Algorithmic Trading - Coursework 1

Eirik A Baekkelund

1 Introduction

In 2022, the stock market faced significant challenges due to various factors, including heightened interest rates, increased inflation, uncertain economic conditions, geopolitical events, and the ongoing impact of the COVID-19 pandemic. Many tech stocks, including Apple, experienced a considerable decline in value. Despite this, Apple's diversified product portfolio, high profit margins, and consistent revenue growth have made it one of the world's most profitable companies. To gain a better understanding of the past and potential future performance of Apple's stock, this study proposes using time series analysis techniques such as moving averages and auto-correlation, along with auto-regressive moving average (ARMA) models. These models can identify patterns and changes in the stock over time and predict its future behavior. To ensure accurate analysis results, the study highlights the importance of conducting stationarity tests and Gaussianity tests. By examining these properties, the aim is to provide valuable insights into Apple's stock performance over the past 300 trading days, enabling investors to make informed decisions based on their risk appetite and investment goals.

2 Data

The study utilize daily data of Apple's stock gathered from yfinance, an easy to use open source data package providing market data from Yahoo! Finance. Only the daily close price has been used in the analysis as it represents the last price of the trading day. Hence, it is a stable measure of a stock's value compared to other prices throughout the day and represent the equilibrium point where buyers and sellers agreed to trade the stock at the end of the trading day. We look at the daily closing price for the past 300 trading days.

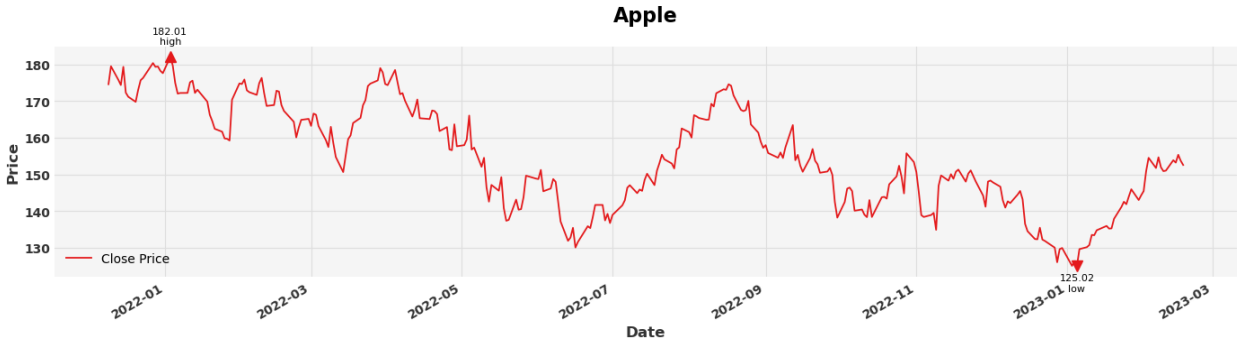


Figure 1: The Apple stock closing price for the past 300 trading days. We observe a loss from year high ($182.01) to year low ($125.02) of -31.33%.

3 Moving Averages & Returns

3.1 Moving Averages (MA)

The moving average of a time series $\mathbf{y} = [y_1, \ldots, y_T]$ with a lookback window of length τ is defined as

$$y_t = \frac{1}{\tau}\Sigma_{i=0}^{\tau-1} y_{t-i} \tag{1}$$

so at time t, y_t is the rolling average of its $\tau - 1$ past values and the value of the time series $\mathbf{y}$ at t. Thus, $\forall i < \tau - 1$, we do not define a moving average as we do not have $\tau - 1$ past values. Thus, if $\hat{\mathbf{y}}$ is a moving average of $\mathbf{y}$ with a lag τ then $|\hat{\mathbf{y}}| = |y| - \tau + 1$ (cardinality of $\hat{\mathbf{y}}$).

When applying MA of an arbitrary window τ to the Apple stock, we look at the average price of the stock for the past $\tau - 1$ days, including the current price y_t. This smooths out short term fluctuations, and provide a insight to trends for the past τ days (Figure 2). This can help identify trends, support, and resistance levels for potential trading opportunities. The slope of the MA can help leverage insight to the strength of the trend. If a price is above the MA, it indicates that the stock is currently trading at a price higher than the MA over the time frame τ. Thus, can suggest that the stock is at an upward trend, and vice versa if the stock is below the MA.

We apply three separate time windows for the moving average to the Apple stock with $\tau = 10, 20$, and 30, denoted MA10, MA20, and MA30, for the past 300 trading days. MA10 follows the stock price more closely than MA20 and MA30. Thus, MA10 provide short-term trends, whereas MA20 and MA30 provide insight to more long-term trends. We observe that the Apple stock's price for the past 300 days has had a downwards trend with fluctuations. Thus, it is an indication that the time series may not be non-stationary given. This will show importance in subsequent sections of the analysis.

