**Do more than six stocks matter for the NASDAQ?**

**David Whitney (Mentor: Sameera Poduri)**

**Introduction and Background**

Hi my name is David Whitney. I’m a postdoctoral scientist at the Max Planck Institute for Neuroscience who recently took Springboard’s [Data Science Intensive](https://www.springboard.com/workshops/data-science-intensive/) course. As a self-taught data scientist, scientific inquiries traditionally drive what I learn in data science. But recently I wanted to fill-in conceptual gaps, and became attracted to Springboard’s Data Science Intensive course. What differentiates Springboard from many other data science courses is that the curriculum strikes a nice balance between teaching both *hard* (analytics) and *soft* (social) skills. While most online courses appropriately emphasize the analytic side (i.e. statistics, algorithms, etc.), other essential skills can often get overlooked, such as meticulously thinking about data storytelling, data visualization, and good practices for accessing/organizing data. This is where the comprehensive Springboard curriculum shines in my opinion, as the team at Springboard want to prepare students to navigate data science projects from start to finish.

For the capstone project, I had the pleasure to be mentored by Sameera Poduri, the head data-scientist at Amino Health. Early on, both Sameera and I decided to try something not related to my expertise in systems neuroscience, and instead challenge myself with an unconnected topic in finance. I was struck by a pretty recent Wall Street journal article that reported just six giant technology/ pharmaceutical-related companies accounted for over half of the NASDAQ’s growth back in 2015: Facebook, Amazon, Apple, Netflix, Gilead, and Google [1]. Not surprisingly, these six high-performing stocks have received considerable attention from business analysts [1-3]. Jim Cramer from Mad Money even went so far to dub four of these fast-growing NASDAQ stocks (Facebook, Amazon, Netflix, and Google) with the acronym ‘FANG’ to emphasize the privileged position that these companies enjoy [3].

While investors understandably are drawn to invest in popular large-cap stocks, I wondered how much growth in the NASDAQ Composite Index over the last six years was attributable just to these six stocks. Furthermore, as someone with a portion of my own 401k savings invested in these high-profile companies (and I suspect many other readers might too), I was motivated to develop the quantitative tools to examine whether stock growth in these companies is tightly interrelated together. While strong, consistent growth is always preferred, modern portfolio theory predicts increased risk associated with investments lacking diversification [4]. So, let’s look at some historical stock data to determine how risky these six companies might be as investments relative to other NASDAQ companies.

**Summary**

* For the last six years (2011-2017), FAANG stocks have shown extraordinary high growth rates as a group (+55.85±9.39% per year, n=6 stocks) relative to other Top-100 NASDAQ stocks (+26.33±3.86% per year, n = 95 stocks) and the bulk of ordinary NASDAQ stocks (+9.22±1.05% per year, n = 2,982 stocks).
* However, a closer inspection of individual NASDAQ stocks reveals that ~20% of ordinary NASDAQ stocks display growth rates comparable to the fast-growing FAANG stocks (65.87±4.14% per year, n = 555). These results suggest that recent growth in the stock market is not driven by a small handful of privileged stocks, but rather reflects a broader advance by American companies.
* Under a modern portfolio theory approach, an investment portfolio containing just FAANG stocks is a less risky investment than a portfolio including either all other Top-100 NASDAQ stocks or ordinary NASDAQ stocks.
* Nonetheless, day-to-day fluctuations in both FAANG and other Top-100 stocks equally predict NASDAQ movements (r ~= 0.55). In contrast, day-to-day fluctuations in the bulk of ordinary NASDAQ stocks more weakly predict market swings (r ~= 0.3). In a regression model, day-to-day fluctuations in either FAANG or other Top-100 NASDAQ stocks more strongly predicted the market by a 3x margin relative to ordinary NASDAQ stocks indicating that larger cap stocks overall better predict market movements (r ~= 0.5).

**Analysis**

For this project, I made an early decision to use Python. While alternatives are available, Python boasts a powerful collection of scientific computing packages, such as sci-kit, scipy, and pandas. Conveniently we can use the urllib Python package to interact with the Google Finance API (see **Note A** about choice of using the Google Finance API versus Yahoo Finance API). And from the Google Finance API, historical information on American stocks can be freely downloaded, such as the adjusted closing stock price for each trading day. That alone made the initial process of collating recent historical stock data more manageable. Because I was interested in relative changes of stock prices, the raw historical stock traces were normalized by the initial reference day (either January 3, 2011 or, if not available, the value at the company’s IPO date).

