Diversification and Market Volatility: Exploring the Role of Concentration in Portfolio Returns*

Insights from the Herfindahl-Hirschman Index and Performance Tiers Across High and Low Volatility Markets

Onon Burentuvshin

November 27, 2024

This paper investigates the role of portfolio diversification in shaping returns under varying market conditions, with a particular focus on the Herfindahl-Hirschman Index (HHI) as a measure of portfolio concentration. Using regression models, we analyze the interaction between diversification and market volatility, represented by the extremes of the VIX index from the past decade. Our findings reveal a nuanced relationship: while less diversification generally enhances returns in low-volatility markets, it significantly inhibits performance during periods of high volatility. Notably, the effects of diversification vary across portfolio performance, with high-performing portfolios benefiting from concentration and low-performing portfolios gaining from increased diversification. These results provide critical insights for portfolio managers, suggesting that diversification strategies should be tailored to both market conditions and portfolio performance levels. The study highlights the limitations of HHI as a diversification metric and proposes methodological improvements for future research, including the incorporation of cross-asset correlations, stratified sampling, and dynamic volatility analysis.

1 Introduction

In portfolio management, diversification is primarily considered a tool for mitigating unsystematic risk. Current literature deeply explores the concept of reducing risk without affecting portfolio performance. However, while the relationship between diversification and risk reduction has been extensively studied, there is limited understanding of how diversification directly influences portfolio performance. In other words, while diversification is thought to stabilize returns by mitigating unsystemic risk, it is less clear whether this stabilization leads to a

*Code and data are available at: https://github.com/ononburen/portfolio-diversification-analysis.

positive or negative impact on portfolio returns overall. This paper aims to address this gap by examining how diversification, measured through the Herfindahl-Hirschman Index (HHI), influences portfolio returns under varying market conditions.

This paper addresses these gaps by analyzing portfolio returns through regression models incorporating interaction terms for diversification and market volatility. Our results reveal that less diversification enhances returns during low-volatility periods but imposes significant losses during high-volatility periods. Furthermore, the effects of diversification vary dramatically across performance tiers: high-performing portfolios benefit from concentration, while low-performing portfolios gain from diversification. These findings supplement the conventional wisdom of diversification and highlight the importance of tailoring strategies based on market conditions and portfolio characteristics.

Our estimand is the effect of diversification on portfolio returns, with particular focus on how this relationship changes during periods of high and low market volatility. We also explore the heterogeneous effects of diversification across performance tiers, investigating whether high-performing portfolios are impacted differently from low-performing ones. This incorporation allows us to evaluate diversification not merely as a risk management tool but as a determinant of portfolio performance, providing a new perspective on its role in portfolio construction.

This paper's objective is to provide investors and portfolio managers more insight into the dynamic of diversity and performance by shifting the focus from diversification as a risk-reduction strategy to diversification as a determinant of performance. Additionally, actionable insights for optimizing portfolio construction can be gained from this analysis.

The remainder of this paper is structured as follows. Section 2 describes the data and methodology, including the use of HHI and the construction of portfolios. Section 3 outlines the regression models and provides justification of the model the paper employs. Section 4 presents the results, highlighting the interaction between diversification, volatility, and performance tiers with which Section 5 discusses the implications of these findings, situates them within existing literature, and identifies the limitations of this study.

2 Data

2.1 Overview

The data used in this analysis consists of both financial stock data and the S&P 500 company list. The analysis focuses on understanding the relationship between diversification and portfolio performance during periods of high and low market volatility, with data sourced and processed as follows:

2.2 Data Sources

1. S&P 500 Company List: The list of companies in the S&P 500 index was scraped from Wikipedia (Wikipedia contributors 2024) using the `rvest` (Wickham 2024) and `dplyr` (Wickham et al. 2023) packages in R. From this list, a random sample of 50 companies was selected to construct a company basket. Randomization ensured a representative subset of the index, mitigating potential biases in company selection.
2. Stock Price Data: Historical daily stock price data for the selected companies was fetched from Yahoo Finance (Yahoo, Inc. 2023) using the `quantmod` (Ryan and Ulrich 2023) package. The data that was acquired spans from January 1, 2010, to January 1, 2023. This time range allows for calculating annualized returns for the selected companies across periods of high (2020) and low (2017) market volatility.

2.3 Portfolio Construction

An inherent challenge of uncovering the relationship between diversification and portfolio performance during different market conditions was the availability of specific portfolio performance and investments made by firms and investors. A workaround to this was using the `dplyr` package (Wickham et al. 2023) to artificially simulate portfolios with random stocks selected from our basket of 50 companies to represent a diverse sample of portfolios with different levels of diversification. This artificial sampling frame enabled us to analyze diversification effects using realistic data from Yahoo Finance for performance metrics. For transparency and clarity, the complete list of company tickers that were used in portfolio construction is provided in Appendix A.2.

2.4 Data Cleaning Process

We employed the R programming language (R Core Team 2023) to manage, clean, and analyze the dataset. The `tidyverse` (Wickham et al. 2019) package facilitated efficient data manipulation, filtering, and reading the large dataset. For visualizations, we relied on `ggplot2` (Wickham 2016) to create descriptive plots and charts that communicate key findings in the analysis. To facilitate saving files and referencing them across files, the `readR` (Wickham, Hester, and Bryan 2024) package was extensively used.