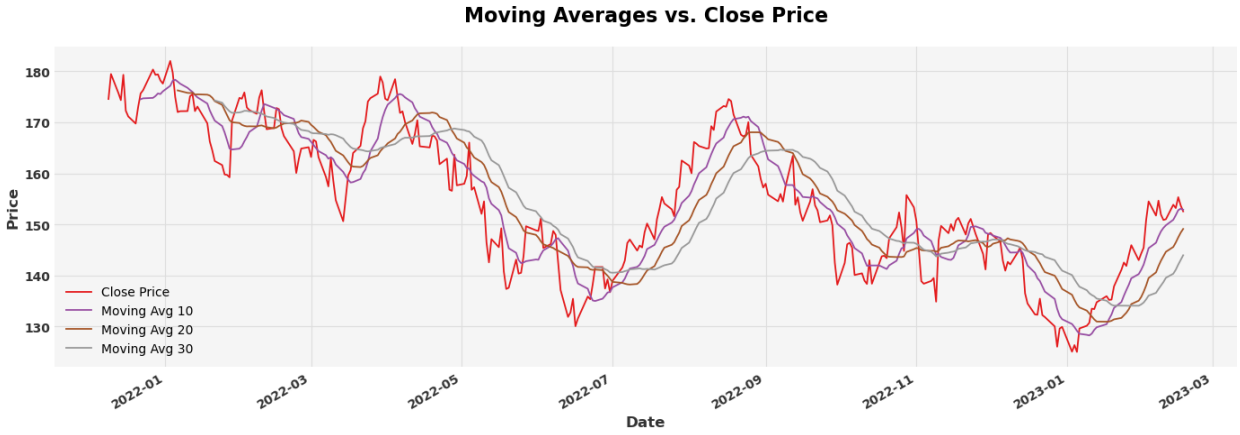


Figure 2: The moving average of the Apple Stock for the past 300 trading days. The actual close price is seen in red, MA10 in purple, MA20 in brown, and MA30 in gray.

3.2 Returns - Linear- & Log Returns

When looking at a stock price's value from day to day, it tend to be relatively stable. Investors look for deviations in price to encapsulate profits or loss. Thus, we look to the return of the stock's time series. Given $\mathbf{y}$ as defined earlier, the linear return is given by

$$r_{linear}(t) = \frac{y_t - y_{t-1}}{y_t} \ , \forall t = \{2, \dots, 300\} \tag{2}$$

and the log return is given by

$$r_{log}(t) = log(\frac{y_t}{y_{t-1}}) \ , \forall t = \{2, \dots, 300\} \tag{3}$$

Inspecting the daily returns can provide insights to a stock's volatility, which helps investors understand the level of risk associated with the stock (See section 6). When comparing several stocks, an investor can consequently quantify which stocks are historically performing better relative to others over a certain time frame. Having this leverage can help an investor's decision to hold or sell assets within their portfolio. Furthermore, analyzing the returns can help in making predictions of its future performance as the historic returns can inform on how it is likely to perform in the future (See section 5). When we calculate the returns of the Apple stock price, its consequent time series $r_{linear}(t)$ and $r_{log}(t)$ look to have a more defined mean $\mu \approx 0$, which may indicate that these exhibit stationarity. However, it is difficult to determine stationarity solely by visualizing, which motivates section 6.

