Facultad de Ciencias Económicas y Empresariales

Behavioral Biases of Retail Investors:

Quantitative Behavioral Finance Analysis and Application

Autor: Pablo Ballesteros Suárez

Director: Leandro Escobar Torres

MADRID | Junio 2021

Behavioral Biases of Retail Investors

Pablo Ballesteros

June 2021

Quantitative Behavioral Finance Analysis and Application

Abstract

Behavioral finance deviates from financial academia mainly through the absence of one of the main assumptions in financial and economic models, rational decision making. By entertaining the idea that our human nature can, at times, interfere with our ability to make rational decisions. This investigation attempts to analyze and understand where a real investor would be at a disadvantage in financial markets when compared to a perfectly rational actor. With statistical models this paper will search for evidence of areas where humans are particularly vulnerable and then try to explain the deviation from rationality through examples of natural or learned behavior.

*I would like to thank Tyler Shumway, PhD for sharing this database with me and first introducing me to this topic at the Ross School of Business, University of Michigan.

Table of Contents

List of Figures

1. Introduction

The Efficient Market Hypothesis (Fama 1965b) is based on the assumption that rational decision making combined with rational self-interest and availability of information result in the ability of investors to exploit deviations in market prices moving markets back to equilibrium.

The field of behavioral finance deviates from these assumptions and studies situations in which investors can be driven to exhibit irrational behavior. The irrationality can have different natures and affect the way investors react to prices, new information and previous experiences. The abandonment of the idea that humans are rational actors allows for a better study of actual behavior exhibited and the recognition of predictable patterns.

Research has shown that individual investors routinely underperform benchmarks (low-cost index funds) (Barber and Odean 2011). While individual investors have many disadvantages when compared to institutional investors such as availability of information and transaction costs. However, grizzled veteran Warren Buffett (1999) believes these disadvantages are more than compensated by the advantages of being a small investor, “It’s a huge structural advantage not to have a lot of money”. In this statement Buffett highlights, smaller investors have no restrictions on what they can invest in, given most funds have many rules and cannot pick small and risky companies and even if they did the large sums of money required to have an impact on an institutions portfolio would result in strong movements in stocks price. This paper studies why individual investors continuously underperform benchmarks by studying the factors involved in the decision-making process.

1.1 Objectives

The first objective of this investigation is to conduct an in-depth analysis of the academic research previously conducted in the field of Quantitative Behavioral Finance. Within this field the investigation will focus on behavioral biases that affect individual financial investors with the purpose of identifying objective and measurable metrics that can be used to compare investors. Once a subset of behavioral biases that present a thorough and complete representation of areas that affect Investors has been identified, the next step is to adapt statistical models to ensure compatibility with the database to be analyzed. With data analysis complete, conclusions from this dataset can be drawn and compared to previous findings to validate claims and study correlations between investors bias metrics.

Using only easily obtainable data from investors trading activity objective statistical models that allow for a range of behavioral biases to be measured accurately and compared to returns, a portfolio analytics tool is to be created. Under the belief that investors improve as they learn about their own ability this tool should be valuable to Investors (Seru 2009). The tool will place individual investors behavioral metrics to create a distribution of the observed population and therefore give a reference of the gravity of these biases.

Finally, relationships within the trading activity of the observed group in the dataset will be analyzed. Through thorough data analysis and exploration new findings and correlation to the value of outside funds invested will be tested. In the reviewed literature behavioral biases are considered as affecting independent events (each trade) this paper will study elements of a portfolio. “Many apparently uninformed investors trade actively, speculatively, and to their detriment. And, as a group, individual investors make systematic, not random, buying and selling decisions” (Barber and Odean 2011). Understanding these systematic movements will provide valuable insights for investment decision making.

1.2 Structure

The investigation will begin with a literary review of the current widely adopted theories present in today's financial ecosystem in the fields of both classical finance and behavioral finance. The purpose of this is to identify major trends and biases that can be prevalent and detrimental to individual investors. Additionally, setting formal definitions for basic concepts that will be used throughout the investigation and later built upon to arrive at theories and finally conclusions.

Subsequently a review of Quantitative Behavioral Finance will identify accepted metrics that can be used to describe the effect of these biases in the decision making of individual investors. Measurable metrics ad adaptable models will allow for the proposal of analytical methods to be applied to the dataset that will be studied during the majority of the investigation.

Having identified areas of vulnerability in the human decision-making process through a literary review of behavioral finance, and successively pairing them with Quantitative Behavioral Finance studies data analysis can be conducted. The dataset will be thoroughly analyzed using R-language for the replication of statistical models derive metrics and look for correlations among the studied biases. The findings will be then used to support conclusions and map distributions of the behavioral bias-derived metrics.

Finally, having mapped and worked the dataset a portfolio analytics tool will be proposed with the intent of helping investors reduce their erratic and irrational behavior, consequence of their human nature. This tool will be created with the purpose of improving the performance of Individual Investors.

2. Literature Review

2.1Behavioral Biases in Finance

In finance there are two types of behavioral biases that affect financial decision making, according to the CFA institute; “behavioral biases may be categorized as either cognitive errors or emotional biases. A single bias may, however, have aspects of both with one type of bias dominating” (CFA Institute). The nature of these biases in fundamental to the understanding of the drivers affecting an investors decision-making process.

Cognitive errors can be described as those involved in the flawed reasoning when making decisions in active investing or when drawing conclusions of ones investing ability. These biases lead to holding undiversified portfolios, excessive trading, and speculative investing with little or no information. Some of the most prevalent biases are as follows.

Overconfidence: Investor’s inclination to overestimate their trading ability, the quality of their information and an underestimation of risk. Overconfidence includes Illusion of control, where a person believes they have control over events that will affect the outcome of their investment. Planning fallacy, the failure to correctly estimate the time and resources available for a plan to reach the desired outcome, this in finance is often seen as a person’s inability to behave in accordance with their long-term goals. Wishful thinking, the tendency to overestimate the probability of positive events happening.

Self-Serving Bias: Investor’s tendency to attribute successful investments to their own ability while blaming “bad luck” for unsuccessful investments, therefore reinforcing a false believe of adequacy in their investment ability.

Reinforcement Bias: Investors tend to overweight learnings from their own experience, favoring behaviors that was rewarded with a positive outcome and repressing actions that were met with negative consequences in the past, regardless of whether the process used to make the decision was fundamentally sound and vice versa.

Emotional Biases are a consequence of investors acting on emotions caused by fluctuations in market prices or in the processing and evaluating the quality of information. The best and latest example can be seen in crypto currency markets and is known as Fear of Missing Out or FOMO, this refers to speculators buying into the peak of bull cycles afraid to miss out on stocks with abnormal returns. More common and accepted emotional biases are listed below.

Disposition Effect: First described by Kahneman and Tversky in 1979 it is the tendency of investors to realize gains prematurely as to monetize the returns whilst keeping losses in their portfolio for longer amounts of time. The disposition effect has been shown to have strong negative correlations to returns (Seru 2009) (Barber and Odean 2011). More recently the creation of momentum funds has reinforced the idea that in markets "winners" continue to outperform while "losers" continue to underperform. This cancels the idea of a reversion to the mean of stock returns and in turn explains why holding on to losers while cutting winners is detrimental to a portfolio. In depth analysis of momentum funds has shown that there are intangible factors such as brand strength and good management that allow winners to stay winners and losers to stay losers.

Attention Based Buying: Refers to the practice of buying stocks that are featured in the news and television broadcasts. Since the meteoric rise of reddit forum r/wallstreetbets this phenomenon has been shown to be all too common, the term for this has been coined as "meme stocks" nevertheless the idea remains the same. This bias has been shown to be detrimental to returns as stocks with mentions in the news and therefore abnormal trading volume were shown to be overpriced in the short run (Barber and Odean 2008). This bias leads to investors paying a premium over a stock and investing in something they didn't arrive at through sound reasoning and analysis.

Prospect Theory: Phrased as "losses loom larger than gains" (Kahneman and Tversky 1979), this human phenomenon is seen as the consequences of psychological pain endured from financial losses exceed the psychological reward form obtaining the same financial gains, in magnitude.

2.2 Selecting Biases

With the purposes of conducting an in-depth quantitative analysis of the dataset three of these biases will be selected and further studied. From the Cognitive errors' subcategory, overconfidence is perhaps the easiest to analyze as metrics on portfolio diversification and trading frequency are objective and easy to quantify. For the Emotional Biases subcategory, the Disposition effect was selected due to its high correlation with returns and the compatibility of models with the dataset. Additionally, Attention Based Buying was selected as it has both a correlation to returns and an objective and comprehensive way of measurement. Furthermore, attention-based buying may prove indicative of speculation and a lack of research by the investor. Finally, overconfidence was selected because it is easy to measure with the LDB dataset and has objective insights that are symptomatic or irrational behavior.

2.3 How do investors react to the returns of their investments? The Disposition Effect.

This section will look into how investors react to positive and negative price movements in the stocks they hold. Traditional financial models like the Markowitz model do not differentiate between buying and selling stocks, conversely, in practice retail investors rarely sell assets they did not previously buy. The decision to treat buys and sells separately allows for the analysis of emotional biases that affect investors. The disposition effect, generally described as the tendency to capitalize on paper gains in an investor's portfolio and refusing to accept losses leading investors to hold negative positions. Simply stated as selling winners and holding losers in one's portfolio. The behavior has been documented by numerous academic papers and shown to be detrimental to investor performance. Evidence has shown the disposition effect affects all

investors, not only retail investors, making it one of the most prevalent behavioral biases in financial markets.

The first study of the different treatment of gains and losses is shown in Kahneman & Tversky's (1979) paper on Prospect Theory. This paper challenges the Expected Utility Theory that states risk neutral individuals will weigh utility from gains and losses equally yielding a straight expected utility curve (1). Kahneman & Tversky find that humans overweigh losses or the possibility of losses and instead propose an adjusted expected utility curve consistent with the findings of their research (2).

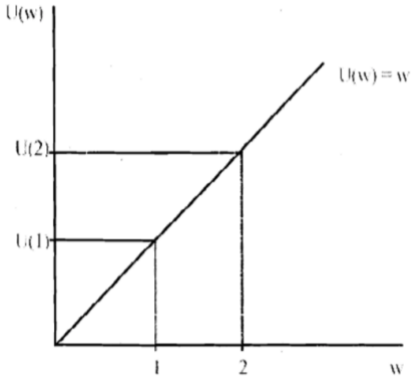


Figure 15: Prospect Theory Expected Utility Curve

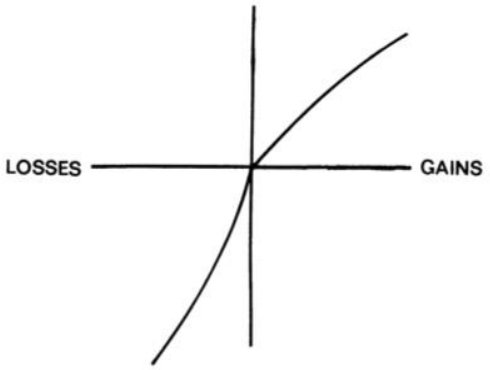


Figure 16: Risk Neutral Expected Utility Curve

Kahneman & Tversky conduct an in-depth analysis of preferences and conclude that humans treat losses differently than they do gains. Research concludes that emotions in financial decision making are inevitable and suggests "that a person who has not made peace with his losses is likely to accept gambles that would be unacceptable to him otherwise. The well-known observation that the tendency to bet on long shots increases in the course of the betting day provides some support for the hypothesis that a failure to adapt to losses or to attain an expected gain induces risk seeking." The emotional pain associated to losses is not only one of the root beliefs of behavioral finance but allows for further research to be built upon this assumption that leads to the disposition effect studied by Shefrin and Statman (1985).