Our data cleaning process involves a multitude of steps that retroactively updates our company basket by excluding company tickers for which data could not be retrieved. After obtaining stock price data, there were two tickers for which data was unavailable. After omitting these two tickers, annualizing daily price data in accordance with the return formula outlined in the following measurement section. However, some dates might have missing data and thus generate non-entry returns in some rows. To reconcile for this, companies with more than 10 missing entries in the annualized data for 2017 and 2020 were flagged as problematic and subsequently removed from the data set. Consequently, two more companies were omitted

from our company basket, resulting in a final basket of 46 companies. After the company basket was updated, portfolios were reconstructed using the tidyverse package to create a new set of 1,000 portfolios that reflected the refined selection of stocks. We then calculated annualized stock returns for 2017 and 2020 by computing the percentage change in prices from the beginning to the end of each year, facilitated by dplyr such as filter and mutate. The exact decisions for certain omissions of data will be discussed in the AppendixAppendix A.2.

2.5 Measurement

In the development of our understanding of the interplay between diversification and stock performance, standard measurements become necessary to capture these ideas.

1. Stock Performance (R_i) Stock performance is a relatively easy metric to measure quantitatively through percentage returns. Our analysis makes use of annual returns of a market with low volatility and high volatility. The **annualized stock return** is calculated as follows:

$$R_i = \frac{P_{\text{end}} - P_{\text{start}}}{P_{\text{start}}}$$

Where P_{end} is the price of the stock at the end of the year and P_{start} is the price at the start of the year.

2. Portfolio Diversification (HHI) The idea of diversification is a difficult measurement to quantify as there are many layers of complexity to assert diversification. This paper uses the The Herfindahl-Hirschman Index (HHI) value to estimate diversification in a portfolio using stock weights as its input. The HHI is a widely used measure of market concentration. Originally developed to evaluate the level of competition within a market, the HHI is calculated as the sum of the squared market shares of all firms in a given market. In our context, we repurpose this metric to capture the level of diversification within a portfolio. Our HHI measure for diversification implies that the higher the value of HHI, the less diversification is present within a given portfolio. The **HHI** is calculated as:

$$\text{HHI} = \sum_{i=1}^{n} w_i^2$$

Where w_i is the weight of the stock in the portfolio.

3. Market Stress (Year2020): This paper uses market volatility to determine the state of the market stress. As a measure of overall market volatility, our analysis uses the VIX index to extract a period of low market volatility and a period of high market volatility to determine market state. In the ten year range of our data, peak market volatility coincided with the year 2020, while the lowest point aligned with 2017. As such our variable for volatility is a binary, representing peak volatility with 1 and low volatility with 0.

2.6 Outcome variables

Thus far, our measurements enable us to calculate stock performance, but not portfolio performance. In each observation (a particular portfolio of our data set), the objective is to extract how that portfolio performed under two different market conditions. This portfolio performance is the outcome variable that we measure in each observation, which we will calculate in the following way:

$$R_P = \sum_{i=1}^{n} w_i R_i$$

where R_i is the annual return of stock i and w_i is the weight of stock i in the portfolio.

Figure 1 displays performance on a sample of the first ten portfolios of our generated data. Among these two years, none of the ten sample portfolios exhibited negative growth. For a comprehensive overview of portfolio performance, Figure 2 demonstrates the general distribution of portfolio returns through a box plot of both years through a density plot. Interpreting these two figures together serves as a rough indication that there are differences in performance of portfolios between 2017 and 2020.

2.7 Predictor variables

The predictor variables that are used for analysis is outlined below:

1. HHI: A set of number ranging from 0 to 1 with lower values indicating greater diversification.
2. Market Volatility: Binary variable indicating high (2020) or low (2017) market volatility.
3. Interaction Term: The interaction between HHI and Market Volatility to examine whether the effect of diversification on portfolio performance varies across different market conditions.

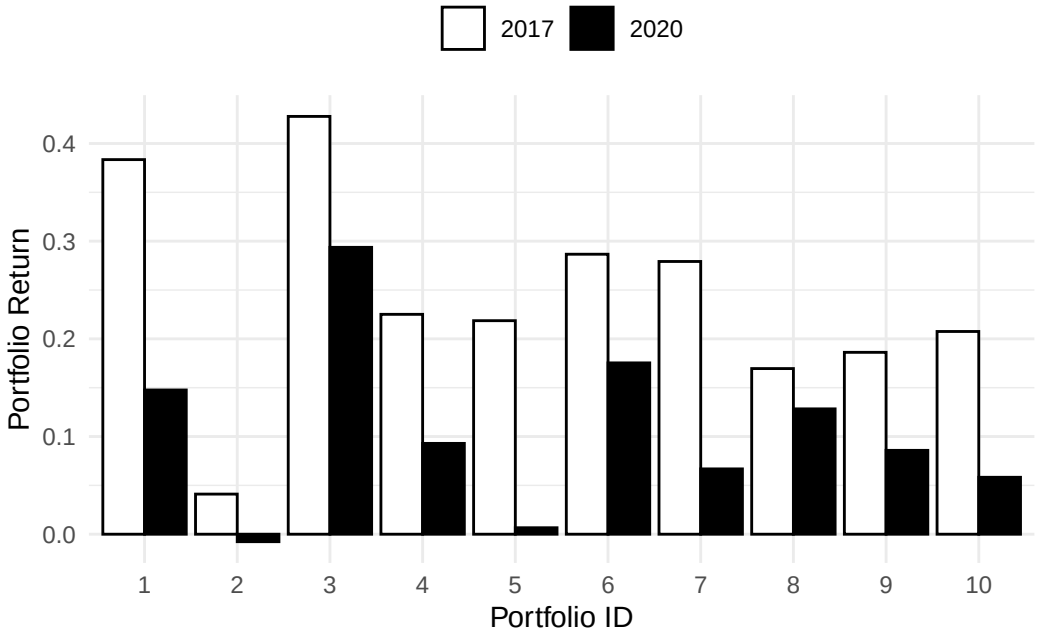

Figure 1: Portfolio Performance for the first 10 Portfolios (2017 vs. 2020)