Next to limit the project’s scope, I focused on recent historical stock data from the last six years, exclusively considering the ~3,000 companies that fell into two specific categories: a.) listed on the NASDAQ exchange and b.) publicly traded for at least one business year. The rationale was to focus on stocks most likely to have contributed to NASDAQ growth over the last six years. Also, because the precise definition of ‘FANG’ leaves out Apple and Gilead, the six major companies that serve as the blog’s focus will be denoted here by the slightly longer acronym ‘FAANG.’

Now that we have the historical stock data loaded into Python, what kind of growth have the FAANG stocks recently displayed? Let’s look at a summary stock chart for the performance of FAANG stocks from 2011 to 2017 (**Figure 1**, red trace). First off, we can immediately establish that the FAANG stocks have consistently grown during this period. Furthermore, day-to-day stock fluctuations consistently covaried with day-to-day market changes in the NASDAQ (**Table 1**). These positive indicators strongly imply that recent market growth could principally be attributed to FAANG stocks.

But what about other NASDAQ stocks? Some of the largest and most influential NASDAQ companies, including FAANG, are listed on the NASDAQ-100 Index. And when a similar analysis is performed on this larger group of stocks, we find that most other Top 100 NASDAQ stocks displayed strong growth rates (**Figure 1**, blue trace). Also, day-to-day stock fluctuations were often highly correlated with stock market fluctuations. Why is this important? Because so much of the absolute growth of the NASDAQ (**Figure 1**, black trace) has recently emanated from FAANG stocks, one might worry whether market growth implies the success story of a few thriving companies, rather than a broader advance by most American companies.

Unfortunately a closer look reveals that many NASDAQ companies have struggled to grow recently. Indeed the bulk of ordinary NASDAQ stocks not associated with the NASDAQ-100 Index have collectively stagnated as a group (**Figure 1**, green trace), and their day-to-day fluctuations weakly predicted market changes. In comparison, both FAANG and Top 100 NASDAQ stocks have grown significantly more than ordinary NASDAQ stocks. An important caveat is that these are summary statistics across three-different groups, and do not imply that every ordinary NASDAQ stock showed lowered growth. In fact, the upper quintile of ordinary NASDAQ stocks (~20%) displayed growth rates comparable to that of FAANG stocks (**Figure 1**, yellow trace). We can therefore importantly conclude that a large subset of NASDAQ stocks has been robustly growing at rates roughly comparable to FAANG stocks. For investors seeking portfolio diversification this is welcome news!

While FAANG stocks are not unique sanctuaries of growth within a sea of stagnating stocks, these stocks do have three desirable properties: high growth, high visibility, and manageable group size. For busy professionals lacking the time to pay close attention to every stock, these properties alone are attractive qualities. But would investments in these stocks be considered risky? Modern portfolio theory states that a conservative investor will select his/her investments based upon striking a careful balance between maximizing the overall expected rate of return and mitigating risk [4]. Under this relatively straightforward model, an optimal investment portfolio spanning a diverse collection of stocks is predicted to deliver minimal risk, so long as the selected stocks are not correlated with each other. One straightforward way to assess this tradeoff of return versus risk is computing the ratio of the expected portfolio return versus the portfolio return variance, where higher values indicate a more optimal investment. Comparing theoretical portfolios, FAANG stocks have a much higher ratio than theoretical portfolios containing either Top-100 stocks or standard NASDAQ stocks. These results unequivocally indicate that FAANG stocks, under this admittedly simplified framework, constitute less risk than broadly investing across the other stock groups.

Nonetheless we need a different analytic to address the originally posed question whether the NASDAQ’s movements are better explained by day-to-day changes in FAANG stocks than by other stocks. As many thousands of stocks comprise the NASDAQ, we have little reason to expect that stocks, even when they are equally correlated with the NASDAQ, will drive market growth equivalently. For instance, the market capitalization of stocks is decidedly lopsided, where FAANG and other NASDAQ-100 stocks respectively make up ~25% and ~35% of the NASDAQ’s market value. These numbers suggest that larger cap stocks might have a disproportionately larger influence on NASDAQ growth.

