Mini-Project: Stock Market Indices

Esteban Lopez

School of Information, San Jose State University

INFM 203: Big Data Analytics & Management

Dr. Glen Mules

May 9, 2022

Overview

This data science mini project focuses on stock exchanges or stock market data. The primary interest is in the concept of indexes. Historically, and currently, indexes are used to gauge how markets are performing. The United States Securities and Exchange Commission (SEC) definition is the following: “A market index tracks the performance of a specific ‘basket’ of stocks considered to represent a particular market or sector of the U.S. stock market or the economy” (2012). In the U.S. there are the traditional indices that are always reported on in the news, S&P 500, Nasdaq, Dow Jones. In truth, there are many more indexes than ones that just track companies in the Unites States. In 2020, it was reported that there were “2.96 million indices worldwide” (Phillipps, 2020).

There are a massive number of indices. There are products called index funds and exchange traded funds (ETF). “As a hypothetical portfolio of holdings, indexes act as benchmark comparisons for a variety of purposes across the financial markets” (Young, 2022). While the traditional indices like the S&P 500 is used as a benchmark for reporting, index funds and ETFs try to track indices whether they are the traditional types or specialized indices. Index funds and ETFs are not hypothetical but hold actual stocks. With about 3 million in existence there are many combinations possible. Some track the traditional well-known markets like the S&P 500 and Nasdaq, others have gimmicky ticker symbols.

One ETF is named UFO which tracks U.S. space companies involved in businesses like GPS imagery, telecommunications, and space tourism (Hicks, 2022). A related ETF called YODA is presented by the same company to track similar companies operating out of Europe (Eckett, 2021). There are ETFs with the tickers WEED, YOLO, TOKE, and MJ which track cannabis companies (Clark, 2022). Another set of ETFs called TGIF and WKLY which pay weekly dividends (Curry, 2021). There are many, many different ETFs that track many different indices based on sectors or markets. It is fun to look up curious ETF names. Of course, there are ETFs and index funds that track traditional well-known markets. With about 3 million different indices there is just an overwhelming amount to manually analyze.

Motivation

The primary motivation for this mini project is to gather stock market information in order to analyze indices. A desire to measure, primarily, ETFs versus the market or specialized indices they claim to follow. Also, a desire to perhaps create custom, possibly personal indices based on different desired outcomes. In fact, one common market investment strategy is called asset allocation. This strategy “aims to balance risk and reward by apportioning a portfolio’s assets according to an individual’s goals” (Chen, 2022). Asset allocation goes hand in hand with diversification and rebalancing (United States Securities and Exchange Commission, 2009). Really anyone who invests has some sort of personal index if they are investing with any sort of strategy. It might be illustrative to list some questions this mini project will attempt to answer when fully carried out.

Foundation Questions

1. Which stocks are the different market indices composed of?
2. ETF and Index Funds track different market indices, how closely do the holdings match the specific market indices they follow?

Questions Built Upon Answering Foundation Questions

1. Based on existing indices, can custom indices be created depending on differing goals, such as dividend investing or capital gains? Or any of the other sorts of fund goals? Of

course, custom indices can be created, but the goal is to measure them against existing indices.

2. How often should a custom index be rebalanced?

These are a few questions to guide this project and these questions will be analyzed a bit more in the following sections. This process is iterative and will be shown below.

Problem Statement

The basic challenge is how to get the data. I have chosen to mainly use Yahoo Finance since many tutorials and articles use it and recommend it because it is free. There is the question of which indices to use as the foundation. Ideally all indices would be to follow and compare. One problem with indices is that companies are added and dropped depending on different factors. Many well-established companies like Coca-Cola, General Electric, and AT&T have been added and dropped numerous times from the Dow Jones Industrial Index (Dierkling, 2022). Companies that belong to any specific index can be dynamic. For this reason, only a select few indices will be used to begin with.

Methodology

Those questions above are the basic goals for this big data mini project. Setting goals if the first step in the first phase of the data science process according to Brian Godsey (2017). The methodology of this mini project is based on the Godsey approach mentioned in the last sentence. There are 3 phases, but this mini project will mainly be concerned with the first phase called the Prepare phase and consists of the following 4 components: Set goals, Explore, Wrangle and Assess (Godsey, 2017, p.6 fig. 1.2). The 2nd phase called the Build phase, begins with a component called Plan (Godsey, 2017, p.6 fig. 1.2). That is where this mini project will end for this course but will be carried out to completion beyond this course.

Phase I: Prepare

Set Goals

The goals are basically set up in the foundation questions section above. While the questions might seem simple, they are not. On the question of which stocks are the different indices composed of? There are around 3 million indices. It is probably not possible to retrieve all the holdings each index is composed of. One reason is that indices are dynamic, and the holdings change according to various criteria. Since the stock market is basically continuous and volatile, the data will always be changing. One boundary that would need to be set is the time frame. Basically, when selecting which indices to measure against, how often should those be updated. In the future it might be desirable to continuously check for changes, but for this project perhaps just check and update every quarter. It is the same issue for the question of tracking index funds and ETFs.

On the second set of questions of building a custom index and rebalancing, there are related questions. These next set of issues are for both the foundation set of questions as well as these custom indices. What are the indices tracking, what are their purposes? There are also investment strategies of dividend/value or income investing versus growth investing. There are different indices to follow. The S&P 500 has not only that main well-known index, but also have specialized indices called the S&P 500 Growth Index and the S&P 500 Value Index (Levy, 2022). Each individual ETF has a different index they try to track or even create their own index. This is where domain expertise is needed, and hopefully this mini project will be a step in that direction.

There are even more questions that these. While it would have been nice for this component of phase I to be completed in one step, the truth is that it was iterative after going through other steps to try to refine this goal setting step. More questions and issues come up in the next steps.

Godsey has two concepts in the goal setting stage highly relevant to this project. These are the concepts of possibility and efficiency, with value as a third parameter (Godsey, 2017, pp. 34-35). With the question of the holdings of market indices, index funds, and ETFs what is possible? It is possible to find the holdings of all these funds, which stocks are in each of these different indices. Efficiency is the real question. Can this information be wrangled programmatically, or is sifting through each prospectus of each fund necessary? Then that would impact efficiency in a negative way if each individual stock must be programmed manually. Both the set up and the dynamic updating as holding shift. It is possible but the analysis of different funds and ETFs would have to be limited. It needs to be limited due to there being 3 million different indices in any case. Over time it this project can turn into an ambitious and robust dashboard that attempts to track all indices.

Explore

Godsey puts the questions of relevant data and previous work on the data in the Set Goals portion of phase I (2017, pp. 31-32). The iterative nature of this process caused this mini project to answer these questions in the explore stage of phase I. It does all go back to refining the Goal Setting leg of phase I, this underscores the importance of setting goals. Initially data was found already collected by various contributors to Kaggle and Data World (as well as other sites), but the limitations where that the data was all over the place and had various levels of granularity. Always in the form of readily available CSV files from various sources. Exploration was initially conducted by using spreadsheets and Excel. When researching how others had managed stock market data, it was made apparent that many people were accessing stock market data from

Yahoo Finance. Even looking at the initial CSV sources, many contributors to Kaggle and Data World were using Yahoo Finance. Since all the various sources had their data set up differently, it was worth looking into how they get their data.

The Yahoo Finance API is not stable and can be changed at any time, but there are some libraries to easily access Yahoo Finance with Python and those libraries are called yfinance and Yahoo_fin and they are free to use (Bland, 2021). The unofficial support of the API is likely why many users have the data in CSV files once they access it. It is a clever idea, to make sure the data is available at least to the date collected. Since libraries are used it is unclear what format the original data is in, but it can all be saved as CSV files. It does come from Yahoo Finance website, so it is JSON or XML, but those are just wild assumptions. Using Jupyter Notebooks brings them in as lists and dictionaries, the dictionaries do look like JSON, but that might just be how the Python libraries convert them. They can be saved and converted to very manageable CSV files. Not only can they be used in Jupyter but Excel or Google Sheets once converted.

Wrangle

This third leg of Phase I is an enjoyable process and can be frustrating as well. Godsey says of data wrangling and expecting problems: "*Double-check everything*" (2017, p. 73). Becoming familiar with the yfinance library, testing was done looking for dividend information. Noticing that certain ETFs had not given dividend at certain times did not raise any alarm bells. Then evaluating some ETFs owned by this author showed that yfinance was not collecting up to date information, there were dividends missing using that library that had been shown as paid on Yahoo Finance. Running the dividend information retrieval commands in Jupyter showed that many stocks and ETFs had missing dividend information. The next day the exact same commands were run again, and the missing year of dividends showed up. The commands were saved in a Jupyter Notebook and were not changed in any way between those days.

Going through thousands of stocks to analyze possibly (once project has matured enough) millions of indices, having missing data can be a problem. Especially if it is an index focused on measuring dividends in this case. While using Yahoo Finance is great for this mini project. This does help to gain domain knowledge and helps to ask questions of any future paid and stable API for stock information. On that note there is a library or package called Pandas Datareader that can read stock data from many different sources using Python, including Yahoo Finance, Google Finance, Morningstar, and Robinhood to name a few (Kharwal, 2021).

Wrangling the data from Yahoo Finance using yfinance is interesting. One thing that can be done is that a lot of the data from a single stock can be downloaded indiscriminately. One potential issue also related to dividends is that dividends are not paid daily but the standard information shown always has a column for dividends, on most days it is 0.00. Looking at a few stocks and streaming in that information is probably harmless on its' face. Not all stocks pay dividends and many that do only pay 4 times a year. There are 4 days where that dividend column is not 0.00. For this mini project it might not be a huge deal, but if working with big data storing a column with 0.00 might take up too much space. There are options to retrieve only the dividend information just for the dates they are paid out.

It is possible to go to Yahoo Finance directly and download historical data and select daily, weekly, or monthly frequencies, along with the time period going back to 1961 if a particular stock has been around that long. One would have to go to each stock and click preferences and click a download button to retrieve CSV files. It is much easier to download the desired data using a Python library and save it as a CSV. There is differing levels granularity allowed, for example historical data can be downloaded with minute-to-minute interval detail up to 3-month interval. This is apart from the total time period one would like to review for example from 1980 to 2000 or something like that. For this project max time period and daily intervals will be used for stock prices. Just to be clear, going directly to Yahoo Finance does not allow for minute-to-minute download of historical data, but it does allow for selecting time periods. The only way to get the interval desired is by using yfinance or other libraries.

As can be seen, without domain expertise, these sorts of questions on time intervals are important. Again, with the iterative approach, this part of the question should have been part of the Goal Setting leg but was not discovered until this portion. It is nice to learn while doing this process methodically and writing this down is illuminating. So documenting is important as Godsey explains (2017, p. 14). But it is not just important for software documentation, it is important to see what nuances are overseen. The nuance of the possibility of intervals and different time periods brings up what probably should have been brought up at first, domain expertise or not. One assumption that was not expressed at the beginning is the different types of investors or players in the stock market. There are basically day traders who buy and sell over seconds and days and investors who buy and sell over years (Mitchell, 2022). This project is assuming that the target audience is investors and not day traders. So, the necessity of any interval shorter than a day is not needed for this mini project.

Assess

This final leg of Phase I was only slightly worked on, but enough to gauge the data. Using Jupyter Notebook and a Python library called matplotlib it was very simple with the help of tutorials to programmatically plot some stock data on a graph. The veracity of the data collected via Python was confirmed when juxtaposed against traditional news graphs of the stock market. In the initial assessment using a library called Pandas, nicely formatted tables were also useful in comparing the numbers pulled with Python against the numbers on traditional brokerage websites. The tables are what revealed that there are wasted columns regarding dividends and stock splitting, it would waste storage space in the future of a big data project. This whole process is iterative and seems as volatile as the stock market. The graphing can be customized with labels and other things. This is probably not part of the assessment leg or even the Prepare phase of data science. But it is something to note. It is also in this leg that it was initially found that retrieving information from Yahoo Finance can be glitchy as in the example of missing dividends for about a year.

Phase 2: Build

Plan

The Plan leg of the beginning of Phase 2 is interesting because Godsey explains that a big part of it is adjusting goals (2017, p. 116). Adjustments were, and remain, constant during this beginning portion of the project. One practice suggested by Godsey is to "Make lists, create flow charts, or write logical if-then statements showing the new information directly affecting future results" (2017, p. 115). The next iteration of this project will incorporate that, because there will be many more iterations as this project is followed through and adjusted after this course. Though this is a personal project, this might expand to a public project or business project. Refining questions, even formally stating and recording assumptions is useful. Godsey writes that, "An explicit and formal evaluation phase . . . can help you organize your progress, your goals, your plan, and your knowledge of the project" (2017, p. 128). The leg reinforces the idea that this framework is meant to be iterative. This framework useful of course for professional projects with clients but is also useful for personal projects.

Reflections and Future Work

The Plan leg of the Build Phase II is meant as a formal checkpoint and is the final checkpoint for this course, but not this project which will continue. This framework was useful to help not only learn about the data science process, but also to inadvertently (or maybe it was by design) gain a slight amount of domain knowledge (above and beyond what was brought into this course). With this framework in hand any data project can be taken on. It would be nicer to come in with domain knowledge, but this framework can force a net gain of domain knowledge in any subject. That is not to say that some questions and assumptions would have been formally framed and stated at the beginning with previous domain knowledge and data science experience. The next iteration of the Phase I will make use of the formal flow charts, and what-if statements recommended by Godsey. The confidence that the rest of the process will be illuminating is high.

Future work on this mini project will continue with Jupyter and Yahoo Finance. Eventually a personal dashboard will be created with this information. As a mini project Yahoo Finance is fine but if this turns into a professional project I will have to look elsewhere. Concerning Yahoo Finance, “Using the Public API (without authentication), you are limited to 2000 requests per hour per IP (or up to a total of 48,000 requests a day)” (Barney, 2020). Plenty for this mini project, but by using the data science process framework I will be able to know which questions to ask and which data I need if I convert this to a professional live product. In a way by taking this mini project through all Godsey’s phases, I will be completing the prepare phase for a live product. I plan on using this project for my own investing, but I also intend to create a web portfolio with this project in there. This is the sort of framework and skill that will be used for personal projects, and hopefully professional projects in the future. It is something I want to use and develop further.

References

Barney, H. (2020, March 21). *Free Stock Data for Python Using Yahoo Finance API.* Towards Data Science. Retrieved May 10, 2022, from https://towardsdatascience.com/free-stock-data-for-python-using-yahoo-finance-api-9dafd96cad2e.

Bland, G. (2021, January 11). *Yahoo Finance API – A Complete Guide.* Algo Trading 101. Retrieved May 9, 2022, from https://algotrading101.com/learn/yahoo-finance-api-guide/.

Chen, J. (2022, March 1). *Asset Allocation.* Investopedia. Retrieved May 8, 2022, from https://www.investopedia.com/terms/a/assetallocation.asp.

Clark, M. (2022, May 6). *Brand-New Weed ETF Deep Dive.* Money & Markets. Retrieved May 8, 2022, from https://moneyandmarkets.com/brand-new-weed-etf-deep-dive/.

Curry. R. (2021, February 22). *WKLY Versus TGIF: Which SoFi Weekly Payout ETF is Better?* Market Realist. Retrieved May 8, 2022, from https://marketrealist.com/p/sofi-weekly-income-wkly-vs-tgif-etc/.

Dierking, D. (2022, January 21). *Four Iconic Companies Dropped From the Dow.* Investopedia. Retrieved May 8, 2022, from https://www.investopedia.com/articles/investing/113015/4-famous-companies-dropped-dow-jones.asp.

Eckett, T. (2021, June 3). *Product Panel: Procure's Space ETF Hits Europe: Europe's First Space ETF.* ETF Stream. Retrieved May 8, 2022, from https://www.etfstream.com/features/product-panel-procures-space-etf-hits-europe/.

Godsey, B. (2017). *Think Like a Data Scientist: Tackle the Data Science Process Step-By-Step.* Manning Publications Co.

Hicks, C. (2022, April 27). *Space Stocks: How Russia is Changing the Game: Russia's Invasion of Ukraine has had a Number of Far-flung Effects, Including Shifting Fortunes in Space Investing.* Kiplinger. Retrieved May 8, 2022, from https://www.kiplinger.com/investing/stocks/604606/space-stocks-how-russia-is-changing-the-game.

Kharwal, A. (2021, March 22). *Pandas Datareader Using Python (Tutorial).* TheCleverProgrammer. Retrieved May 9, 2022, from https://thecleverprogrammer.com/2021/03/22/pandas-datareader-using-python-tutorial/.

Levy, A. (2022, March 23). *Value vs. Growth Investing: Which Should You Buy?* The Motley Fool. Retrieved May 8, 2022, from https://www.fool.com/investing/stock-market/types-of-stocks/growth-stocks/value-vs-growth-stocks/.

Mitchell, C. (2022, March 2). *Day Trading vs. Investing: What's the Difference? It's more than just buying and selling stocks.* The Balance. Retrieved May 10, 2022, from https://www.thebalance.com/day-trading-versus-long-term-investing-4139868.

Phillipps, J. (2020, June 4). *A Brief History of Indices.* CityWire. Retrieved May 8, 2022, from https://citywireusa.com/professional-buyer/news/a-brief-history-of-indices/a1363301.

United States Securities and Exchange Commission. (2009, August 28). *Investor Publications: Beginners' Guide to Asset Allocation, Diversification, and Rebalancing.* Retrieved May 8, 2022, from https://www.sec.gov/reportspubs/investor-publications/investorpubsassetallocationhtm.html.

United States Securities and Exchange Commission. (2012, October 15). *Fast Answers: Market Indices.* Retrieved May 8, 2022, from https://www.sec.gov/fast-answers/answersindiceshtm.html.

Young, J. (2022, February 2). *Market Index.* Investopedia. Retrieved May 8, 2022, from https://www.investopedia.com/terms/m/marketindex.asp.