Statistical Analyses of ETFs
An in-depth comparison of SPY and QQQ ETFs

Anantajit Raja

February 2024

1 Introduction

This paper aims to compare two exchange traded funds (ETFs) through in-depth visualisations and statistical analyses of time series data.

Why trade ETFs? As the name suggests, ETFs are a basket of securities, i.e. an investment fund, that are traded on exchanges. They provide easy market accessibility and are generally less risky/costly than directly investing in traditional stocks and bonds (unless you know the company really well). They can be used to build a diversified portfolio, and a relatively cheap one at that.

Why compare two ETFs? Comparing ETFs involves looking at returns, volatility, a bunch of other statistical properties, and seeing how each ETF performs during market downturns or upswings. This allows investors to identify which funds best complement their existing portfolio to achieve diversification, hence enabling them to choose the ETFs that best match their investment and risk management strategies.

2 Time Series

For this comparison we use the exchange-traded funds 'SPDR S&P 500 ETF Trust' and 'Invesco QQQ Trust' over the 439-day period from November 18th 2022 to January 31st 2024 which has 300 data points. The ETFs trade under the symbols 'SPY' and 'QQQ' respectively. SPY and QQQ represent different sectors and industries, with SPY tracking the broader S&P 500 index and QQQ focusing more on technology and innovation companies [Zhu and Bao(2019)]. Analyzing these ETFs over this period could reveal sector-specific trends, such as the tech sector's performance relative to the broader market.

We download real-time historical data using the Application Programming Interface (API) provided by Alpha Vantage. It's highly beneficial to download data using APIs because it can provide real-time, up-to-date data enabling us to make timely investment decisions. It's also automated and so it saves time since we won't need to manually download the data every time we want to access its latest version. Consistency and reliability are also ensured given that we get this data from a reputable source.

We store the API key for future use. For now we use it as an input for our TimeSeries object, specifying that we want the output to be a pandas file (to allow for ease of analysis) rather than the default json output. We first try to get a broad overview of the statistical properties of these ETF prices (initially disregarding the time element) by using the describe function from the pandas Python package.

Table 1: Descriptive Statistics of SPY and QQQ Closing Prices

Statistic	SPY	QQQ
Count	300	300
Mean	427.6	344.1
Standard Deviation	27.5	43.5
Minimum	376.7	260.1
25% Quantile	406.0	305.8
Median (50%)	427.2	357.3
75% Quantile	448.8	375.3
Maximum	491.3	428.15

Note that SPY has a higher mean closing price (427.6) compared to QQQ (344.1). This suggests that SPY, on average, traded at a higher price level than QQQ during the period. QQQ exhibits a higher standard deviation (43.5) than SPY (27.5), indicating that QQQ's prices were more volatile, which can imply greater risk but also potential for higher returns for QQQ. Its greater volatility might reflect its tech-heavy composition, as technology stocks tend to have higher price fluctuations [Sadorsky(2003)].

Now incorporating the time element, we plot our two price time series (that are their closing prices) on the same graph for a common time window of T = 300 (we'll use the 300 data points until the end of January for this analysis). We get the plot in **Figure 1**.

From the plot we can deduce through visual inspection that the general trend for both ETF prices is upward, making them seemingly profitable investments. Let's look into some more interesting statistical analyses to confirm this...

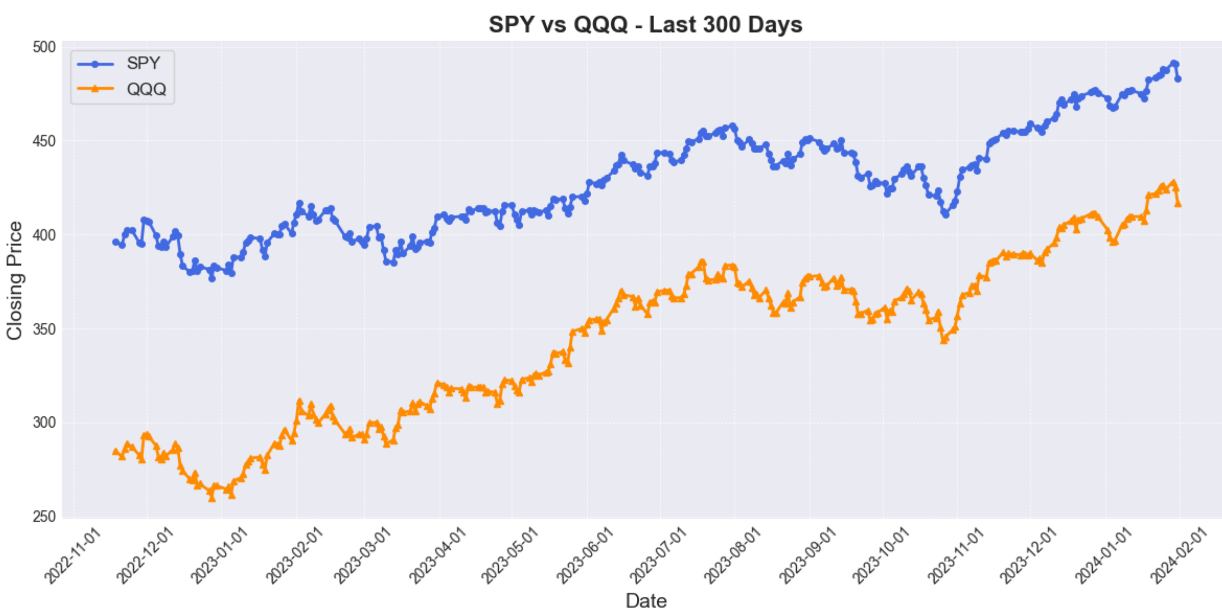

Figure 1: Comparison of SPY and QQQ ETFs over the last 300 business days up until the end of January. Closing price measured in dollars.

3 Moving Averages

To get a better idea of the longer-term trends in our data, we compute and plot the moving averages of our price time series data for both ETFs. In practice, we could potentially use the intersection of the price crossovers of the current price and these moving average prices as signals that indicate when we should buy/sell the ETFs [Huang and Huang(2020)] (when the current price drops below the moving average we sell and when the price goes above the moving average we buy, since these suggest potential upward and downward trends respectively).

Definition 1. *(Moving average) The moving average of a price time series with window τ is the average of its current price and its τ - 1 neighbours in the past. Mathematically, let $P(t)$ be the price at time t and τ our chosen time window. The moving average at time t is computed as the mean of a quantity over the last τ periods, namely:*

$$MAv\ (t) = \frac{1}{\tau} \sum_{i=t-\tau+1}^{t} P(i)$$

We try to find successively longer-term trends in both our ETF price time series by using moving averages of window lengths $\tau = 5, 20, 60$ business days, noting that the longer the time window used for the moving average the less sensitive the data is to recent price movements. We get the plots in **Figure 2**.

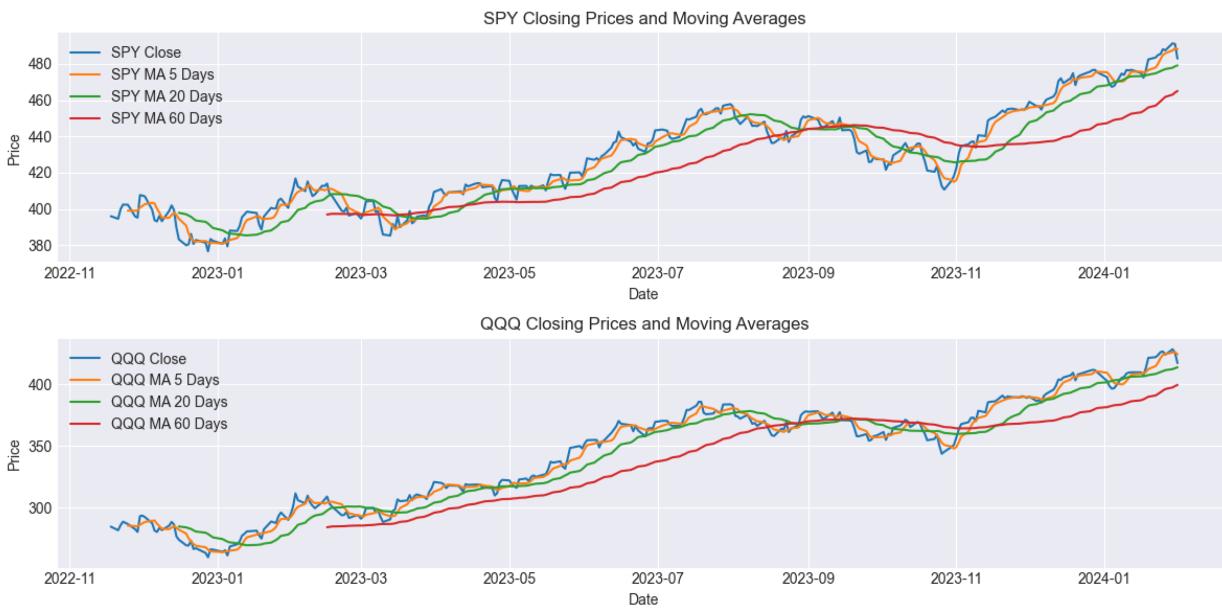

Figure 2: Moving averages of SPY and QQQ ETFs over the last 300 business days up until the end of January. Price is the closing price measured in dollars.

Note that for an n-day moving average we need n data points, which is why the moving average curves start on the n^{th} day. To make a simple prediction of future prices we could use the moving average as an indicator. For more short term predictions we could use the 5-day moving average trend, and for longer term predictions perhaps we could use either the 20-day or 60-day average depending on how far into the future we want to predict. We can see that the longer the time window used for the average, the smoother the curve.

Going deeper into our analyses of the time series, we can encapsulate the actual relative profit or loss one may have incurred when holding the ETF in a given time interval using a linear return time series.

Definition 2. *(Linear return) The linear return is defined as*

$$r(t) = \frac{P(t) - P(t-1)}{P(t-1)}$$

where $P(t)$ is the current value of the price time series and $P(t-1)$ is its value one time step before that. Linear returns represent the percentage change in the value of an investment over a period.

Alternatively one can use log-returns for analysing returns over multiple periods. They can be easier to use due to their additive property (you can simply add log returns to get the cumulative log return) making them more natural for compounding compared to linear returns. Log returns are also better than linear returns for statistical modelling under the assumption of normality.

Definition 3. *(Log return) The log return is defined as*

$$r_{log}(t) = log\left[\frac{P(t)}{P(t-1)}\right]$$

where $P(t)$ is defined as before.

Computing and plotting the linear return and log return time series on the same graph for both ETFs we get the plot in Figure 3. The linear and log returns are both derived from the same underlying changes in price so they tend to exhibit similar movements as can be seen on our plot. Mathematically this is a result of the approximation $ln(1+x) \approx x$ where $x = P(t)/P(t-1) - 1$ (this can be proved using a Taylor expansion, it works for x small) [Engsted et al.(2012)]. Daily data points do not exhibit significant differences between linear and log returns, making them appear almost identical on a chart. This phenomenon is particularly noticeable when the price changes are relatively small, as is often the case with daily market movements. Note in this paper we will use log returns when analysing statistical properties (they have almost the same underlying pattern as linear returns but are better for modelling).

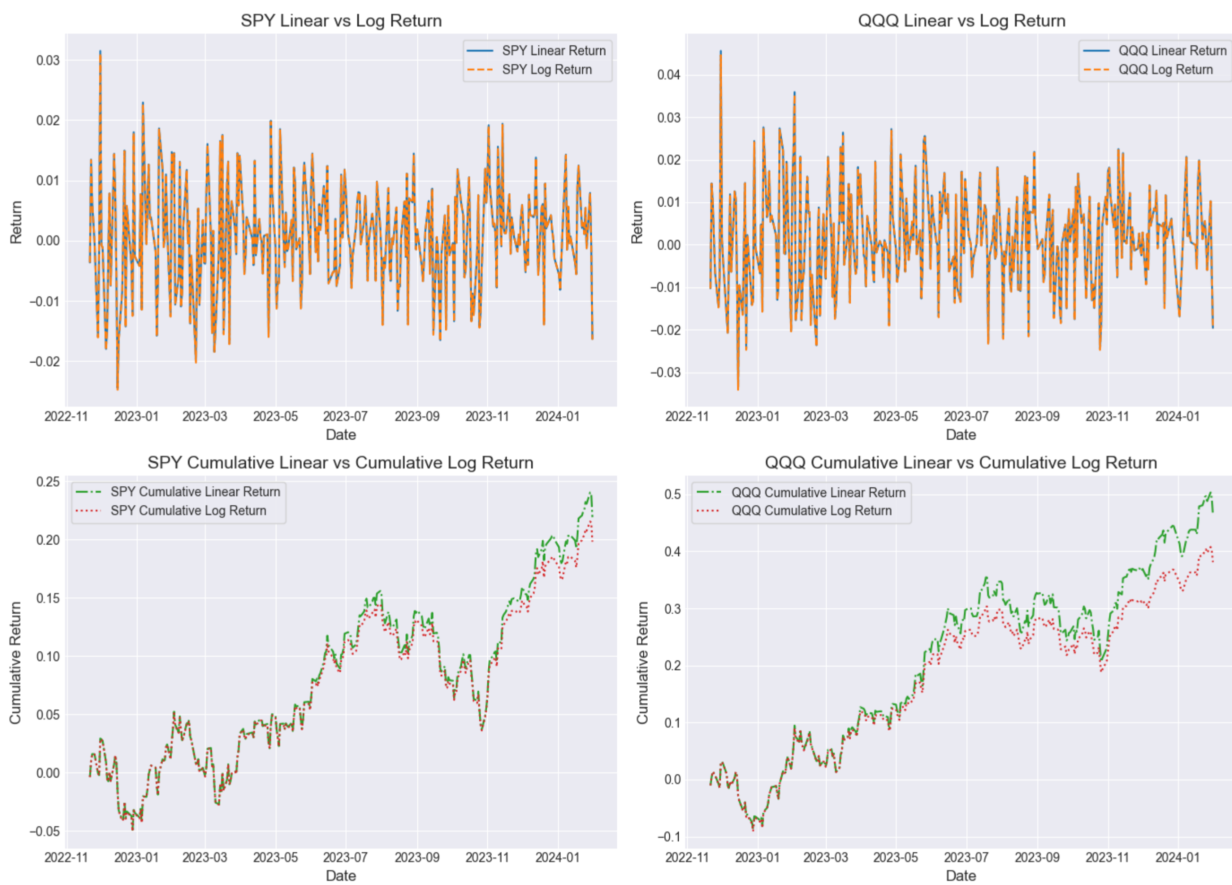

Figure 3: Linear, Log and Cumulative Linear and Log Returns of SPY and QQQ ETFs over the last 300 business days up until the end of January.

Note that returns on both ETFs generally oscillate about the zero value and seem to be mean-reverting, indicating stationarity (though we will formally confirm the stationarity of the returns data later in our stationarity tests).

We also included a plot for both cumulative linear returns and cumulative log returns to see how profitable each investment would be if we held the ETFs over the 300 day period. These plots are known as 'account curves'. Note that they resemble the plots for the prices of the ETF time series over time - the account curve quite literally tells you how much return you will earn for your account given that you invest 1 dollar at the start of the period. It seems that investing in both ETFs is quite a profitable strategy, though QQQ seems to be more profitable with a cumulative return of 0.5 (meaning you earn 50% on your original investment) compared to SPY which has a return on investment of about 23% over the period.

Now that we have an idea of how profitable an investment in each ETF could be, we try to see if there's any correlation between their current and past values.

4 Correlation analysis

We can use correlation analysis on both ETF price and return time series to see how the time series is related to its past values. In particular this can help capture seasonality, temporal structures of the data that may not have been captured using the moving average, how much 'memory' each time series has etc.

Definition 4. *(Autocorrelation) Autocorrelation is a measure of the degree of similarity between a time series and a lagged version of itself over successive time intervals. Let X_t be a time series. For each lag k, it calculates the sum of the product of the deviations of X_t and X_{t-k} from the mean $\bar{X}$ and normalizes it by the sum of squared deviations of X_t from the mean:*

$$R(k) = \frac{\sum_{t=k+1}^{T}(X_t - \bar{X})(X_{t-k} - \bar{X})}{\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

where T is the number of observations of the time series.

Significant autocorrelations at specific lags can indicate seasonality in the data. For example, a strong autocorrelation at a lag of 30 days in the data may suggest a monthly seasonal pattern. Moreover, the rate at which the autocorrelation values decrease as the lag increases can indicate the "memory" of the process. A slow decay in autocorrelation suggests a long memory, meaning past values have a prolonged effect on future values. Plotting the price ACF for both ETFs we get **Figure 4**.

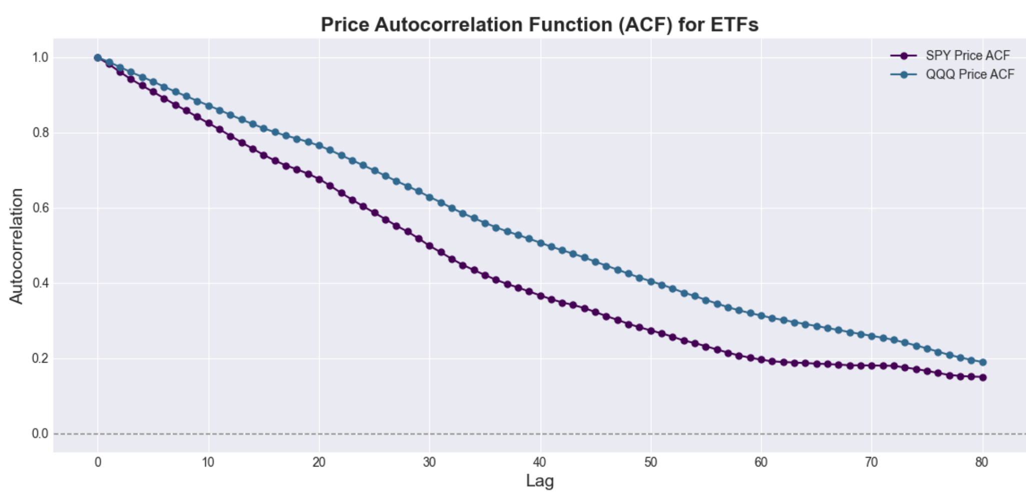

Figure 4: Autocorrelation functions of SPY and QQQ ETFs over the last 300 data points up until the end of January for different lags k

Note that the price ACF decreases fairly slowly with lag for both ETFs and autocorrelation doesn't even reach zero even for a large lag of 80. This shows that prices are indeed autocorrelated and hence have long memory. We can try to narrow down exactly how prices at different lags are directly related through analysis of the partial autocorrelation function (PACF).

Definition 5. *(Partial Autocorrelation) Partial autocorrelation measures the direct relationship between an observation and its lag, removing the effect of intermediate lags. This provides a clearer picture of the relationship between observations at different times. Given a time series X_t, the partial autocorrelation of lag k, denoted $\phi(k)$, is the autocorrelation between X_t and X_{t+k} with the linear dependence of X_t on X_{t+1} through to X_{t+k-1} (inclusive) removed:*

$$\phi(1) = corr(X_{t+1}, X_t) \text{ for } k = 1$$

$$\phi(k) = corr(X_t - \hat{X}_t, X_{t+k} - \hat{X}_{t+k}) \text{ for } k \geq 2$$

where $\hat{X}_t$ and $\hat{X}_{t+k}$ are linear combinations of the smaller lags that minimise the mean squared error with respect to X_t and X_{t+k}

The PACF is especially useful for identifying the order of the autoregressive component in time series models. A sharp cut-off in the PACF after a certain lag suggests the value of that lag as the order of the autoregressive (AR) process. PACF can uncover the underlying dynamics of the data by revealing the extent of direct linear relationships across time, independent of intervening values. Plotting the price PACF both ETFs we get **Figure 5**.