In their paper “The Disposition to Sell Winners Too Early and Ride Losers Too Long” Shefrin and Statman are the first to introduce the disposition effect and its incongruencies with US tax code. They define the effect as a combination of “prospect theory; mental accounting; regret aversion; and self-control…In addition, we introduce a fifth element, the potential gain to be had from exploiting Constantinides strategy.” Prospect theory refers to the different emotional assimilation of losses versus gains. Mental accounting in regard to the failure to exploit the tax advantage of incurring a capital loss, as the position is mentally pegged to the acquisition price and the global implications on the portfolio are ignored. Regret aversion explains why investors refuse to take a loss based on the belief they may regret not holding the stock to break even or seeking the pride of “winning” on a position. Self-control as the succumbing to the urge of realizing a gain and adding a “win” to the “score”. Finally, Shefrin and Statman explain that the US tax code treats short term losses differently that long term capital gains. Long term capital gains are taxed at a lower rate than short term losses which are taxed as income, this is an incentive for investors to act against what the disposition effect dictates; however, their research observes evidence investors fail to act rationally.

Terrance Odean in his publication “Are Investors Reluctant to Realize Their Losses?” (1998) further analyzes the disposition effect using quantitative methods and finds that investor behavior “does not appear to be motivated by a desire to rebalance portfolios, or to avoid the higher trading costs of low-priced stocks, nor is it justified by subsequent portfolio performance. For taxable investments, it is suboptimal and leads to lower after-tax returns.” Furthermore, Odean studies the consequences of the disposition effect on investors’ portfolios and observes “winning investments that investors choose to sell continue in subsequent months to outperform the losers they keep.” This finding adds to the benefits of correcting this adverse behavior.

Odean (1998) looks at trading records for 10,000 accounts from 1987 through 1993. For his quantitative analysis he uses two ratios calculated by the following formulae, where realized

gains/losses refer to positions that have been sold and paper gains/losses refer to positions still in an investor's portfolio.

$$\frac{\text{Realized Gains}}{\text{Realized Gains} + \text{Paper Gains}} = \text{Proportion of Gains Realized (PGR)} \quad (1)$$

$$\frac{\text{Realized Losses}}{\text{Realized Losses} + \text{Paper Losses}} = \text{Proportion of Losses Realized (PLR)} \quad (2)$$

Using these two ratios, Proportion of Gains Realized (PGR) and Proportion of Losses Realized (PLR), he compares investors tendency to liquidate positions at profit or loss. His findings support the disposition effects hypothesis.

	Entire Year	December	Jan.–Nov.
PLR	0.098	0.128	0.094
PGR	0.148	0.108	0.152
Difference in proportions	−0.050	0.020	−0.058
t-statistic	−35	4.3	−38

The table above shows how PLR exceeds PGR by 5% and is statistically significant, this relationship holds in the months of January through November and is reversed in December or at the end of the fiscal year, which supports Constantinides strategy mentioned in Shefrin and Statman (1985). This relationship can be observed in the following graphic.

Are Investors Reluctant to Realize Their Losses?

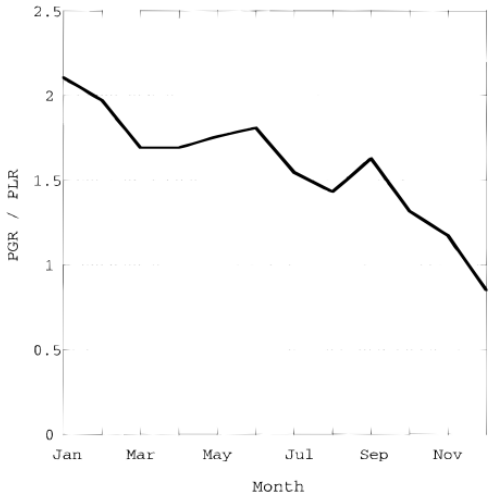


Figure 3 illustrates the effects of capital loss harvesting for tax benefits on the measurement of the disposition effect, however the proportion of paper gains realized still is significantly higher than the proportion of losses realized. To conclude the literary review of Odean's "Are Investors Reluctant to Realize Their Losses?" we can answer yeah, they are.

Figure 17: The effects of capital loss harvesting

In 2009 Amit Seru, Tyler Shumway and Noah Stoffman study the disposition effect by analyzing a robust dataset of Finnish investors from 1995 to 2003 with over 22 million trades made by households. The main objective of their research is to learn whether investors improve as they gain experience from trading activity. They come to find that investors with poor performance cease trading as they learn about their own ability or lack of it, they also come to find that on average, "An additional year of experience increases average 30-day post-purchase returns by 41 − 4 = 37 bp, or approximately 3 percent at an annualized rate. An additional 100 trades increases returns at slightly over one-fourth of this rate." Saru, Shumway, and Stoffman (SSS) attribute this in part to observing a declining disposition coefficient as traders gain experience. This paper is set apart by the decision to account for investor attrition[1], avoiding survivorship bias.

In their analysis SSS study each account and trading year individually to isolate trader's performance. They estimate the disposition effect using a hazard model that calculates probability of an investor selling a position at time t (30 days after the purchase) given a dummy variable, 1, for when purchase price < price at time t and 0 otherwise. Hazard models are well suited for this analysis and the accurate method for measuring the disposition effect.

In their paper The Behavior of Individual Investors (Brad M. Barber and Terrance Odean 2011) they study why individual investors routinely underperform benchmarks and make irrational financial decisions detrimental to their wealth. Barber and Odean continue to study the disposition effect using hazard models to measure it. They improve on SSS model by using daily observations and total returns. For this study Barber an Odean use both the Large Discount Brokerage dataset to be analyzed in this paper and the Finnish dataset used by SSS. Barber and Odean conclude that reinforcement learning is to blame for the disposition effect, the positive emotions from realizing a successful trade conditions investors to repeat that action while the

[1] Investor Attrition: Refers to investors who stop trading after realizing their ability is poor.

emotional pain from closing or accepting a losing trade conditions investors to avoid making that decision.

Review of previous research regarding the disposition effect shows a strong prevalence of this bias among individual investors. However, Andrea Frazzini (2006) challenges the notion that institutional investors do not show signs of behavioral biases in their investment decisions, "[He] also document[s] the extent of the disposition effect among mutual fund man- agers and show[s] that it adversely affects returns. Loser funds tend to be as dis- position prone as retail investors." Frazzini uses the PGR, PLR method used by Odean (1998) and finds "The magnitude of the aggregate difference (PGR − PLR) is around 3%, which is smaller than the average 5% reported by Odean (1998) for retail investors, but still of the same order of magnitude." This finding demonstrates the clear and unavoidable tendency of humans to hold on to losers and sell winners.

The disposition effect, derived from prospect theory, is a consequence of the internalization of loss by individuals. It is evident, despite tax incentives to act differently, and in line with other behavioral biases. It is hazardous to investment returns leading investors to underperform benchmarks and can form negative feedback loops that reinforce the behavior. Investors need only to let winners run and cut looser early, however, even hedge fund manages seem to be unable to escape the disposition effect.

2.4 What drives an individual investor to buy a stock? Attention based buying.

"Investors have time to weigh the merits of only a limited number of stocks. Why do they consider some stocks and not others?" (Odean 2006)

This section will analyze what factors drive individual investors to trade and more specifically the decision to buy any given stock. Individual investors must decide from thousands of options what firms' stock to purchase. Understanding the factors that affect the process of investors selecting, analyzing, and deciding to buy a stock while removing the assumptions of rationality and ability to acquire and process limitless information help better understand the behavioral components that drive individual investors.

One of the first papers to analyze what drives trading volume and its relation to events in the stock market is "A Theory of Trading Volume" by Jonathan M. Karpoff in 1986 where he proposes that informational events such as relevant news or earnings reports drive volume as they generate disagreements between informed traders as an "exchange occurs when market agents assign different values to an asset" (Karpoff 1986). The relation between earnings and trading volume is proven to be more extreme when there are 'surprises' (Brown and Han 1992). In a later publication Bamber, Barron and Stober (1997) confirm that disagreement leads to increased trading volume while acknowledging that liquidity trading[2] contributes to trading volume. While Karpoff (1987), Brown and Han (1992) and, Bamber Barron and Stober (1997) all agree earnings reports and informational events (news) lead to increased trading volume they do not analyze the decision to buy. In markets for every transaction there is both a buyer and a seller, furthermore, their research fails to distinguish between sophisticated and retail investors.

Individual investors are those who believe in their ability to invest actively in the stock market, however they do not have teams of analysts running complex models and calculations using most or all of the information available. Individual investors are limited by; the time they can spend analyzing a stock, their cognitive ability and, the quality of the information available to them. With thousands of stocks listed in the various exchanges across the world and "unable to evaluate each security, [individual] investors are likely to consider purchasing securities to which their attention has been drawn" (Odean 1999). Consequently Grullon, Kanatas, and

[2] Liquidity Trading: selling a position to free capital and not based on an assumption of future prices.

Weston (2004) find that “firms that spend more on advertising attract a significant larger number of both individual and institutional investors. We also find that advertising improves stock liquidity by reducing trading costs.” This finding further reinforces attention-based buying as a commonplace bias present among both individual and sophisticated investors.

Attention based buying refers to the tendency of investors to buy stocks that have been mentioned recently in the news. This bias is framed as a search problem referencing the hundreds of thousands of investment alternatives and within that the thousands of stocks available to retail investors. Odean (1999) synthesizes this problem saying, “Investors do not buy all stocks that catch their attention; however, for the most part, they only buy stocks that do so.” Logically investors who chose to buy a certain stock will do so based on some sort of reasoning and for that to happen they have to have previously acquired information; this predominantly happens when reading the news. However, once this assumption is made investors are still faced with a choice to make from the subset of stocks that have been mentioned or ‘trending’ in the news. They may vary as “Contrarian investors, for example, will tend to buy out–of-favor stocks that catch their eye, while momentum investors will chase recent performers.” (Odean 2006) nonetheless individual investors in aggregate will be buying stocks in the news rather than selling.

While a sophisticated investor will see the choice to buy or sell a particular stock equally and will in turn be just as likely to sell (short) a stock they don't like as they are to buy a stock that they do like. However, Odean (2006) observes less than 1.0% of retail investors hold short positions. Understanding why retail investors do not regularly sell stocks they do not own, is due in part to the fact that they are constrained to selling the stocks that make up their portfolio. Further analysis of individual investor portfolios by Odean (2006) shows the average investor holds 4.3 different stocks in a period of a month, naturally the subset of stocks available for them to buy is far larger than the subset of stocks available for them to sell. Consequently, when reading or watching the news investors will be learning about stocks that they do not own and be faced with the decision whether to buy or not rather than to buy or to sell. Finally, Odean finds that institutional or sophisticated investors do not show the same levels of attention-based

buying given, they do not rely on the news for information on stocks rather 'analysts' and the fact that they tend to own larger and more diversified portfolios.

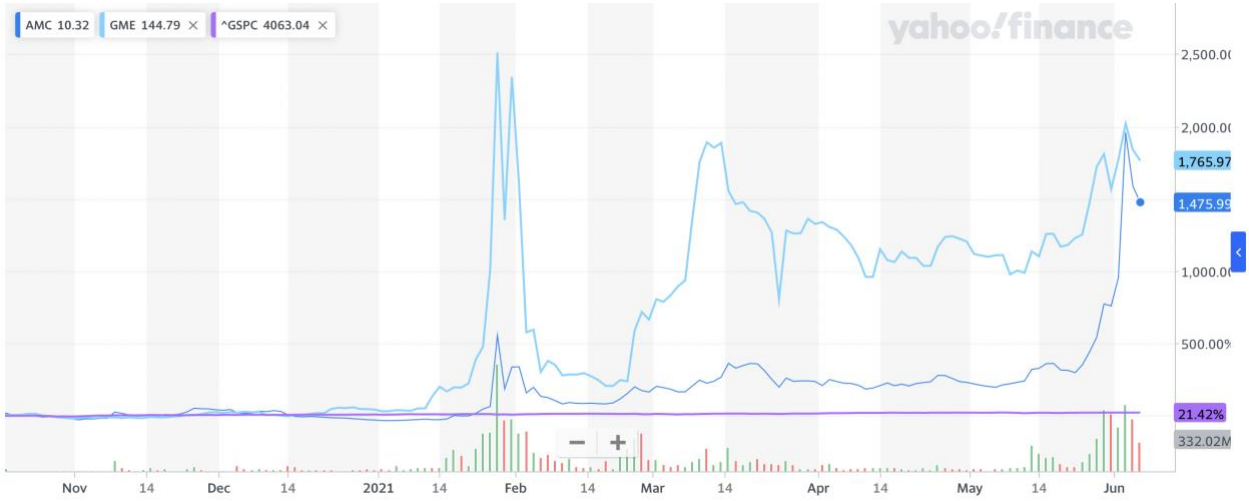
There is ample evidence informational events (news) drive trading volume (Bamber, Barron and Stober 1997) and that retail investors seem to be net buyers of attention-grabbing stocks (Odean 2006). Furthermore, coupled with evidence that increased media presence has noticeable effects on stocks "by increasing the breadth of ownership and the improvement in liquidity, advertising may increase firm value" (Grullon, Kanatas, and Weston 2004). There is a clear impact consequential to investors who decide to buy firms as a result of having heard of them in the news.

In the publication "All That Glitters" Odean (2006) analyzes the discount brokerage firm database and assigns three measure to determine whether an investor was paying attention to a stock, trading volume, previous one day return and whether the related company appeared in the news. Odean's intention is to determine if an investor is a net buyer of a stock as opposed to if the investor simply traded on the given date. Odeans methodology is particularly relevant to this paper as the large discount brokerage database (LDB database) has data for the trading volume and monthly average volume for the traded stocks, allowing for the analysis of attention-based buying in the data analysis. When analyzing volume Odean (2006) finds "investors at the large discount brokerage make nearly twice as many purchases as sales of stocks experiencing unusually high trading volume (the highest five percent) and nearly twice as many purchases as sales of stocks with an extremely poor return (lowest five percent) the previous day." Evidence suggests traders are likely to be driven by the same informational events leading to abnormally high trading volume this mass effect is the result of the behavioral bias known as attention-based buying.

Further research on attention grabbing events and its role a a behavioral bias for individual investors is shown by Seasholes and Wu (2007). They find evidence that attention grabbing stocks are overpriced in the short-term leading investors who buy as a result of

attention-based bias at a loss. Their analysis of the Shanghai Stock Exchange finds “The day after an attention-grabbing event, individual investors are net buyers and prices appear to be “pushed” upward. Between dates t + 1 and t + 6 prices mean-revert back to pre-event levels.” Furthermore, they hypothesize that there are actors who anticipate and profit from attention grabbing events, “and earn an average daily profit of 1.16% (0.71% net of transaction costs).” This analysis is in line with Odean 2006 and suggests individual investors who trade as a result of informational events or abnormal previous day returns are providing liquidity for well-informed rational institutional investors. These conclusions show the dangers of attention-based buying as a behavior detrimental to returns.

Attention based buying is a bias that has been exacerbated by the rise of mass media. Through the Covid-19 Pandemic social media platform Reddit has seen groups of individual investors band together and ‘pump’ what has come to be known as meme stocks. Two of these stocks stand out, the first GME which has returned more than +5200% from June 1st, 2020, to June 1st, 2021. While extremely volatile and risky many traders have joined the ranks of the reddit forum r/wallstreetbets, which has sustained a heavy inflow of purchases and allowed the stocks to balloon. The second-best performing meme stock is AMC which has returned almost +1500% from November 2020 to June 2021 also blowing the S&P 500 out of the water. While there are many reasons why these stocks have enjoyed abnormal returns there is little evidence to suggest this behavior would have happened without the heavy news coverage and social media mentions. The reddit forum r/wallstreetbets saw an increase in members of +1000% through the pandemic and news articles mentioning these meme stocks saw a similar pattern to the prices of these stocks.



(Source: Yahoo Finance June 2021)

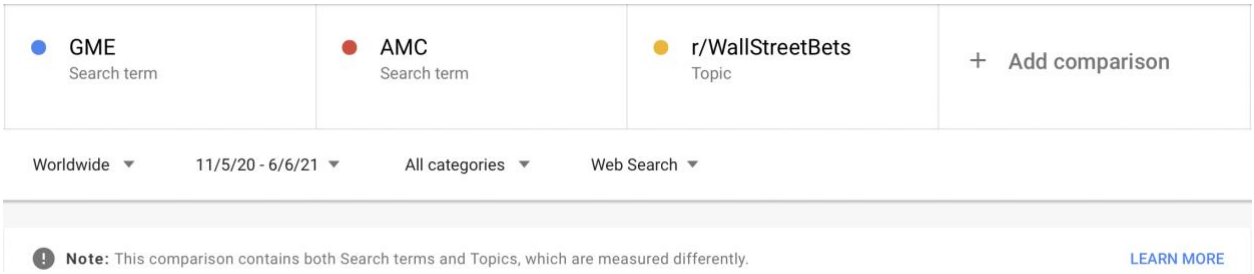


Figure 18: Meme stock vs SP500 returns Nov 2020 - June 2021

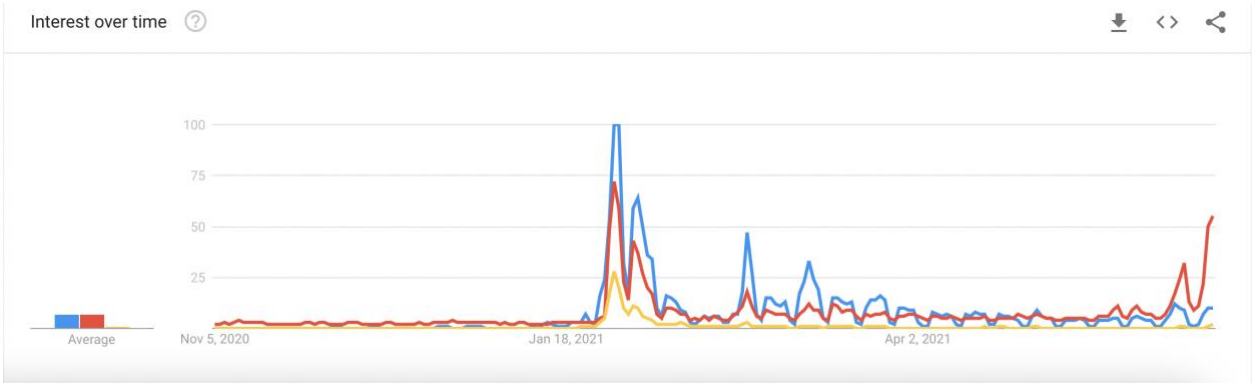


Figure 19: Meme stock and r/wallstreetbest mentions Nov 2020 - June 2021

(Source: Google Trends June 2021)

2.5 Are investors aware of their own ability?

"Two aspects of knowledge are what one believes to be true and how confident one is in that belief …While it is often not difficult to assess the veridicality of a belief … evaluating the validity of a degree of confidence is more difficult." (Fischhoff, Baruch 1977)

Investors seem eager to trade and test their skill (luck?), despite evidence that **"t**he average individual investor underperforms a market index by 1.5% per year. Active traders underperform by 6.5% annually" (Odean 200). Tan investors perception of their ability to understand the market and select stocks they believe will outperform may be warped. Studies have shown humans are particularly bad at evaluating the credibility of the information available but rather focus on the "extremeness" of it (Griffin and Tversky 1992). Accurate appraisal of one's own ability and the quality of the information is critical for the success of active investors, for individual investors mistakes in these areas will lead to poor performance.

Studies have shown humans tend to overestimate their own predictive ability (Fischhoff, Baruch 1977), this is particularly evident when difficult tasks are involved such as selecting securities (Griffin and Tversky 1992). An investor who decides to invest actively must believe in his ability to outperform the market. Investment decisions are made by the combination of information and the investors ability to correctly interpret it and estimate its impact on markets. Trading is driven by disagreements (Karpoff 1986) and in markets (excluding fees) are a zero-sum game, meaning that for every investor that buys a 'winner' another investor must sell it to him. Growing numbers of retail investors who actively and aggressively manage their portfolios (Lust et. al. 2021) suggest strong confidence in their own abilities, however, in aggregate individual investors underperform the market, meaning many over-estimate their ability.

Overconfidence bias extends beyond finance and is present among professionals in most fields. In finance it has been documented to affect corporate financial executives finding “realized market returns are within the executives ’80% confidence intervals only 36% of the time.” (Ben-David, Graham and Harvey 2013) the effect has also been found to influence investment bankers and professional traders (Glaser, Langer and Weber 2013). In finance professionals seem to display lower levels of behavioral bias, however they cannot seem to escape overconfidence, with advanced metrics and extensive feedback sophisticated investors are undoubtedly more prepared to avoid these hazardous biases than retail investors.

Investors both sophisticated and unsophisticated are routinely overconfident. Overconfidence leads investors to trade too much (Graham et al. 2006) because they believe they possess special knowledge or abilities. Overconfident investors make larger volume trades (Glaser and Weber 2003), underestimate risk (Pompian 2006) and hold riskier portfolios (Odean 1998).

Excessive trading is widely documented and has been shown to reduce investor returns. Barber and Odean (2000) measure the detriment of excessive trading finding “households that trade frequently earn a net annualized geometric mean return of 11.4 percent, and those that trade infrequently earn 18.5 percent” a sizable and compounding cost associated to overconfidence. The difference in net annualized returns is mostly attributed to transaction costs, while markets have evolved to reduce fees and increase liquidity mitigating this effect overconfidence remains noxious to investment performance. In a controlled study of university students that had taken financial courses but had no experience high overconfidence students earned an average return of -5.54% opposed to low overconfidence students who earned an average return of -0.53% (Trinugroho and Sembel 2011). While inexperienced students in a controlled study should not be used to estimate the effects on individual investors the study shows that excessive trading even without large fees has a negative impact on performance.

While overconfidence has been observed across the financial industry it is particularly evident in certain demographics and festers in the right market conditions. Barber and Odean (1998) find excess volume (market turnover rate) in equity markets that cannot be explained by rational investor behavior. Long periods of strong market performance have been linked with higher levels of overconfidence. Statman, Thorley, Vorkink (2006) conclude “biased self-attribution causes the degree of overconfidence to vary with realized market outcomes”. Furthermore, there are risk factors associated to gender and marital status, Barber and Odean (2011) find “men trade 45 percent more than women. Trading reduces men’s net returns by 2.65 percentage points a year as opposed to 1.72 percentage points for women.” The effect is even more evident when single men are studied. The strong correlation between overconfidence observed by excessive trading is indicative of underlying human characteristics.

3 Method and Data Analysis

3.1 Dataset Description

The dataset named LDB contains 207497 trades made by 4174 Charles Schwab discount brokerage accounts from 1991 to 1996. The trades analyzed were exclusively of publicly traded companies standard issue stock. Discount brokerage services do not include financial advisory. The dataset was shared by Tyler Shumway Ph. D. who at the time was a professor in the financial department at the Ross School of Business. This dataset is referenced in many publications by Brad M. Barber Ph. D. and Terrance Odean Ph. D.

R language was chosen to process the dataset, Microsoft Excel was not an option as the size of the dataset exceeded its capabilities.

Variables analyzed are as follows:

1. Account number (“account”) -> Categorical, identifies the account that executed a trade.
2. Ticker (“ticker”) -> Categorical, ticker of the stock that was traded.
3. Buy/Sell (“bs”) -> Categorical, indicates whether the trade was for a purchase or sale of stock.

4. Number of Shares (“shares”) -> Numeric, Indicates the number of shares bought or sold in the transaction.
5. Price (“price”) -> Numeric, price at which the trade was executed, not including fees or bid ask spread.
6. Price at close (“pclose”) -> Numeric, price at which the traded equity closed the day the trade was executed.
7. Price 5 days (“p5days”) -> Numeric, price at close 5 days after the transaction was executed.
8. Price 20 days (“p20days”) -> Numeric, price at close 20 days after the transaction was executed.
9. Volume (“vol”) -> Numeric, trading volume of the day in which the trade was executed.
10. Average Volume (“avgvol”) ->Numeric, arithmetic mean of daily trading volume for the calendar year in which the trade was executed.
11. Count (“cnt”) -> Numeric, number of trades made by each account.
12. Sex (“sex”) -> Categorical, sex of account owner.
13. Income (“income”) -> Categorical, decile of income unrelated to investment activities.
14. Median duration (“mduration”) ->Numeric, holding period in days for which the share was held, 9999 indicates the position was never sold.

Additionally, a vector with the variable Transaction value (“tradevalue”) was created by multiplying variables Price and Shares. Other variables that provided further information related to the account owner such as home ownership, marital status, and zip code were deleted as the information used is meant to be purely financial.

3.2 Data Processing

	X7109	X910118	CNR	B	X100	X27.25	X27.37500	X24.75	X25.875	X455400	.	X54	M	X6	X40	CA	X92666	O	X1	M.1	X13	X0	X5589(
1	7109	910131	CNR	S	-100	27.25000	26.7500	27.75	24.625	558900	385710	54	M	6	40	CA	92666	O	1	M	.	.	.
2	7109	910201	ATM	S	-100	24.87500	24.6250	23.25	27.125	102500	48590	54	M	6	40	CA	92666	O	1	M	.	.	.
3	7109	910201	QUIK	S	-100	9.25000	10.2500	11.25	12.5	190800	71621.2	54	M	6	40	CA	92666	O	1	M	.	.	.
4	7109	910315	ASTA	B	100	22.25000	22.5000	26	30.25	2038068	731149.3	54	M	6	40	CA	92666	O	1	M	1007	0.0561	121
5	7109	910315	SRR	B	100	36.87500	36.7500	38.25	43.75	87700	126990	54	M	6	40	CA	92666	O	1	M	679	-0.461	307
6	7109	910404	CNR	B	100	24.00000	24.0000	26.125	24.5	1826500	850930	54	M	6	40	CA	92666	O	1	M	671	-0.026	502
7	7109	910819	AGN	B	100	21.75000	21.7500	21.5	21.375	166500	173730	54	M	6	40	CA	92666	O	1	M	534	0.0344	159
8	7109	920210	MIKE	B	100	19.62500	19.8750	20.75	23.25	321929	607171.7	54	M	6	40	CA	92666	O	1	M	9999	.	.
9	7109	930121	SYNC	B	100	27.00000	26.8750	26.5	25	13470	21048.4	54	M	6	40	CA	92666	O	1	M	587	-0.2962	753

Figure 20: Top ten rows of Large Discount Brokerage (LDB) firm dataset

Figure 6 shows the first 10 rows of the dataset, naturally the first step to the analysis is to add column names in order to identify and manipulate the database, to do so a column was added in order to avoid losing any data.

To continue the data cleaning process various variables were set to null, this was done with two reasons in mind. The first reason is the eliminated variables do not contribute relevant or necessary information for the behavioral bias metric calculations. The second is the intention to use minimal information to allow for scalability when a tool to analyze investors is created. Data from trading activity has many shapes and forms but usually contains related ticker, price, quantity, buy or sell, and date with this information, the rest of the relevant variables can be found like closing price, trading volume on transaction date, price and volume at times t and t±1.

The following lines of code were used to find the mean number of trades placed by the investors whose trades populated the database. The resulting mean was 49.71 which over the five-year period is roughly equal to 10 trades per year per account. The formula: [print(mean(att$cnt))] would have printed the arithmetic mean of the variable count which is skewed returning a value of 56.53 due to higher count values appearing more often, as the variable count is constant for each account and equals the total number of trades executed during the observed period.

In order to determine whether a trade happened in attention-based buying conditions, two new variables were created from the (“volume”) variable. The first called (“vol.ratio”) represents the quotient of trading volume on transaction date over average volume on the month of the transaction. The second variable (“vfour”) is a dummy variable which returns a 1 for trades placed on days where (“vol.ratio”) exceeds four and 0 for when it doesn’t. The cut off of four was selected as just over 10% of the trades placed exceed a quotient of four which shows that on the day the trade was placed abnormal volume occurred.

With the purpose of finding new conclusions in a dataset thoroughly worked by Ph. D. s and behavioral finance giants like Brad M. Barber and Terrance Odean, a decision was made to generate a new variable and to test its relation to the behavioral bias metrics. The new variable created called (“value”) was calculated by going into each account and adding the created variable (“tradevalue”) when the trade corresponded to a buy and subtracting (“tradevalue”) when the investor sold a position. The rationale behind this variable is that it allows for the analysis of the amount invested procured from non-investment activities, in other words it measures the “skin in the game” every investor adds to their accounts. To calculate this variable the following for loop was used to iterate through the 4174 accounts and each of the 207497 trades in the database.

Next, iterating through all accounts and transactions to create a new data-frame which consists of the accounts with distilled metrics for the significant variables.

The new data-frame contains one row with:

- The maximum value of the variable (“value”) which measures the point in which the amount invested less the amount sold reached its maximum during the time period.
- The arithmetic mean of the variable (“vol.ratio”) which measures the average volume multiple at which the investor executed his trades.
- The number of trades executed at a volume multiple of more than four.
- The number of trades placed by each account, measured by the amount of row entries corresponding to each account number.
- The number of tickers or stocks invested in the time period by each account.
- The arithmetic mean holding period for each account.
- The median holding period for each account.
- The arithmetic mean 20 day returns of trades where shares were bought.
- The arithmetic mean 20 day returns of trades where shares were sold.

- The arithmetic mean 5 day returns of trades where shares were bought.
- The arithmetic mean 5 day returns of trades where shares were sold.
- The arithmetic mean 1 day returns of trades where shares were bought.
- The arithmetic mean 1 day returns of trades where shares were sold.

The created data-frame named ("uaccounts") aggregates processed data organized by account number with a single row entry for every account. The columns as mentioned above allow for a apples-to-apples comparison of trading activity.

In order to calculate the disposition effect calculating the two ratios, Proportion of Gains Realized (PGR) and Proportion of Losses Realized (PLR) is necessary. For these calculations it was necessary to separate the trades and create a subset of only the stock purchases, it was also necessary to separate the trades into paper gains and paper losses. With the distinctions made it was possible to look at all trades in binary depending on whether the trade was realized or not and at a gain or not. For this analysis PGR and PLR were calculated two ways the first using the number of trades, for example, one portfolio with one unrealized (paper) gain and one realized gain would yield a PGR, the same applies for PLR. The second way of calculation weights the positions based on their value at date t+20 rather than all trades equally, in other words a dollar weighted PGR/PLR. Code can be seen below.

Once the data has been thoroughly worked the last step is to assign all accounts to a decile based on the variable ("max") and variable ("r20db"), this allows us to compare the top 10% of investors by returns and net amount invested. Using these quantiles comparisons can be drawn between the best and worst performing and the investors who invest the most.

Additionally, a second measure of the disposition effect can be calculated by creating a linear regression to predict the variable ("Realized") this variable is binary and returns a 1 for a position (buy trade) that has been sold inside time t+20 and 0 for positions that were held through 20 days. This regression encountered a challenge given the small number of trades for

each account, to get around this the regression was ran aggregating the accounts into deciles for both variables (“max”) and (“r20db”).

To measure the level of overconfidence in investors the number of trades placed over the period was used, to study how overconfident investors measure up against the sample mean in different metrics the data was normalized using min-max normalization. The normalized data was then split into quantiles based on number of transactions as a measure of overconfidence.

The final step of this data processing journey was to get quantile averages for each of the ten deciles, again an iterating for loop was used to further distill the dataset coming from over 200,000 rows to just 10 in a data frame called Dec1 and Dec 2. Then the means are compared using ANOVA means comparison for statistical significance. The Uaccounts data frame, Dec1 and Dec2 data frames are exported as CSV’s so that they can be graphically analyzed on excel.

3.3 Results

- Descriptive statistics

Figure 7 shows the descriptive statistics for the generated array uaccounts, these values are the calculated metrics and averages for each of the 4174 accounts present in the database.

	20-day Returns	5-day Returns	PGR	PLR
Mean	0.91%	0.50%	11.93%	8.06%
Standard Error	0.07%	0.05%	0.26%	0.20%
Median	0.83%	0.39%	4.35%	0.00%
Standard Deviation	4.37%	2.97%	17.09%	12.79%
Minimum	-12.41%	-8.08%	0.00%	0.00%
Maximum	199.18%	151.24%	100.00%	100.00%
Count	4174	4174	4174	4174
Median	$ 50,756.25	1.98	4.00	44.00
Standard Deviation	$ 197,652.03	2.54	4.11	18.42
Minimum	$ -	0.00	0.00	30.00
Maximum	$ 9,348,437.90	43.49	32.00	129.00
Count	4174	4174	4174	4174

	Diversification	Holding Period Mean	Holding Period Median	tperiod
Mean	19.74	666.23	457.89	4.56
Standard Error	0.14	3.11	7.72	0.01
Median	18.00	634.65	290.75	5.00
Standard Deviation	8.98	201.11	498.89	0.86
Minimum	1.00	1.00	1.00	0.00
Maximum	93.00	1667.00	1667.00	5.00
Count	4174	4174	4174	4174

Figure 21: Descriptive Statistics

Some takeaways from the descriptive statistics of this array are the mean Volume Ratio which is 2.64 this is indicative of retail investor suffering attention-based bias regularly. Furthermore, the mean of the variable (“divr”) or diversification as seen on the table was almost 20, while this doesn't imply the 20 different tickers were held at any one time it is a good indication of low diversification amongst retail investors, pairing this average with a median holding period of 458 days or 1.25 years and an average trading period of 4.56 years for each account we can assume most accounts never held more than 20 different stocks. Holding 20

randomly selected stocks is considered enough in terms of the marginal benefits achieved by diversification. (Evans, John L., and Stephen H. Archer, 1968)

```
mtx <- subset.data.frame(uaccounts, select = c(trann, divr, vr4, max, PGR, PLR))
mtx <- mtx[is.finite(rowSums(mtx)),]
corrplot(cor(mtx, use= "complete.obs"))
```

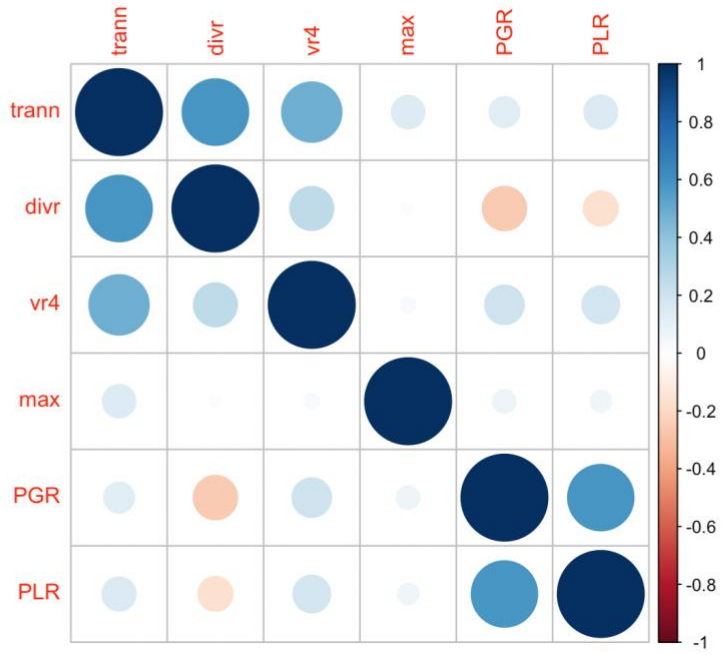


Figure 22: Correlation Matrix

relationships to each other can be observed. One of the most interesting findings in this correlation matrix is the negative correlation between paper gains realized (PGR) and diversification (divr) which counts all the tickers traded during the observed period. Low diversification is one of the main symptoms of overconfidence and it is interesting to see the potential link there may be between overconfidence and the disposition effect which is measured by PGR. The second, yet lesser, interesting correlation is the one between PGR and vr4. With vr4 being a metric that counts the number of trades that happened when there was abnormal trading volume (> 4 times the average monthly volume). The relation is interesting given that a positive correlation indicates that investors that suffer from the disposition effect also suffer from attention-based buying. Other than this the correlations can be, for the most part, logically explained. The correlation between number of transactions and tickers traded (trann and divr) is

in all likelihood explained by the idea that as an investor trades more frequently he is more likely to have the opportunity to buy different stocks. The same principle can be applied to the correlation between number of transactions and volume multiples greater than 4 as the more trades place the more likely some of those trades happened under abnormal volume.

The correlation seen between the three biases studied might be explained by the lack of education and resources available to retail investors. It is also likely that the correlation is due to some traders being more prone to behavioral biases than others, this meaning that they allow their emotions and cognitive biases to influence their trading decisions. An investor who acts irrationally in some facets of the trading process is likely to be irrational in all of them and vice versa, perhaps explaining this correlation.

4 Findings

4.1 The Disposition Effect

The first bias analyzed was the disposition effect to calculate metrics that could accurately indicate this bias the PGR and PLR method used by Odean 1998. Additionally, 2 linear regressions were used to calculate the disposition effect. The dependent variable trying to be predicted was binary, whether the trade was realized (the bough stock was sold) within 20 days or not. The first regression used a dummy variable that returned a one if the price of the stock at time t+20 was greater than the price of the stock at time t or purchase date. The second regression looked to predict the same dummy variable, whether the stock was realized (sold) within 20 days, for this second regression the variable of 20-day returns was used.

For the 4174 accounts the PGR was 11.93% while the PLR was 8.06% this yields a difference, (PGR-PLR) of 3.87%. This metric shows that, as a whole, the retail investors studied in the discount brokerage dataset suffered from the disposition effect. The PGR and PLR analysis shows that investors were more prone to selling stocks at a profit than selling stocks at a loss. Furthermore, running a regression using the 20 day returns of stocks bought to predict whether a position would be sold within 20 days or not yielded the following output.

Furthermore, a logistic regression was used to determine if 20-day returns increased the probability of investors realizing their trade at or before 20 days. Then the regression was run two more times, the first time for the for the top 10% of investors on the calculated metric; value, a proxy for new funds invested. Interestingly the regressions find a much higher coefficient for 20-day returns meaning that investors who invested more outside funds into their portfolios had higher disposition coefficients than the bottom 10% of investors in the value metric. The coefficients in both cases were significant at a level of <.001.

Figure 9 shows the strong correlation between the disposition coefficient (PGRa- PLRa) and the value metric of investor account funding deciles. Strengthening the conclusion that investors who add more 'outside money' to their accounts display more disposition effect.

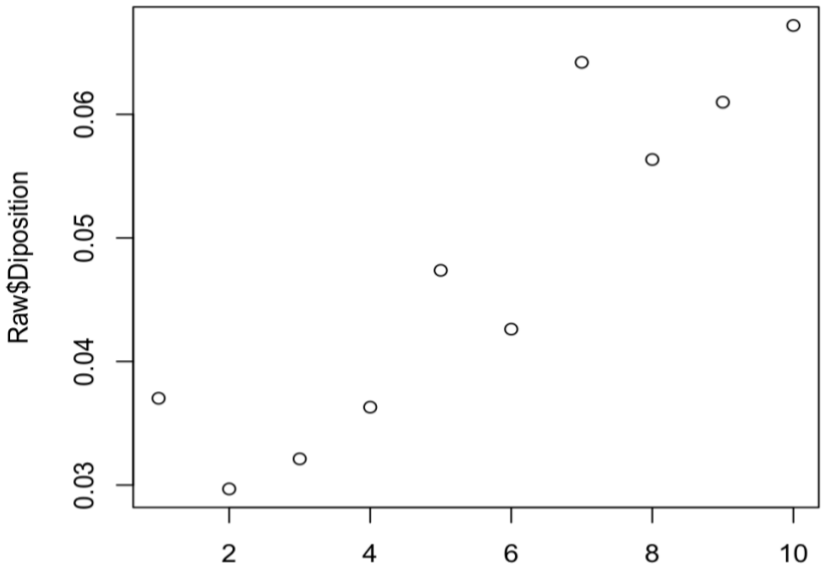


Figure 23: Disposition coefficient vs by value deciles

Interestingly the graph above finds a strong positive trend between the value deciles and the 5-day returns. This indicates that while investors who add funds to their accounts regularly display higher disposition coefficients and therefore bias, they are good at finding stocks that

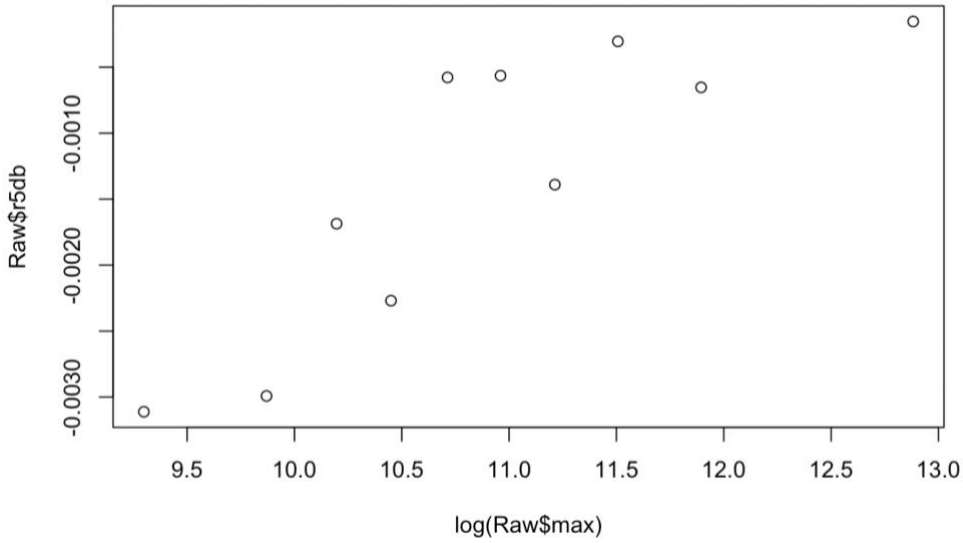


Figure 24: 5 day returns vs log(max)

perform well over a 5-day period. The trend does not hold when looked at from a 20 day returns perspective which may be indicative of a spurious correlation.

The disposition effect is an emotional bias, the sensations or feeling investors experience because of the unrealized gains/losses influence their trading behavior. Investors that succumb to the urge to act emotionally experience worse performance (Odean 2000). In this paper the value metric adds to Odean's finding s by proving the hypothesis that those who add more funds to their accounts behave more emotionally, possibly due to 'having more skin in the game' than investors who don't add as many outside funds. It is a possibility that investors who route more funds to their investment accounts have to not spend those funds on leisure or debt repayment and therefore place higher mental weight on them which in turn makes them more emotionally exposed to fluctuations in the value of their investments.

On a final note, on the disposition effect there was no significant correlation found between disposition coefficient and 20-day returns, while other authors have found links between returns and disposition effect the data in this case finds no correlation regarding investor performance.

4.2 Attention Based Buying

Attention based buying refers to the cognitive bias leading investors to buy stocks they have seen mentioned on the news or as of late on social media. This bias, derived from the search problem individual investors face, readily explained in this quote by Merton (1981) "a potential investor must at least be aware of a firm before deciding whether to acquire additional information and deciding whether to buy the firm's stock." In order to translate this bias into data analysis trading volume, a readily available data source, was used. In this analysis the volume on the day a stock was purchased is compared to the average daily trading volume of the month the trade was placed in. Trades placed under abnormal daily volume were considered to be as a

consequence of attention-based buying. The threshold for abnormality was set at 4 as it represents a cut off close to 90%.

Retail investors studied had an average transaction day volume ratio (volume on transaction day / average daily volume of the month the trade was placed) of 2.64, for the 5-year period the average number of trades placed on a day with abnormal volume (volume ratio > 4) was 5.24. On average the accounts of the LDB dataset made 50 trades in the 5-year period. While the aggregate information is not a strong indicator (outside of the high average volume ratio) of attention based buying further analysis is needed to arrive at significant conclusions.

When volume ratio and volume ratio multiple are plotted against the value max and separated in to ten quantiles, we get the following bar graph (Figure 10).

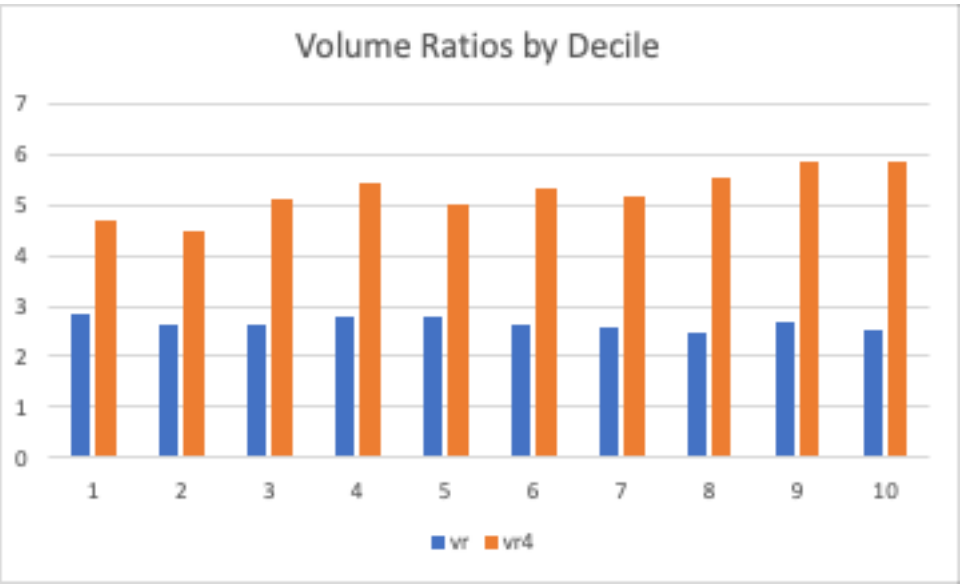
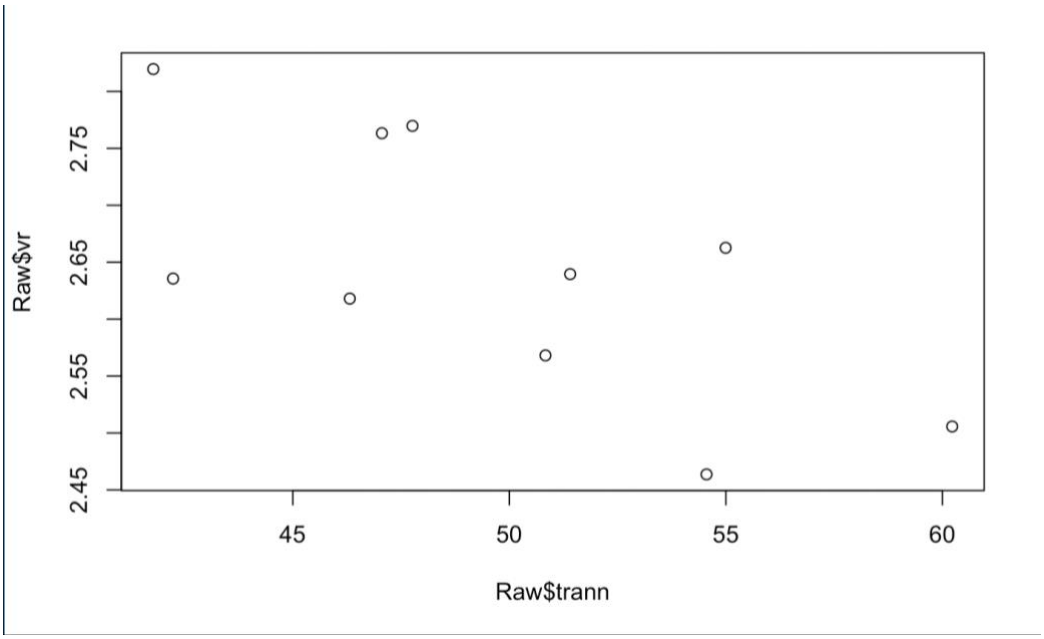


Figure 11: Volume ratios and volume multiples by value invested quantile

This graph shows no particular relationship between volume ratio and value quantiles, however, there is a slight relationship seen when looking at volume ratio multiples greater than four. This is only natural as investors who invest more outside funds also place more trades

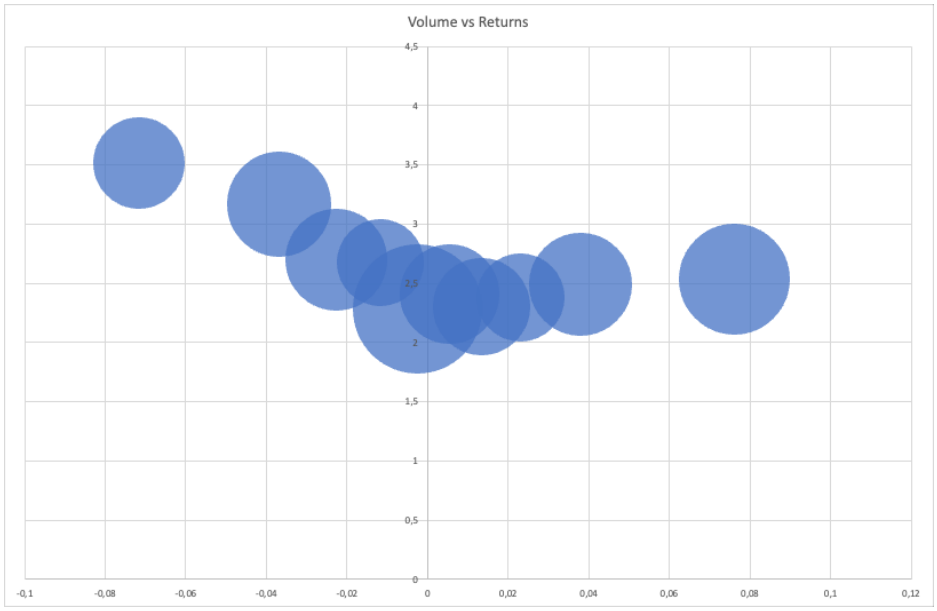
leading to more opportunities to make a stock purchase on a day with abnormal trading volume. In conclusion, no correlation between invested value and volume on trading date is found in this dataset.



Analysis of volume ratio uncovered a relationship to the number of transactions; this relationship can be seen in Figure 12.

Figure 25: Volume ratios vs number of transactions

Number of transactions (x -axis) has a negative relationship with volume ratios (y-axis), this relationship is likely explained by the amount of time individual investors spend researching stocks. More active retail investors who trade more are likely to spend a greater amount of time



analyzing the market and the options available and less likely to depend on news headlines for trading initiatives. This trend is evident even when analysis of volume ratio multiples found a greater occurrence of trades being placed on days with abnormal volume further strengthening the conclusion of no relationship between attention-based buying and value invested.

In contrast when volume ratios are compared to 5- and 20-day returns, we see some correlation, the following graph maps average 20 day returns for stocks bough by each account and then separated into deciles of 20-day returns. In the chart 20-day returns are in the x axis, while volume ratios are in the y axis. Every bubble represents a quantile of investors and the size of the bubble measures 5-day returns.

Figure 26: 5 and 20 day returns vs volume ratios

Figure 13 shows a negative correlation between 20 day returns and volume ratios for quantiles earning below negative returns, suggesting that on days where stocks traded with abnormal volume they were overpriced. The size of the bubbles not having a similar trend represents a reversion to the mean happening in a period of more than five days. Furthermore, analysis of volume ratio multiple in the dataset as a whole found that stocks purchased with abnormal volume (volume ratio > 4) had -1.18% returns over a 20-day period versus stock purchased on days without abnormal trading volume (volume ratio < 4) which returned 0.07% over a 20-day period. These returns if annualized are -19.45% for abnormal volume trades and 1.2% for no abnormal volume days. While 20-day returns are not the best indicator of total returns it is safe to say that attention grabbing stocks, in the short term at least trade at a premium. Literary review studied how some institutions are aware of this phenomenon and trade to take advantage of individual investors suffering from attention-based bias.

Attention based buying is harmful to performance, stocks that are featured on the news or traded excessively seem to be trading at a premium in the short term. This is not to say that good companies cannot be found on the news it simply serves as a warning for retail investors to avoid relying solely on the news when searching for new investments. Attention based buying, however, seems to be growing and changing with the new reality of markets. The discount brokerage dataset is quite old starting in 1991 and ending in 1995 some of the trades analyzed are more than 30 years old. Attention based buying seems to have adopted an entirely new meaning as retail investors have banded together on online forums and have even coordinated short squeezes. Communication seems to have closed a gap between large institutional investors and individual investors. In 2021 after a sustained period of handsome market returns and the growing adoption of cryptocurrency some tickers in both markets have been driven on sentiment alone. Attention based bias could be one of the most relevant factors in the irrational behavior driving markets away from rational equilibrium.

4.3 Overconfidence

The literary review revealed that individual investors who ‘suffer’ from overconfidence tend to trade excessively and hold under-diversified portfolios (Odean 1999). Looking at the number of trades placed during the observed period will determine whether investors in the Large Discount Brokerage firm dataset showed signs of overconfidence, additionally studying the number of different stocks that were bought and sold will reveal the level of diversification in their accounts. On average each account placed 49.71 trades on 19.74 different tickers over the 5-year trading period observed. While placing placing 10 trades per year on average is not a strong indication of overconfidence, trading only 20 tickers in 5 years is indicative of under-diversification at an aggregate level. Moreover, it is important to note the large standard deviation of the number of trades placed, 18.42, indicative of the presence of active and overconfident investors.

In order to measure overconfidence, the dataset was subsetted into two groups, the first containing the 10% of investors who placed the most trades during the period, these were the most active investors. The second group was populated by the bottom 10% of accounts by trades placed during the period. The figure below displays the difference of means of the two groups. To graphically represent this data on one graph the results were normalized using min/max normalization technique. The data shows overconfident investors might have a reason to be confident in their ability as they were able to find better 5 and 20 day returns on average. Additionally, the top 10% most active investors displayed lower disposition coefficients and lower volume ratios. This comparison of both extremes of the dataset could be failing to account for investor experience, the most active investors trade more and have more opportunities to learn, it is also likely that the most active spend the most time on acquiring and evaluating information which in turn results in better performance. This comparison does not account for fees.

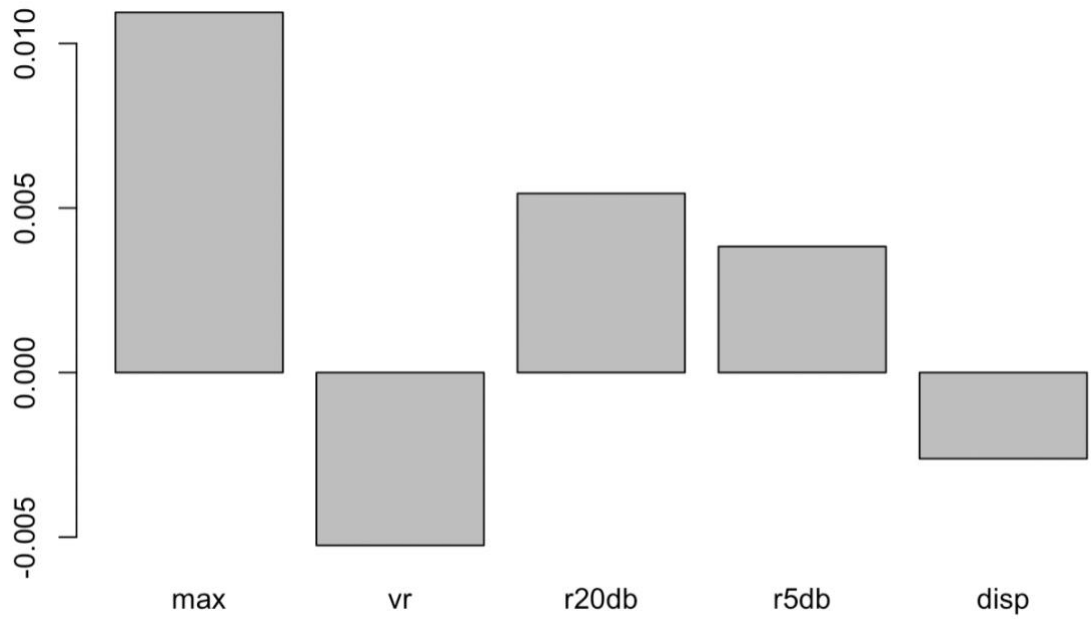


Figure 27: Normalized difference top 10% - bottom 10%

While it may seem, overconfident investors may not be so overconfident after all further comparison against the means of the normalized dataset as a whole tells a different story. The chart below shows overconfident investors underperform the average investor in the dataset in terms of returns and have higher disposition coefficients on average however these differences are almost zero.