To see if this is true, I setup a Ridge regression model where the input features are the day-to-day fluctuations in z-scored stock data for the 3,000+ NASDAQ stocks with the goal of predicting day-to-day changes in the NASDAQ (See **Notes B** **and C** respectivelyfor more information on either Z-Scoring or model choice). The model was independently trained on quarterly data from the last six years using 5-fold cross-validation, resulting in a unique weighting coefficient for each stock for every trained market quarter (i.e. 24 coefficients for each stock). A higher weighting coefficient intuitively indicates that a stock better predicts market movements relative to other stocks. To summarize how well every stock nominally predicted the NASDAQ’s movements over the *full* six years, we can compute the median weighting coefficient across the 24 quarterly weighting coefficients.

When the median weighting coefficient is compared across groups, FAANG and other Top 100 stocks are indistinguishable from each other. However, both groups have a ~3-fold larger weighting than most ordinary NASDAQ stocks, including the upper quintile of ordinary stocks. Furthermore, there is a clear connection between a stock’s market cap value and median weighting coefficient (r = 0.51) indicating that larger-cap stocks tend to be better market predictors.

In summary, we’ve taken advantage of the analytical framework developed here to apply basic data science principles to investigate the recent performance of FAANG stocks, including their impact on the stock market. While the analyses confirm FAANG stocks are attractive investments, they are not singular. If an investor is willing to do the research, he/she is likely to be able to search for a set of stocks that outperforms these FAANG stocks. In fact, a major reason I became attracted to this capstone project is that I wanted to develop quantitative tools to help me become a better investor. After all, the best tools are those that ultimately lead to actionable intelligence for executing profitable trades in the stock market. Here this relatively simple analytical framework has potential to help intrepid investors identify prospective investment options through automated data mining approaches. In the future, I personally hope to use these tools to analyze the performance of stocks in my own investment portfolio, and then relate their competitiveness to other stocks I might be interested in purchasing.

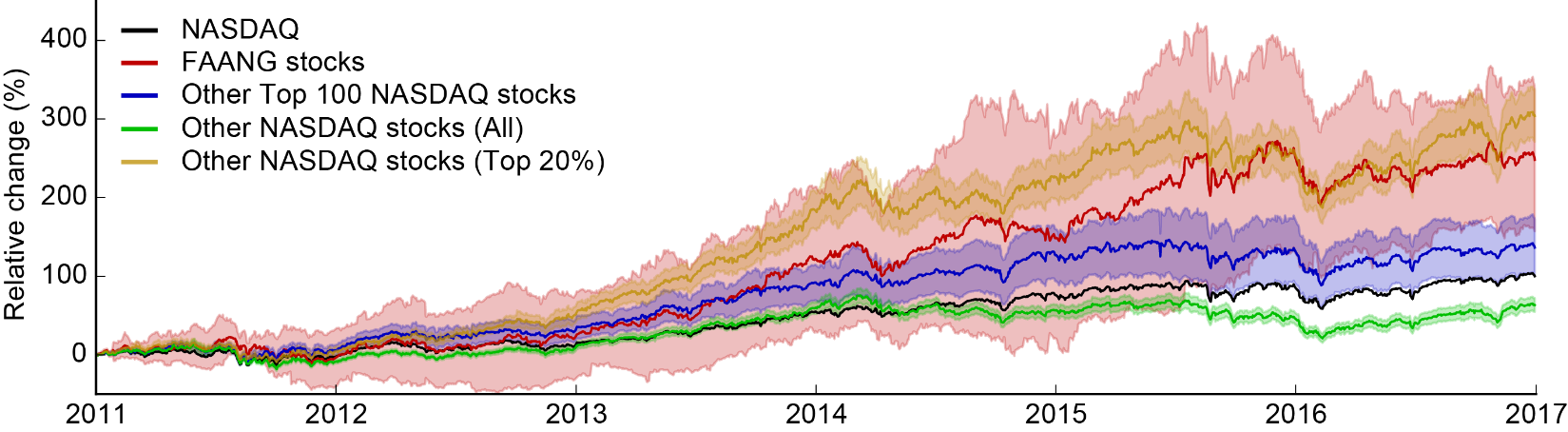
Thanks for taking the time to read the article. Any comments or suggestions would be welcome. And happy investing!

