UNIVERSITY COLLEGE LONDON

DEPARTMENT OF COMPUTER SCIENCE

Algorithmic Trading Coursework 1

Student ID:
23075753

Lecturer:
Dr Paolo Barucca
Dr Nick Firoozye

February 19, 2024

Page Count: 10
Word Count: 1242

1 Time Series Analysis

ETF Time Series

An Exchange-Traded Fund (ETF) is an investment fund that aggregates a portfolio of assets, such as stocks, commodities, or bonds and is traded on stock exchanges. ETFs enable investors to collectively invest in a diversified basket of securities. This analysis concentrates on two prominent ETFs: the SPDR S&P 500 ETF Trust (Ticker: SPY) and the Invesco QQQ Trust (Ticker: QQQ). The former tracks the performance of the S&P 500 Index, which consists of 500 leading companies in the United States. Conversely, the latter reflects the NASDAQ-100 Index, encompassing 100 of the largest non-financial companies listed on the NASDAQ stock exchange.

For this analysis, 300 days of time-series data for both ETFs were collected using the Yahoo Finance API. Acknowledging that financial markets are closed on public holidays and weekends, data spanning the previous 450 days were initially downloaded into a Pandas DataFrame. This dataset was subsequently refined to retain only the most recent 300 days of trading data, from 18th November 2022 to 31st January 2024.

ETF Time Series Visualisation

The time series plots of the selected ETFs are shown in two formats. Figure 1 is the candlestick graph that captures information of the 'open', 'high', 'low' and 'close' prices. Figure 2 on the other hand, is a standard line chart that only shows the 'close' price of both ETFs. The 'close' price will be continuously used in future analysis as it is the last price at which the ETF is traded during the regular trading hours on a given day.

