, 82.3% forecasted values were in 95% prediction interval, which indicates that ARIMA(2,0,2) was best fit.

4. Conclusion & Recommendations:

- Based on analysis of Apple and Microsoft stock prices, it can be concluded that prices may increase in near future. As more than half of the values are positive indicating positive trend.
- The model can not be used to real time investment decisions, but it gives user idea about how stocks will perform in near future.
- The forecast ability of model can be tested using forecasted returns and actual returns. But calculating returns and comparing them with actual returns is beyond scope of this course. Thus, method used in coursework is applied to check forecast ability.
- Same dataset can be tested using seasonal trend model and seasonal exponential model. Comparison can be made based on forecast ability to check better model.
- Above three models can be tested and optimised based on different prediction intervals.
-

References

1.) Introduction to Time Series Analysis and Forecasting, Montgomery, Jennings, and Kulahci, 2015, 2nd ed, Wiley
2.) ARIMA model in R https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/
3.) https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
4.) https://medium.com/@aaronyen/https-medium-com-aaronyen-arimaproject-ab892486dc84