Coursework 1 - COMP0051 Algorithmic Trading

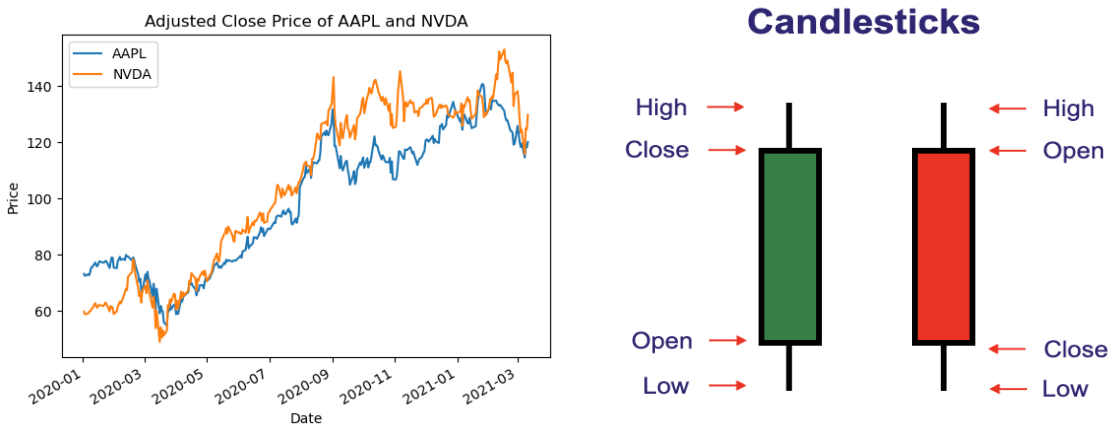
1 Time Series [10 Points]

In this question, two ETF time series from Apple ($AAPL$) and Nvidia ($NVDA$) have been downloaded using the API of Yahoo Finance. As the length is required to be 300 data points, the dates selected as the start and the end are 2020-01-01 and 2021-03-12 respectively. This guaranteed that there would be 300 points in the time series. The resolution is daily. The downloaded data including several columns: Adjusted Close Price, Close Price, High Price, Low Price, Open Price and Volume.

(a) Candlestick of AAPL (b) Candlestick of NVDA

Figure 1: Candlestick of Each Time Series

Figure 1 shows the candlestick chart of each stock, and this chart is the most commonly used in the financial market showing the trend of stocks. The candlestick chart fully used the properties of the stock price listed above and more details can be found in figure 2(b). Finally, by setting the adjusted close price as the final price of the stock, the plot of AAPL and NVDA is shown in figure 2(a).



(a) Adjusted Close Price for AAPL and NVDA (b) Candlestick Explanation

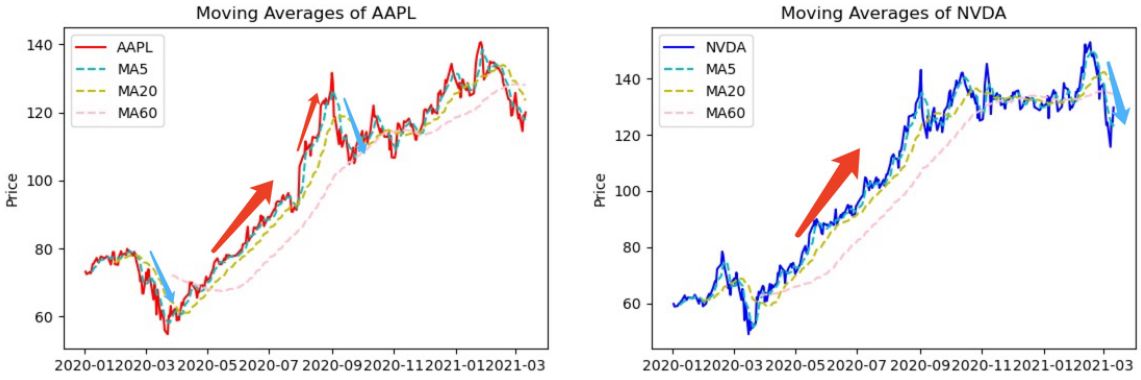
Figure 2: Price Plot and the Candlestick

