# AI6123 - TIME SERIES ANALYSIS


**G2303513K**

Maheswaran Rohin Kumar

rohinkum001@e.ntu.edu.sg

**Dataset :**

The dataset includes daily historical prices of Apple stock (open, high, low, close, and adjusted close) from February 1, 2002, to January 31, 2017, sourced from Yahoo Finance. It tracks the stock's prices every day, comprising open, close, low, high, and adjusted close prices over a 15-year period. The project aims to uncover intriguing trends in Apple stock prices over the past 15 years (3775 data points) and create the most effective model for forecasting.

**Data Analysis :**

To gain insights into the dataset and prepare for forecasting, we conducted a preliminary analysis. Here are some key statistics derived from the data:

- Minimum Adjusted Closing Price: 0.198346
- Maximum Adjusted Closing Price: 29.863834
- Mean Adjusted Closing Price: 9.45998066357616

The provided statistics offer a glimpse into the range and distribution of the Adjusted Closing Prices over the 15-year period. The minimum value signifies the lowest recorded price, while the maximum represents the peak value observed during the dataset's timeframe. The mean value indicates the average Adjusted Closing Price over the entire period, serving as a central measure of the dataset's tendency.

Additionally, we plotted the original data to visualize the trends and patterns present in the Adjusted Closing Prices. The resulting plot, referred to as Fig. 1 in our analysis, provides an initial understanding of the stock's price movements over time.

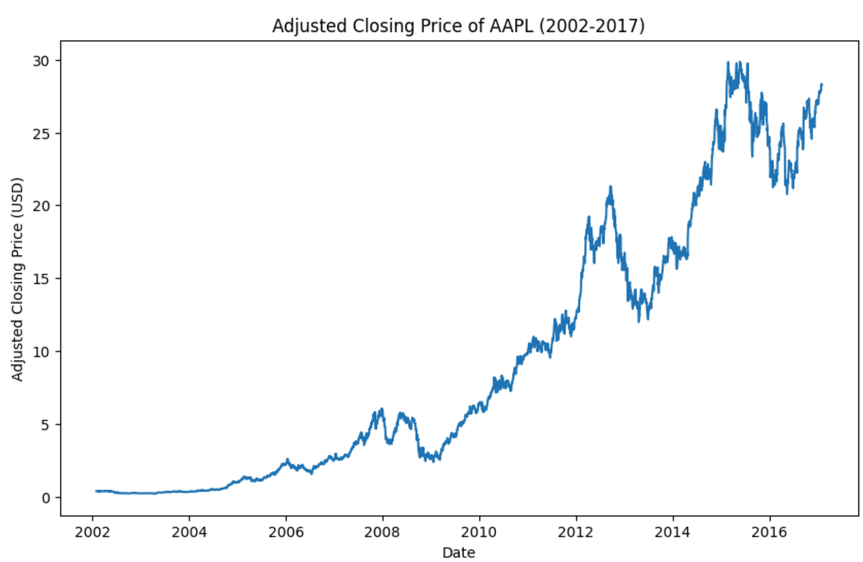
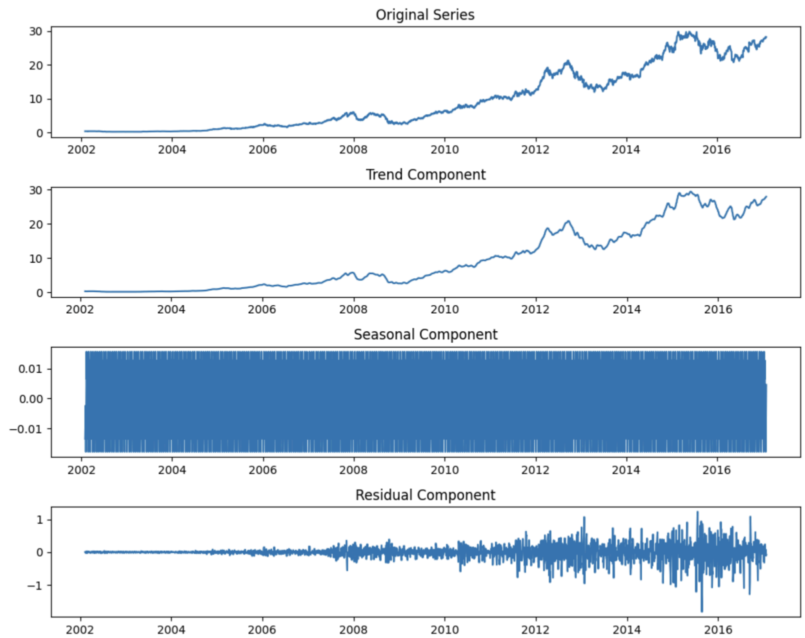



