TIME SERIES ARMA PROJECT

Contents

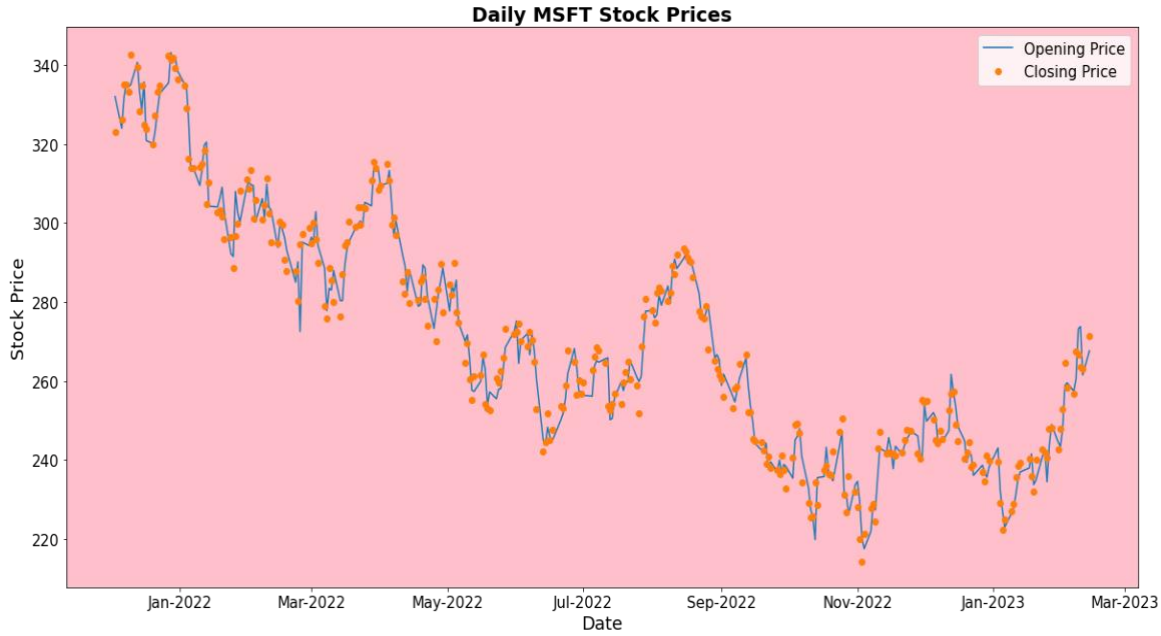
A. Time Series

1. Download Microsoft (MSFT) stock price data for T= 300 business days.

The initial code for this project (until personalised ARMA models) has been designed to be user-friendly, with user-defined functions that allow easy analysis of potentially all stocks from the yfinance library. By simply changing the stock_name variable, users can automatically update the resulting plots and tables to personalize their analyses for any other stock recorded in yfinance.

For this coursework, we will focus on analysing 300 business days of Microsoft (MSFT) stock from the yfinance library from 3rd December 2021 to 14th February 2023.

2. Plot the price time series.



Graph 1: Daily Opening and Closing Prices of Stock MSFT over 300 days

Over the course of nearly 14 months, as shown in Graph 1, the MSFT stock prices display an overall decreasing trend, dropping from approximately $340 in January 2022 to reach an annual record low of around $215 in January 2023. This decline was followed by a subsequent upward price movement, with prices rising until the mid of February 2023 and recouping roughly half of the annual losses.

B. Moving Average

3. Define mathematically the moving average of the price time series with an arbitrary time window t.

Moving average is a widely employed approach in the financial industry to smoothen the daily price fluctuations to have a better long-term view of the stock prices. The mathematical formulation for calculating the moving average of a price time series with an arbitrary window size of t, can be expressed as:

$$\mathrm{MA}_t = \frac{1}{t}\Sigma^{t}{}_{i=0}\,\mathrm{P}_i$$

where MA_t is the moving average at time t and P_i is the stock price at time i. The (1/t) factor normalises or averages the summation of stock prices from 0 to time t.

For instance, we are supposed to find a moving average at time t=4 and if we want to average using close neighbouring values, we can use a 3-point moving average, which is:

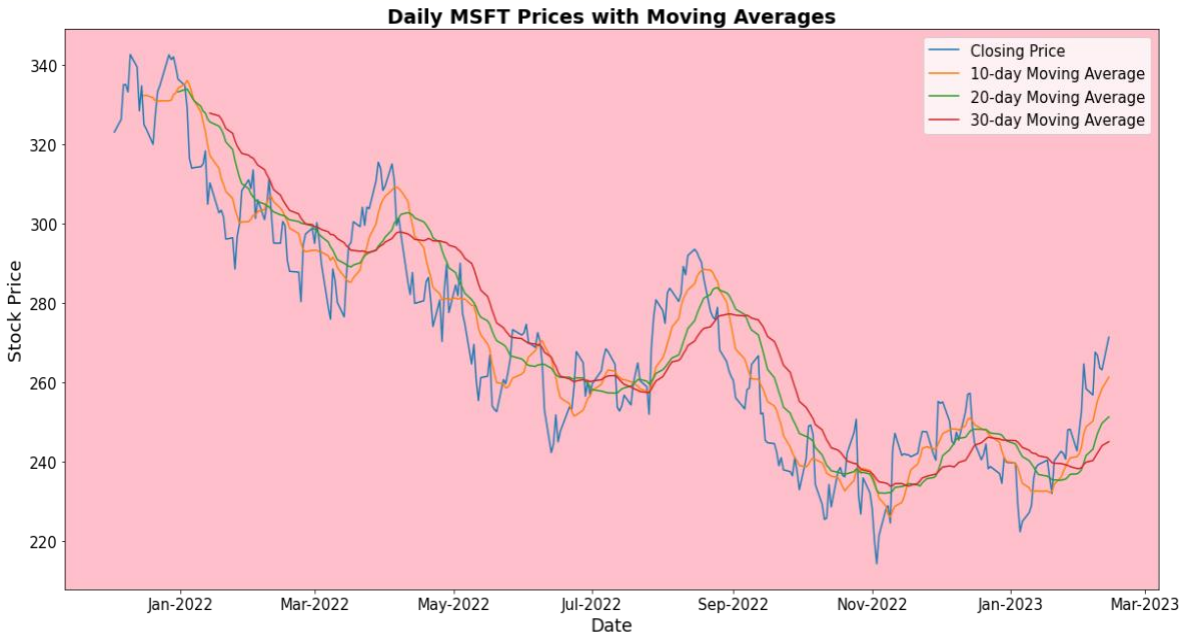
$$\mathrm{MA}_4 = \frac{1}{3}\{\mathrm{P}_2 + P_3 + P_4\}$$

This gives us the average of the previous 3 prices (P_2, P_3, P_4), which is used to smooth the time series and remove noise or short-term fluctuations. The moving average can be calculated for different time windows to adjust the level of smoothing and capture different patterns in the time series.

4. Compute three moving averages of the price time series, with time windows t = 10, 20, 30.

Used the `rolling(window).mean()` function in the code to generate the short, mid and long-term moving averages at time windows t = 10,20,30.

5. Plot the moving averages against the price time series.



Graph 2: MSFT Closing Stock Prices against T=10,20 & 30 Moving Averages

Based on the plot shown in Graph 2, it can be observed that during the months of Jan-Mar and Sept-Nov of 2022, the moving averages, particularly the 30-day MA, were consistently higher than or equal to the stock price. This is evidenced by the steep decrease in stock prices during these brief periods, as compared to the longer durations covered by the latter moving averages. Moreover, during the peak periods in April and August 2022, the 10-day MA takes precedence over the mid and long-term averages, with the 10-day MA moving closer to the stock price. This suggests higher volatility of the stock during these periods in the financial market.

6. Compute the linear and log-return of the price time series.

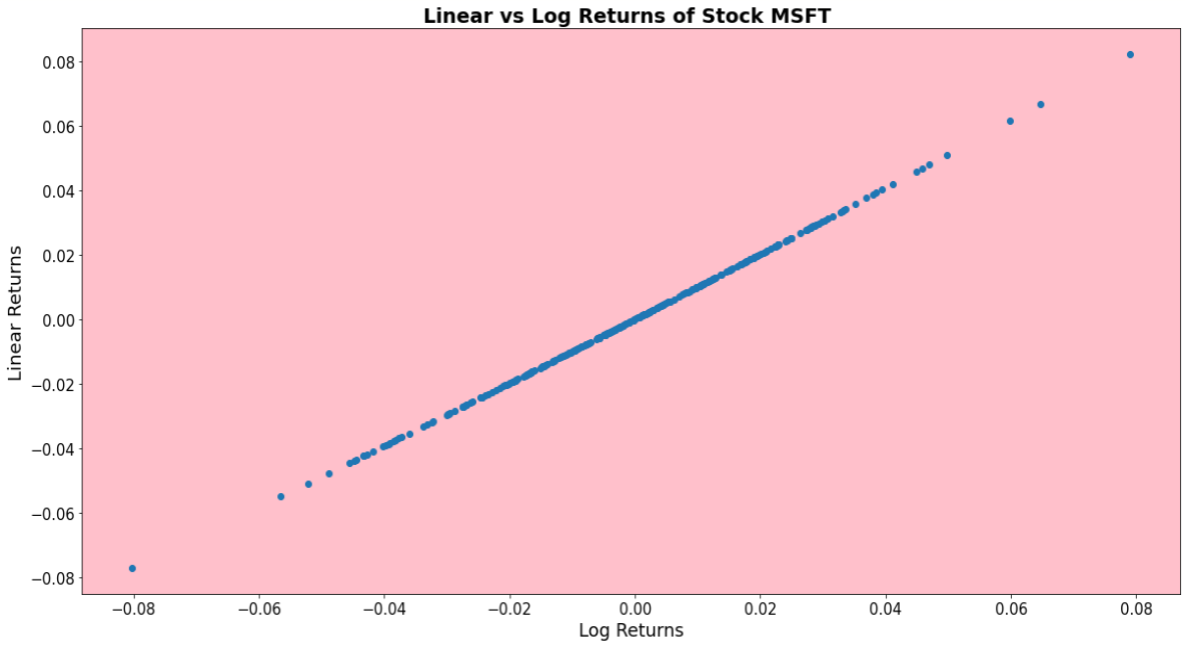
We based our code on these mathematical equations of linear and log-returns of stock prices:

$$Linear\ Return\ R_T = P_r/P_{r-1} - 1$$

$$Log\ Return\ X_T = \log\left(P_r/P_{r-1}\right)$$

where P_r is the closing stock price at time T.

7. Plot the linear return against the log-return time series.



Graph 3: Linear Returns Against Log Returns of Stock MSFT

Analysis of Graph 3 reveals that linear returns and log returns move in the same direction, which suggests a positive correlation between the two. This correlation may also suggest that the stock's volatility is relatively consistent over time.

C. Time Series Analysis

8. Define the auto-correlation function (for a stationary time-series).

The autocorrelation function (ACF) of a stationary time series measures the correlation between the series at different lags, or time lags. It is a statistical tool that allows us to analyse and model the temporal structure of a time series. The autocorrelation function (ACF) of a stationary time series $\{P_t\}$ at lag h can be written as:

$$\rho(h) = \frac{\gamma(t+h,t)}{sqrt\left(\gamma(t+h,t+h)\gamma(t,t)\right)} = \frac{\gamma(h)}{\gamma(0)} = Corr(P_{t+h}, P_t)$$

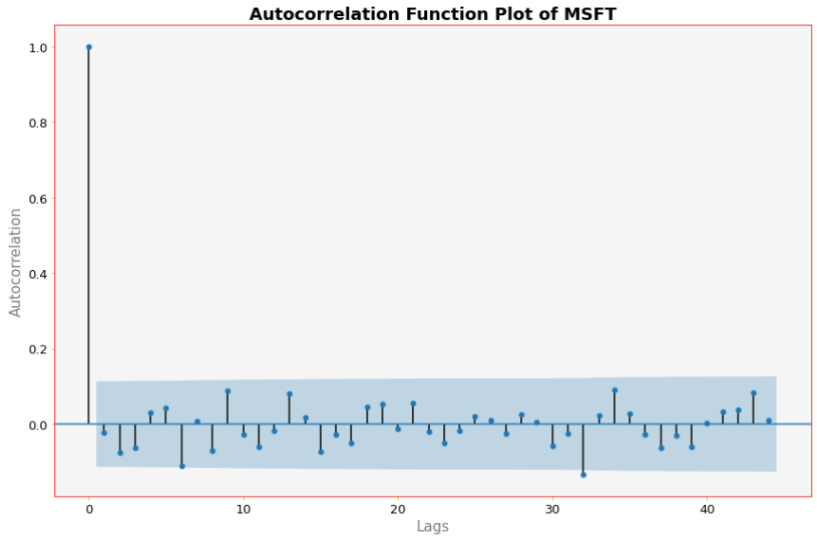
where ρ(h) is the ACF of our time series, $\gamma(h)$ or $\gamma(t+h,t)$ is the autocovariance function (ACVF) of $\{P_t\}$ at lag h and $Corr(P_{t+h}, P_t)$ is the correlation between P_{t+h} and P_t. If a time series is weakly stationary,

the ACF depends only on the time lag h and not on the current time t. The ACF not only benefits in finding the presence of seasonality or trend in a time series, but also provides support in finding the best time series model for our data, something which we will be exploring in the next section.

Note: As our original price time series was non-stationary, we used differencing in order to make it a stationary price time series, which is required for accurate ACF, PACF values for model predictions.

9. Compute the auto-correlation function (ACF) of the price time series.

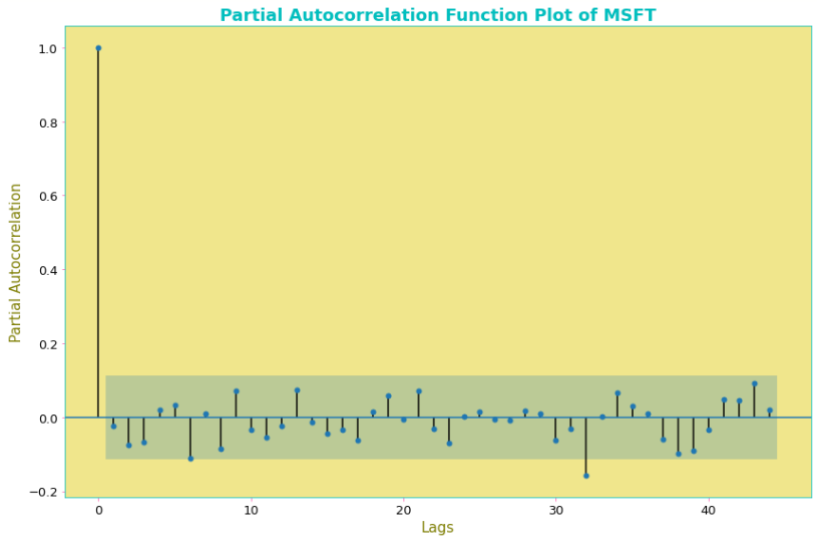
10. Plot the price ACF.