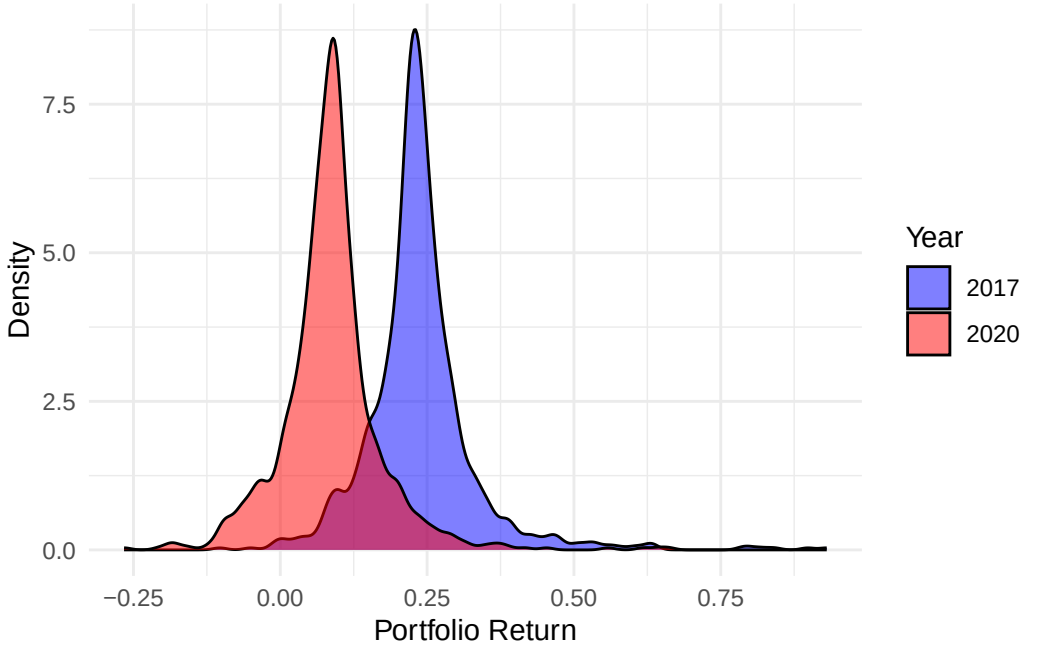

Figure 2: Average Performance of Portfolio 2017 vs 2020

Figure 3 provides information on the distribution of HHI values among our portfolios. Since the weights of the portfolios do not change throughout years, it satisfies our expectation that the distribution of HHI levels remain constant in our data. However, the HHI levels are not evenly distributed and mainly concentrated over lower levels of HHI, which may induce bias in our data set over representing highly diversified portfolios. However, the possible causes and properties of this distribution are discussed in the Appendix of this paper, expanding on the idea of different types of measurements for diversification. To grasp a general idea of the relationship between portfolio HHI levels and portfolio performance, Figure 4 outlines a heat map that plots portfolio returns against HHI levels for both 2017 and 2020. The heat map uses color intensity to represent the density of observations, with darker areas indicating a higher concentration of portfolios. Observing the distribution patterns across the two years, there is no clear trend or pattern among levels of distribution and performance. One key observation that Figure 4 communicates is that as HHI levels increase, the variability of returns also increase, highlighted by the broader ranges of portfolio returns past an HHI of 0.6.

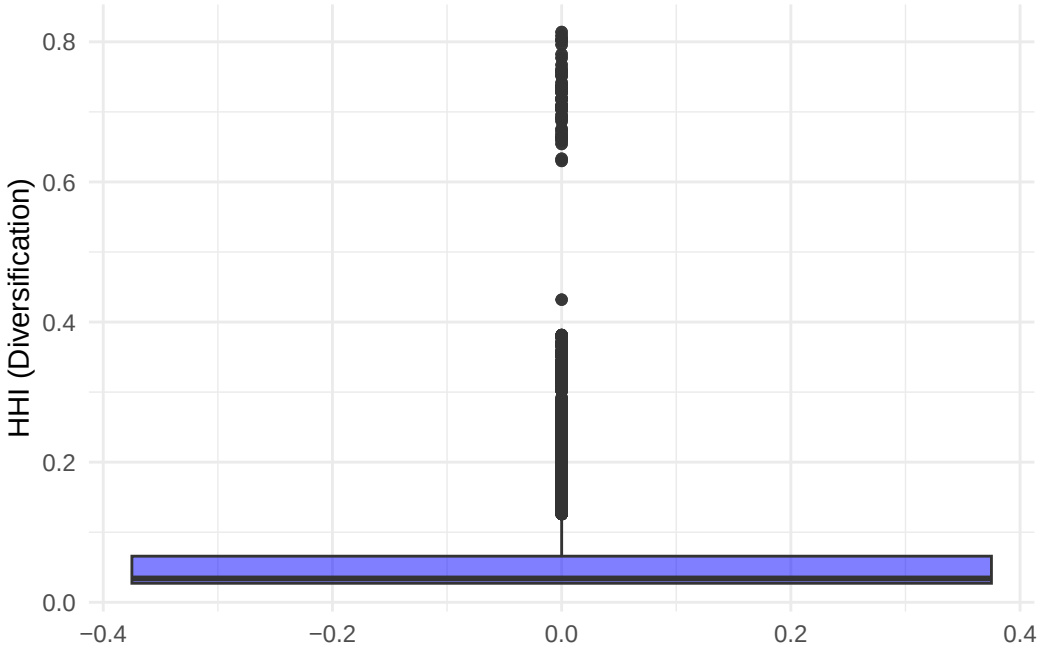

Figure 3: Boxplot of HHI distribution among 1,000 Portfolios

3 Model

The model employed in this paper contains a two-fold goal outlined below:

1. Model the treatment effect of diversification and market volatility on portfolio returns

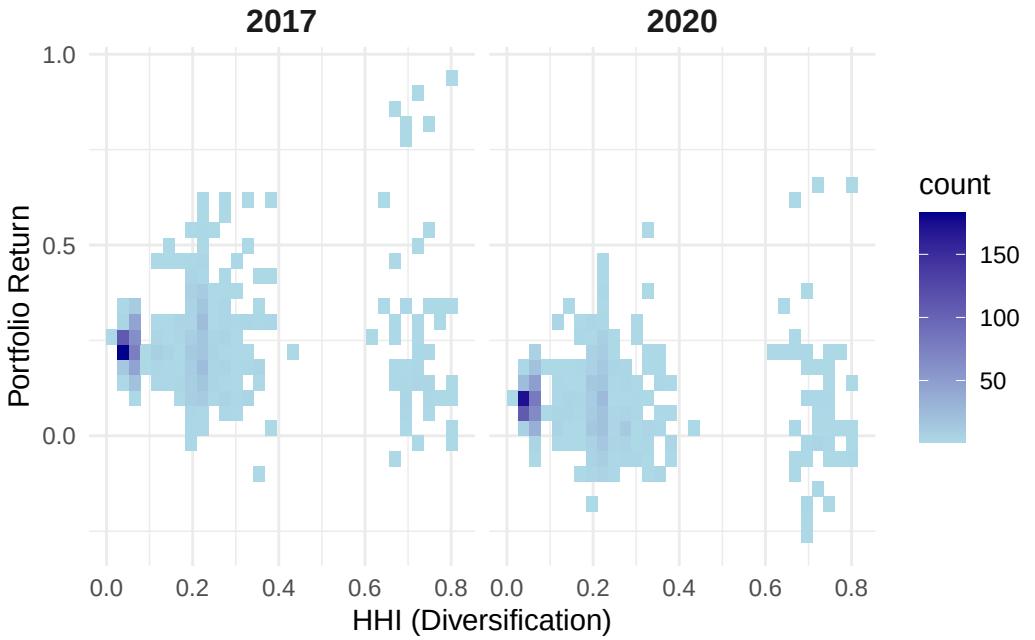

Figure 4: Heat Map of Portfolio Returns at Different Levels of HHI in 2017 and 2020

2. Extract heterogeneous effects of treatment across different portfolio performances

3.1 Base Model

This paper explores the relationship through the following functional form of simple linear regression model:

$$R_P = \beta_0 + \beta_1 \cdot \text{HHI} + \beta_2 \cdot \text{Year2020} + \beta_3 \cdot (\text{HHI} \cdot \text{Year2020}) + \epsilon$$

where:

β_0 is the intercept, β_1 captures the effect of HHI on portfolio returns, β_2 captures the effect of extreme market volatility on portfolio returns, β_3 captures the interaction effect between HHI and market volatility, and ϵ is the error term

The linear regression model was implemented in R (R Core Team 2023), with results processed and cleaned by the packages `tidyverse` (Wickham et al. 2019) and `broom` (Robinson, Hayes, and Couch 2024) with which `kable` (Xie 2014) was used to visualize tables.

3.2 Heterogeneous Effects

Our simulated portfolios contain randomly selected stocks and our outcome variable measures the aggregate returns of these stocks within specified time periods of volatility and non-volatility. Thus, different portfolios may respond differently to market volatility based on their composition, size, or level of diversification. To account for this possible heterogeneous effects of our predictor variables, our model will also stratify the data into two groups: top-performing portfolios (above the 75th percentile of returns) and bottom-performing portfolios (below the 25th percentile of returns). This stratified analysis enables us to determine whether the role of diversification is amplified or subverted for portfolios at opposite ends of the performance spectrum and account for heterogeneous effects of diversification and volatility on portfolio return.

3.2.1 Model justification

Analysis of market dynamics with variables like diversification and volatility is complex as there a multitude of variables that can affect portfolio performance. Despite this complexity, this linear regression model provides a valuable framework for estimating the effect of simple portfolio diversification and market volatility on its returns through quantifiable measures. Moreover, coefficients of our base model describe isolated effects of our predictors with all else held constant, uncovering specific dynamics of market performance. Although a Bayesian regression model provides a more compelling analysis with its considerations of uncertainty, the challenge of specifying appropriate priors in this context would be challenging due to the limited availability of prior studies quantifying the effects of diversification (HHI) and market conditions on portfolio returns, rendering any assumptions made to be precarious. Although there is literature on portfolio theory that outlines specific dynamics and expectations of returns and diversification, it is prudent to assume that we have no knowledge about how HHI measures of diversification might interact with portfolio returns. Therefore, this paper elects to incorporate linear regression for its simplicity and practicality of analysis. From our model, β_1 and β_2 capture the effects of diversification (HHI) and market conditions (Year), respectively, on portfolio returns. The interaction term, β_3, is particularly important as it reveals whether the effect of diversification on portfolio returns varies between high- and low-volatility market conditions. A significant interaction term would indicate that the relationship between HHI and portfolio returns is dependent on the state of the market, highlighting the interplay between diversification and market volatility. This interaction provides a nuanced understanding of how portfolio strategy may need to adapt under different economic conditions.

