

Project Proposal

Analyzing the Correlation Structure and Time-Series Behavior of Four Major ETFs (VOO, QQQ,
ARKK, MAGS)**

Authors: Connor, Ty, and Mike

Course: Data Science

Final Project Proposal

Introduction to what we are trying to research

Exchange-traded funds (ETFs) represent a significant component of contemporary investment strategy. They facilitate broad exposure across various asset classes, including the overall stock market, specific industry sectors like technology, or specialized investment methodologies. A fundamental analytical approach involves examining their correlation. Given that ETFs frequently exhibit co-movement due to underlying constituent holdings, prevailing market dynamics, or significant macroeconomic events, assessing their correlations provides crucial insights into portfolio diversification potential, systemic risk identification, and the primary drivers of return.

This project focuses on four widely followed ETFs:

- **VOO** – Vanguard S&P 500 ETF (broad U.S. market benchmark)
- **QQQ** – NASDAQ-100 ETF (mega-cap tech concentration)
- **ARKK** – ARK Innovation ETF (actively managed, high-volatility, disruptive innovation strategy)
- **MAGS** – Roundhill Magnificent 7 ETF (seven largest technology companies driving market returns)

These ETFs, offering varied diversification and growth exposure, are key for analyzing co-movement, risk clustering, and changing correlations across economic cycles. Grasping their interconnected movement, particularly during macro uncertainty, is vital for portfolio and risk management. We will explore static and time-varying correlations, preparing for future modeling with methods like VAR, rolling-window analysis, and volatility modeling.

- **Static Correlations:** Analyzing the relationships between variables over the entire data set, assuming the correlation structure remains constant across the time period.
- **Time-Varying Correlations:** Investigating how the relationships between variables change over time, acknowledging that market conditions, economic factors, or other influences can alter these relationships.
- **VAR (Vector Autoregression):** A time-series model used to capture the linear interdependencies among multiple time series. It is often used for forecasting and impulse response analysis.
- **Rolling-Window Analysis:** A method of repeatedly estimating a model or statistic (like correlation) on a fixed-size moving window of data, allowing for the observation of how the results change over time.
- **Volatility Modeling:** Techniques (such as GARCH) used to model and forecast the changing variance (volatility) of a financial time series, which is crucial for risk management and asset pricing.

Background

The core concept for this project originated from a simple observation: examining the correlation between silver and gold, metals whose value is clearly linked due to their status as some of the most precious on Earth. Furthermore, during periods of heightened fear of war, these valuable metals consistently experience a spike in value. This project was a team effort; we thought of combining a computer science major's skills with the insight of two business majors. It could make for an interesting proposal, especially looking into Quant statistics to make good insights into their personal stock investments.

Motivation and Examples

Financial markets are interconnected systems in which assets rarely move independently. Traditional diversification theory suggests that holding a combination of stocks or ETFs with low correlations reduces risk. However, correlations are not constant over time, especially during market turbulence. For example:

- During the COVID-19 crash (March 2020), cross-asset correlations sharply increased.
- During the inflation-driven drawdown of 2022, growth-heavy ETFs like ARKK diverged from broad-market ETFs such as VOO.
- In 2023–2024, the “Magnificent 7” stocks dominated index returns, influencing the behavior of QQQ and MAGS.
- During the 2024 election and the volatility of stocks and ETFs based on who was going to win the election

These dynamics motivate our research questions. Since **VOO**, **QQQ**, **ARKK**, and **MAGS** represent distinct market segments, they allow us to investigate:

- How diversification breaks down or strengthens in different regimes
- How innovation-focused ETFs behave compared with large-cap tech
- Whether structural shifts in the market lead to persistent correlation changes

This analysis also prepares us for a future capstone project centered on forecasting, factor modeling, or machine learning applications to ETF behavior.

Methods

This section outlines the core statistical methods relevant to our analysis. Some methods will be used directly in this project; others are included because they provide the theoretical foundation for possible expansions in the capstone.