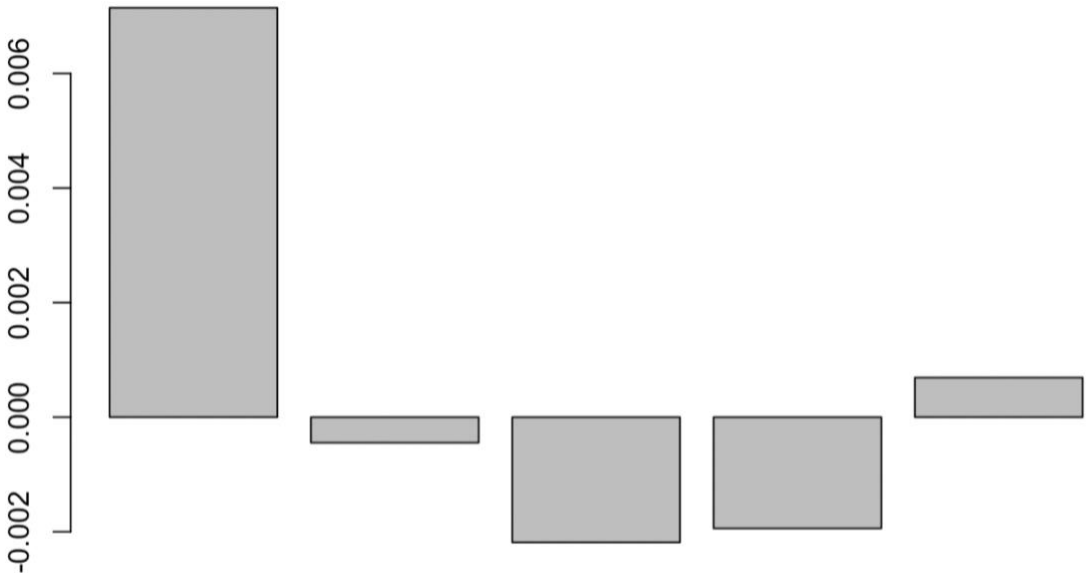


Figure 28: Top10% - sample means, normalized.

The bar plot below shows the average number of tickers traded and the average number of transactions made by each value decile. The accounts with the highest value invested showed higher number of transactions, an investor needs to have confidence in his ability to trade and believe he will have positive returns in order to decide to route funds to an investment account. There is a positive trend with money invested or value and trading volume. Interestingly, there seems to be less of a correlation between value deciles on the diversification metric, and value deciles perhaps a consequence of the search problem referenced in the attention-based buying section. The dataset average holding through the value deciles at around 20 could be a consequence of the ability to keep track of only a limited number of tickers.

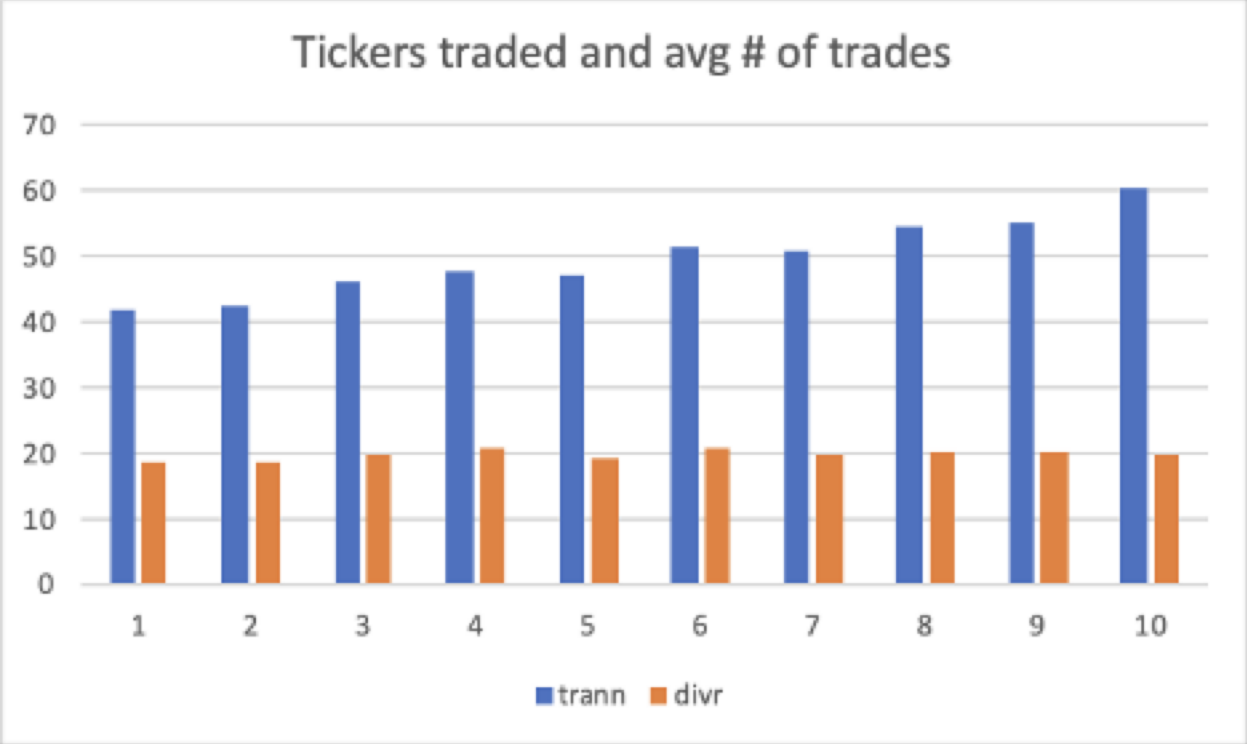


Figure 29: Number of transactions by decile and diversification

Overconfidence in the literary review is linked to subpar returns, the main contributor to underperformance being excessive trading. Fees erode returns, however, in 2021 as opposed to 1991 retail investors enjoy 0% fees on platforms like Robinhood. While investors still have to pay bid ask spreads, high liquidity makes fees in 2021 negligible when compared to what was the

market standard in 1991. While fees may not be as detrimental in 2021, overconfident investors incorrectly assessing their ability or the quality of their information and with under-diversified portfolios will in aggregate fail to beat the market.

Comparing the most active investors and the least active investors may have shown a light on a less discussed behavioral bias, under-confidence. Investors who traded the least underperformed in terms of returns and measured worse in disposition and attention-based buying. Further research would have to be conducted to reach well founded conclusions on the dangers of being an under-confident investor, however it is completely rational to assume these investors exist. Lack of confidence in one's own ability could lead individual investors to place excessive weight on journalists' opinions on stocks, friends' recommendations and to ignore their own deductive reasoning leading to bad performance. This bias is likely to feed itself as bad performance leads to less confidence and could be linked to the investor attrition mentioned by Seru et al. (2009).

4.4 Tool Creation

One of the objectives of this paper is to create an educational tool capable of diagnosing behavioral biases and their severity measured by returns of individual investors. To work this tool needs an input in the form of a table like the one below. Ideally a CSV file that is uploaded to a webpage. From these variables the remaining necessary information (volume, price at t+n, holding period) to calculate behavioral bias metrics can be called automatically using an API.

ID	Date	Time	Ticker	Price	Quantity
###	YYMMDD	246060	TKR	1$	1

The next step for tool to work would be to slightly modify the models to adapt to newer and richer data, having continuous returns would allow for daily portfolio value at close calculations amongst other advantages. The disposition effect could be calculated using the

superior hazard model for more accurate metric calculation, google trends API could be called to enrich the attention based bias metric and even a potential implementation of the most common brokerage platforms for a fee erosion estimation to be assigned to overconfidence driven excessive trading.

This tool true value would be unlocked by populating a database of active investors which would allow a combined bias metric, which could be a weighted average of bias measures, to be mapped against investment returns derived from the daily close portfolio valuations. This comparison would allow for the monetization of behavioral biases. Further data analysis of the population would also allow for bias metric distribution where the tool could tell its users where they rank amongst the other individual investors in the behavioral bias metrics and what the least biased investors are earning through alpha adjusted models. While it may seem fairly complex, the hardest part of creating this tool would be populating the database.

With additional time and further research in the area more models could be added to enhance the tool. Potential use cases go further than simply telling individual investors how they are doing compared to their peers. With trading platform integration notifications could be used to alert users of potential biases as they occur; for example, an alert that notifies an investor when a negative position exceeds his average holding period or a notification telling a user the trade he is trying to make is happening under abnormal volume. Another and slightly more controversial use case could be as an evaluation tool for financial advisors. This use case would help demonstrate how effective financial advisors are at getting their clients to behave rationally.

Furthermore, as waves of young investors pick up trading and investing earlier on in their lives, they stand to benefit the most from a tool designed to help them improve and reduce the cost of learning. This in turn could lead to a decrease in investor attrition which would translate to less churn rate for investment accounts, higher performance for retail investors and higher overall volume which institutions would profit from.

5 Conclusion

5.1 Key Takeaways

1. Literary review revealed the existence of some of the **most prevalent behavioral biases** in finance. **The disposition effect**, an **emotional bias**, leads investors to treat losses and gains differently, an irrational behavior driven by the feeling of satisfaction/regret experienced from a winning/losing position. **Cognitive biases** reviewed were **attention-based buying**, a consequence of the search problem faced by individual investors. Attention based buying happens when individual investors look at the news for investment initiatives, not surprisingly mass media gave its large audience the same ideas leading to overvalued stocks in the short term. Finally, o**verconfidence** which leads retail investors to actively manage their portfolio under the belief they can over-perform even when they are faced with a steep uphill battle

against institutional investors who have hundreds of analysts, premium information and complex computer models.

2. **Finance is a zero-sum game** and all of three of the previously mentioned **behavioral biases** have been linked to **subpar performance**, meaning nonretail investors, or institutions, benefit from this under-performance. While it has been shown humans are predisposed to act irrationally when investing, there is **hope for retail investors**. Evidence finds that an investor **learning about their own ability** is one of the main sources of **improvement**. The three biases can be easily observed when looking at an investors trade history, making it easy for a tool to be created that can diagnose behavioral biases and the severity of them helping educate traders about their own ability and in this way helping them improve.

3. The **Large Discount Brokerage dataset** allowed for the recreation of some of the models used to measure behavioral biases. **The disposition effect** was shown to be **widespread throughout the dataset**; conversely, it had little correlation to 20-day returns. Having no relation to 20-day returns, nonetheless, does not conclude the bias is not detrimental to investor performance. The disposition effect would perhaps have been **better measured** at a **longer timeframe**; however, the data analysis was constricted by what was available in the dataset. **Attention based buying** is a bias that is present in the **short term** and in turn was found to be **negatively correlated to 20-day returns**. Finally, **overconfidence** too was shown to have **negative correlation with 20 day returns** as individual investors who traded the most had worse average 20 day returns than the sample mean. Overconfidence, as opposed to disposition effect and attention-based buying has implication in both the long run and the short run, while the **dataset could only measure the short-term** effect excluding fees the long-term effects of overconfidence are likely even greater as fee erosion compounds.