Fig 1 : Adjusted Closing Price of AAPL (2002-2017)

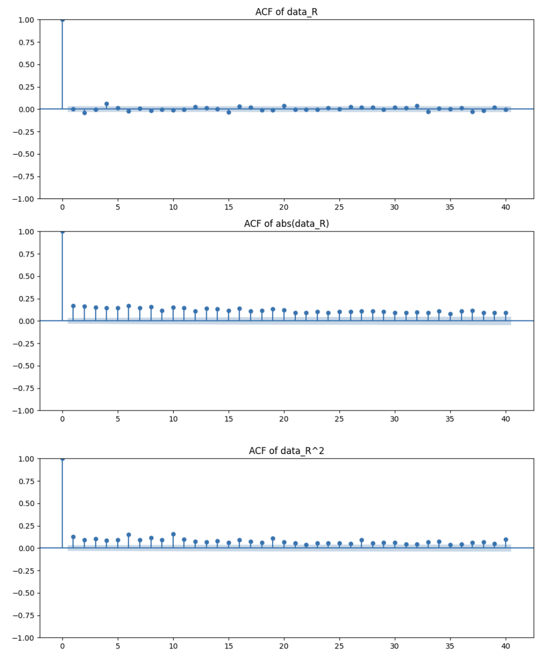


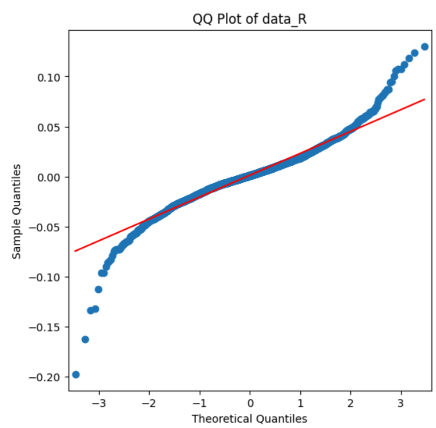

The data plots and remainder plot indicate a significant increase in seasonal variance, prompting consideration of implementing the Box-Cox transformation. Additionally, the Augmented Dickey-Fuller Test yielded a p-value of 0.9779, indicating non-stationarity in the data. The Box-Cox transformation, which includes a hyper-parameter called lambda, aims to stabilize variance fluctuations. In this project, a Box-Cox transformation with a lambda value of zero will be applied. Furthermore, first-order differencing will be used to eliminate trends and enhance data stability. Subsequently, the ADF test resulted in a p-value of 0.01, signifying data stationarity.

- ADF Statistic: 0.3142401652361502
- p-value: 0.9779884661475815

The ADF Statistic yielded a value of 0.3142, while the p-value stood at 0.9780.

To understand the clustering of volatility in log-return data, it's important to examine ACF/PACF plots of absolute and squared log-return data. By analyzing these plots, we can confirm that the returns do not follow an independent and identically distributed pattern.




We also use QQ-plot analysis to investigate the distribution of returns. The QQ-plot indicates that the returns distribution has a thicker tail and is somewhat skewed to the left, which is known as a heavy-tailed distribution. Additionally, we employ kurtosis and skewness tests on the returns data, resulting in values of 5.4356 and -0.1900, respectively. These values confirm our earlier observation: the positive kurtosis value signifies a heavy-tailed distribution, while the negative skewness value indicates a left-skewed pattern.

To summarize, the log-returns data show no serial correlation and exhibit a heavy-tailed, left-skewed distribution. Given these characteristics, the ARCH/GARCH model would be suitable for modeling and analyzing such time series data.

**Splitting the Dataset :**

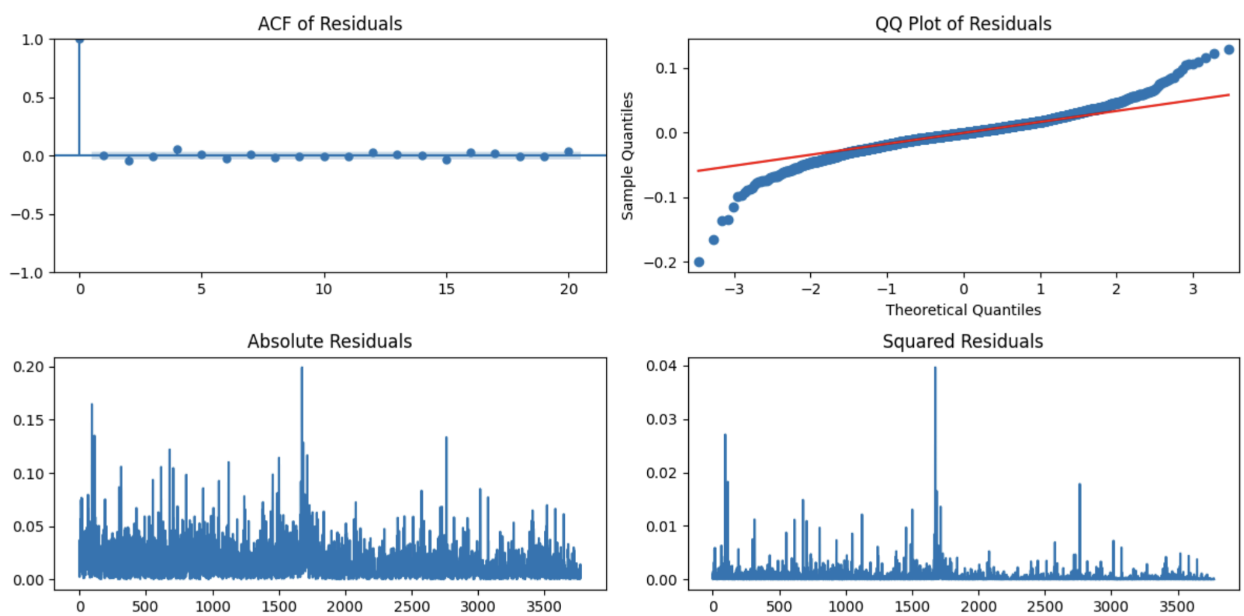
Before we start choosing the model, we divide the data into training and validation sets. Because of the volatility clustering in the time series data, we don't need to predict stock prices beyond one month. So, we set the validation period to 30 days instead of using a fixed ratio for the training and validation sets.

After splitting the data, we have 3745 data points in the training set and 30 data points in the validation set. We won't make any changes to the validation data except for evaluating the prediction results.

**GARCH MODEL :**

Firstly, we use the extended autocorrelation function (EACF) to find the best parameter setting for the ARMA model based on different aspects of the daily returns. The EACF analysis suggests using parameters of (4,0) for daily returns, (1,1), (2,2), or (3,3) for absolute returns, and (1,1) for squared returns. Considering all these suggestions, we select (1,1) as the optimal parameter setting for our data, indicating the GARCH(1,1) model.

Next, we perform a diagnostic check on the GARCH(1,1) model to assess its characteristics and suitability for our analysis.



The p-values above 0.05 in Fig. 11 indicate that the squared residuals are likely uncorrelated over time, suggesting that the standardized residuals could be independent.

We utilized the arch library in Python to identify the best-fitting GARCH(1,1) model for both data distributions and various sub-types of GARCH models. To select the most suitable distribution, we tested several sGARCH(1,1) models with different data distribution settings, assessing their likelihood (higher is better) and information criterion (lower Akaike is better) to determine the optimal choice.

| **Distribution** | **Likelihood** | **Akaike** |
|---|---|---|
| Normal Distribution | 9296.009 | -4.9229 |
| Skew Normal Distribution **T-Distribution** | 9296.242 | -4.9225 |
| **Skew T-Distribution** Generalized Error | **9472.314** | **-5.0158** |
| Distribution Skew Generalized Error | **9472.777** | **-5.0155** |
| Distribution Normal Inverse Gaussian | 9449.263 | -5.0036 |
| Distribution Normal Inverse Gaussian | 9450.581 | -5.0038 |
| Distribution Generalized Hyperbolic | 9468.302 | -5.0131 |
| Distribution Johnson's SU Distribution | 9471.499 | -5.0148 |

Based on the table results, we see that the (Skew-) T-Distribution outperforms other distributions. Therefore, we choose the T-Distribution as the distribution setting.

To select the most suitable sub-model under the fGARCH model (or other GARCH models), similar to the distribution selection process, we conduct a series of tests to identify the optimal sub-model.

| Sub-Model / Model | Likelihood | Akaike |
|---|---|---|
| fGARCH | 9472.314 | -5.0158 |
| fGARCH - TGARCH | 9494.400 | -5.0270 |
| fGARCH - AVGARCH | 9494.357 | -5.0264 |
| fGARCH - NGARCH | 9482.028 | -5.0204 |
| fGARCH - NAGARCH | 9487.081 | -5.0231 |
| fGARCH - APARCH | 9494.421 | -5.0264 |
| fGARCH - GJRGARCH | 9482.289 | -5.0206 |
| fGARCH - ALLGARCH | 9495.035 | -5.0262 |
| eGARCH | 9496.097 | -5.0279 |
| gjrGARCH | 9482.289 | -5.0206 |

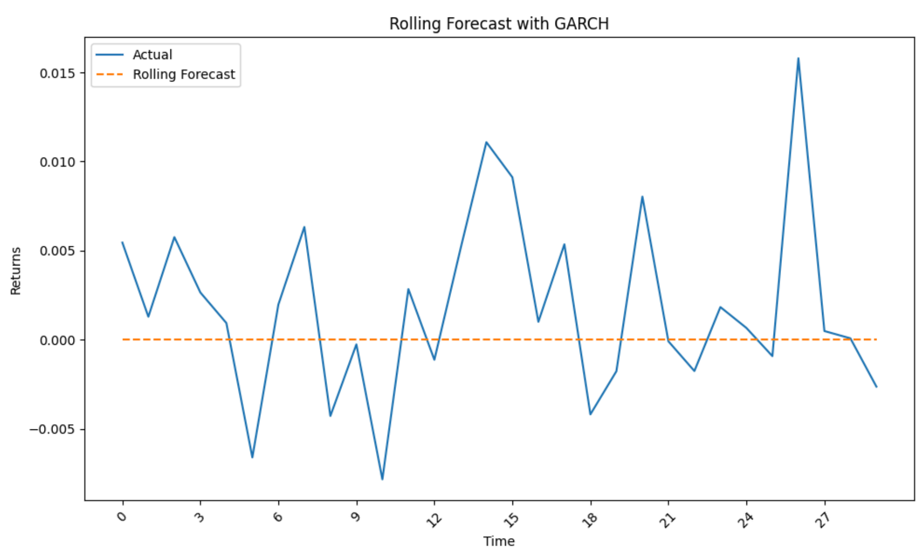
| | | |
|---|---|---|
| **apARCH** | **9494.421** | **-5.0264** |
| **iGARCH** | **9471.993** | **-5.0162** |
| **csGARCH** | **9486.041** | **-5.0220** |

The table results suggest that the best model for fitting the time series data returns is eGARCH(1,1). Therefore, we selected eGARCH(1,1) as the final model for forecasting AAPL. Furthermore, all the models mentioned above have successfully passed diagnostic and Ljung-Box testing.

**Forecasting :**

Moving forward, our focus shifts to the training set data rather than the entire dataset. We redo the model fitting process to assess the likelihood and Akaike score specifically for the training data. The results indicate a likelihood of 9385.934 and an Akaike score of -5.009311.

We can effortlessly create forecasting charts using the N-Roll configuration.



(1) Let

$$X_t = \begin{cases} Y_t & \text{if t is even} \\ Y_t + 1 & \text{if t is odd,} \end{cases}$$

where $Y_t$ is stationary time series. Is $X_t$ stationary ?

**Solution :**

Let $Y_t$ be a stationary time series. Define $X_t$ as follows:

$$X_t = \begin{cases} Y_{t+1} & \text{if } t \text{ is odd} \\ Y_t + 1 & \text{if } t \text{ is even} \end{cases}$$

Since $Y_t$ is stationary, its statistical properties such as mean, variance, and autocovariance remain constant over time.

When $t$ is odd, $X_t = Y_{t+1}$. Therefore, $X_t$ at odd time points inherits the stationarity of $Y_t$.

However, when $t$ is even, $X_t = Y_t + 1$. The stationarity of $X_t$ at even time points depends on how adding 1 affects the statistical properties of $Y_t$. If adding 1 does not change these properties, $X_t$ will be stationary at even time points; otherwise, it will not be stationary.

(2) Suppose that

$$X_t = (1 + 2t)S_t + Z_t$$

where $S_t = S_{t-12}$. Suggest a transform for $X_t$ so that the transformed series is stationary.

**Solution :**

To suggest a transformation for $X_t$ so that the transformed series becomes stationary, we can look at the structure of the original series:

$$X_t = (1 + 2t)S_t + Z_t$$

Here, $S_t = S_{t-12}$, which indicates some sort of seasonal pattern or periodicity in the series $S_t$.

One common transformation to remove such seasonal patterns is differencing. Specifically, we can take the seasonal difference by subtracting the value of $S_t$ at the same season in the previous year:

$$\Delta S_t = S_t - S_{t-12}$$

Applying this transformation to $X_t$, we get:

$$\Delta X_t = X_t - X_{t-12}$$
$$= (1 + 2t)S_t + Z_t - [(1 + 2(t - 12))S_{t-12} + Z_{t-12}]$$
$$= (1 + 2t)S_t + Z_t - (1 + 2t)S_{t-12} - Z_{t-12}$$

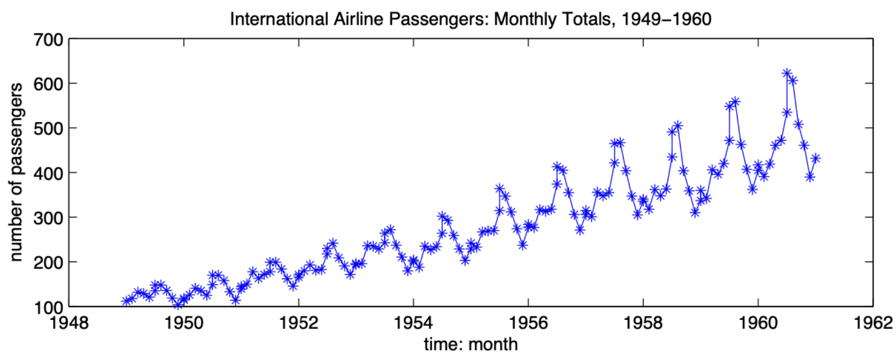
Now, let's simplify this expression:

$$\Delta X_t = (1 + 2t)(S_t - S_{t-12}) + (Z_t - Z_{t-12})$$
$$\Delta X_t = (1 + 2t)\Delta S_t + \Delta Z_t$$

Now, the transformed series $\Delta X_t$ consists of the seasonal difference of $S_t$ and the difference of $Z_t$. Depending on the properties of $S_t$ and $Z_t$, this transformed series may exhibit stationarity, especially if the seasonal pattern in $S_t$ was the main source of non-stationarity in $X_t$.

(3) Based on the time plot of International Airline Passengers. Answer the following questions.

- ▹ (a) Is it stationary ? Justify it .
- ▹ (b) What kind of time series components do the data contain ?
- ▹ (c) Suggest a transformation so that it may equalize the seasonal variation.




**Solution :**

a) No, the time series plot of international airline passengers is not stationary. A stationary time series means that the statistical properties of the series, like the mean, variance, and autocorrelation, are constant over time. In the plot, we can see that the number of passengers increases over time (upward trend). This trend indicates that the mean is not constant over time. Also, the plot shows a seasonal pattern, with higher passenger numbers in the summer months and lower numbers in the winter months. This seasonality suggests that the variance is not constant over time.

b) The time series plot of international airline passengers contains the following components:

- **Trend:** There is an upward trend in the data, which means the average number of passengers is increasing over time.
- **Seasonality:** There is a seasonal pattern in the data, with higher passenger numbers in the summer months and lower numbers in the winter months.
- **Irregularity:** There is also some irregularity in the data, which is the random fluctuations that we see around the trend and seasonal components.

c) There are a couple of transformations that can be used to equalize the seasonal variability in the time series data of international airline passengers. One common transformation is seasonal differencing. This involves taking the difference between the current value and the value from the same season in the previous year. For example, we could calculate the difference between the number of passengers in January 1949 and the number of passengers in January 1948. This would remove the seasonal component from the data and make it more stationary.

Another possible transformation is deseasonalization. This is a statistical method that removes the seasonal component from the data by fitting a seasonal model to the data and then subtracting the fitted model from the original data.Deseasonalization can be more complex to implement than seasonal differencing, but it can be more effective at removing the seasonal component from the data.

By applying one of these transformations, we can equalize the seasonal variability in the time series data and make it more suitable for further analysis, such as forecasting.