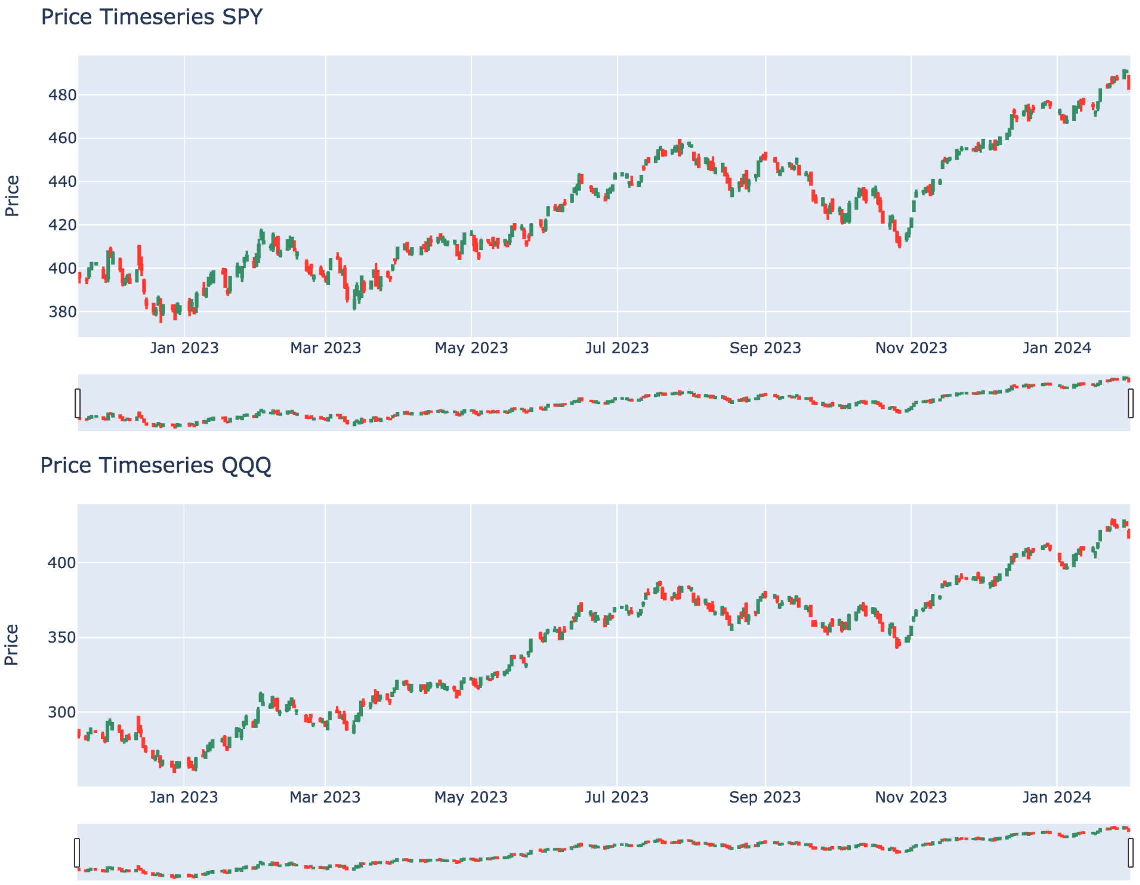

Figure 1: Candlestick graph of SPY and QQQ

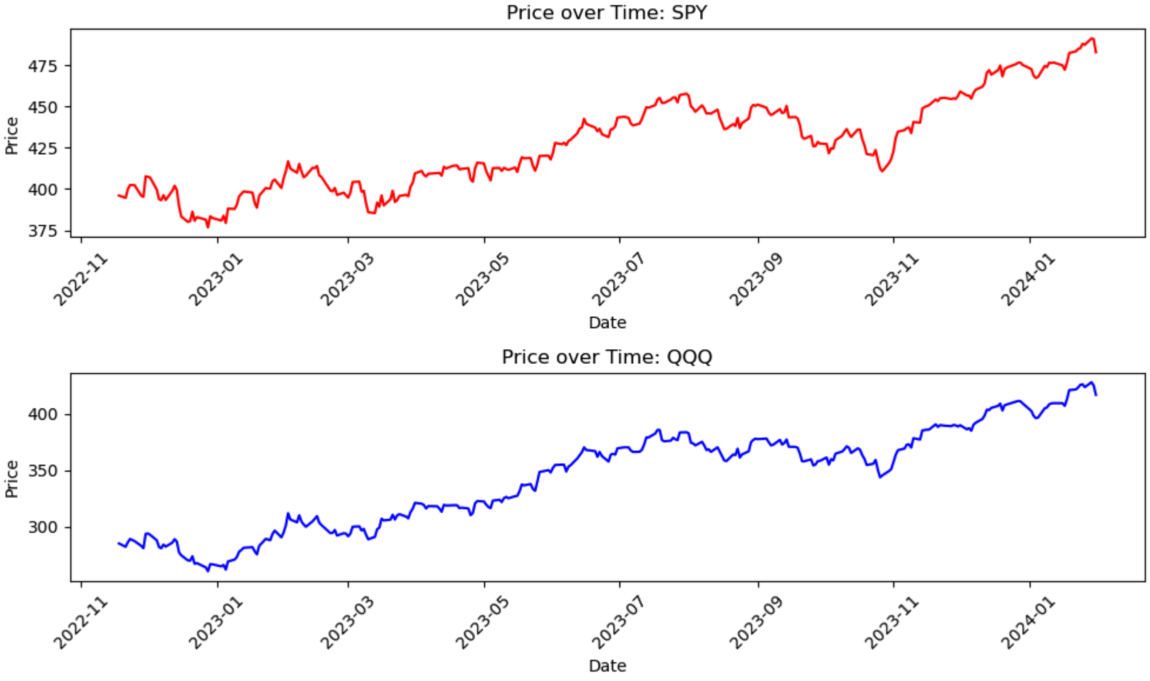

Figure 2: Time-series plot of SPY and QQQ

2 Moving Average

Definition of Moving Average

Equation 1 shows the mathematical definition of the moving average, where MA_t refers to the moving average value at time t, τ is the time window and x_i is the quantity of interest at time step i.

$$MA_t = \frac{1}{\tau} \sum_{i=t-\tau+1}^{t} x_i \tag{1}$$

Visualisation of Moving Average

Based on Figure 3, the choice of τ affects the sensitivity of the moving average to changes in the data. A larger τ will result in a smoother moving average that is less responsive to short-term variations, effectively highlighting longer-term trends. A common trend-following trading strategy is to go long when a short-term moving average exceeds the long-term moving averages.