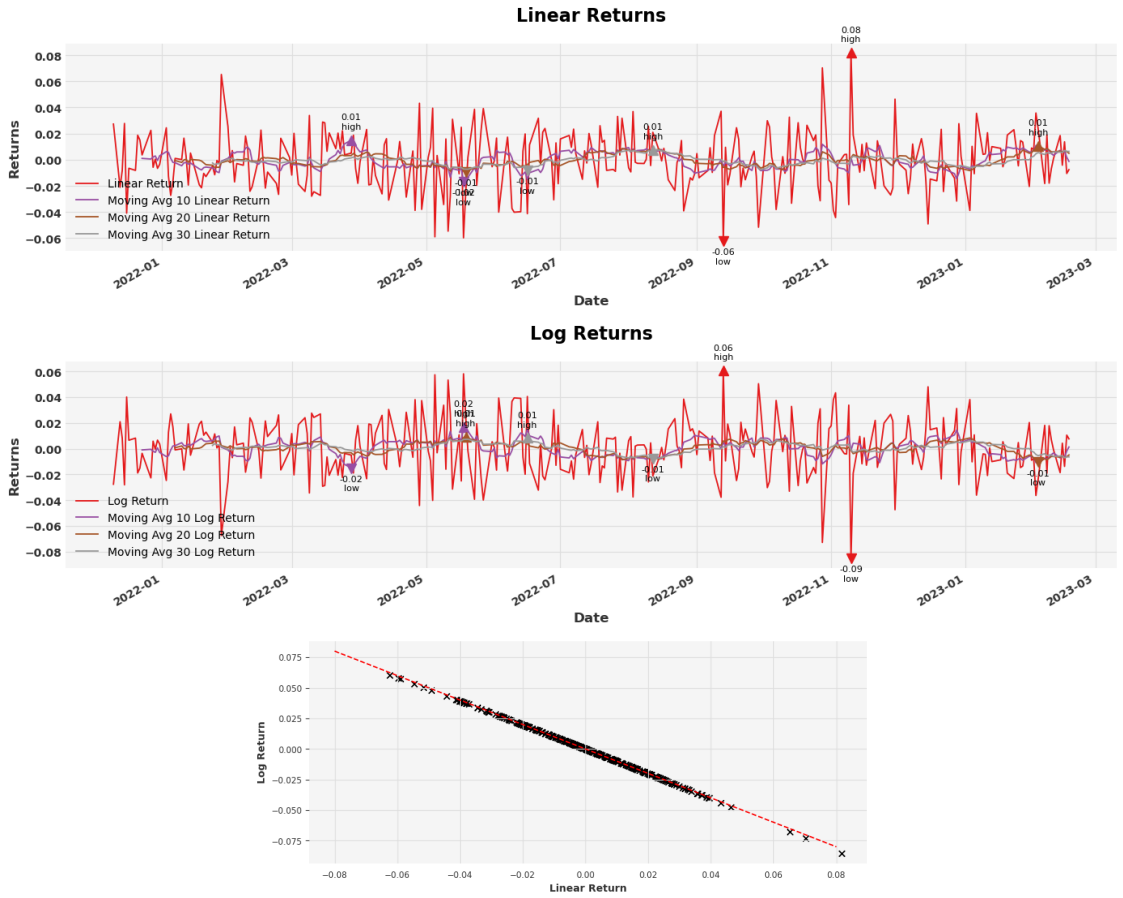


Figure 3: Top: Linear returns of the series close price, MA10, MA20, and MA30. Middle: Log returns of the same series as at the top. All series trend around zero, and look more stochastic and stationary compared to the original closing price time series. Bottom: By plotting the linear return against the log return we see that there is a strong linear relationship. They look directly proportional to each other, which can indicate a normal distribution of the stock returns evenly distributed around some mean.

4 Autocorrelation Function (ACF) & Partial Autocorrelation Function (PACF)

4.1 Autocorrelation Function

The autocorrelation function (ACF) measures the linear predictability between two arbitrary points in time i and j. Let $\mathbf{y} = [y_1, \ldots, y_{300}]$ be the time series, then the ACF is defined as:

$$ACF(y_i, y_j) = \rho(i, j) = \frac{Cov(y_i, y_j)}{\sqrt{Cov(y_i, y_i)Cov(y_j, y_j)}} \forall i, j \in \{1, \ldots, 300\} \tag{4}$$

It is an easy proof to show $ACF \in [-1, 1]$. If we can perfectly predict y_i from y_j, and vice versa then $ACF = \pm 1$. In section 6, we show that the time series is indifferent to its

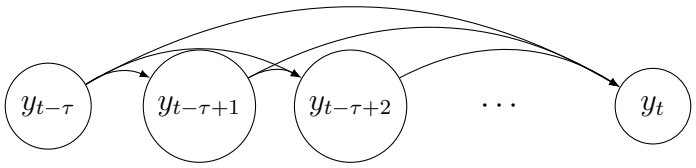
indices i and j, but depend only on the absolute difference $|i-j|$. Considering the indices i and $i+s$, we are effectively computing the ACF for a lag of s for the univariate time series $\mathbf{y}$. Then the ACF function takes the values y_i and y_{i+s}, and is defined as:

$$\gamma(s) \equiv Cov(y_{i+s}, y_i) = \mathbb{E}[(y_{i+s} - \mu)(y_i - \mu)] = \frac{1}{N}\Sigma_{t=1}^{N-s}(y_{t+s} - \bar{y})(y_t - \bar{y}) \tag{5}$$

for a stationary time series where $\bar{y} = \frac{1}{N}\Sigma_{i=1}^{N} y_i$. From here, we can see that $\gamma(i+s, i) = Cov(y_{i+s}, y_i) = Cov(y_s, y_0) = \gamma(s, 0)$ (from the definition of stationarity). Consequently, we can the we get the following for our ACF, given that our time series is stationary:

$$ACF(s) = \rho(s) = \frac{\gamma(i+s, i)}{\sqrt{\gamma(i+s, i+s)\gamma(i, i)}} = \frac{\gamma(s, 0)}{\sqrt{\gamma(s, s)\gamma(0, 0)}} = \frac{\gamma(s)}{\sqrt{\gamma(0)^2}} = \frac{\gamma(s)}{\gamma(0)} \tag{6}$$