Graph 4: ACF Plot of MSFT Stock Prices

11. Compute the partial auto-correlation function (PACF) of the price time series.

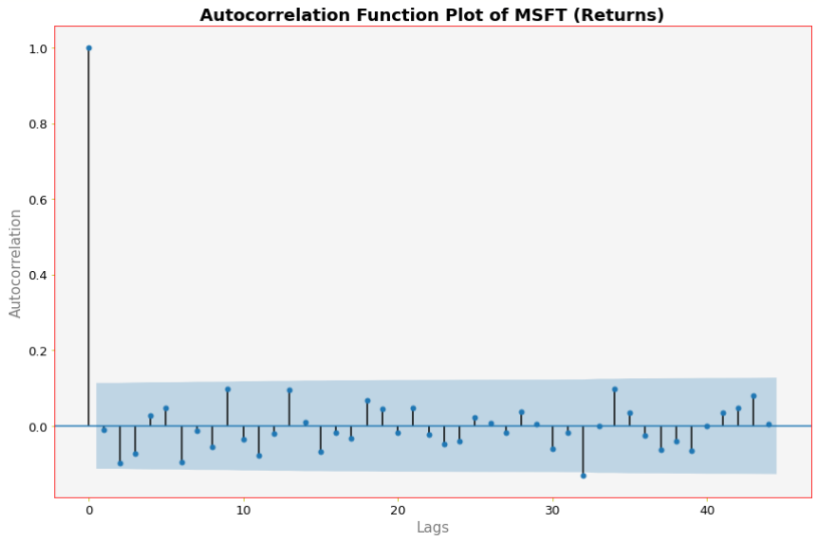
12. Plot the price PACF.



Graph 5: PACF Plot of MSFT Stock Prices

13. Compute the auto-correlation function (ACF) of the return time series.

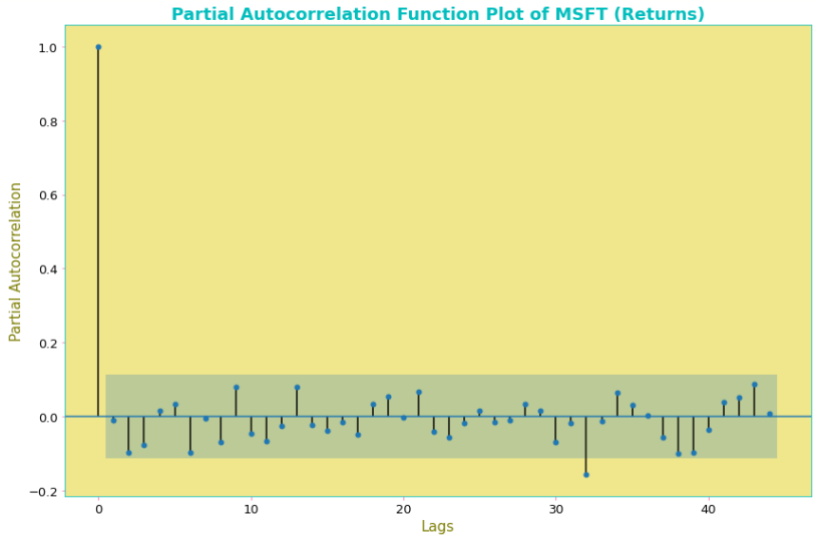
14. Plot the return ACF.



Graph 6: ACF Plot of MSFT Stock Returns

15. Compute the partial auto-correlation function (PACF) of the return time series.

16. Plot the return PACF.



Graph 7: ACF Plot of MSFT Stock Returns

ANALYSIS: The autocorrelation function (ACF) and the partial autocorrelation function (PACF) of MSFT stock prices show a unique behaviour through their values but especially through Graphs 4 and 5. Both the ACF and PACF plots seem to have a cutoff at lag 0, which indicates the significance of the randomness of the MSFT stock prices. While it maybe also influenced from making our price time series stationary, it is very unlikely to have a real-life stock to follow an ARMA(0,0) model or white noise process. We will be exploring the best ARMA models in the next section.

Through the Graphs 6 and 7 of the ACF and PACF plots of log returns, we find ourselves with a similar observation which is also based upon the MSFT prices being significantly hard to predict because of its randomness. Further, log returns stabilize variance of time series and transform values towards being more normally distributed than its original stock prices.