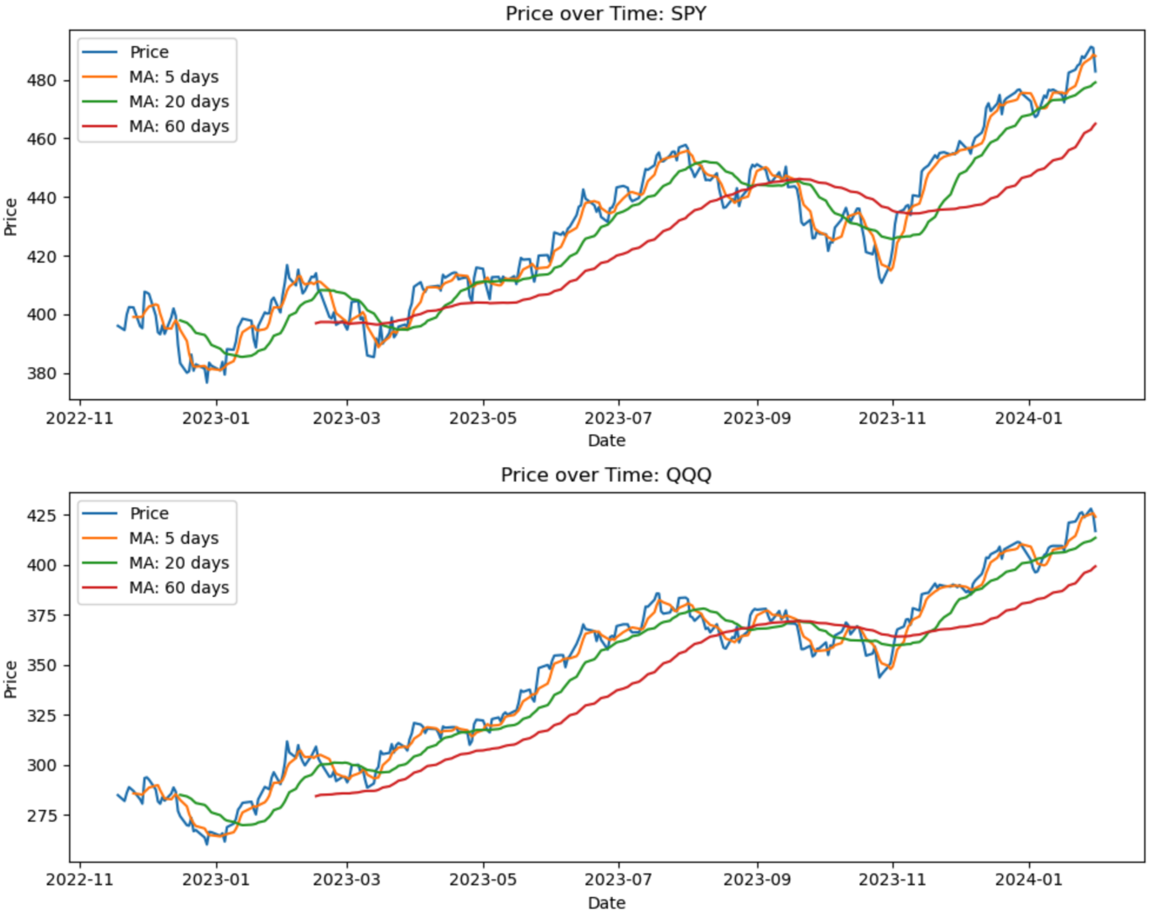

Figure 3: Moving average of SPY and QQQ at three time windows $\tau = 5, 20, 60$

Definition of Return Series

The linear return and log return are defined in Equations 2 and 3, where $p(t)$ represents the price at time t.

$$r(t) = \frac{p(t) - p(t-1)}{p(t-1)} \tag{2}$$

$$r_{log}(t) = \log \frac{p(t)}{p(t-1)} \tag{3}$$

Visualisation of Return Series

As illustrated in Figure 4, the volatility of QQQ's returns appears to be higher compared to SPY's. Additionally, the shape of the return time series closely mirrors that of the log return series, as depicted in the figure below.