3.3 Model weaknesses

Despite its effectiveness for the purpose of this paper, modelling through simple linear regression imposes the unfounded assumption of a strict linear relationship between our predictor

and outcome variables. In financial markets, the relationships between diversification, market conditions, and returns may be nonlinear or influenced by external, unobserved factors. Although our model is expanded to incorporate heterogeneous effects, it still does not capture the non-linear relationship between our predictor and outcome variable. Furthermore, the model is limited to HHI and market volatility as predictors, omitting other variables. Consequently, factors, such as sector performance, macroeconomic indicators, or specific stock characteristics, are omitted, potentially leading to omitted variable bias. Beyond the scope of confidence intervals, linear regression provides limited measures of uncertainty to our coefficients. Section 4.3 provides model validation by performing tests that verify underlying assumptions assumed of a linear regression model.

4 Results

4.1 Base Model Interpretation

Our initial regression results are summarized in Table 1, outlining estimated statistically significant coefficient values. The coefficient for β_1 suggests that a portfolio that transitions from absolute diversification (HHI=0) to complete concentration in a single stock (HHI=1) stands to realize a 4.77% return on average. However, since HHI is not a binary variable, it is more appropriate to interpret this coefficient as gaining, on average, a 0.477% return for every 0.1 increase in HHI level in the absence of volatility. Our coefficient for β_2 estimates that during volatility like those observed in 2020, our portfolios lose 14% of their value on average. Most importantly, when we inspect the interaction term coefficient, β_3, it relays that 7.9% of our portfolio value is lost when we transition to complete concentration or a 0.79% loss in returns when we increase the HHI value of our portfolio by 0.1 when we are observing a volatile market.

Table 1: Regression Analysis Summary

		Regression Results		
Explanatory Variable	Unit of Measurement	Coefficient	t-value	P-value
Intercept	Constant	0.2314282	60.982982	0.0000000
HHI	HHI (0–1)	0.0476709	2.803709	0.0051007
Year (2020)	Year (0 for 2017, 1 for 2020)	-0.1400588	-26.096823	0.0000000
Interaction (HHI x Year)	Interaction Term	-0.0792550	-3.296031	0.0009978

Note: P-values < 0.05 are considered significant.

4.2 Model Incorporation of Heterogeneous Effects

Table 2 shows the regression results when we stratify our data set into high-performing and low-performing portfolios specified by the model in the earlier section. We find that the effect of less diversification is much more pronounced with a β_1 value at 36.37% among portfolios that average returns above the 75th percentile. However, the estimate for β_3 indicates that volatility greatly reduces this estimate back down to 3.77%. On the other hand, among low-performing portfolios, less diversification actually yield negative coefficients of β_1 and β_3, indicating that there are heterogeneous effects of diversification.

Table 2: Regression Results for Top and Bottom Performers

Performer_Group	Explanatory Variable	Coefficient	Standard Error	t-Statistic	P-Value
Bottom Performers					
Bottom Performers	(Intercept)	0.0643	0.0155	4.1379	< 0.001
Bottom Performers	HHI	-0.0631	0.0348	-1.8148	0.07015
Bottom Performers	Year2020	-0.0066	0.0158	-0.4153	0.67811
Bottom Performers	HHI:Year2020	-0.1201	0.0365	-3.2921	0.00107
Top Performers					
Top Performers	(Intercept)	0.2503	0.0048	52.3413	< 0.001
Top Performers	HHI	0.3637	0.0223	16.3419	< 0.001
Top Performers	Year2020	-0.0377	0.0229	-1.6497	0.09963
Top Performers	HHI:Year2020	0.0377	0.0693	0.5436	0.58696

4.3 Model Validation

We validate our model through a series of tests according to the standards outlined by **DataDrive** (DataDrive 2021). These standards provide a framework for assessing linear regression assumptions, including tests for linearity, independence of predictors, and constant variance of residuals. However, our model validation results are not always clear, necessitating discretion and interpretation in certain cases.

4.3.1 Linearity test

Based on the linearity test results in Figure 5, we observe that the residuals plot is approximately horizontal and near zero, presenting a reasonable case of linearity between our outcome and predictor variables. However, there is a concerning trend of clustering of residuals around both ends of the plot. Although it is not a case against linearity of variables, it raises concerns about the measurement of our data that will be discussed further in Appendix B.

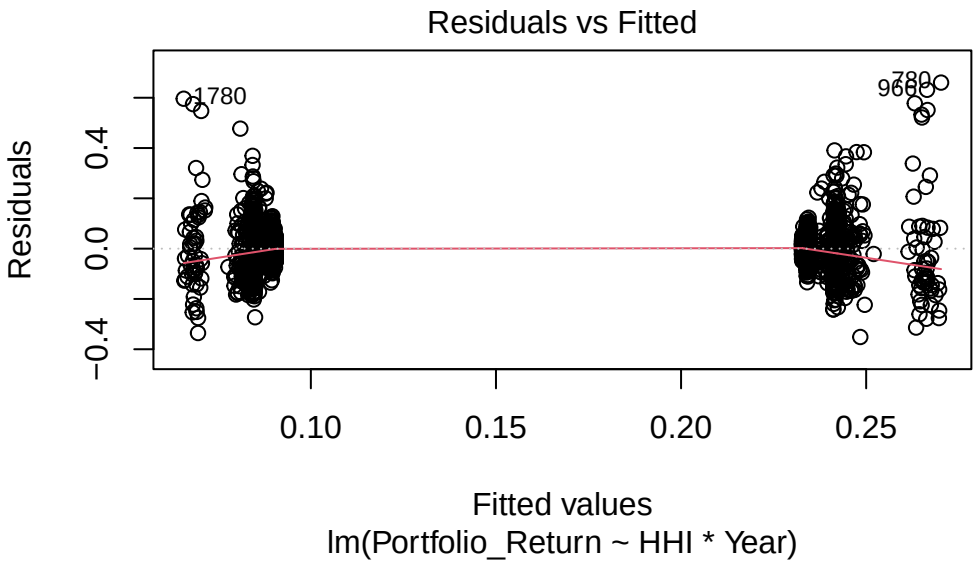

Figure 5: Residuals vs Fitted Plot

4.3.2 Predictors are Independent

Using the `car` (Fox and Weisberg 2019) package, a Durbin-Watson test was conducted on our regression model, which yielded a p-value greater than 0.05, rejecting the null hypothesis: our predictors are dependent on each other. The test results suggest that our predictors are reasonably independent of one another, supporting the validity of this assumption.

4.3.3 Residuals have mean value of 0

This assumption is easily observed from Figure 5 by confirming that our residual line closely tracks the value 0 across the entire plot. This observation supports the assumption that the residuals have a mean value of zero, supporting that the model's predictions are unbiased.

4.3.4 Residual errors have constant variance

Through the `lmtest` (Zeileis and Hothorn 2002) package, we tested for heteroscedasticity in our model by conducting a Breusch-Pagan test. Our regression model clearly violates the assumption of homoscedastcity with a p-value significantly smaller than 0.05. This violation is not a surprisingly result as the variation in portfolio returns is not solely dependent on

diversification and volatility. Reconciling for such omitted variables requires financial theory and speculation that is beyond the scope of this paper.

5 Discussion

5.1 Implications of Model Coefficients

Our estimates for our β_1, β_2, and β_3 values in both models have important implications. Although in our aggregate data set, less diversification led to positive portfolio returns, in the volatile market of 2020, this relation was reversed. The negative value of β_3 in Table 1 indicates that, in the presence of market volatility, less diversification leads to negative returns. A central motivation behind this paper is inspecting whether diversification impacts portfolios differently during different periods of market stress. Therefore, an implicit assumption of our paper is that there are differing magnitudes of diversification effects under different market conditions. Our model captures this idea by measuring diversification effects with little volatility through β_1 and, by the same token, measuring diversification effects with extreme volatility through β_3. The expanded model of heterogeneous effects expands on this narrative through Table 2.

1. **Bottom Performers**: The coefficient values for all our predictor variables are estimated to be negative among the low-performer group. Although β_2's negative prescription is not important for our analysis, negative values for both β_1 and β_3 tells us that more diversification within a low-performing portfolio yields positive returns regardless of market volatility. In presence of volatility, the average returns from total concentration of assets to absolute diversification are magnified from 6.31% to 12% approximately.
2. **Top Performers**: A stark contrast can be observed when observing coefficient estimates among top performers. More diversification hinders portfolio returns reflected by the positive values of β_1 and β_3. In fact, our model estimates that in the presence of low volatility, a mere 0.1 movement down the HHI index toward more diversification leads to 3.64% loss in returns.

5.2 Strategic Implications for Portfolio Management

Our model will not predict optimal levels of HHI levels, constrained by the structure of our HHI distribution. However, it does offer portfolio managers into some insight regarding decisions to diversify. From our base model in Table 1, our coefficient for our binary variable of market volatility gives a value of -14%. However, absolutely diversifying can offset those losses by a predicted amount of 7.9% through our coefficient for the interaction term. According to the results of our model, the decision to diversify should depend on the classification of whether a portfolio is a top-performer or bottom-performer. A portfolio that generates returns above the 75th percentile provides little justification to diversify according to our data. In fact, when

the volatility is low, diversifying can be the worst possible strategy, resulting in a 36.37% loss of portfolio returns in the case of absolute diversification. According to our model's predicted results, moving up the HHI index is the portfolio manager's best strategy if the portfolio classifies as a top-performer. On the other hand, a portfolio that generates returns below the 25th percentile has all the justification to diversify. A bottom performing portfolio's loss in a high volatility market (estimated to be 0.66% in our Table 2) can be completely offset by a 12% return if the decision of absolute diversification is taken. In a low volatility setting, this absolute diversification decision would net a 6% return. Therefore, moving down the HHI index is the portfolio manager's best strategy if the portfolio is a bottom-performer.

5.3 Alignment with Current Portfolio Theory

Our findings that diversification benefits low-performing portfolios but hinders high-performing ones expands upon prior research that often examines diversification effects in aggregate. Studies like Goetzmann and Kumar's in 2008 suggest that diversification benefits are not uniform across investors and depend on factors like risk tolerance and investment strategy (Goetzmann and Kumar 2008). Our results substantiate and refine this claim by offering empirical evidence that portfolio performance tiers moderate the impact of diversification regardless of volatility. Modern Portfolio Theory would not consider that diversification has a direct impact on portfolio returns, but it has the effect of reducing unsystemic risk. However, certain studies like the one conducted by Kritzman et al. (2010) partially discusses the aggregate economic stabilizing effects of diversification during significant market stress periods, though the author uses a different measure for diversification (Kritzman et al. 2011). This stabilizing effect could be seen in our discussion of offsetting losses brought by a volatile environment. Our result of the ubiquitous negative effects of diversification on top-performing portfolios lends credence to the ideas of the trade-off between risk reduction and return potential when diversification levels increase excessively.