**Analytic notes:**

**Choosing Google Finance over Yahoo Finance**: Pandas previously had a built-in remote data access package (pandas.io.data or pandas-datareader) that could be conveniently configured to interact with the Yahoo Finance API to access historical stock data. However, as of early May 2017, Yahoo discontinued offering their free API service. Fortunately, the Google Finance API offers comparable access to historical stock data. As an important note, the original capstone project was based on historical stock data acquired using the Yahoo Finance API, whereas data in this blog post was acquired using the Google Finance API.

**[B]** **Rationale for z-scoring:** While the relative change in a stock’s price is a straightforward normalization metric for revealing actual growth, we need to z-score so that each stock is weighted equally by the regression analysis. This normalization step is desirable, because intuitively the model weights from the regression analysis can then be interpreted as the relative strength that each stock predicts the NASDAQ’s movements.

**[C] Regression model**: Different regression models were considered and ultimately Ridge regression (L2-norm) was selected. But why? When performing multiple regression analysis, the independent variables (i.e. features) should be independent. However, many of the stocks in this dataset showed positive correlations with each other. This situation unfortunately led to a classic case of the problem of multicollinearity, where it becomes challenging to disentangle the predictive effects of highly correlated independent variables with standard regression models. One option is to use L1-norms (like Lasso regression), in which the most highly predictive of the correlated features is selected by the regression model. But ideally we would like to weight highly correlated features equally. Fortunately, dimensionality reduction techniques like principal component analysis can save the day, because the regression model can instead be trained on the principal components (or repeating patterns in the data) rather than the actual data itself. Ridge regression is a popular regression technique that is loosely related to this general dimension-reduction approach [5].



**Figure 1:** FAANG (blue), other Top 100 stocks (red), and the top 20% growing ordinary NASDAQ stocks (yellow) consistently show more growth than the NASDAQ composite index (black) and the bulk of ordinary NASDAQ stocks (green). For the five traces shown, the relative change in stock price since 2011 is shown as the dashed/solid lines, and the 95% confidence intervals are the shaded areas. Summary data for analyses in this blog are available in **Table 1**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Stock group:** | **FAANG stocks** | **Other Top-100 stocks** | **Ordinary stocks**  **(All)** | **Ordinary stocks (Top 20%)** |
| **Number of stocks** | 6 | 95 | 2,982 | 555 |
| **Annual growth rate (%)** | +55.85±9.39 | +26.33±3.86 | +9.22±1.05 | 65.87±4.14 |
| **Correlation with NASDAQ** | 0.56±0.03 | 0.53±0.01 | 0.29±0.01 | 0.32 ±0.01 |
| **Ratio of portfolio return versus portfolio variance (x10-4)** | 40.90 | 5.93 | 0.38 | 0.46 |
| **Median weighting coefficient (x10-4)** | 7.39±1.33 | 6.47±0.28 | 1.97±0.04 | 1.92±0.06 |

**Table 1:** **Analytics summary for FAANG stocks, other Top-100 stocks, ordinary NASDAQ stocks not associated with the Top-100 stocks, and the Top 20% growing ordinary stocks**. Annual growth rate was determined by taking the average relative change in stock price over the last six years. The correlation between day-to-day stock fluctuations with the NASDAQ were computed using Pearson’s correlation coefficient (r). The last two analytics are detailed in the blog text. Each analytic is summarized as the group mean ± SEM.

**References**

**[1]** D. Strumpf. Wall Street Journal. “The Only Six Stocks That Matter”. <http://www.wsj.com/articles/the-only-six-stocks-that-matter-1437942926> (July 26, 2015).

**[2]** A. Mirhaydari. CBS Moneywatch. “Uh oh -- just 8 stocks are propping up the market”. <http://www.cbsnews.com/news/uh-oh-just-8-stocks-are-propping-up-the-market/> (November 12, 2005)

**[3]** C. Ciaccia. “What Are FANG Stocks and Why Does Jim Cramer Love Them?”. <https://www.thestreet.com/story/13230576/1/what-are-fang-stocks-and-why-does-jim-cramer-love-them.html> (July 24, 2015)

**[4]** "Modern Portfolio Theory." *Investopedia*.  Investopedia. Web. 21 April 2017.

**[5]** Tibshirani R. Regularization: Ridge Regression and the LASSO [PDF document]. Retrieved from Lecture Notes Online Web site: <http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf>