TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Project 1 - Advanced Corporate Finance

Humphrey Hui

Peyton Gibbs

Jack Feen

Overview

In this project, we pulled CRSP stock data to create three different stock portfolios and analyzed their returns compared to industry-standard ETFs.

Methodology

First, we pulled data from the CRSP Monthly Stock File covering January 2015 through December 2024. We only kept regular U.S. stocks (share codes 10 and 11) that trade on the three major exchanges (codes 1, 2, and 3).

Next, we cleaned the raw data to make it usable. We made all stock prices positive by taking absolute values. Then, we converted the shares outstanding numbers from thousands to actual amounts. We calculated market capitalization by multiplying price times shares outstanding which was required for ranking stocks by size.

Then, we dealt with survivorship bias, which is when only the “survivors” in a dataset are looked at while the ones that dropped out are ignored. To avoid this, we added delisting returns from CRSP using the formula: $r_{eff} = (1 + r_{CRSP}) \times (1 + r_{delist}) - 1$. This helps capture the full picture, including the negative returns when companies fail instead of pretending they never existed.

Afterwards, we created three different index types to see how weighting matters. The equal-weighted (EW) index gives every stock the same weight in the index. The value-weighted (VW) index works like the S&P 500, where bigger companies matter more. The price-weighted (PW) index works like the Dow Jones, where stocks with higher prices get more weight regardless of company size.

Each month, we picked the top 100 stocks by market cap and set the appropriate weights based only on the previous month's information. Then, we used those choices to calculate the current month's returns. All indices start at 100 and grow from there. By doing so, we avoid any data leakage.

Key Implementation Choices

We chose 100 securities as our cutoff because it gives us enough diversification without including small companies that don't matter much to investors. This keeps our indices focused on stocks that truly matter to the general market.

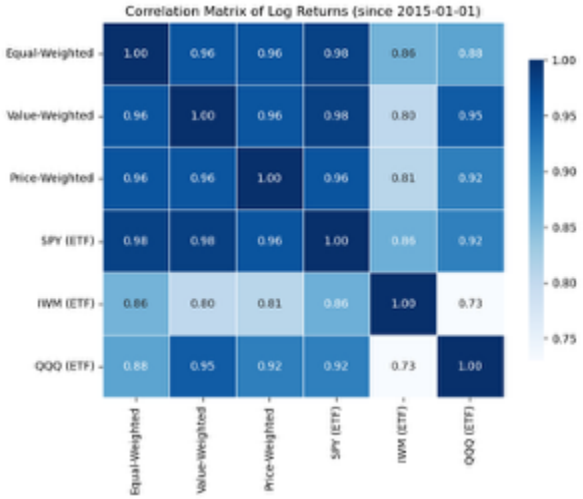
We ranked stocks each month by market cap using the previous month's data, then picked the top 100 for the next month's index. When selected stocks were missing return data, we dropped them and adjusted the weights among the remaining stocks. By doing so, we avoid using any future information and any potential data leakage.

When we encountered missing data, we carried forward the previous index level. This approach helps keep our indices stable instead of having data gaps that create volatility that is not actually there.

We compared our custom indices to SPY, IWM, and QQQ using their ticker symbols. These ETFs are industry standards showcasing the 500 largest US companies by market cap, the 1000-3000 largest publicly traded US companies, and the 100 largest nonfinancial companies, respectively. We used log returns for correlation analysis because they are additive, making it easier to plot stock correlations over time.

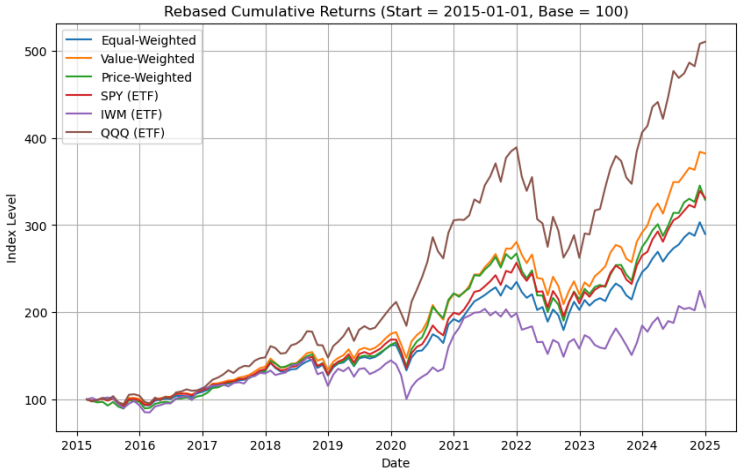
Analysis

Plots and Correlation Matrix Results: The correlation shows us that the closest correlations with SPY and QQQ come from the top 100 value-weighted stocks. Logically, this makes sense due to the fact that it contains the top 100 companies from SPY, a similar set to that in QQQ, and it holds the same weighting schema. The top 100 stocks typically share a much lower correlation with the Russell 2000 (IWM). This is likely due to the fact that the Russell 2000 holds an entirely different basket of stocks than the other portfolios and ETFs.

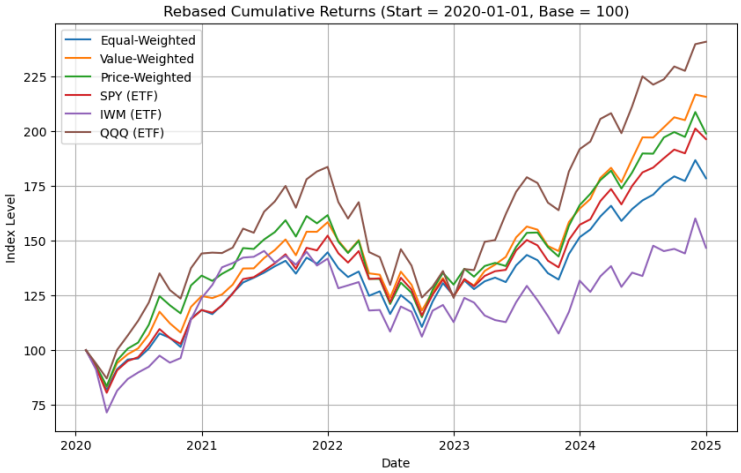


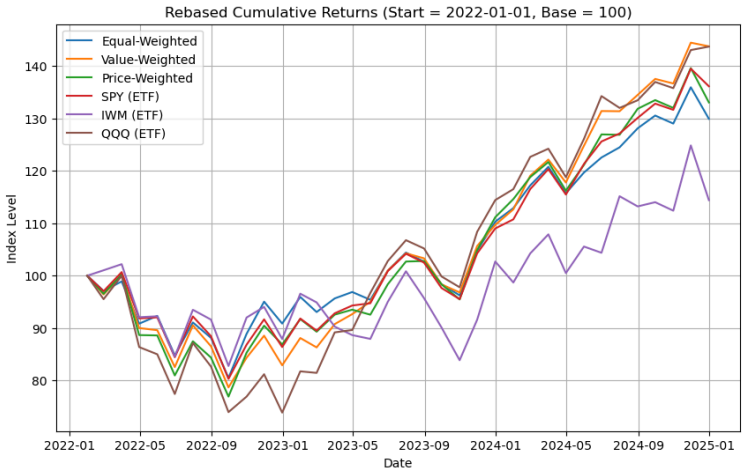
Performance Differences: Plotting all of these portfolios from 2015, starting at a base value of $100, we can observe that all of these portfolios tend to rise and fall together, with QQQ having the most

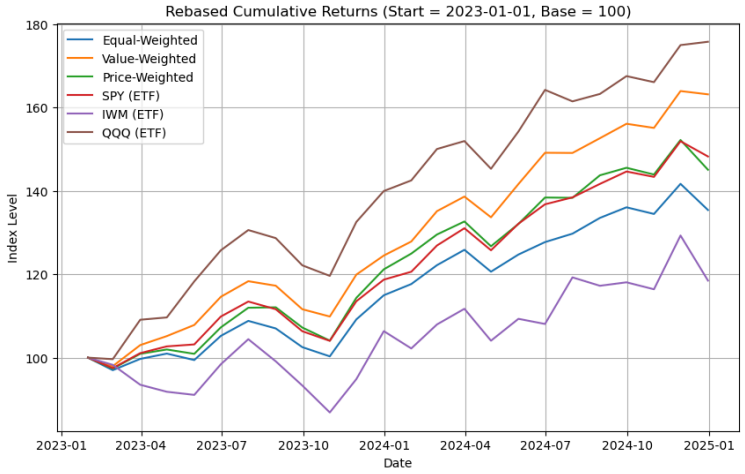
exaggerated returns and IWM being the least variable. The equal, value, and price weighted indices have similar performances to each other and are also closely correlated with the S&P 500.



Robustness: One issue we wanted to investigate was the relationship between the different indices after initial spreads, specifically in response to market booms and crashes. We reset the index of each of our portfolios in 2020, 2022, and 2023 to see how portfolio performance rankings would change.







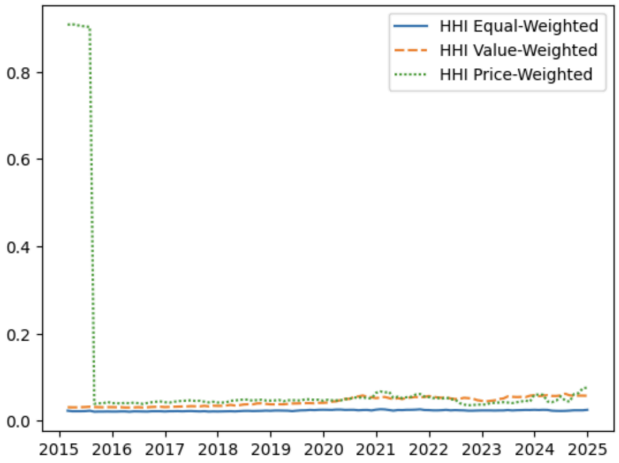
Creating these new performance timelines yields similar results to the initial 2015 portfolio returns, with the most surprising result coming from observing a crash first, then recovery. The higher variance portfolios fall beneath the lower variance portfolios in value, and don't have time to fully recover their gains. Return lines cross, and value-weighted returns win out as the most profitable portfolio.

Diversification: To find how diversified each index was, we grouped by the sectors using the “hsiccd” code to classify it into a certain sector. With the weights per sector, we calculated the HHI (sum of squared sector weights), where a lower score indicates higher diversification and a higher score indicates lower diversification.

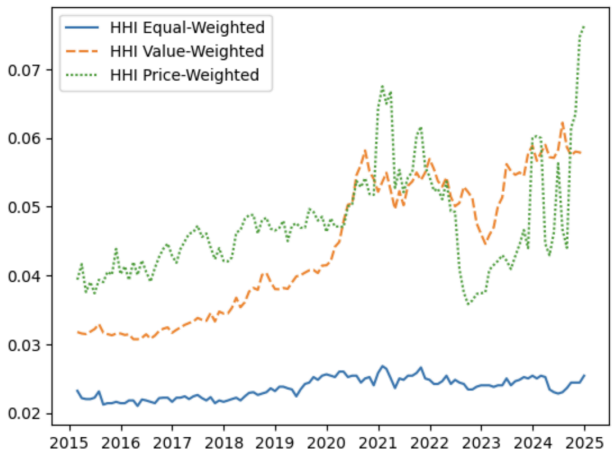
Our HHI scores are low for our Equal Weighted and Value Weighted indices; both indices are below .1 in their HHI score for all the months, showing that one sector is not dominating the index. The biggest sector weight in the Equal-Weighted index was Pharmaceutical Companies for most months, except for a few where the National Commercial Banks sector had the largest weight in the index. For Value Weighted, the biggest sector was the Electronic Computers sector for most of 2015 but around 2017, this index’s biggest sector was Business Services related to computers and data processing. Even though these were the biggest sectors, they did not dominate the weight share of the index.

The price-weighted index had the most surprising findings. Berkshire Hathaway dominates the index (due to Warren Buffett's famous stance against stock splits leading to BRK.A being worth over $700,000 today), but we only have 7 months of data for that company. After those 7 months, the weights are much more equal, like the other two indices. Below are graphs showing the price-weighted index, one with Berkshire Hathaway data and one without Berkshire Hathaway data.

HHI scores over time INCLUDING Berkshire Hathaway:



HHI scores over time WITHOUT Berkshire Hathaway:



Challenges Faced

We faced some challenges that we had to work around at each step of our index creation and visualizations of returns, mainly during our data cleaning, as this sets up the rest of our index calculations and comparison. One of the challenges we had to account for was the fact that data was split between multiple tables in the WRDS database. We had to make sure we used the correct join on the correct parameters to ensure our data quality. Once we had our data loaded into Pandas data tables, there was cleaning we needed to do to make sure we could create our indices correctly. The first cleaning problem was standardizing a few of our parameters. The first parameter we standardized was the last day of the

month. Then we had to standardize the price and the shares outstanding by taking the absolute value and multiplying by 1000, respectively.

Another challenge during the data cleaning part of this script was dealing with Null values. In calculating effective return, we filled the normal returns and delisting returns with 0 where there were NA values. Once we got the effective return, we dropped any values that were NA between market capitalization and effective return for a specific month. We also did not include any market capitalizations that were less than 0. We also had to deal with NA values when we were building our indices. We were looking at the prior month for our top indices of the current month, which led to some NA values, so we created a max to filter only rows that were populated with the effective return for that prior month. We did this for all three of the indices we built. The effective returns are also how we dealt with survivorship in our data. We included the delisting returns in these calculations to account for any stocks being delisted.

During the comparisons and evaluations of our indices, we faced an issue of comparing the correlations after the indices initially started to spread in our time series data. The way we accounted for this was to recompute correlations and portfolio returns during multiple different time periods: Full, prior to a market crash, during a market crash, and after a market crash. This allowed us to see that the correlations between the portfolios stay relatively constant over time, but each faces a different level of risk that may impact the portfolio's value at the end of a given period.