2 Moving Averages [20 Points]

A moving average is a technical indicator that market analysts and investors may use to determine the direction of a trend (Hansun, 2013). To mathematically define the simple moving average of the price time series with a time window τ, let $P(t)$ represent the price at time t, where t is the index of the time series data points. The moving average at time t, denoted as $MA(t)$, is calculated by taking the average of the prices over the previous τ time points

$$MA(t) = \frac{P(t-\tau) + P(t-\tau+1) + ... + P(t-1) + P(t)}{\tau} \tag{1}$$

By using the `rolling().mean()` in Python, the moving averages can be easily calculated with different time windows $\tau = 5, 20, 60$, which are corresponding to one week, one month and one quarter. Then the corresponding plots are in Figure 3.



(a) MA Against the Price of AAPL

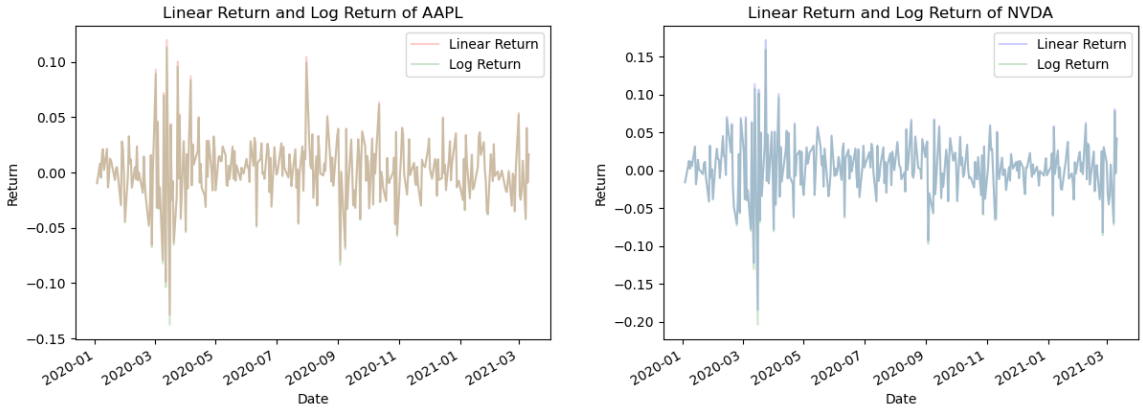
(b) MA Against the Price of NVDA

Figure 3: MA of Different Window Size Again the Adjusted Close Price

The 5-day moving average is one vital criterion of short-term trading. If the 5-day moving average rises sharply, above the monthly and quarterly lines, it is a bullish move, indicating that the stock price may double (See the red arrows in Figure 3). In contrast, when the weekly line is below all monthly and quarterly lines, it shows a short arrangement, indicating that the short trend will continue for some time (See the blue arrows in Figure 3).

Additionally, two criteria can be applied to evaluate the return of the price time series: linear return R_{linear} and log return R_{log}

$$\begin{aligned} R_{linear} &= \frac{P(t) - P(t-1)}{P(t-1)} = \frac{P(t)}{P(t-1)} - 1 \\ R_{log} &= \ln \frac{P(t)}{P(t-1)} = \ln\left(1 + R_{linear}\right) \end{aligned} \tag{2}$$



(a) Return for AAPL

(b) Return for NVDA

Figure 4: The Linear Return Against The Log Return for Price Time Series

These two methods have been applied to calculate the return time series in two scales for the two ETF time series fetched above. The plots are shown in Figure 4. The trend of the two scales is similar while the numerical results for the returns are different due to the different calculations.