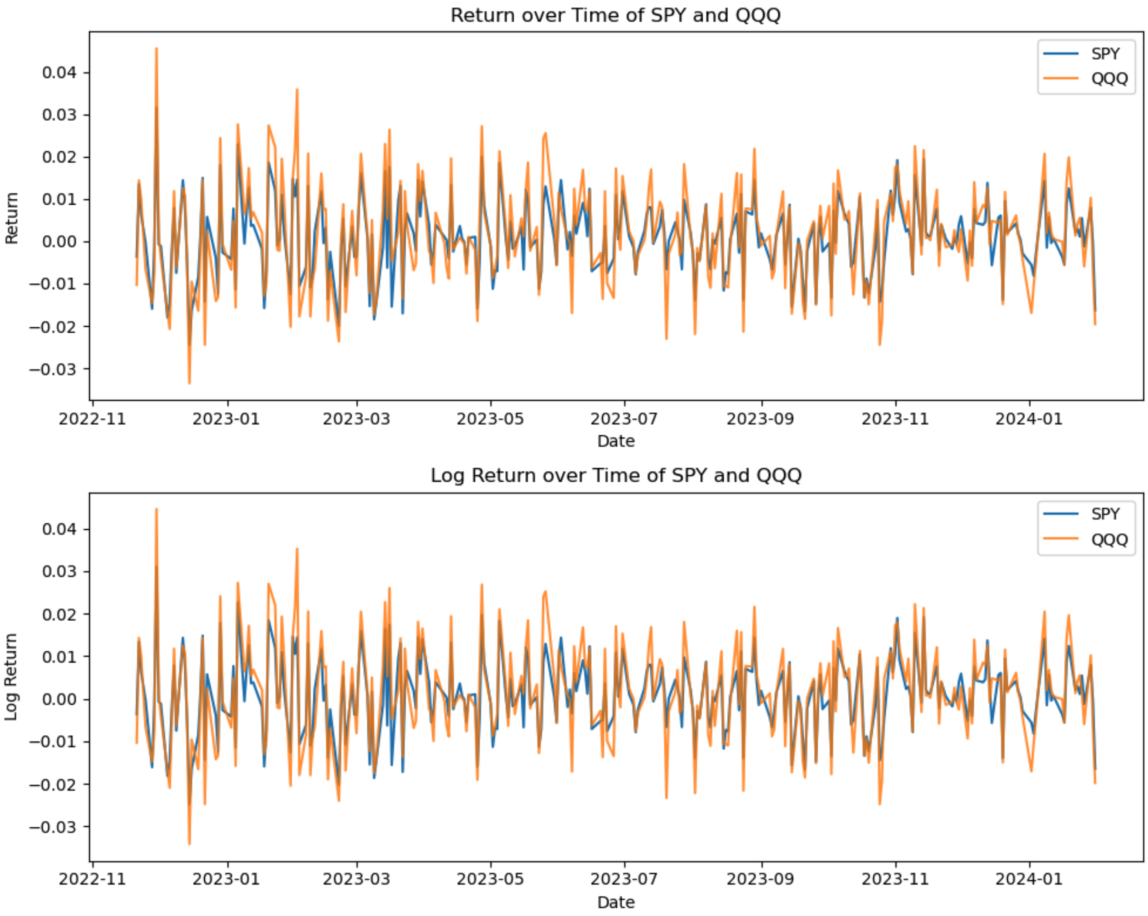

Figure 4: Return and log return of SPY and QQQ

3 Correlation Analysis

Definition

Let X_t be a strongly stationary time-series sequence, the auto-correlation function (ACF) is defined in Equation 4. It measures the correlation between observations in a time series at different times, separated by lag k. In the below equation, T is the total number of observations, x_t refers to observation at time t and $\bar{x}$ is the average observation.

$$ACF(k) = \rho(k) = \frac{\sum_{t=k+1}^{T}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2} \tag{4}$$

Similarly, the partial auto-correlation function (PACF) measures the correlation between observations in a time series while removing the effects of previous time lags. It captures the direct effect of past data on future data at lag k, excluding any indirect correlations. Its mathematical definition is shown in Equation 5 where $\hat{x}_x$ is the predicted value of x_t based on the linear combination of smaller lags.

$$PACF(k) = corr(x_t - \hat{x}_t, x_{t-k} - \hat{x}_{t-k}) \tag{5}$$

Visualisation of ACF and PACF

Figures 5 and 6 illustrate the ACF and PACF coefficients of the price series, while Figures 7 and 8 focus on the return series. The time lag k is set to 50. In these figures, the blue shaded area denotes a 95% confidence interval, indicating that correlations at lags within this area are not statistically significant.