5.4 Weaknesses and next steps

5.4.1 HHI as a Measure for Diversification

A reoccurring weakness of our analysis is the limited scope and inherent problems with measuring diversity with the Herfindahl-Hirschman Index (HHI).

5.4.1.1 Ambiguity of Actionable Steps to Improve Portfolio Diversification

Although HHI offers utility both in its quantifiable measurement and calculation, it does not inherently prescribe how to achieve a desired level of diversification. A portfolio manager may struggle to determine how to to achieve a targeted change in HHI. For instance, to reach a desired HHI index, weights can be adjusted with the current asset profile, assets can be

added, or both. HHI does not directly indicate which assets to add, remove, or adjust to reach a diversification goal, only an abstract value that neglects the performance of specific stocks and assets already within a portfolio. Furthermore, there is also the consideration that the HHI index involves theoretical measures of diversity that is impossible to achieve. While total concentration in a single asset (HHI = 1) can be achieved, our principle of absolute diversification (HHI=0) is impossible to acquire as this would instill an assumption of infinite assets with infinitesimally small weights in each of them.

5.4.1.2 Correlation of returns within the portfolio

The HHI index focuses on concentration within the portfolio but does not account for cross-asset correlations. Within the context of measuring returns, if two theoretical assets yield roughly the same amount of returns then HHI indices fail to explain any variation between them. In the real world, there are a multitude of asset classes that move in roughly the same magnitudes and directions. Within this framework of asset correlations, it becomes possible that a portfolio with a low HHI value is, in fact, not diversified at all. Two portfolios with the same HHI can have vastly different risk profiles if one spans multiple asset classes while the other does not. In fact, in our process of creating a company basket from which to construct portfolios, no consideration was given to the specific asset classes to choose from. Consequently, this introduces a sampling bias in our analysis that may overrepresent or underrepresent certain asset classes that have different trends in its returns. Investors likely consider a variety of assets other than stocks such as bonds, commodities, cryptocurrency, etc.

5.4.1.3 Healthy distribution of HHI values

Figure 3 showed the distribution of HHI levels among the 1,000 portfolios in our data set. Our HHI values were not evenly distributed, diverging from the conventions of a normal distribution that is expected when we randomly sample companies to create our portfolios. Most of our portfolios were concentrated on the lower end of the spectrum of HHI. This is due to our process for constructing portfolios without constraining the amount of stocks that can be used. With a high amount of stocks in any given portfolio, our HHI values come out naturally inflated, misrepresenting the sampling population.

5.4.2 Volatility represented by two extremes

Market volatility is also a complex idea that our paper simplifies for analysis. Essentially, this paper reduces the idea of market volatility to its two extremes: high volatility and low volatility as measured from the VIX index by its peak and trough from the last decade. While this binary classification facilitates clearer comparisons and allows us to focus on stark contrasts, it oversimplifies the inherently continuous nature of volatility. By limiting the analysis to only the two extremes of the spectra, we lose information that can be gained from analyzing

moderate levels of volatility in the market transitional behaviors: how portfolios respond as volatility shifts from low to moderate or from moderate to high levels.

5.4.3 Next steps for deeper analysis

5.4.3.1 Stratification of companies

To fully capture the idea of diversification, further research could refine the sampling process of selecting companies by stratifying by sector, correlation, and asset type. By creating these strata and then randomly sampling from each of them, future studies can ensure more variation in portfolio returns and a more accurate representation of the sampling population of portfolios. A robust diversification metric should account for how assets within the portfolio co-move. For instance, the following elements can be accounted for:

1. **Correlation-Adjusted HHI**: Portfolios with similar HHI values but divergent correlation structures could then be evaluated to better explain variations in returns. Portfolios with identical HHI values but divergent correlation structures (e.g., one containing highly correlated assets and another with less correlated assets) can behave differently. By analyzing these differences, future studies could better quantify the relationship between diversification and returns.
2. **HHI as concentration of asset type**: Instead of assigning weights to individual stocks, weights can be assigned to asset types, sectors, and classes. Incorporating the number of stocks in each portfolio as an additional predictor variable could further enhance the accuracy of models measuring diversification.

Such stratification could also allow for exploring sector-specific or asset-class-specific trends that are otherwise hidden in aggregated data. This has the potential to offer insight into the variation nature of performance among different asset types, allowing portfolio managers to reach more conclusive decisions.

5.4.3.2 Dynamic Volatility Analysis

Building on the limitations of reducing market volatility to two extremes, future research should incorporate a dynamic analysis of volatility. For instance, stratifying VIX levels into low, moderate, and high volatility tiers to examine how diversification impacts portfolio returns across a broader range of conditions is a natural progression of the discussion of this paper. Furthermore, investigations into specific volatility inducing events like extreme market downturns, interest rate shocks, or geopolitical events can also be incorporated to supplement the idea of volatility.

5.4.3.3 Addressing Portfolio Bias

The uneven distribution of HHI values in the current data set highlights the need for more balanced portfolio construction methods. Future studies could introduce constraints during portfolio generation to ensure

1.**Balanced HHI Ranges**: Enforcing a target distribution of HHI values that mirrors a normal distribution. This would ensure that portfolios are representative across a wide spectrum of diversification levels.

2.**Fixed Portfolio Sizes**: Standardizing the number of stocks or strata in each portfolio to avoid natural inflation of HHI values. By capping the number of assets within portfolios or introducing minimum thresholds for specific strata, researchers can mitigate natural inflation or deflation of HHI values.

Introducing these constraints would reduce sampling biases that arise when certain sectors or asset classes are overrepresented or underrepresented in the sample.