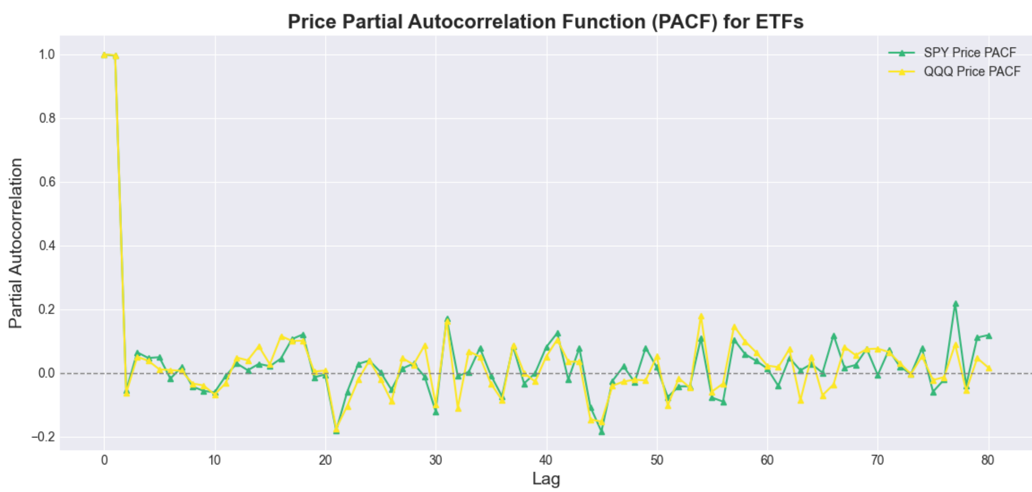

Figure 5: Partial autocorrelation functions of SPY and QQQ ETFs over the last 300 data points up until the end of January for different lags k

Now looking at the price PACF for both ETFs, we can see that there's a sharp cutoff after the first lag - a strong indication that this is an AR(1) process. This means that the current value of the price time series is highly dependent only on its first previous value and not on the values before that (it's a Markov process).

We can also try to see if there's autocorrelation or partial autocorrelation amongst the return series for both ETFs. Plotting the return ACF and PACF for both ETFs we get **Figure 6**.

Given the immediate sharp cutoff for both SPY and QQQ, neither autocorrelation nor partial autocorrelation seem to be present in either ETF return time series. In addition to this, both ACF and PACF seem to oscillate with a small amplitude about the zero line indicating that the return time series seem to be purely random (or follow a random walk if you will). This supports the fact that linear autocorrelations of asset returns are often insignificant, aside from very small intraday time scales for which market microstructure effects play a bigger role [Cont(2001)]. Given that each return value is independent of every past return value, returns of the ETFs are essentially unpredictable using the historical data.

We next try to assess the normality and stationarity of the historical log return data for both ETFs to gain better insights into the underlying distribution at play here.

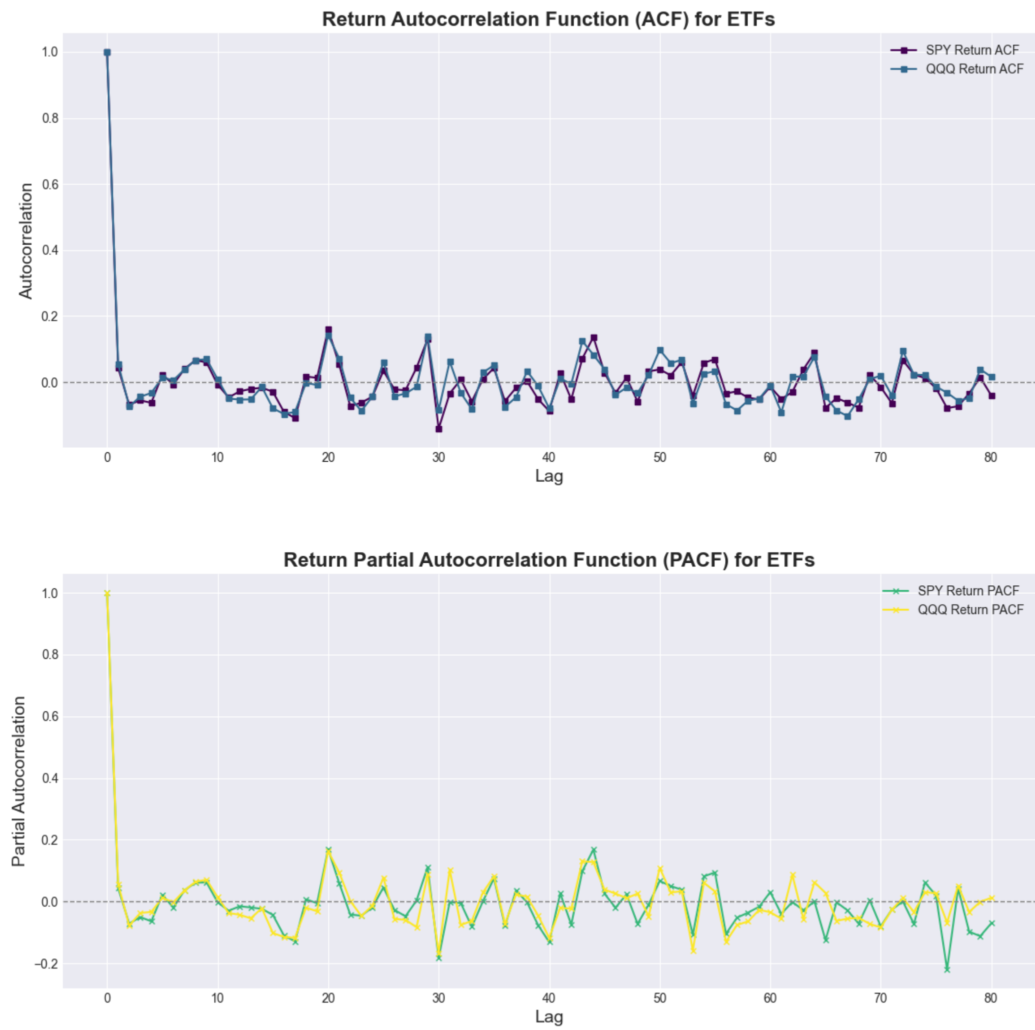

Figure 6: ACF and PACF of SPY and QQQ ETF log returns over the last 300 business days up until the end of January for different lags k

5 Gaussianity and Stationarity Tests

A Gaussianity test is used to determine whether a dataset $\{x_i\}_{i=1}^n$ follows a normal distribution $N(\mu, \sigma^2)$ with a given mean and variance. We use here a Kolmogorov-Smirnov Distribution Test to see if there's evidence to support the fact that returns follow a normal distribution.

Definition 6. *(Kolmogorov-Smirnov Distribution Test) The Kolmogorov-Smirnov (KS) test is a non-parametric test that measures the goodness of fit between a sample distribution and a reference probability distribution (i.e. whether $\{x_i\}_{i=1}^n$ was sampled from a given cumulative distribution function (CDF) $F(x)$). Given we define the empirical CDF $F_n(x)$ according to*

$$F_n(x) = \sum_{i=1}^{n} \theta(x - x_i) \text{ where } \theta(x - x_i) = \begin{cases} 1 & \text{for } x \geq x_i \\ 0 & o/w \end{cases}$$

and the Kolmogorov–Smirnov statistic for a given cumulative distribution function F(x) is:

$$D_n = sup|F(x) - F_n(x)|$$

which converges to zero if the variables $\{x_i\}_{i=1}^n$ are indeed sampled from $F(x)$, the null and alternative hypotheses of the KS test are:

H_0: the variable $\sqrt{n}D_n$ is compatible with the supremum of a Brownian bridge (a Brownian motion with constrained initial and final conditions), and follows the Kolmogorov distribution P(K)

H_1: $\sqrt{n}D_n$ does not follow P(K)

The null hypothesis is rejected at significance level α if $\sqrt{n}D_n > K_\alpha$ where K_α is the test statistic we use for the KS distribution i.e. when the value of $\sqrt{n}D_n$ is larger with respect to a value of K such that $P(K < K_\alpha) \leq \alpha$ (valid for large n too).

In our example for a test of normality of returns, we apply the Kolmogorov-Smirnov Test with a highly accurate library function approximation for the normal distribution CDF. We get the results

		SPY	**QQQ**	**SPY conclusion**	**QQQ conclusion**
Gaussianity test (KS test) on returns	**Statistic**	0.04	0.032	Gaussian	Gaussian
	p-value	0.706	0.917		

We get that since the p-value isn't small enough for either SPY or QQQ at the 5% level, there isn't enough evidence to reject the null hypothesis that the returns distributions is normal. Usually we expect to see fat-tailed returns in a financial market [Cont(2001)], but that doesn't seem to be the case here. There could be several reasons for this: the period of our data has not experienced many significant market events or volatility, the choice of daily data frequency (higher frequency data might tend to show more fat tails), or maybe if we looked at data over a longer period it might seem that the distributions for both ETFs may be more heavy-tailed. However we omit this further analysis of returns distributions for the sake of brevity of the paper.

Now onto the stationarity analyses of the time series:

Definition 7. *(Strong Stationarity) Strong stationarity is such that every joint-distribution is time-invariant i.e. $(y_1, y_2, ...y_n)$ is equal in distribution to $(y_{1+\tau}, y_{2+\tau}, ...y_{n+\tau})$ where τ is a particular length of time.*

We can use the Augmented Dickey-Fuller test [Cheung and Lai(1995)] to see if a time series has a unit root (i.e. if any of the solutions of the characteristic equation is 1). If such a root is present, this would be indicative of non-stationarity (and absence will indicate strong stationarity). A unit root indicates that the time series will have a persistent change after a shock (since the lag B becomes 1 where $y_t = By_{t-1}$), meaning that it will not revert to a long-term mean, violating the stationary condition. (Note the definition of weak stationarity too for some background info: if first and second order moments of the time series are time-invariant the time series is weakly stationary). We introduce a unit root test (which is a test for stationarity):

Definition 8. *(Augmented Dickey-Fuller Test) Given a time series X_t and the model*

$$\Delta X_t = a + \rho X_{t-1} + c_1 \Delta X_{t-1} + c_2 \Delta X_{t-2} + ... + \epsilon_t$$

where Δ is the difference operator, a, ρ, c_1, c_2 etc. are constants, and given the test statistic used for the ADF test

$$\frac{\hat{\rho}}{SE(\hat{\rho})} \text{ where } \hat{\rho} \text{ is the estimate of the } X_{t-1} \text{ coefficient and SE its the standard error}$$

our null and alternative hypotheses are

H_0: $\rho = 0$ (giving unit root, implying non-stationarity)

H_1: $\rho \neq 0$ (which would imply stationarity).

The extremity of this test statistic is evaluated by comparing the statistic to critical values for the ADF test. If the test statistic is more extreme than the critical value (i.e. more negative), the null hypothesis of the presence of a unit root is rejected, suggesting evidence that the series is stationary (equivalently we can obtain a very low p-value under significance level α to reject the null hypothesis since this means that the observed data would be very unlikely under the assumption that the null hypothesis is true)

		SPY	QQQ	SPY conclusion	QQQ conclusion
(prices) ADF test for unit root	**Statistic**	-0.544	-0.461	There is a unit root	There is a unit root
	p-value	0.883	0.899		

		SPY	QQQ	SPY conclusion	QQQ conclusion
(returns) ADF test for unit root	**Statistic**	-16.362	-16.199	There is no unit root	There is no unit root
	p-value	∼0	∼0		

Presense of a unit root implies that the process in question is integrated of order 1 ($I(1)$) whilst its absence implies $I(0)$. Using the ADF test on the price time series for both ETFs we get that the p-value is 0.883, suggesting that the evidence is not strong enough to support both ETF price time series being stationary at the 5% level (cannot reject H_0). This implies non-stationarity of the price series. Using the ADF test on the return time series for both ETFs we get that the p-value is 0 (or extremely small), suggesting that there's strong evidence to support both ETF return time series being stationary at the 5% level. Now onto seeing how series that fall under the same category are possibly related in the long-term:

6 Cointegration Tests

One can apply cointegration tests to see if there's a long-term relationship between two time series that are individually non-stationary. In general it's used to help make informed decisions based on the underlying dynamics that govern the movements of pairs or groups of time series.

Definition 9. *(Cointegration) Two time series are said to be cointegrated if a linear combination of these series results in a stationary time series, i.e. they are cointegrated if a linear combination of them has a lower level of integration.*

Cointegration ensures that any deviation from the equilibrium is temporary. Mathematically it entails the ability to find a coefficient, i.e. a cointegrating vector, $\beta = (\beta_1, \beta_2, ..., \beta_n)$ such that the linear combination

$$\beta Y_t = \beta_1 X_{1_t} + \beta_2 X_{2_t} + ... + \beta_n X_{n_t}$$

of non-stationary series $\{X_{1_t}, X_{2_t}, ..., X_{n_t}\}$ is stationary (ensuring that the linear combination does not randomly wander over time). Informally, cointegrated series describe 'shared forces' that ensure the series are not too far adrift.

In our example, we found strong evidence (through the ADF test) to support the fact that the price series have a unit root, implying the non-stationarity of the price series. We try to see if a linear combination of these is stationary using the theory outlined below, which would in turn imply that the series are cointegrated.

Cointegrating vectors aren't unique, so to make them interpretable, we set rules such as picking one of the β_j's to have a value of 1. WLOG we set $\beta_1 = 1$ and take the negative of the other coefficients, giving for our earlier system (called 'common normalisation'):

$$\beta Y_t = X_{1_t} - \beta_2 X_{2_t} - ... - \beta_n X_{n_t}$$

which is $I(0)$ if they are cointegrated. The reason we choose this common normalisation is that it can be written conveniently in regression form:

$$X_{1_t} = \beta_2 X_{2_t} - ... - \beta_n X_{n_t} + \gamma_t$$

where γ_t is some residual, which if stationary, models short-lived deviations in the long term. This idea is used for the Engle-Granger test for cointegration [Engle and Granger(1987)].