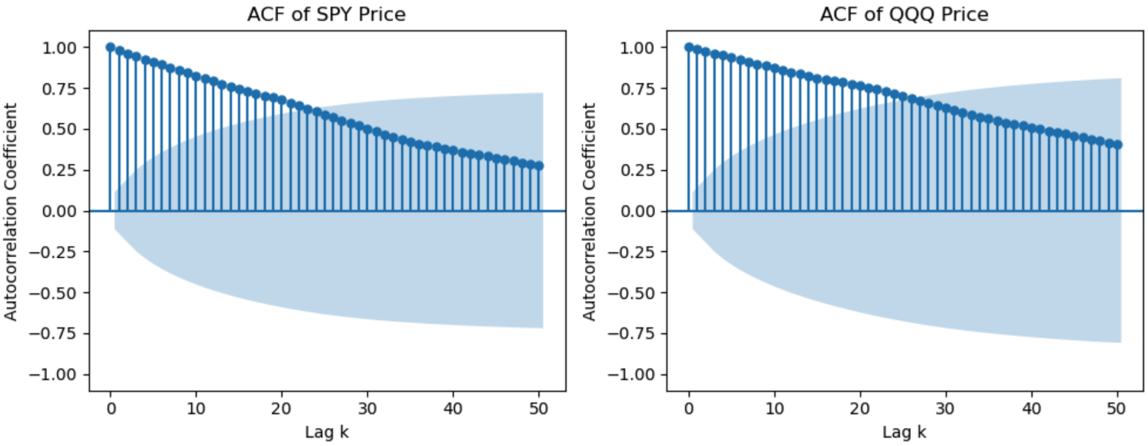

Figure 5: Price ACF of SPY and QQQ

The ACF diagrams presented in Figures 5 and 7 capture both direct and indirect effects of preceding lags on the current price or return. In the context of the price series, it is observed that lags up to approximately 22 days exhibit a robust correlation with today's price. Conversely, the correlation depicted in the return series within Figure 7 diminishes markedly over time.

As highlighted in Figure 6, there is a notable direct impact of the previous day's price on the current price, marked by a lag of 1 day, whereas the influences from other lags appear to be minimal. However, from the return PACF graph in Figure 8, the return on the previous day no longer has a strong correlation with today's return.