3 Time Series Analysis [20 Points]

In time series analysis, the understanding of the past is important to predict the future. The autocorrelation function (ACF) is one way to discover, which reveals how the correlation between any two values of the signal changes as their separation changes (Nounou & Bakshi, 2000; Heilbronner, 1992; Flores et al., 2012). The correlation coefficient indicates how strong the relationship is between two variables. If the value is 1, variables are completely positively related; if the value is -1, then completely negatively related; if the value is 0, then completely unrelated. Given a time series y_t, the auto-correlation of lag k, denoted $\rho(k)$ can be expressed

$$\rho(k) = \frac{\mathrm{Cov}(y_t, y_{t-k})}{\sqrt{\mathrm{Var}(y_t) \cdot \mathrm{Var}(y_{t-k})}}, \tag{3}$$

where

- $\mathrm{Cov}(y_t, y_{t-k})$ is the covariance between y_t and y_{t-k} at lag k,
- $\mathrm{Var}(y_t)$ and $\mathrm{Var}(y_{t-k})$ are the variances of y_t and y_{t-k} respectively.

The ACF is a useful tool in time series analysis for identifying patterns, seasonality, and trends in the data. Besides, to directly measure the relationship between the time series and its lagged values, the partial auto-correlation function (PACF) has been given when controlling for the values of the variable at all shorter lags. It is the auto-correlation between y_t and y_{t+k} with the linear dependence of y_t on y_{t+1} through y_{t+k} remove. Given a time series y_t, the partial auto-correlation of lag k, denoted $\phi(k)$ can be expressed

$$\begin{aligned} \phi(1) &= \mathrm{Corr}(y_t, y_{t+1}), \\ \phi(k) &= \mathrm{Corr}(y_t - \hat{y}_t, y_{t+j} - \hat{y}_{t+k}), \end{aligned} \tag{4}$$

where $\hat{y}_t$ and $\hat{y}_{t+k}$ are linear combinations of the smaller lags that minimize the mean squared error for $\hat{y}_t$ and $\hat{y}_{t+k}$.