D. ARMA Models

17. Define mathematically an ARMA(p, q) model.

$A\ time\ series\ \{P_t;\ t = 0, \pm1, \pm2, \ldots.\}\ is\ ARMA(p,q)\ if\ it\ is\ stationary\ and$

$$P_t = \phi_1 P_{t-1} + \cdots + \phi_p P_{t-p} + \omega_t + \theta_1\, \omega_{t-1} + \cdots + \theta_q\, \omega_{t-q}$$

$with\ \phi_p \neq 0, \theta_q \neq 0\ and\ \sigma_w^2 > 0.$

$The\ moving\ average\ and\ autogressive\ orders\ are\ our\ parameters\ q\ and\ p, respecively.\ \{P_t\}\ as\ we\ know\ from\ earlier\ is\ our\ time\ series\ and\ \{\phi_t\}\ and\ \{\theta_t\}\ are\ our\ other\ AR\ and\ MA\ parameters, respectively.\ \omega_t \sim wn(0, \sigma_w^2)\ are\ common\ parameters\ between\ both\ ARMA\ components.$

$If\ P_t\ has\ a\ non-zero\ mean\ \mu, we\ set\ \alpha = \mu(1 - \phi_1 - \cdots - \phi_p)\ and\ write\ the\ model\ as$

$$P_t = \alpha + \phi_1 P_{t-1} + \cdots + \phi_p\, P_{t-p} + \omega_t + \theta_1 \omega_{t-1} + \cdots + \theta_q\, \omega_{t-q}$$

$We\ can\ also\ write\ our\ ARMA(p,q)\ model\ in\ a\ concise\ form\ as:$

$$\phi(B) P_t = \theta(B) \omega_t$$

$Autoregressive\ Moving\ Average\ (ARMA)\ model\ is\ one\ of\ the\ fundamental\ time\ series\ model\ that\ combines\ two\ components:\ an\ autoregressive\ (AR) component\ and\ a\ moving\ average\ (MA) component.\ The\ AR\ component\ models\ the\ dependency\ of\ the\ current\ value\ of\ the\ time\ series\ on\ its\ past\ values,\ while\ the\ MA\ component\ models\ the\ dependency\ on\ past\ error\ terms.\ When\ q\ equals\ 0, the\ model\ is\ called\ an\ autoregressive\ model\ of\ order\ p, AR(p), and\ when\ p\ equals\ 0, the\ model\ is\ called\ a\ moving\ average\ model\ of\ order\ q, MA(q).$

NOTE: We will utilize the stationary price time series again, as we did in the previous section for computing and plotting our autocorrelation function (ACF) and partial autocorrelation function (PACF) values.

18. Define a training and test set and fit an ARMA model to the price time series.

Typically, a training set that covers 70-80% of the sample data is considered optimal. For our analysis, we utilized 230 out of 300 data points, representing approximately 76.6%, for the training set, while the remaining 70 data points were allocated to the test set. Following this, we identified the ARMA(1,1) model to fit our price time series data.

Dep. Variable:	first_diff	**No. Observations:**	230
Model:	ARMA(1, 1)	**Log Likelihood**	-733.781
Method:	css-mle	**S.D. of innovations**	5.849
Date:	Sat, 18 Feb 2023	**AIC**	1475.563
Time:	22:45:29	**BIC**	1489.315
Sample:	0	**HQIC**	1481.110

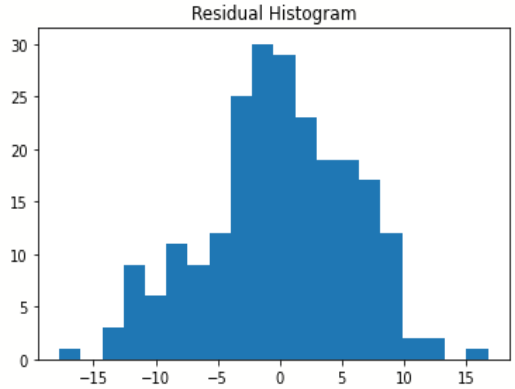
	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.3960	0.060	-6.597	0.000	-0.514	-0.278
ar.L1.first_diff	0.9157	0.028	32.524	0.000	0.861	0.971
ma.L1.first_diff	-1.0000	0.012	-82.259	0.000	-1.024	-0.976

Roots

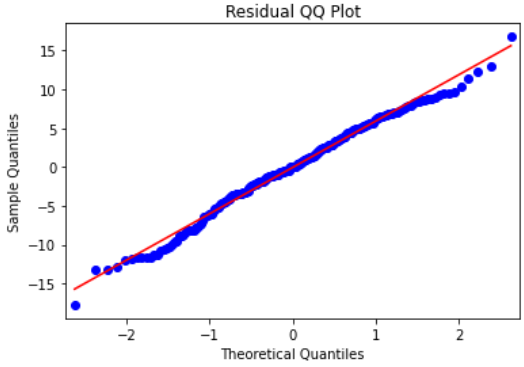
	Real	Imaginary	Modulus	Frequency
AR.1	1.0921	+0.0000j	1.0921	0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000

Table 1: ARMA(1,1) Model Results For Our Price Time Series

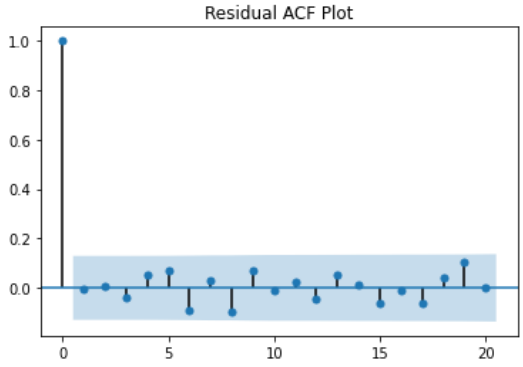
Table 1 displays the AIC, BIC, and HQIC values for our ARMA(1,1) model, which are (1475.563, 1489.315,1481.110) respectively. Given our considerably large sample size, greater emphasis should be placed on the AIC and BIC values, despite the HQIC of our chosen model being the lowest among all the models considered. Notably, the ARMA(1,2) and ARMA(2,1) models yield identical information criterion values, which are (1477.551,1494.741,1484.485), and both exceed the values of our chosen model. Moreover, the ARMA(2,2) model produces inferior AIC and BIC values of 1476.325 and 1496.953, respectively. For confirming the selection of ARMA(1,1) model, we also look at the other model evaluation techniques such as the residual histogram, Q-Q plot and ACF plot, which are below:



Graph 8: Residual Histogram for ARMA(1,1) Model



Graph 9: Residual Q-Q plot for ARMA(1,1) Model



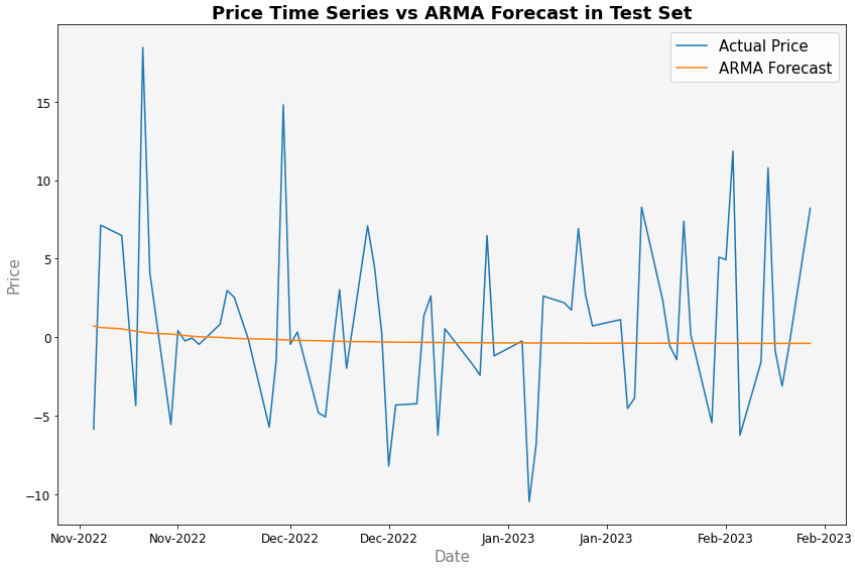
Graph 10: Residual ACF plot for ARMA(1,1) Model

Based on our model selection criteria, we choose the ARMA(1,1) model to fit our price time series. This decision is supported by our diagnostic plots, including the histogram residual plot shown in Graph 8, which reveals a bell-shaped normal-like distribution of the residuals. The Q-Q plot in Graph 9 shows that the residual points lie on or near the red line, indicating a good fit to a normal distribution. Finally, Graph 10 shows that the residual values are uncorrelated, as indicated by the cutoff at lag 0. Overall, these diagnostic plots provide evidence that the ARMA(1,1) model is an appropriate choice for our price time series analysis.

19. Display the parameters of the model and its Mean Squared Error (MSE) in the training set and in the test set.

The parameter values for the constant, autoregressive (AR), and moving average (MA) components are presented in Table 1, with values of -0.3960, 0.9157, and -1.000, respectively. Additionally, the mean squared error (MSE) values for the training and test sets are 34.5733 and 29.0904, respectively.

20. Plot the price time series vs the ARMA forecast in the test set.



Graph 11: MSFT Stock Prices Against the ARMA Forecast

From Graph 11, we can observe that our ARMA(1,1) forecast is a slightly curved line that captures the trend of the stock price time series, highlighting the significant volatility in the data. The forecast indicates

that the stock price is expected to continue its downward trend in the short term, with some slight fluctuations.

21. Fit an ARMA model to the return time series.

Similar to the price time series, we choose an ARMA(1,1) model for our log returns data. This is supported by the evaluation statistics displayed in Table 2 below where the AIC, BIC, and HQIC values for our ARMA(1,1) model are (-1106.574, -1092.822 and -1101.027) respectively. This is less as compared to the identical information criterion values of ARMA(1,2) and ARMA(2,1) model which are (-1104.578, -1087.388 and -1097.644) and ARMA(2,2) model which are (-1105.714, -1085.085 and -1097.393).

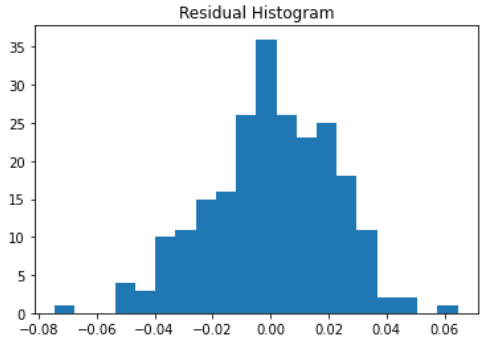
Dep. Variable:	Log Return	**No. Observations:**	230
Model:	ARMA(1, 1)	**Log Likelihood**	557.287
Method:	css-mle	**S.D. of innovations**	0.021
Date:	Sun, 19 Feb 2023	**AIC**	-1106.574
Time:	09:37:27	**BIC**	-1092.822
Sample:	0	**HQIC**	-1101.027

	coef	std err	z	P>|z|	[0.025	0.975]
const	-0.0014	0.000	-6.652	0.000	-0.002	-0.001
ar.L1.Log Return	0.9138	0.029	31.693	0.000	0.857	0.970
ma.L1.Log Return	-1.0000	0.012	-82.888	0.000	-1.024	-0.976

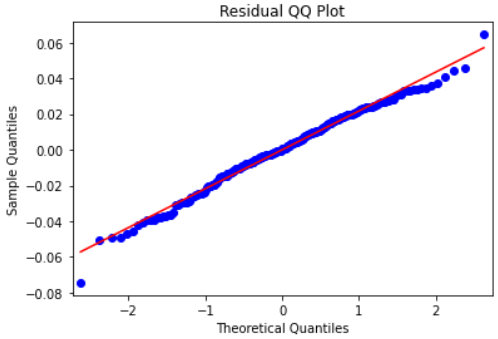
Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.0943	+0.0000j	1.0943	0.0000
MA.1	1.0000	+0.0000j	1.0000	0.0000

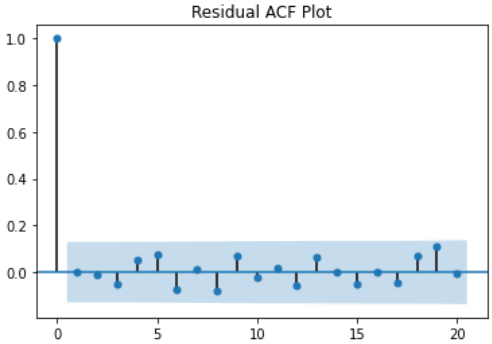
Table 2: ARMA(1,1) Model Results For Our Log Return Time Series



Graph 12: Residual Histogram for ARMA(1,1) Model



Graph 13: Residual Q-Q plot for ARMA(1,1) Model



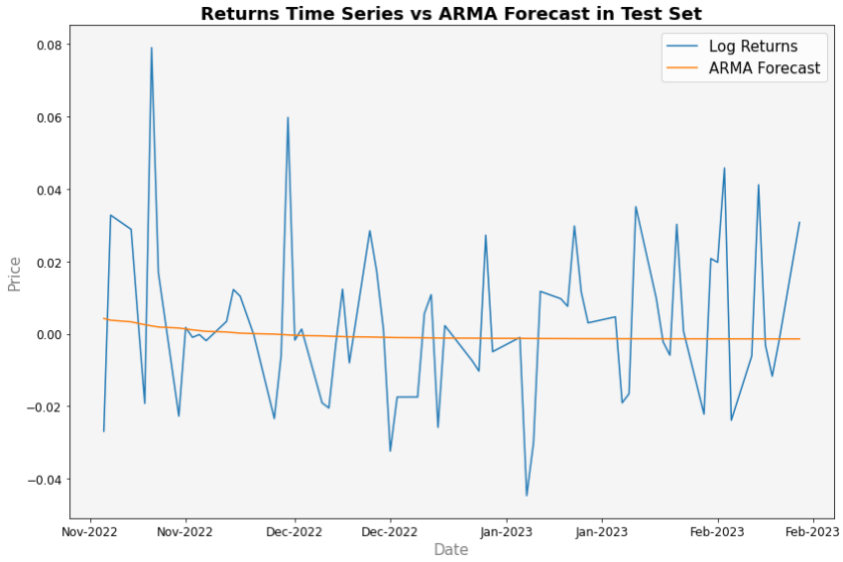
Graph 14: Residual ACF plot for ARMA(1,1) Model

Upon analysing the model selection using the diagnostic plots above, we observe that the residual histogram in Graph 12 may appear to be less normally distributed than the price time series we analysed earlier. However, the Q-Q plot in Graph 13 indicates that more tails lie on the line, which suggests a better fit of the residuals to the normal distribution. Overall, these diagnostic plots provide compelling evidence that the ARMA(1,1) model is a suitable choice for our log return time series analysis.

22. Display the parameters of the model and its Mean Squared Error (MSE) in the training set and in the test set.

The parameter values for the constant, autoregressive (AR), and moving average (MA) components are presented in Table 2, with values of -0.0014, 0.9138, and -1.0000, respectively. Additionally, the mean squared error (MSE) values for the training and test sets are both equal to 0.0005.

23. Plot the return time series vs the ARMA forecast in the test set.



Graph 15: MSFT Log Returns Against the ARMA Forecast

Graph 15 displays a plot that is nearly identical to the one in Graph 11, as both depict changes in price over time, albeit in different forms. This similarity is due to the close match in magnitudes between linear and log returns, which was observed in Graph 3. The trend differencing we conducted on our price time series

reinforces this correspondence. Consequently, the ARMA forecast line in Graph 15 exhibits a similar downward trend to the one in Graph 11, indicating a consistent pattern in the stock's behaviour over time.

E. Gaussianity and Stationarity Test

24. Introduce mathematically a Gaussianity test.

A Gaussianity test, or better known as Normality test, as the name describes is a method to determine whether if a set of data follow a normal or Gaussian distribution. In order to understand Gaussianity tests, it is vital to first understand normal distribution fundamentally. A normal distribution is characterized by two parameters: the mean, denoted by mu (μ), and the standard deviation, denoted by sigma (σ). The probability density function (PDF) of a normal distribution is given by:

$$f(x) = \frac{1}{sqrt(2\pi\sigma^2)} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where x is our random variable.

One of the widely used normality tests is Shapiro-Wilk test, which is based on the comparison of the data to the expected values of a normal distribution. The test statistic is calculated as follows:

$$W = \frac{\left(\Sigma\left(a_i * x_{(i)}\right)^2\right)}{\Sigma(x_i - \bar{x})^2}$$

where $\bar{x}$ is the sample mean, x_i are the data points, $x_{(i)}$ i.e. x_i with parentheses is the i^{th} order statistic and a_i are constants derived from the sample size and the estimated mean and variance of the data. The test statistic W is then compared to critical values derived from tables or calculated using Monte Carlo simulations.

25. Perform a Gaussianity test of the return time series.

Confirming the normality of a distribution through multiple tests is generally considered good practice. So, we performed two tests: the Shapiro-Wilk Test and the Jarque-Bera Test. The p-values obtained from the Shapiro-Wilk Test and Jarque-Bera Test were 0.34 and 0.062, respectively. Thus, both the p-values are higher than the significance level 0.05, rejecting the null hypothesis that the distribution is not normally distributed. Thus, it is reasonable to say that the return time series follow Gaussianity.

26. Introduce mathematically a stationarity test.

Stationarity Tests are methods to check whether a time series data is stationary or not. A stationary time series display unique properties of a constant mean, variance, and autocovariance over time. It is important to make a time series stationary when trying to fit models as the opposite could lead to model parameters being biased or incorrect. When we usually use the term stationary, it usually refers to mean weakly stationary. So, we say a time series $\{P_t\}$ is weakly stationary if

(i) $the\ mean\ value\ function\ of\ \{P_t\},\ \mu(t)\ is\ independent\ of\ t,$
(ii) $the\ covariance\ function\ of\ \{P_t\}\ with\ lag\ h,\ \gamma(t+h,t)\ is\ independent\ of\ t\ for\ each\ h$

The Augmented Dickey-Fuller (ADF) test, which determines if a time series is stationary by looking at how the series' first differences behave, is a popular stationarity test. We calculate the ADF test statistic through:

$$\begin{aligned} \Delta P_t &= P_t - P_{\{t-1\}} \\ \Delta P_{t-1} &= P_{\{t-1\}} - P_{\{t-2\}} \\ &\dots \\ \Delta P_2 &= P_2 - P_1 \\ \Delta P_1 &= P_1 - P_0 \end{aligned}$$

where P_t is the time series, and ΔX_t is the first difference of X_t. The null hypothesis of the ADF test is that the time series is non-stationary, and the alternative hypothesis is that the time series is stationary. The ADF test statistic is then compared to critical values derived from tables or calculated using Monte Carlo simulations.

27. Perform a stationarity test of the return time series.

To determine whether our log returns are stationary, we performed two tests: the Augmented Dickey-Fuller (ADF) test and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test. The ADF Statistic we obtained was -11.403, which is less than the 1% critical value of -3.453 and significant at the 0.05 level. This result indicates that we can reject the null hypothesis of the ADF test, which states that the time series is non-stationary.
Further, our KPSS test yielded a p-value greater than 0.1, which is above our significance level of 0.05. As a result, we failed to reject the null hypothesis and cannot conclude that the time series is non-stationary. Based on the results of both tests, we can conclude that our log returns follow stationarity.

References:

1. Robert H. Shumway, David S. Stoffer (2010): Time Series Analysis and Its Applications: With R Examples. Springer
2. Brockwell, P. J., & Davis, R. A. (2016). Introduction to time series and forecasting. Springer.