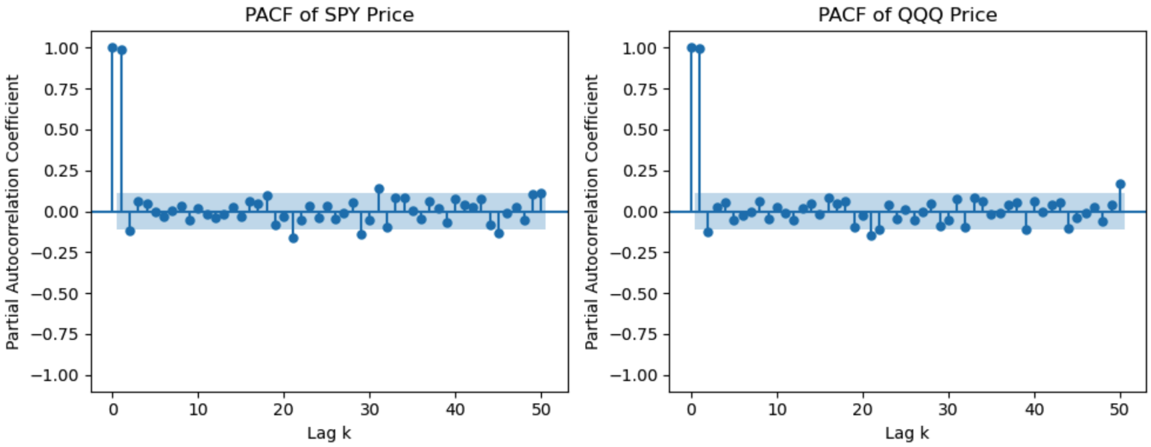

Figure 6: Price PACF of SPY and QQQ

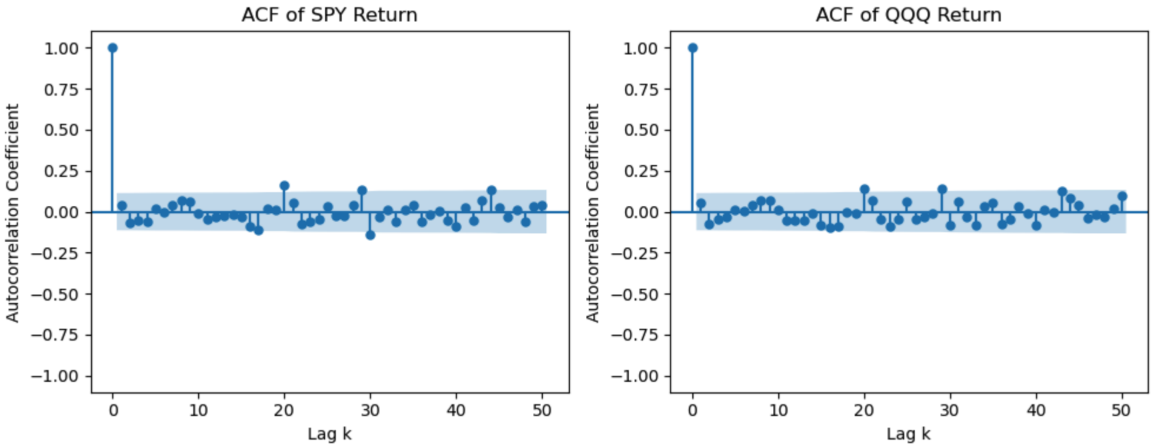

Figure 7: Return ACF of SPY and QQQ

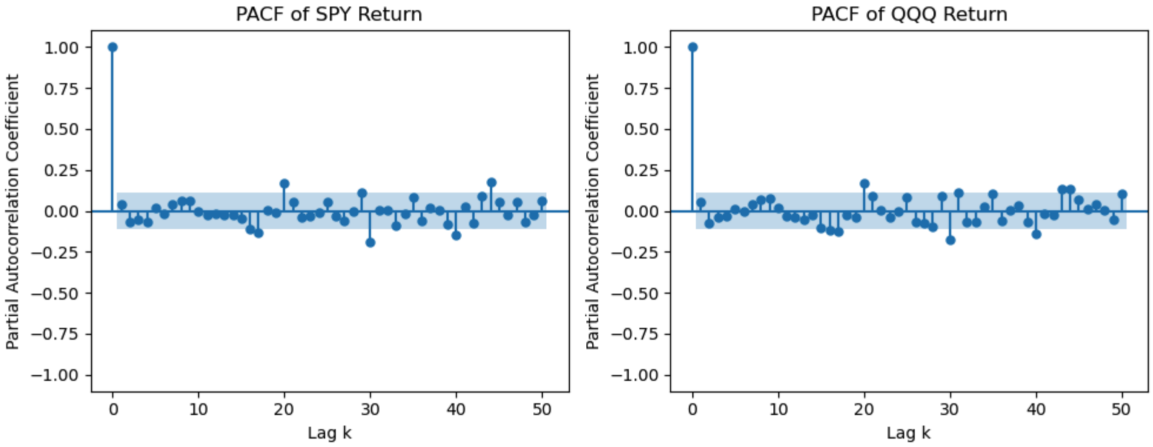

Figure 8: Return PACF of SPY and QQQ

4 Gaussianity and Stationarity Test

Gaussianity Test

The Shapiro-Wilk test is a popular method for evaluating the normality of a sample distribution. This test specifically assesses the null hypothesis that a sample $x_1, x_2, ..., x_n$ comes from a normally distributed population. The test statistic is defined in Equation 6, where a_i is a constant generated from covariances, variances, and expected values of all the order statistics in a Gaussian distribution. Besides, $r_{(i)}$ refers to the i-th smallest value in the sample and $\bar{x}$ is the sample mean.

$$W = \frac{(\sum_i^T a_i r_{(i)})^2}{\sum_i^T (r_{(i)} - \bar{x})^2} \tag{6}$$

Using the Scipy library, the test statistic and p-values are computed as shown in Table 1. Since both p-values are greater than the threshold of 0.05, thus the null hypothesis fails to be rejected, meaning that the distribution of the ETFs' returns is similar to Gaussian.

	test statistic	p-value
ETFs		
SPY	0.994	0.271
QQQ	0.995	0.454

Table 1: Shapiro-Wilk test results

The Gaussianity test result can be double-checked by a Kolmogorov-Smirnov Distribution (KS) test, which compares the empirical distribution function of the sample data with the cumulative distribution function of the Gaussian reference distribution. The KS test reached the same conclusion that both time series are similar to Gaussian. The QQ plot in Figure 9 provides a visual check of Gaussianity.