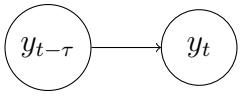
If we have a lag τ for calculating the ACF, then the ACF cares about all indirect and direct effects for all intermediary values between y_t and $y_{t-\tau}$ (given $\tau > 1$) illustrated as a directed acyclic graph (DAG) below



Now, even if the correlation $y_{t-\tau}$ is insignificant with respect to y_t, it can become significant indirectly through preceding values. As the ACF measures the linear correlation between the time series and its lags, it directly assumes that there is correlation between each lag, and that the time series is independent of other lags. However, it may be there may be implicit dependency orders in the time series. For example, if the autocorrelation between y_{t-1} and y_t is high but the autocorrelation between y_{t-2} and y_t is low, the ACF can give misleading results because of the implicit dependency between y_{t-2} and y_{t-1}. This is why we consider the the Partial Autocorrelation Function ($PACF$).

4.2 Partial Autocorrelation Function

Here, we remove any indirect effect of $y_{t-\tau}$ on y_t through intermediary values and only consider the direct effect $y_{t-\tau}$ has on y_t. By removing these indirect effects, we obtain a clearer picture of the direct relationship between $y_{t-\tau}$ and y_t

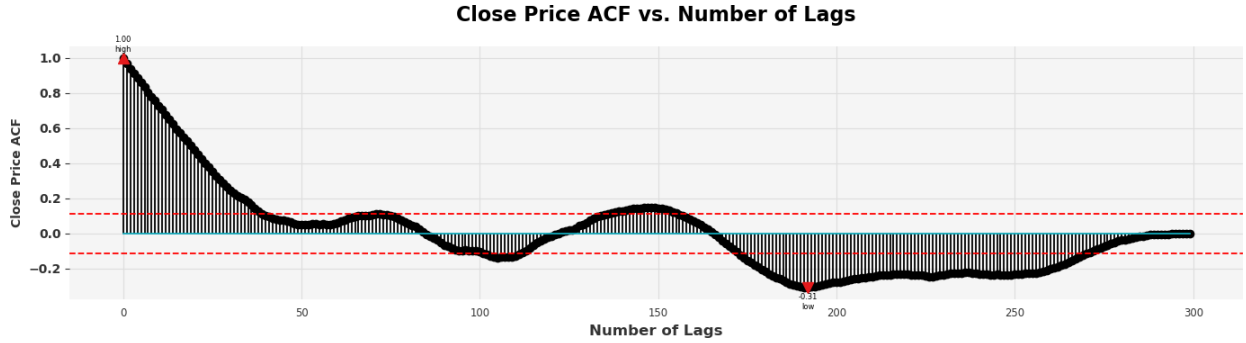


Given a stationary process $\mathbf{y} = \{y_t\}_{t=1}^{T}$, the $PACF$ is defined by

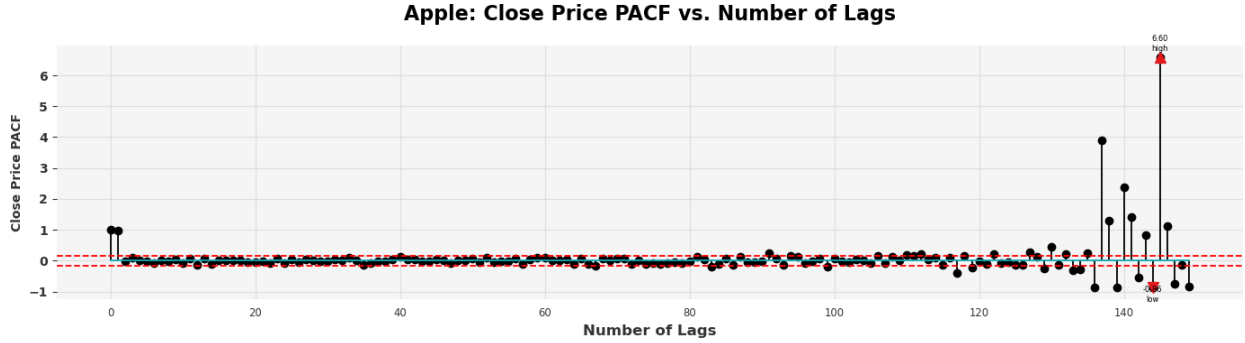
$$\phi_{1,1} = Corr(y_{t+1}, y_t) \text{ , for } k = 1, \text{ and} \tag{7}$$

$$\phi_{k,k} = Corr(y_t - \hat{y}_t, y_{t+k} - \hat{y_{t+k}}) \text{ , } 2 \leq k \leq T-1 \tag{8}$$

where $\hat{y}_{t+k} = \beta_1 y_{t+k-1} + \cdots + \beta_{k-1} y_{t+1}$ and $\hat{y}_t = \beta_1 y_{t+1} + \cdots + \beta_{k-1} y_{t+k-1}$. In other words, $\hat{y_{t+k}}$ and $\hat{y}_t$ are linear combinations of smaller lags. The parameter set $\{\beta_1, \ldots, \beta_{k-1}\}$ is determined by minimizing the mean squared error between $\hat{y_{t+k}}$ and $\hat{y}_t$. It is important to note however that (7) and (8) assume stationarity. Thus, it can produce misleading results whenever $\mathbf{y}$ is non-stationary.



(a)



(b)

Figure 4: The AFC (a) and PACF (b) of the close price of the Apple stock plotted with a red dotted lines at $\frac{\pm 1.96}{\sqrt{300}}$ (ACF) and $\frac{\pm 1.96}{\sqrt{149}}$ $(PACF)$to show a 95% confidence interval.

From Figure 2, the downward trend can suggest that the stock price for Apple is non-stationary as it looks like it is downward trending and with time varying fluctuations. They can stem from seasonality, even though the underlying effect of the fluctuations are not discussed in this study.

In Figure 4, we see that the ACF and $PACF$ have several values that fall above or below the horizontal red lines. They represent a threshold for a 95% confidence interval with critical value $\alpha = 0.05$. The region between the upper and lower bounds of the interval represents the range of values that would be expected if the ACF or $PACF$ values were random and unrelated to each other. The values outside of the region indicate a significant statistical correlation or pattern between data points. The higher values in later lags of Figure 4 suggest a long term memory effect of the ACF and $PACF$, even though there may not be any correlation. When we look at the ACF plot in 4(a), we see that several later

values fall in the statistically significant region. The same is true for the $PACF$. As this rarely happen by chance, it can imply seasonality or other underlying factors is the reason for high correlations occurring in the ACF and the $PACF$ at later lags. This can imply non-stationarity in the Apple stock's time series, which will be discussed in section 6. Note that the values of the number of lags in all PACF plots are set to $\frac{N}{300} - 1$, which is because of the default setting in statsmodels Python package. This is because it is expected that stationary time series are expected to approach zero for all values greater than this.

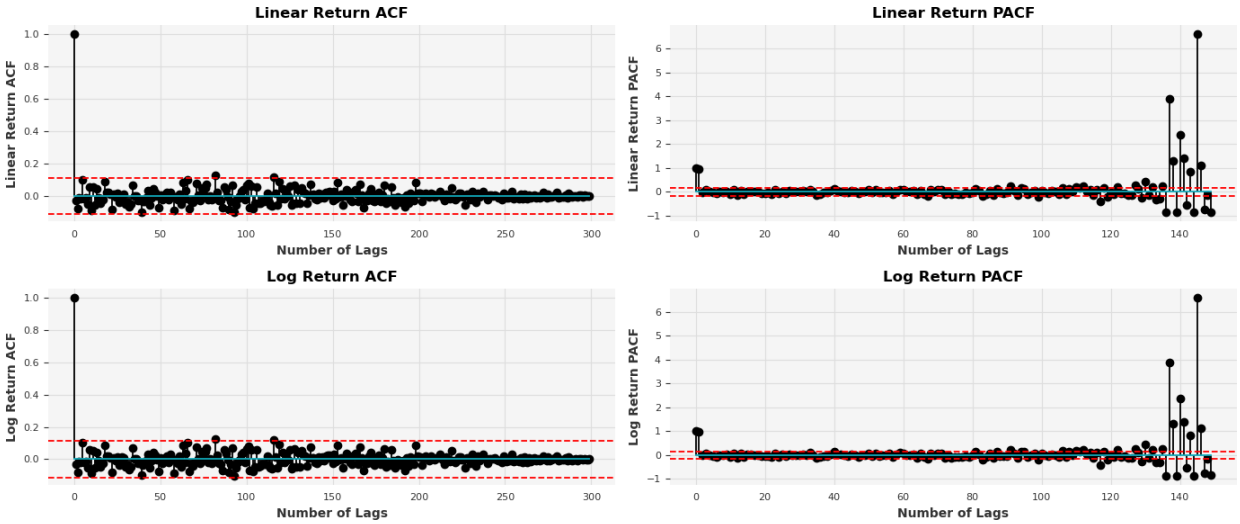


Figure 5: Left: Displays the ACF for the linear return (top) and the log return (bottom) against number of lags. Right: Displays the $PACF$ in the same format. The red dotted lines are as before $\frac{\pm 1.96}{\sqrt{300}}$ (ACF) and $\frac{\pm 1.96}{\sqrt{149}}$ $(PACF)$to show a 95% confidence interval.

Time series of the linear-and log returns (Figure 5) of a stock are more likely to be stationary than the actual stock time series. Stocks tend to carry trends and seasonality over time. Returns, however, remove the trend component making them more likely to be stationary. In Figure 3, we see that there is less indication of trend when applying returns to the time series. In Figure 5, we see that the AFC log- and linear returns quickly converge within the 95% confidence interval of statistical insignificance. Thus, the linear- and log returns are more likely to be stationary. This means we have a more desired auto-regressive behavior of lower order. According to the figure, it can look like it is $AR(2)$ for both returns based on the $PACF$. However, the $PACF$ of log-and linear returns still have higher values at later lags ($t \geq 140$). Thus, it suggests that the correlation between the time series values and their past errors is statistically significant. This may stem from the seasonality/trends in the time series.

5 Auto-Regressive Moving Average (ARMA)

5.1 Defining the Model

An $ARMA$ model is a combination of the aforementioned Moving Average (MA) model and an Auto-Regressive (AR) model. The AR is defined as

$$y_t = \mu \Sigma_{i=1}^{p} \phi_i y_{t-i} + \epsilon_t \tag{9}$$

where $\phi_1, \ldots, \phi_p$ are its parameters, μ is the expectation of the time series (often the maximum likelihood estimator), and the ϵ_t is white noise drawn from $N(0, \sigma^2)$. The model outputs y_t based on its linear dependence of its previous values and an unpredictable noise term. The MA component is defined as

$$y_t = \epsilon(t) + \Sigma_{i=1}^{q} \theta_1 \epsilon(t-i) \tag{10}$$

where $\theta_1, \ldots, \theta_q$ are the parameters of the MA model and ϵ_{t-i} is the error term $\forall i = 1, \ldots, q$ and $\epsilon(t) \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ $\forall t$. Thus, the MA describes the influence of past error terms. Combining (9) and (10), we manage to capture the overall output y_t as a linear dependence on the past values and errors. The $ARMA(p, q)$ is defined as

$$y_t = \mu + \epsilon(t) + \Sigma_{i=1}^{p} \phi_i y_{t-i} + \Sigma_{i=1}^{q} \theta_1 \epsilon(t-i) \tag{11}$$

In (11) the AR and MA components may have common factors. To avoid parameter redundancy we can add restrictions to the $ARMA$ model imposed as:

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p$$
$$\theta(z) = 1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q$$

where z is a complex number. The restrictions on the roots of $\phi(z)$ and $\theta(z)$ are that they must share no common factors and that all the roots of both polynomials must lie outside the unit circle, i.e., $|z| > 1$. These restrictions ensure that the AR and MA components of the model do not overlap and that each has a unique effect on the time series.

It is important to note that the $ARMA(p, q)$ model assumes stationarity (discussed in section 6) and that the relationship between the errors and past values and the output is linear. Violation of these assumptions can lead to biased or inefficient parameter estimates, and consequently unreliable forecasts.

5.2 Fitting the Model

The preceding discussion on ACF and $PACF$ impacts the choice of hyperparameters, p and q, to the $ARMA$ model. p represents the number of past values included in the model. q represents the degree of dependence on past values. Therefore, the ACF can determine

q as it has a significant correlation to the past values to consider as we saw through its dependence on implicit relations. The $PACF$ can determine p as it will have a significant correlation with the direct dependence on past values.

The ACF for the close price have statistically significant lags around values ≈ 50. The $PACF$ have 2 clear statistically significant values, which indicate direct statistical significance come from the values y_{t-1} and y_{t-2}. Thus, we do a grid search over $p = [1, 2]$ and $q = [1, 2, 50]$ on the close price of the Apple stock. For the log return time series, we $p = [1, 2]$ and $q = [1, 2]$ for the linear and log return of our $ARMA(p, q)$ model as indicated by the plots of the ACF and $PACF$. We split the series in a proportion 90/10 of train/test, and use the mean squared error (MSE) to evaluate performance on the out-of-sample predictions. The MSE is defined as $\epsilon = \frac{1}{N}\Sigma_{i=1}^{N}(y_i - \hat{y}_i$ where y_i is the ground truth and $\hat{y}_i$ is the point estimate.