Definition 10. *(Engle-Granger Test) If y_{1t} and y_{2t} are each I(1) and we find a suitable θ, e.g. with OLS, such that $z_t = y_{1t} - \theta y_{2t}$, then a unit-root test on z_t can tell us if y_{1t} and y_{2t} are cointegrated. The absence of a unit root will indicate stationarity of the residual, and hence that y_{1t} and y_{2t} are cointegrated.*

H_0 : *The residual series* z_t *has a unit root*
H_1 : *The residual series* z_t *does not have a unit root*

Note that this just uses the common normalisation applied to a beta of length $n = 2$ and the residual we test here would be z_t.

We use the Engle-Granger test on both our ETF price time series and get the results:

			Conclusion
Engle Granger Test for cointegration	**Statistic**	-2.082	The residual series has a unit root
	p-value	0.252	

It turns out that the p-value is too big to be significant at the 5% level so we fail to reject the null hypothesis, meaning that the evidence is not strong enough at the 5% level to suggest that the residual series is stationary, thereby implying that the ETF price time series are *not cointegrated*.

The SPY and QQQ ETFs are fundamentally different in their market focus, which likely impacts their potential for being cointegrated. SPY tracks the S&P 500, which offers exposure to the stock market across many different sectors in the US, reflecting the overall market performance. In contrast, QQQ is much more weighted towards the technology sector, tracking the Nasdaq-100 Index. This sector-specific focus subjects QQQ to different market drivers than SPY, such as technological innovations or regulatory changes affecting tech companies. While both ETFs may move together in response to general market trends in the short-term, their long-term movements are influenced by distinct factors. These together contribute to the fact that SPY and QQQ prices are not cointegrated.

Applying the cointegration test to returns is not be appropriate since it's primarily designed for non-stationary series, whereas returns are typically stationary, as we saw earlier. Given that the returns series are already stationary to begin with, the process essentially finds that the residuals are stationary as well, leading to the rejection of the null hypothesis of the ADF test (indicating no unit root is present) and falsely suggesting cointegration. This misuse arises because cointegration is a concept meant to detect a kind of "hidden" stationarity in relationships between non-stationary series, not to analyse series that are already stationary on their own. Here's a table illustrating its misuse for the returns series:

			Conclusion
Engle Granger Test for cointegration	**Statistic**	-17.315	The residual series has no unit root
	p-value	~ 0	

This would imply at the 5% level that the returns series are *cointegrated*, however this result has no meaning because the conditions for using the cointegration test were not satisfied in the first place (the returns time series were stationary already).

7 Possible further analyses

Our cross-comparison of SPY and QQQ ETFs provides a foundational understanding of their market behavior for the given period. To build upon this groundwork, further research could significantly benefit from exploring more advanced methodologies. Volatility modeling, particularly through GARCH models, would offer deeper insights into the risk dynamics inherent to these ETFs, highlighting periods of heightened market uncertainty. Additionally, employing machine learning techniques could uncover non-linear patterns and predictive relationships within the data. These factors together can be used to refine investment strategies in these ETFs by accurately identifying periods of opportunity and risk.

References

[Cheung and Lai(1995)] Y.-W. Cheung and K. S. Lai. Lag order and critical values of the augmented dickey–fuller test. *Journal of Business & Economic Statistics*, 13(3):277–280, 1995.

[Cont(2001)] Rama Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001. doi: 10.1080/713665670.

[Engle and Granger(1987)] Robert F. Engle and Clive W.J. Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987.

[Engsted et al.(2012)] Tom Engsted et al. The log-linear return approximation, bubbles, and predictability. *The Journal of Financial and Quantitative Analysis*, 47(3):643–665, 2012. URL `http://www.jstor.org/stable/41653563`. Accessed 16 Feb. 2024.

[Huang and Huang(2020)] Jing-Zhi Huang and Zhijian (James) Huang. Testing moving average trading strategies on etfs. *Journal of Empirical Finance*, 57:16–32, 2020. ISSN 0927-5398. doi: 10.1016/j.jempfin.2019.10.002. URL `https://www.sciencedirect.com/science/article/pii/S0927539819300830`.

[Sadorsky(2003)] Perry Sadorsky. The macroeconomic determinants of technology stock price volatility. *Review of Financial Economics*, 12(2):191–205, 2003. ISSN 1058-3300. doi: 10.1016/S1058-3300(02)00071-X. URL `https://www.sciencedirect.com/science/article/pii/S105833000200071X`.

[Zhu and Bao(2019)] Xiaoyu Zhu and Si Bao. Multifractality, efficiency and cross-correlations analysis of the american etf market: Evidence from spy, dia and qqq. *Physica A: Statistical Mechanics and its Applications*, 533:121942, 2019. ISSN 0378-4371. doi: 10.1016/j.physa.2019.121942. URL `https://www.sciencedirect.com/science/article/pii/S0378437119305126`.