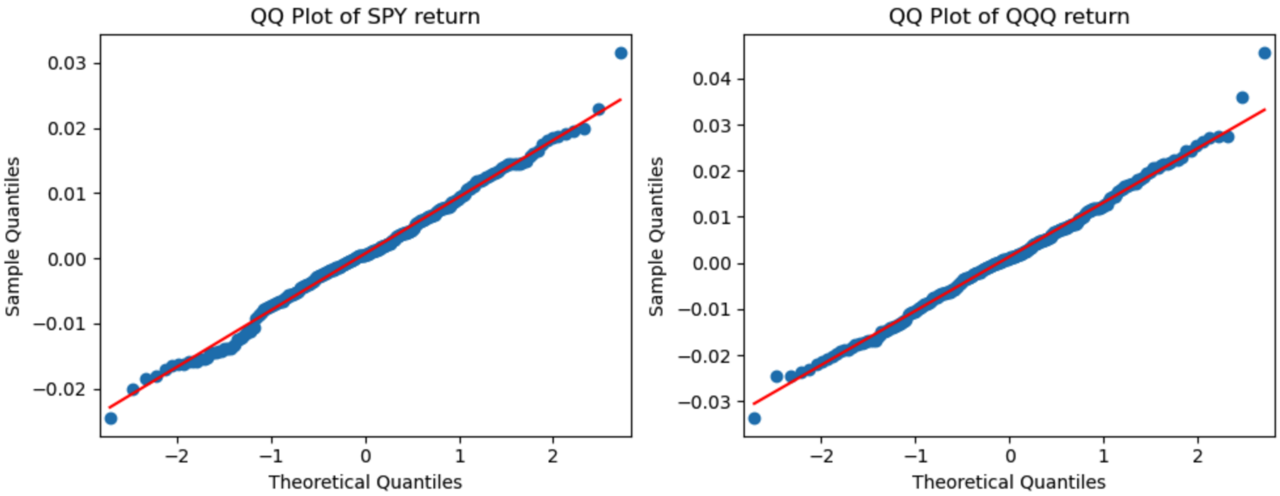

Figure 9: QQ plot of SPY and QQQ

Stationarity Test

A stationary time series is one whose statistical properties such as mean and variance are constant over time. The Augmented Dickey-Fuller (ADF) test is a well-known statistical test for stationarity, specifically designed to test for the presence of unit root non-stationarity. The equation that represents the regression model used in the ADF test is shown in Equation 7.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \epsilon_t \tag{7}$$

- Δy_t is the change of y at time t
- α is the intercept term
- β is the coefficient on a time trend
- γ is the coefficient on the lagged series
- p is the number of lagged differences
- ϕ_i is the coefficient of the lagged differences
- ϵ_t is the error term

The null hypothesis of this test is the presence of unit root ($\gamma = 0$), assuming that the series is not stationary. The test statistics for this test are described in Equation 8, where $\hat{y}$ is the estimated coefficient and SE refers to the standard error. The results of this test are shown in Table 2. Since both p-values are less than 0.05, the null hypothesis is rejected, suggesting that the return series for both SPY and QQQ are stationary.

$$t_p = \frac{\hat{y}}{SE(\hat{y})} \tag{8}$$

	test statistic	p-value
ETFs		
SPY	-16.383	0.000
QQQ	-16.211	0.000

Table 2: Augmented Dickey-Fuller test results

5 Cointegration Test

Definition

Two series are cointegrated when they have a long-term, stable relationship despite being non-stationary on their own. This concept is useful in financial markets for identifying pairs of assets that move together in the long run.

A common cointegration test is the Engle-Granger method. The first step is to check whether both time series (X_{1t} and X_{2t}) have an order of integration of 1 ($I(1)$), meaning they become stationary after differencing once. The second step checks whether there exists θ such that $Z_t = X_{1t} - \theta X_{2t}$ is stationary. In this step, a regression analysis is performed, which utilises the ordinary least square (OLS) to estimate the coefficient θ. A unit-root test, such as the Augmented Dickey-Fuller test, is then conducted on the residual Z_t. If the residuals are found to be stationary, given both series are individually non-stationary, they are cointegrated.

Results on Price and Return Time Series

The cointegration test is conducted on both the price and return time series. The null hypothesis of the test is that the two time series have no cointegration. From Table 3, the p-value is less than 0.05 for the return series and greater than 0.05 for the price series. Therefore, there is no sufficient evidence to reject the null hypothesis on the price time series, meaning that the prices of two ETFs are not cointegrated. On the other hand, the return series are cointegrated, indicated by the minimal p-value.

	test statistic	p-value
Time series		
Price	-2.087	0.483
Return	-17.313	0.000

Table 3: Engle-Granger cointegration test