Correlation and Covariance Analysis

We will begin by computing **pairwise Pearson correlations** among the four ETFs:

- Correlation between VOO and QQQ
- Correlation between QQQ and MAGS
- Correlation between QQQ and ARKK
- Correlation between ARKK and the others

Example:

```
wide_prices <- prices_tbl |>
  select(symbol, date, adjusted) |>
  pivot_wider(names_from = symbol, values_from = adjusted)

correlation <- cor(wide_prices$QQQ, wide_prices$ARKK, use = "complete.obs")

correlation
```

This gives a baseline understanding of co-movement.

We will also compute:

- **Spearman correlations** (robust to non-normality)
- **Covariance matrices** to measure joint variability

Rolling Window Correlations

Because financial correlations are **time-varying**, we will compute rolling correlations over windows such as:

- 30-day
- 60-day
- 90-day

These rolling analyses allow us to identify **regime shifts**, including:

- Market stress periods (e.g., 2020, 2022)
- Divergence between ARKK and QQQ during innovation bubbles
- Convergence between QQQ and MAGS due to tech dominance

Portfolio-Theoretic Metrics

Using tools from *PerformanceAnalytics*, we will calculate:

- Daily log returns
- Annualized volatility
- Sharpe ratios
- Betas of each ETF with respect to VOO

These measures help interpret correlation patterns in the context of risk and return.

R Packages

A critical component of this proposal is identifying R packages that facilitate data retrieval, time-series manipulation, risk analysis, and visualization. Below, we summarize the packages most relevant to our research. The required R packages for this analysis include quantmod, tidyquant, BatchGetSymbols, xts, zoo, PerformanceAnalytics, corrplot, ggplot2, plotly, vars, rugarch, dplyr, tidyr, tibble, and readr. These packages cover a range of financial data retrieval, manipulation, time series analysis, visualization, and general data wrangling capabilities essential for quantitative analysis and modeling in R.

Data Acquisition Packages

quantmod

The primary tool for downloading ETF price data directly from Yahoo Finance.

- `tq_get()` retrieves historical prices
- Automatically structures data as an xts time series

```
StocksETF <- c("VOO", "QQQ", "ARKK", "MAGS")  
  
prices_tbl <- tq_get(StocksETF,  
                      from = "2015-01-01",  
                      to   = Sys.Date())
```

BatchGetSymbols

Efficiently retrieves many tickers at once, handling missing values and alignment issues.

Time Series and Modeling Infrastructure

xts and zoo

- XTS(eXtensible Time Series) This project uses fundamental data structures, primarily based on the object, for financial analysis. The key benefit is treating time-series data like a matrix but with a date-based index. This ensures automatic date-alignment for operations like merging and applying rolling calculations, simplifies managing multiple assets (e.g., VOO, QQQ), and

handles date-related issues (missing dates, weekends) robustly. It integrates seamlessly with popular financial packages like quantmod.

-zoo object (Zero-Observation Object) is a specialized data structure in R for time-series data. It stores data with an ordered index, typically dates, enabling calculations like moving averages (rolling-window calculations) while ensuring the data remains correctly aligned by its index. It's fundamental to many financial time-series analyses in R.

PerformanceAnalytics

The most important package for our project.

Includes:

- Return.calculate()
- chart.Correlation()
- rollapply() for rolling Sharpe, vol, and correlation

Think of the PerformanceAnalytics package as the ultimate R toolkit for financial analysis. Sure, R's built-in functions can handle simple tasks like calculating average return or standard deviation in a pinch. But when you move to sophisticated calculations like a rolling Sharpe Ratio, generating a comprehensive correlation plot using chart.Correlation(), or deep-diving into tail risk with Value-at-Risk (VaR) the standard functions quickly become complex and require extensive custom coding. PerformanceAnalytics solves this by packaging all those complex, standardized financial calculations into one cohesive place. It saves you the major effort of writing and quality-checking custom code, instead providing reliable, industry-standard metrics immediately.

corrplot

A visualization package specifically designed to get correlation metrics. In this project, it will help us see which ETFs are tightly linked and compare correlations over time

vars

Implements Vector Autoregression models, which are multivariate time series models that get the returns of our ETFs and their regressions. It helps us answer questions like "Does past ARKK returns help predict QQQ?"

rugarch

Used for univariate GARCH modeling that helps us model the conditional variance of return. Helps us see volatility clustering and compare volatility regimes across our chosen ETFs

Visualization Packages

ggplot2

The primary tool for custom and publication-quality visualizations.

plotly

Adds interactivity for presentations.