Table 1: **ARMA model results**

$ARMA_{closeprice}(p, q)$	MSE	$ARMA_{log}(p, q)$	MSE
$ARMA(1, 1)$	103.73	$ARMA(1, 1)$	0.0002644
$ARMA(1, 2)$	131.94	$ARMA(1, 2)$	0.0002714
$ARMA(1, 50)$	41.47	$ARMA(2, 1)$	0.0002702
$ARMA(2, 1)$	86.79	$ARMA(2, 2)$	0.0002713
$ARMA(2, 2)$	121.83		
$ARMA(2, 50)$	37.25		

We observe that the $ARMA(p, q)$ models perform poorly on the close price. Based on the moving averages of close price, it looks like the data has seasonality/trends, making an ARMA model inaccurate. The underlying factors effecting a stock are complex and may exhibit non-linear properties, violating the assumption of linear dependence in the ARMA model. We see that the close price time series has better performance with the high lag order q. high-order serial correlation means that the correlation between the time series and its lagged values persists over time. It can suggest a high q capture the autocorrelation more accurately. The log return have very similar error for all values in the grid search, and we see a flat prediction centred at $\mu = 0$, meaning it fails to capture the underlying structure of the returns. In the analysis of ACF and $PACF$ of the close price time series we have seen implications that the series is non-stationary. We expand on the stationarity analysis in 6.2.

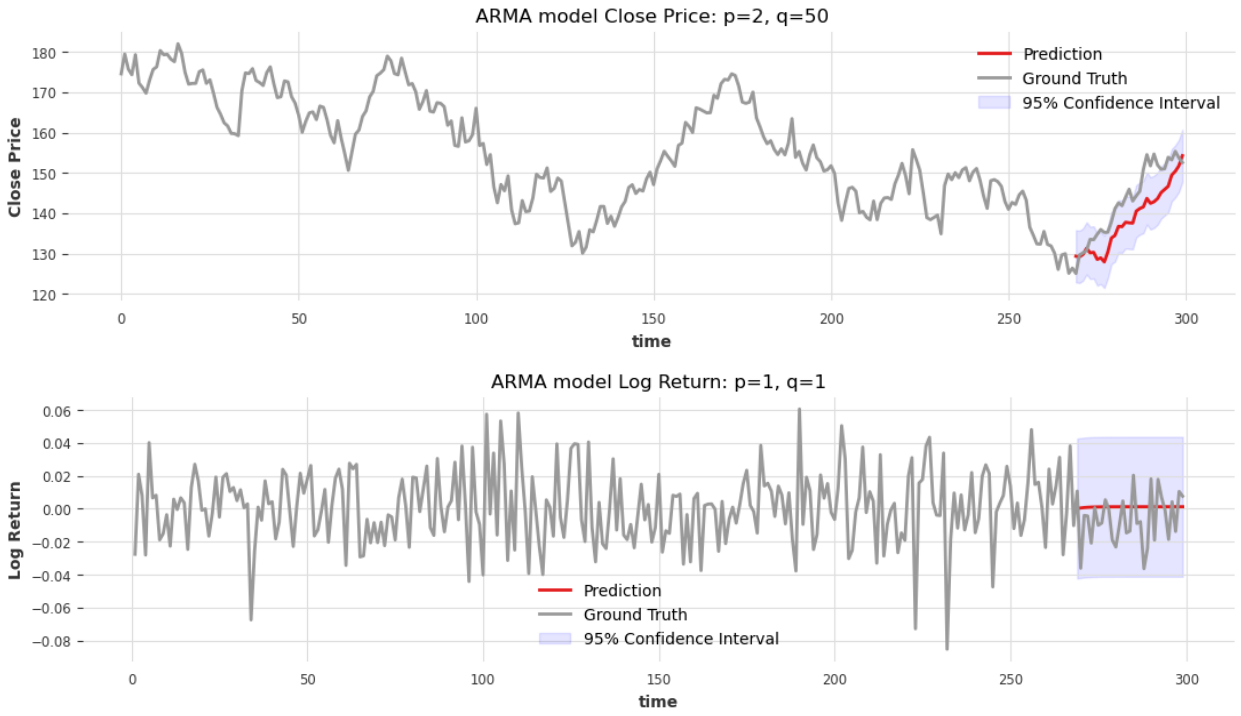


Figure 6: Top: The best model fit for the Close Price. Bottom: The best model fit for the Log Return. The plots shows us that the best fit ARMA models do not perform well. Given $\hat{\mathbf{y}}$ is the prediction estimate, the confidence interval is displayed as the region between $\hat{\mathbf{y}} \pm se$ where $se = \sqrt{\frac{1}{N-1}\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2}$ is the unbiased standard error.

6 Gaussianity and Stationarity Test

6.1 Gaussianity (Shapiro-Wilk Test)

Gaussianity implies that the time series data is centered around some mean value μ with a variance σ^2. We test for the distribution of the close price of the Apple stock and its log returns using the Shapiro-Wilk test. The test statistic W which takes values between 0 and 1 is calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i \times y_{(i)}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{12}$$

where $(a_1, .., a_n) = \frac{m^T \times V^{-1}}{C}$, C is vector norm, m is vector of the order statistics i.i.d. random variable samples from standard normal distribution, V is covariance of the order statistics, $y_{(i)}$ is ith ordered sample of our time series.