A Appendix {A}

A.1 Simulated Portfolios

The data set used in this paper is achieved through a recursive process. Due to the confidential and sensitive nature of finance, acquiring data on existing portfolios and tracking their performance is not feasible. Therefore, this paper employs a simulation-based methodology, generating random portfolios and acquiring stock price data for the assets within these simulated portfolios. While the overall process is straightforward, several nuances and methodological considerations arise in defining the sampling population, frame, and the implications of sampling design.

A.1.1 Sampling Population

For this paper, the sampling population comprises portfolios exclusively invested in company stock. This specific focus excludes other financial instruments such as bonds, commodities, and derivatives, which are often found in portfolios. This decision to restrict the sampling population aligns with the paper's focus on stock-based diversification, ensuring the interpretation of results remains clear. However, this choice narrows the applicability of the findings, as real-world portfolios often contain a mix of asset classes.

A.1.2 Sampling Frame

The sampling frame of our study is not straightforward as there are two main layers to defining it:

1. Initial Stock Selection: S&P 500 company data was scraped from Wikipedia (Wikipedia contributors 2024) using the `rvest` package (Wickham 2024). This generated a list of eligible stocks representing the broader sampling population.
2. Random Sampling: A script was implemented to randomly select 50 company tickers to form the our company basket, a subset of the S&P 500 companies. This basket is then used to generate 1,000 portfolios with random selection from this basket with random weights assigned to each stock.

Thus, in conventional views, our sampling frame is the population of portfolios that exclusively invest in companies that are contained within the S&P 500 list.

A.1.3 Sampling Method

A.1.3.1 Random Sampling

This study's iteration on random sampling is demonstrated through the selection of stocks for our company basket. By randomly sampling companies from the S&P 500 list, it introduces a sampling bias within the context of exploring diversity, talked about in section 4.2.

1. **Sectoral Representation**: By not accounting for the uneven distribution of industries within the S&P 500, the analysis fails to capture diversification effects that depend on sectoral differences since not every sector has an equal chance of being chosen.
2. **Generality of Findings**: The simulated portfolios may not reflect real-world investment practices, where diversification often considers sectoral and regional balance. This affects the external validity of the study's findings.
3. **Correlation of Movement**: Our company basket generation does not differentiate between the correlated movements between assets that exist within them.

Despite these limitations, the random sampling approach provides a manageable and replicable method to explore the relationship between diversification and portfolio performance. As the paper demonstrates, some insight can still be gained through the analysis of data acquired from our sampling methodology.

A.1.3.2 Case for Stratified Sampling

Stratified sampling would have provided a more robust and representative sampling frame for this analysis. Unlike random sampling, which treats all stocks in the S&P 500 as equally likely to be selected, stratified sampling accounts for the inherent structure and composition of the population. In our discussion of measurements for diversification, such composition and structure is important to capture within our data. Stratified sampling would allow us to do so by introducing sectoral balance. Sampling proportionally to sectoral representation, stratified sampling ensures that each sector contributes equitably to the company basket. This minimizes the risk of over-representing sectors that are present when random sampling is employed.

A.2 Reconciliation of two data sets: Omissioins in data

In our process of generating portfolio performance data, we first obtain company basket and then price data of those companies in the basket. Through obtaining stock prices from Yahoo Finance (Yahoo, Inc. 2023), price data on two companies could not be fetched at all, while two other companies contained enough empty entries to warrant dropping them completely and removing them from our company basket completely. Consequently, this has the effect of:

1. **Regenerating portfolios with updated basket**: Dropping companies requires us to update our portfolio generation process to work with a smaller basket. This process makes reproducibility of our results more challenging.
2. **A smaller basket of companies**: Constructing portfolios with a smaller basket of companies without analyzing the effects of sectoral representation could bias our data further.

The original list of 50 companies is outline in Table 3

Table 3: Complete list of companies used in our analysis

ATO	DIS	D	BR	DOC
DVN	ED	ZTS	CPAY	MMC
AMCR	MNST	ON	EIX	CSCO
GOOGL	SBAC	PEP	PAYC	PSA
LULU	WEC	LHX	SNA	CTAS
ACN	SJM	NUE	PCAR	IVZ
MET	LYB	WST	DTE	MTB
RF	ROST	CMI	LYV	TAP
WYNN	V	VZ	FSLR	ABBV
HAS	ATO	DIS	D	BR

References

DataDrive. 2021. "A Basic Guide to Testing the Assumptions of Linear Regression in r." https://godatadrive.com/blog/basic-guide-to-test-assumptions-of-linear-regression-in-r.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://www.john-fox.ca/Companion/.

Goetzmann, William N., and Alok Kumar. 2008. "Equity Portfolio Diversification." *Review of Finance* 12 (3): 433–63. https://doi.org/10.1093/rof/rfn005.

Kritzman, Mark, Yuanzhen Li, Sebastien Page, and Roberto Rigobon. 2011. "Principal Components as a Measure of Systemic Risk." *Journal of Portfolio Management* 37 (4): 112–26. https://doi.org/10.3905/jpm.2011.37.4.112.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/.

Ryan, Jeffrey A., and Joshua M. Ulrich. 2023. *quantmod: Quantitative Financial Modelling Framework*. https://CRAN.R-project.org/package=quantmod.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2024. *Rvest: Easily Harvest (Scrape) Web Pages*. https://rvest.tidyverse.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wikipedia contributors. 2024. "List of S&P 500 Companies." https://en.wikipedia.org/wiki/List_of_S%26P_500_companies.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Roger D. Peng Victoria Stodden Friedrich Leisch. Chapman; Hall/CRC.

Yahoo, Inc. 2023. "Yahoo Finance." https://finance.yahoo.com.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.