Datasets

The central dataset for this project will be daily historical prices for these four massive and liquid US ETFs:

- **VOO** (Vanguard S&P 500 ETF) is a broad US large-cap benchmark that tracks the S&P 500 index, which tracks roughly the 500 largest companies that are publicly traded in the US across all sectors. The VOO's top 10 holdings cover over 40% of their total assets. VOO is very tech-dominant in their top holdings, for example, within their top 10 holdings are NVDA (8.47% of total assets), AAPL (6.88%), MSFT (6.60%), AMZN (4.06%), AVGO (2.98%), GOOGL (2.80%), META (2.41%) GOOG (2.26%), TSLA (2.20%), and BRK-B (1.5%). As you can see the most prominent holdings within VOO are a part of the dominant Magnificent 7 tech companies that have been holding the market over throughout this volatile market. From a pure data sense, VOO is actually a clean proxy for “the U.S. stock market” within our sample because the daily stock prices and the Adjusted Close data series reflect the overall market patterns, not those of specific sectors, so it is the perfect “anchor” stock when examining correlations between exposure to the overall market and growth-focused ETFs.
- **QQQ** (NASDAQ-100 ETF) is a tracker of the Nasdaq-100, which is very heavily tilted to large cap growth and tech stocks. Similar to the VOO the top 10 holdings of the QQQ that represent 53% of the indexes total assets are very similar and very Mag& dominant. The top 10 holdings are NVDA (9.09%), AAPL (8.75%), MSFT (7.73%), AVGO (6.63%), AMZN (5.26%), GOOGL (3.94%), GOOG (3.68%), TSLA (3.31%), META (2.97%), NFLX (2.38%). QQQ has 100-102 holdings, an expense ratio of around 0.20%, and very large assets, which makes the daily data of price and volume very reliable. In our project, QQQ acts as a “mega-cap growth/tech benchmark” that lies between the level of diversification of VOO and the concentrated level of exposure of the remaining funds, which give us the chance to examine the correlations for the different levels of concentration of tech and growth.
- **ARK Innovation ETF (ARKK)** is basically an actively managed fund that's all about disruptive innovation—think artificial intelligence, gene editing, robotics, and digital platforms. Unlike funds that just track an index (like VOO or QQQ), ARKK's holdings are driven by the manager's conviction and include high-growth, high-volatility names such as TSLA, ROKU, CRSP, and SQ (Block). This ETF is seriously volatile, known for its big

drops and sharp rebounds, which makes it a fantastic subject for studying how correlations break down when the market gets stressed. ARKK often moves quite differently from standard tech indices, giving us some great insight into how innovation-focused assets behave compared to the overall market.

- **MAGS** (Roundhill Magnificent 7 ETF) exclusively holds the “Magnificent 7” mega-cap technology companies: AAPL, MSFT, NVDA, META, AMZN, TSLA, and GOOG. Because these companies have accounted for a large portion of recent market gains, MAGS acts as a pure megacap tech exposure vehicle. It is highly concentrated, making it extremely sensitive to movements in just a handful of companies. This ETF is particularly useful for examining whether the dominance of these firms has caused traditional diversification to break down. By comparing MAGS with QQQ and VOO, you can analyze how concentrated tech leadership shapes correlation structures and systemic risk.

We will retrieve:

- Open, High, Low, Close, Adjusted
- Daily volume
- Returns (log and simple)

Data Source: **Yahoo Finance** via quantmod and tidyquant.