Intuitively, the term inside the parenthesis in the numerator of W is interpreted as the slope of observed data versus the expected normal value, normalised to a constant. It is the slope of QQ-plot squared and if the data is normally distributed then the numerator would be an estimate of variance (σ^2). The denominator is also an estimate of population variance

which means $W = 1$ would indicate H_0 to not be rejected. If $W < 1$, then the *p–value* would tell us the approximate probability of observing W as extreme or more extreme than observed. We reject the null hypothesis H_0 that the time series $\mathbf{y}$ is Gaussian if the p-value is less than the given significance level α, and conclude that the distribution is non-Gaussian.

Table 2: **Shapiro-Wilk Test Results**

Time Series	**Test Statistic**	**p-value at** $\alpha = 0.05$	H_0 **Rejected**
Close Price	0.99	0.003893	True
Log Return	0.97	$1.0338 \cdot 10^{-5}$	True

In this table, we have the two time series, Close Price and Log Return, and the results of the Shapiro-Wilk test at significance levels of 0.05. The table shows whether the null hypothesis H_0 (i.e., the time series is normally distributed) is rejected or failed to be rejected based on the p-value obtained from the test. From Table 2 we conclude that the close price and its log returns are not Gaussian.

6.2 Stationarity (Augmented Dickey-Fuller Test)

When we forecast time series, we are assessing certain assumptions to what we expect in the future. Now, time series are non-deterministic so that we cannot assess with certainty what will happen in the future. However, stationarity can help mitigate the difficulty of forecasting. Stationarity implies that the properties of the time series do not change with time. This means the mean, variance, auto-correlation, and so on stay constant with time. We have two types of stationarity. A time series $\{y_t, t \in \mathbb{Z}\}$ is **strictly stationary** if $P(y_1, \ldots, y_T) = P(y_{1+s}, \ldots, y_{T+s})$ for some shift s in time. This means the joint distribution stays the same, regardless of shifting the time stamp (i.e., it only depends on the differences in time). To most applications, this is too strong of an assumption. Hence, we have **weak stationarity**. A time series $\{y_t, t \in \mathbb{Z}\}$ is weakly stationary if:

$$\begin{aligned} \mathbb{E}[y_t^2] &< \infty, \forall t \in \mathbb{N} \\ \mathbb{E}[y_t] &= \mu_t, \forall t \in \mathbb{N}, \text{ and} \\ \text{Cov}(y_i, y_j) &= \text{Cov}(y_{i+s}, y_{j+s}),\ \forall i, j \text{ given some shift } s. \end{aligned}$$

This means the time series must exhibit a finite variance, constant mean, and that the covariance only depends on the difference $|i-j|$ not on the choices of i and j. It is important to see that strict stationarity $\not\Rightarrow$ weak stationarity as it does not assume finite variance. Now, if the time series were Gaussian then weak stationarity would imply strict stationarity as a Gaussian distribution is characterized by its first two moments - μ and σ^2. To test for stationarity on the Apple stock time series and its log return time, we introduce the Augmented Dickey-Fuller (ADF) test. The ADF test is based on the following model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \epsilon_t \tag{13}$$

where $\{y_t\}_{t=1}^{T}$ is the time series, Δy_t is the first difference of y_t, α is a constant, β is a coefficient on a linear trend, y_{t-1} is the lagged value of the time series, ϕ_i are the coefficients on the lagged differences, p is the lag order, and ϵ_t is the error term. The ADF test whether the first order AR model with single root at 1. It includes lags on the dependent variable, accounting for higher-order AR processes. The test statistic is calculated as

$$DF = \frac{\hat{\rho} - 1}{se(\hat{\rho})}$$

where $\hat{\rho}$ is the estimate of the autoregressive coefficient, and se($\hat{\rho}$) is the standard error of the estimate. H_0 of the ADF test is that the time series has a unit root ($\hat{\rho} = 1$, implying non-stationarity. H_1 is that the time series is stationary with $\hat{\rho} < 1$. From Table 3, we confirm

Table 3: ADF Test Results

Time Series	**Test Statistic**	**p-value at** $\alpha = 0.05$	H_0 **Rejected**
Close Price	-2.3088	0.1691	True
Log Return	-17.776	$3.2919 \cdot 10^{-30}$	False

our intuition on the close price being non-stationary and conclude that the log returns of the apple stock is stationary. On the one hand, it strengthens the results of the analysis performed on the log returns. On the other hand, it weakens the results on close price. However, Figure 7 show that its $ARMA(1, 1)$ model predicts what seemingly is the mean of the sequence. This may be due to the insignificant autocorrelation in Figure 5. The model suffers by not being able to capture patterns in the data, resulting in poor predictions. It can also be that the ARMA model is not appropriate for the sequence, or that it exhibits non-linear dynamics that cannot be captured by the ARMA model.