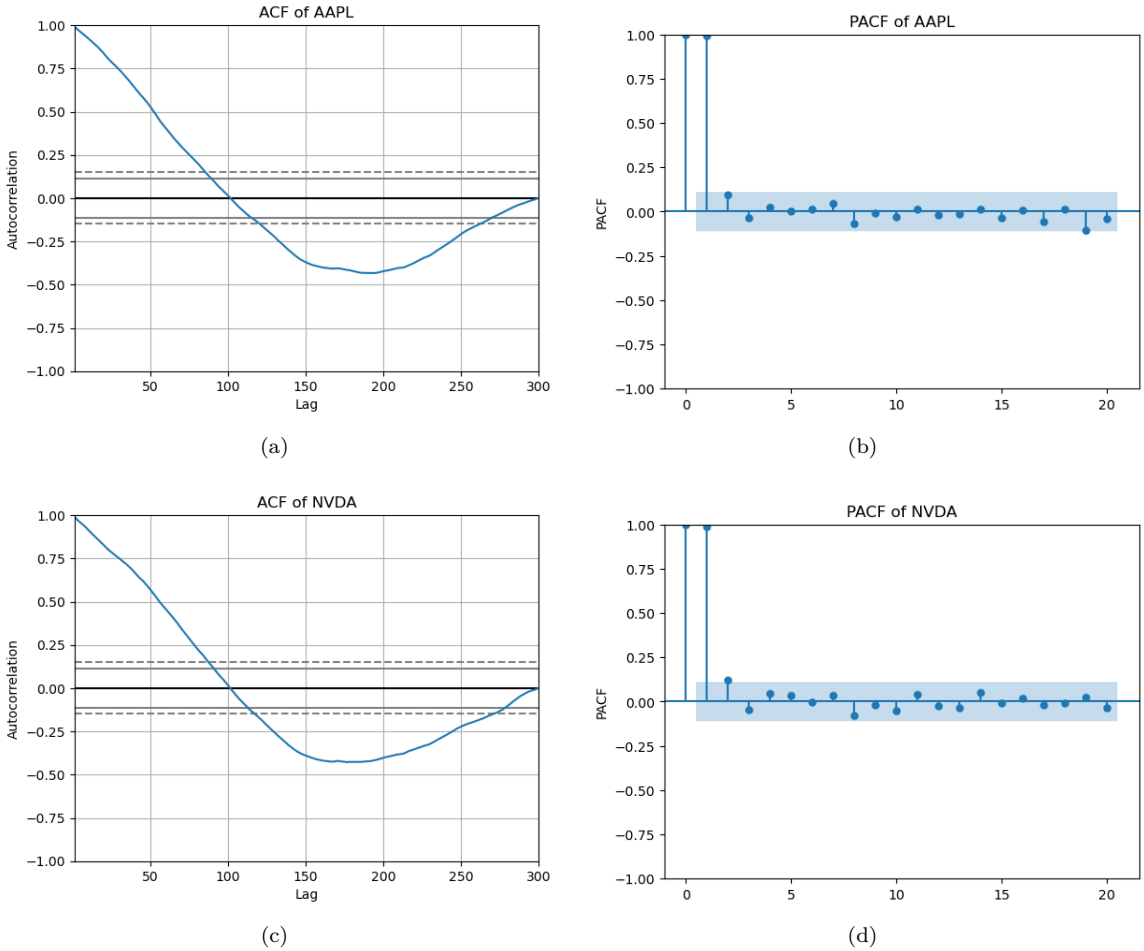


Figure 5: The ACF and PACF for Price Time Series

The ACF and the PACF have been applied to the two ETF price time series for analyzing the relationship. The corresponding results are shown in Figure 5. The trend in Figure 5(a) and Figure 5(c) implies that the correlation coefficient is decreasing with the increase of the lag. Even though when the lag is around 160, the ACF is close to -1, meaning that this part is negatively related, it still converges to 0. For the information given by Figure 5(b) and Figure 5(d), the blue region is the confidence interval which means that the variables here are unrelated. Overall, it is reasonable to demonstrate that the single day's price of AAPL and NVDA does not have a solid relationship with other days.

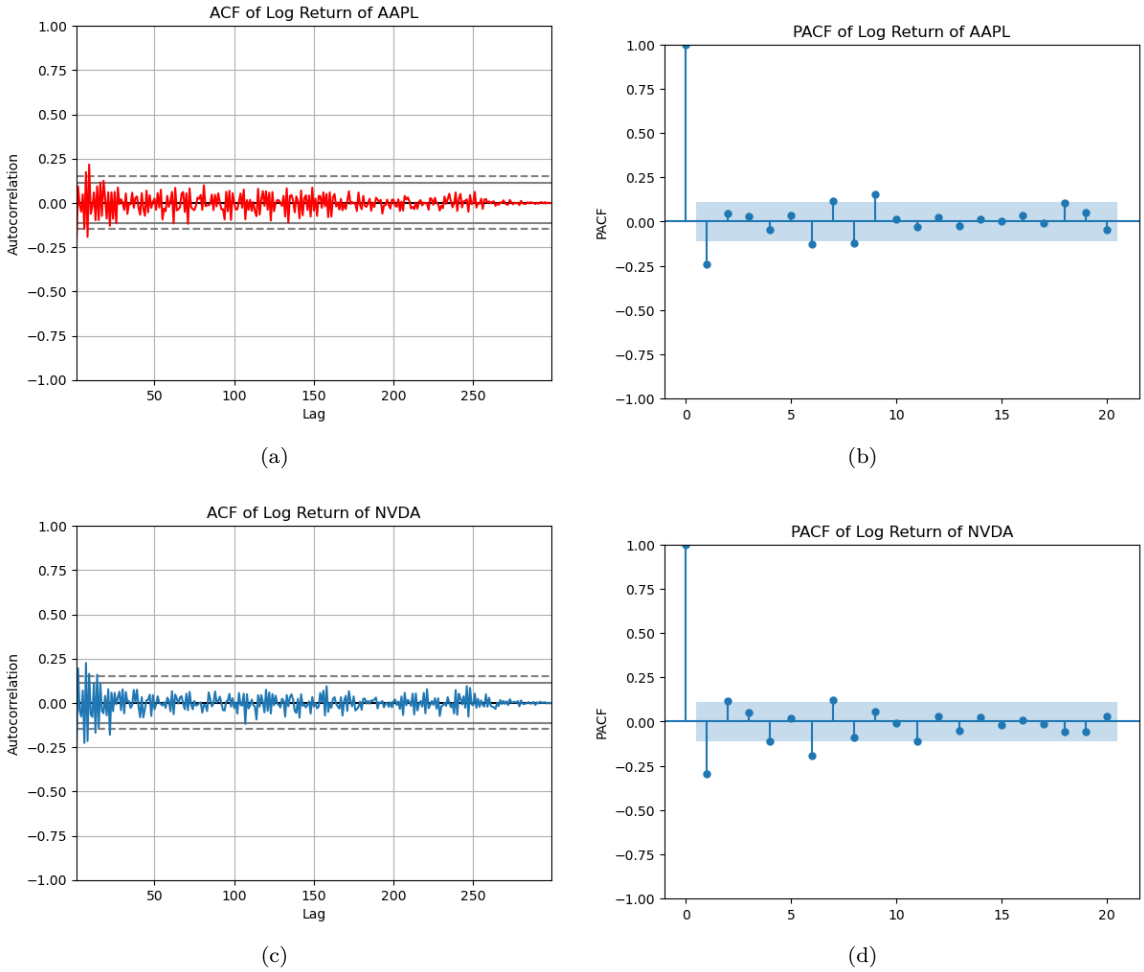


Figure 6: The ACF and PACF for Log Return Time Series

Besides, these two tools were also applied for the log return time series of each price time series. Figure 6(a) and Figure 6(c) show similar results in that the ACF oscillates around 0 while most of them are located in the confidence interval. Results from the PACF of Figure 6(b) and Figure 6(d) also proved the conclusion that the log return of the single day is unrelated to the others log return.

4 Gaussianity and Stationarity Test [20 Points]

One common test for Gaussianity is the **Shapiro-Wilk Test** which is based on a detailed analysis of sample order statistics for the Gaussian variables. It is a statistical test that assesses whether a sample of data comes from a normally distributed population. The Shapiro-Wilk test uses both marginal order statistics and also joint order statistics. It can be mathematically expressed as follows: Given a sample of n ordered data points $x_1, x_2, \ldots, x_n$, where x_i is the i-th smallest value in the set, the quantity that is evaluated in the Shapiro-Wilk test is

$$W = \frac{(\sum_i^n a_i x_i)^2}{\sum_i^n (x_i - \overline{x})^2}, \tag{5}$$

and a_i is given by

$$a_i = \frac{m^n V^{-1}}{\mathcal{N}},$$

where $\mathcal{N}$ is simply a normalization factor, m is the vector of expected values of all the order statistics in a Gaussian distribution, and V is the expected covariance of pairs of order statistics. Its null hypothesis is that

the data is normally distributed, and the test statistic W is compared against critical values to determine whether the null hypothesis can be rejected. If W is less than the critical value and the corresponding $p-$value is larger than the pre-set value, the null hypothesis is not rejected and vice versa.

The Gaussianity test has been utilized for the log return time series of both the AAPL and the NVDA. The acceptable $p-$value is supposed to be larger than 0.05 to support the hypothesis. The result is visible in Table 1.

ETF Log Return	W	$p-$value	**Result**
AAPL	0.946	0.00	Not Normal Distributed
NVDA	0.948	0.00	Not Normal Distributed

Table 1: Shapiro-Wilk Test for AAPL and NVDA Log Return with the Result

Even though the W values are fairly close to 1, the $p-$value is less than 0.05 and therefore two log return time series cannot be considered as normal distributed.

Another aspect of analyzing a time series is to test its stationarity as a stationary series is easier for statistical models to predict effectively. It requires some statistical properties such as mean, variance, covariance and standard deviation do not vary with time, or these properties are not a function of time. One common method to test the stationarity is the **Augmented Dickey-Fuller Test (ADF)** which is also called a unit root test.

The ADF test's null hypothesis is that the time series has a unit root, meaning it is non-stationary. It is typically implemented with the following regression model:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \ldots + \varepsilon_t, \tag{6}$$

where

- y_t is the value of the time series at time t,
- Δy_t is the difference series,
- t is the time trend,
- α is a constant term,
- β is the coefficient of the time trend,
- γ is the coefficient of the lagged value of the series,
- $\delta_1, \ldots, \delta_{p-1}$ are coefficients of the lagged differences,
- ε is the error term.

The ADF test statistic is calculated based on the estimated coefficients in this regression model, and it is compared to critical values to determine whether the null hypothesis of non-stationarity can be rejected. If the test statistic is less than the critical value, we can reject the null hypothesis and conclude that the time series is stationary. Otherwise, if the test statistic is greater than the critical value, we fail to reject the null hypothesis, indicating that the time series is non-stationary.

ETF Log Return	ADF Stat	$p-$value	**Result**
AAPL	-5.278435	0.000006	Stationary
NVDA	-6.863883	0.000000	Stationary

Table 2: ADF Test for AAPL and NVDA Log Return with the Result

Two log return time series from AAPL and NVDA have been tested for their stationarities and results are shown in Table 2. Three critical values are -3.453(1%), -2.872(5%) and -2.572(10%). As the ADF statistics

are less than the critical value, the null hypothesis can be rejected and these two log return time series can be declared as stationary.

5 Cointegration [30 Points]

Cointegration is a statistical property of a collection of time series variables. Two series are co-integrated if a linear combination of them has a lower level of integration, or in other words, the combination is stationary. The concept of cointegration developed the time series analysis by solving the problem of spurious regressions (Granger & Newbold, 1974). The phenomenon of spurious regressions (Figure 7) faded the time series analysis as only the stationary series, which is unusual, can be analyzed.

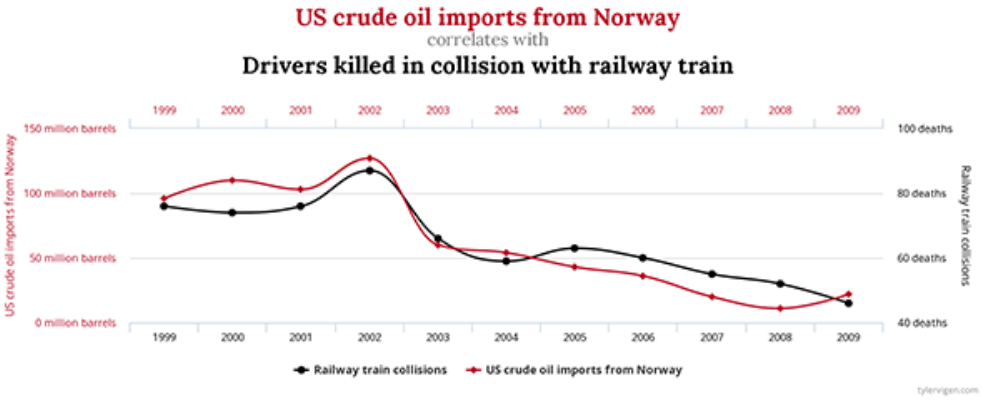


Figure 7: Correlation between US crude oil imports from Norway and drivers killed in a collision with a railway train has a very high correlation coefficient of 0.95, representing a strong positive relationship, while there is no causal relationship between the two

In 1987, Engle & Granger (1987) considered cointegration and typically the meaning of it is the validation of whether the casual relationship described by the regression formula of two series is spurious regression. Consider a group of time series, Y_t, which is composed of 2 separate time series: y_1, y_2, and they are non-stationary time series. Cointegration implies that while y_1 and y_2 are independently non-stationary, they can be combined in a way that their linear combination is stationary

$$\beta Y_t = \beta_1 y_{1t} + \beta_n y_{2t} \ I(0). \tag{7}$$

One famous method test on cointegration is the **Engel-Granger Test**, which follows the very simple intuition that if series variables are cointegrated, the residual of the cointegrating regression should be stationary. This approach mainly has two steps:

1) Forming the cointegrating residual z_t

$$z_t = y_{1t} - \theta y_{2t},$$

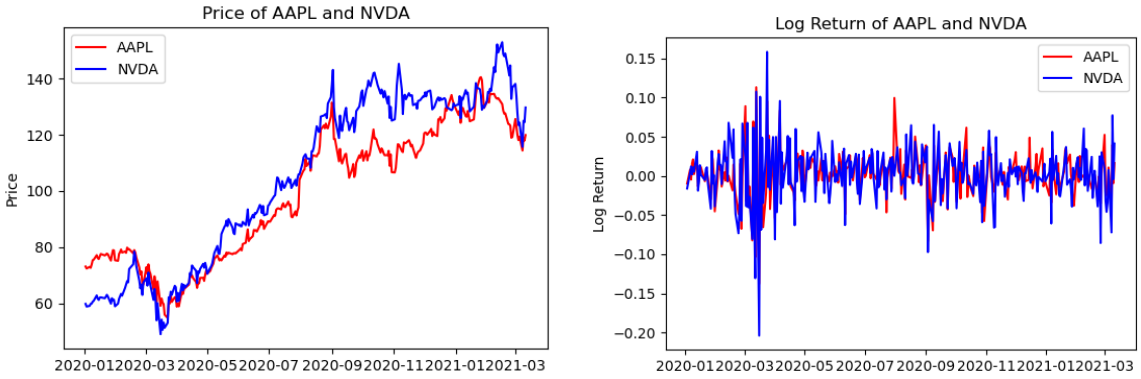
2) Check the stationarity of the residual z_t using the unit root test or ADF test. If z_t $I(0)$, then y_{1t} and y_{2t} are cointegrated.

Up to now, two groups of time series are available to test the cointegration inside each group: the price time series and the log return time series. However, noticeably the group of the log return time series has proved their stationarity previously, which does not fit the appliance of the concept of cointegration. According to Granger (2004), when two series are stationary, they cannot be cointegrated. Therefore, the trial would be applied focusing on the group of ETF price time series.

The results are in Table 3 and the confidence level set to evaluate the $p-$value is 0.05. Besides, two plots in Figure 8 have shown the trend of each one to make the results visible.

Time Series Group	$p-$value	**Result**
Price	0.21422	Not Cointegrated
Log Return	2.104e-29	Not Cointegrated

Table 3: Engle-Granger Test for Price and Return Time Series



(a) (b)

Figure 8: Two Groups of Time Series for Cointegration

Overall, the analysis has proved that the two price time series are not cointegrated even though intuitively they have a similar changing trend. Additionally, even though the $p-$ value satisfies the cointegration condition of the return series, as they are both stationary, they still cannot be regarded as cointegrated.

References

Robert F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913236.

João Henrique Ferreira Flores, Paulo Engel, and Rafael Pinto. Autocorrelation and partial autocorrelation functions to improve neural networks models on univariate time series forecasting. pp. 1–8, 06 2012. ISBN 978-1-4673-1488-6. doi: 10.1109/IJCNN.2012.6252470.

Clive W. J. Granger. Time series analysis, cointegration, and applications. *The American Economic Review*, 94(3):421–425, 2004. ISSN 00028282. URL http://www.jstor.org/stable/3592936.

C.W.J. Granger and P. Newbold. Spurious regressions in econometrics. *Journal of Econometrics*, 2(2): 111–120, 1974. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(74)90034-7. URL https://www.sciencedirect.com/science/article/pii/0304407674900347.

Seng Hansun. A new approach of moving average method in time series analysis. In *2013 Conference on New Media Studies (CoNMedia)*, pp. 1–4, 2013. doi: 10.1109/CoNMedia.2013.6708545.

Renée Panozzo Heilbronner. The autocorrelation function: an image processing tool for fabric analysis. *Tectonophysics*, 212(3):351–370, 1992. ISSN 0040-1951. doi: https://doi.org/10.1016/0040-1951(92)90300-U. URL https://www.sciencedirect.com/science/article/pii/004019519290300U.

Mohamed N. Nounou and Bhavik R. Bakshi. Chapter 5 - multiscale methods for denoising and compression. In Beata Walczak (ed.), *Wavelets in Chemistry*, volume 22 of *Data Handling in Science and Technology*, pp. 119–150. Elsevier, 2000. doi: https://doi.org/10.1016/S0922-3487(00)80030-1. URL https://www.sciencedirect.com/science/article/pii/S0922348700800301.