- VOO: <https://finance.yahoo.com/quote/VOO>
- QQQ: <https://finance.yahoo.com/quote/QQQ>
- ARKK: <https://finance.yahoo.com/quote/ARKK>
- MAGS: <https://finance.yahoo.com/quote/MAGS>

We may optionally incorporate macroeconomic time series from **FRED**, such as:

- **VIX (market volatility index)** is a daily measure of the implied volatility of S&P 500 index options over the next 30 days. It is often referred to as the market’s “fear gauge,” since the level of the VIX tends to soar when investors are fearful of big stock market movements. The data on the VIX begins on January 1st of 1990, and it is measured on a scale of annual percentage volatility (for example, when the level of the VIX is 20, it means that the S&P 500 will probably vary by about 1.25% on a single trading day, given that there are 252 trading days in the year). For our analysis, VIX will proxy for risk sentiment on a high frequency. It is possible to plot VIX on correlations between the ETFs to examine:

- Correlations between VOO, QQQ, ARKK, and MAGS will become stronger when VIX rises, implying that diversification effects will be weaker when fear is present.
 - Highly innovative or concentrated funds, such as ARKK and MAGS, have a higher sensitivity for VIX compared to other market funds, capturing their “risk-on” characteristic.
 - Because the VIX is reported daily, it aligns well with our data on the return of the ETFs, and it lends itself well to detailed descriptive analysis (graphs, state analysis) or very basic modeling.
- **10-year Treasury yield** is a daily data series that reveals the interest rate on U.S. Treasury securities with a 10-year constant maturity. It is derived by the U.S. Treasury by using the yields on outstanding securities adjusted to have a constant 10-year maturity, and is quoted in annual percentage yields (for example, 4.2% per year). In our analysis, we can:
 - Track the relationship between the co-movement of growth-weighted ETFs and VOO during low-rate regimes versus high-rate regimes.
 - Analyze if the correlation spikes are coinciding with steep moves in DGS10, for instance, during the rapid tightening cycles.
- **CPI inflation data** is a comprehensive measure of consumer price inflation within the US economy. It measures the changes that take place over time for the average price that urban consumers pay for a given basket of consumer goods and services, ranging from food and energy, to housing, transportation, and healthcare. This data is published on a monthly basis, starting from the middle of the last century. Firstly, the reason why CPI is less ideal for correlation analysis on a day-to-day basis compared to other variables, like VIXCLS and DGS10, is that it only comes out on a monthly basis:
 - We can make a distinction between the phases of high inflation and the phases of moderate/low inflation and examine how the behavior of the ETF varies between the phases.
 - Continued inflation often causes higher interest rates, hence a higher DGS10, which puts pressure on growth and innovation stocks.
- In the proposed project and possible capstone, the use of CPI will:
 - Describe the environment of inflation that existed throughout the sample period. For example, it was the period of inflation following COVID.
 - Provide context for the patterns that emerge when analyzing the correlations and volatility, for example, realizing that the time of high correlations also sees high CPI and high 10-year yields.

Data for the above FRED Metrics:

- VIX (VIXCLS): <https://fred.stlouisfed.org/series/VIXCLS>
- 10Y Yield (DGS10): <https://fred.stlouisfed.org/series/DGS10>
- CPI (CPIAUCSL): <https://fred.stlouisfed.org/series/CPIAUCSL>

Research Questions

Primary Questions

1. How strongly correlated are VOO, QQQ, ARKK, and MAGS over the last decade?
2. How do correlations change across market regimes (e.g., COVID crash, 2022 drawdown)?
3. Does ARKK behave differently from tech-heavy ETFs like QQQ and MAGS during periods of high volatility?
4. Does market stress increase or decrease co-movement among these ETFs?
5. Are the correlations stable, or do they exhibit structural breaks?

Expected Outcomes

By the end of the analysis, we expect to:

- Produce static and rolling correlation matrices for all ETF combinations.
- Identify periods of divergence between innovation-focused ETFs and broad-market ETFs.
- Understand the influence of technological concentration on ETF co-movement.
- Develop visualizations that clearly show changes in correlation over time.
- Establish a solid foundation for a capstone project involving forecasting or factor modeling.

Project Significance

This research is relevant because portfolio construction and risk management rely heavily on understanding how assets behave together. As the U.S. stock market becomes increasingly driven by a small set of mega-cap technology companies, the correlations among ETFs like QQQ and MAGS have grown more important for investors seeking diversification. Meanwhile, ARKK's volatility presents a compelling contrast, allowing us to examine how an actively managed innovation strategy correlates with more traditional indices.

The findings of this project could be useful not only academically but also for anyone interested in ETF investing, risk management, or quantitative finance.