4. Data analysis also uncovered an interesting yet not surprising **correlation between all three biases**. This correlation could have many underlying explanations and further research could uncover causation. Certain individual investors rank highly on all three bias metrics perhaps due to inexperience or just a general high susceptibility to biases, emotions cannot be the culprit as only deposition has emotional causality.

5. The analysis of the new variable ("max") which measures invested amount from 'outside' funds for each account yielded some interesting conclusions. The most active investors, those who were linked with **overconfidence** seemed to have the **highest ("max") value**, this seems logical as strong belief in one's own ability is a good reason to destine funds to an investment account. The value metric also showed **lower volume ratio multiples** on average for deciles with **higher max value**, likely a consequence of having more 'skin in the game' and thus spending more time researching investment options and relying less on news for stock information. The same rationale can be applied to disposition effect as investors who had invested more outside funds showed higher disposition, an emotional bias. The **higher disposition** measured for investors with the **highest ("max")** value is indicative of a stronger emotional reaction to losses and gains. The analysis of this variable is **something new** and not done before in the field of behavioral finance, while it is not a perfect metric it is in line with all the findings and attributed causes of behavioral biases for retail investors.

In this paper behavioral biases detrimental to performance are studied. To avoid bias when trading an investor should:

1. Not rely solely on the news for information on stocks but rather conduct their own research and use deductive reasoning to evaluate fundamental and technical variables to find stocks to invest in.

2. An investor should not take profits prematurely and should regularly review negative positions and reflect on why they continue to hold that stock. Or simply put cut losers and let winners run.

3. An investor should not trust their perception of their own ability but rather use measures like portfolio alpha and returns over an extended period of time to judge ability.

4. Investors should, for the most part, invest passively unless they can routinely beat the market or treat investing as a hobby in which they value the sensations and investment experience higher than they do the performance against a benchmark.

Understanding one's ability as an investor is important so objective self-evaluation of both the information used and the decision-making process is important. This should be done while keeping in mind common behavioral biases and how they manifest themselves as in most cases this knowledge can be enough to avoid the corresponding behavior. Nevertheless, quoting Daniel Kahneman's newest book; "We know we have psychological biases, but we should resist the urge to blame every error on unspecified 'biases.' Understanding behavioral biases and how to correct them is not enough to beat the market.

5.2 Limitations

The first and biggest limitation is the age of the dataset, with some of the observation being more than 30 years old. The relevance of any findings and conclusions drawn from the data are severely diminished by the different market condition that existed in 1991. 0% trading fees and online forums make markets today extremely different to the status quo in the early 90s. Trends today are moving more and more people to invest which makes research into behavioral biases of individual investors more relevant than ever. New technology such as cryptocurrency

and blockchain make the future of markets even more uncertain and the learnings reviewed in this paper less applicable.

Additionally, the limited measures of returns hindered the ability to find correlations regarding investor performance. 20-day returns used to evaluate a dataset with a mean holding period of almost 2 years is inadequate at best. Using stock price API' san attempt was made to gain weekly stock prices for all the trickers referenced by the transactions, however since 1991 many firms have gone out of business, been acquired, or changed their legal name. This made finding prices impossible without serious survivorship bias or paying for access to higher quality databases. Without time series data for stock prices constructing a hazard model and using it to calculate coefficients for disposition effect is impossible. Using a hazard model is the more recent method of measuring disposition and provides more accurate re

The ("max") variable had limitations in the way it was calculated. To calculate this variable every trade which bought a stock increased the value of the variable by the dollar amount of the trade at the time it was made, and every position sold decreased the value by the amount the stocks were sold for. The limitation of this variable is it holds no reference to an investors net worth or portfolio size, this in turn makes it an imperfect measure of the 'skin in the game' of an individual investor.

Final Note: I first gained access to this dataset when it was shared by Tyler Shumway, a professor at the University of Michigan, Ann Arbor in early 2020. Since then, I have been trying to find a more recent dataset containing transaction made by individual investors. There is very little data available in this area and I was unable to find any public datasets online. My search also led me to request several retail banks for transactions, however all request failed to gain any traction.

6 Bibliography

Ackermann, Carl, et al. “The Performance of Hedge Funds: Risk, Return, and Incentives*Journal of Finance*, vol. 54, no. 3, 1999, pp. 833–874. JSTOR, www.jstor.org/stable/222427. Accessed 6 June 2021.

Barber, Brad M., and Terrance Odean. “The Behavior of Individual Investors.” SSRN Electronic Journal, Sept. 2011, pp. 1–54., doi:10.2139/ssrn.1872211.

Barber, Brad M., and Terrance Odean. "All that glitters: The effect of attention and news on the behavior of individual and institutional investors." *The review of financial studies* 21.2 (2008): 785-818.

Bamber, Linda Smith, Oria E. Barron, and Thomas L Stober. “Trading Volume and Different Aspects of Disagreement Coincident with Earnings Announcements.” *Accounting Review*, vol. 72, no. 4, 1997, pp. 575–597. JSTOR, www.jstor.org/stable/248176. Accessed 5 June 2021.

Barber, Brad M., and Terrance Odean. "Boys will be boys: Gender, overconfidence, and common stock investment." *The quarterly journal of economics* 116.1 (2001): 261-292.

Barber, Brad M., and Terrance Odean. “Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors.” *The Journal of Finance*, vol. 55, no. 2, 2000, pp. 773–806., doi:10.1111/0022-1082.00226.

Ben- David, Itzhak, John R. Graham, and Campbell R. Harvey. “Managerial miscalibration.” *The Quarterly Journal of Economics* 128.4 (2013): 1547-1584.

Brown, Lawrence D., and Jerry C. Y. Han. “The Impact of Annual Earnings Announcements on Convergence of Beliefs.” The Accounting Review, vol. 67, no. 4, 1992, pp. 862–875. JSTOR, www.jstor.org/stable/248328. Accessed 5 June 2021.

Buffett, Warren. Interview with BusinessWeek, 1999.

Evans, John L., and Stephen H. Archer. “Diversification and the Reduction of Dispersion: An Empirical Analysis.” *The Journal of Finance*, vol. 23, no. 5, 1968, pp. 761–767. *JSTOR*, www.jstor.org/stable/2325905. Accessed 12 June 2021.

Fama, Eugene F. 1965b. “Random Walks in Stock Market Prices.” *Financial Analysts Journal* September/October: 55–59.

Fischhoff, Baruch, et al. “Knowing with Certainty: The Appropriateness of Extreme Confidence.” *Journal of Experimental Psychology*: Human Perception and Performance, vol. 3, no. 4, 1977, pp. 552–564., doi:10.1037/0096-1523.3.4.552.

Frank, Jerome D. “Some Psychological Determinants of the Level of Aspiration.” *The American Journal of Psychology*, vol. 47, no. 2, 1935, pp. 285–293. JSTOR, www.jstor.org/stable/1415832. Accessed 5 June 2021.

Glaser, Markus, and Martin Weber. "Overconfidence and Trading Volume." *The Geneva Risk and Insurance Review* 32.1(2007):1-36 doi:10.2139/ssrn.471925.

Glaser, Markus, Thomas Langer, and Martin Weber. "True overconfidence in interval estimates: Evidence based on a new measure of miscalibration." *Journal of Behavioral Decision Making* 26.5 (2013): 405-417, doi:10.1002/bdm.1773.

Graham, John R., et al. "Investor Competence, Trading Frequency, and Home Bias." SSRN Electronic Journal, 2006, doi:10.2139/ssrn.620801.

Google Trends.

Grullon, Gustavo and Kanatas, George and Weston, James Peter, Advertising, Breadth of Ownership, and Liquidity. Available at SSRN: https://ssrn.com/abstract=304240 or http://dx.doi.org/10.2139/ssrn.304240

Henriksson, Roy D., and Robert C. Merton. "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills." The Journal of Business, vol. 54, no. 4, 1981, pp. 513–533. JSTOR, www.jstor.org/stable/2352722. Accessed 12 June 2021.

Kahneman, Daniel, and Amos Tversky. "Prospect Theory: An Analysis of Decision under Risk." *Econometrica*, vol. 47, no. 2, 1979, pp. 263–291. JSTOR, www.jstor.org/stable/1914185. Accessed 27 Mar. 2021.

Karpoff, Jonathan M. "A Theory of Trading Volume." *The Journal of Finance*, vol. 41, no. 5, 1986, pp. 1069–1087. JSTOR, www.jstor.org/stable/2328164. Accessed 5 June 2021.

Lush, Mark, et al. A Collaboration between the FINRA Foundation and NORC at the University of Chicago., 2021, pp. 1–20, *Investing 2020*: New Accounts and the People Who Opened Them. https://www.finrafoundation.org/sites/finrafoundation/files/investing-2020-new-accounts-and-the-people-who-opened-them_1_0.pdf

Odean, Terrance. "Do Investors Trade Too Much?" *The American Economic Review*, vol. 89, no. 5, 1999, pp. 1279–1298. JSTOR, www.jstor.org/stable/117058. Accessed 5 June 2021.

Odean, Terrance. "Are Investors Reluctant to Realize Their Losses?" SSRN *Electronic Journal*, 1998, doi:10.2139/ssrn.94142.

Pompian, Michael. M., "Behavioral Finance and Wealth Management." John Wiley & Sons Inc. New Jersey, 2006.

Seasholes, Mark S., and Guojun Wu. “Predictable Behavior, Profits, and Attention.” *Journal of Empirical Finance*, vol. 14, no. 5, 2007, pp. 590–610., doi:10.1016/j.jempfin.2007.03.002.

Seru, Amit, et al. “Learning By Trading.” SSRN Electronic Journal, 2009, doi:10.2139/ssrn.891694.

Shefrin, Hersh, and Meir Statman. “The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence.” *The Journal of Finance*, vol. 40, no. 3, 1985, pp. 777–790., doi:10.1111/j.1540-6261.1985.tb05002.x

Statman, Meir and Thorley, Steven and Vorkink, Keith, Investor Overconfidence and Trading Volume (March 2003). AFA 2004 San Diego Meetings, Available at SSRN: https://ssrn.com/abstract=168472. or http://dx.doi.org/10.2139/ssrn.168472

“The Behavioral Biases of Individuals.” CFA Institute, www.cfainstitute.org/en/membership/professional-development/refresher-readings/behavioral-biases-individuals.

Trinugroho, Irwan, and Roy Sembel. “Overconfidence and Excessive Trading Behavior: An Experimental Study.” *International Journal of Business and Management*, 6.7 (2011), doi:10.5539/ijbm.v6n7p147.

“Wall Street Reins In Hedge Funds' Short Bets on Meme Stocks.” Yahoo! Finance, Yahoo!, 4 June 2021, finance.yahoo.com/news/wall-street-banks-rein-hedge-190856404.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAIX4sElDEYgndA9-cNwPcUkPS15uDmJO41ToFa05g6vC8CahU5PifUmeHV0iUVofEOgvAWbgE79CVvKKJMG4gqcXd4aMNuZin892-hy3HpGJTWGvbVpSCPj0zTjkOrbVedeR2cM2wVqM8ZoUpmjdV9vQtQ7wAVTg-KNQuHBkdP0B.

Yahoo Finance Charts

TFG

Pablo B

2/13/2021

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## corrplot 0.89 loaded
```

```
att <- read.csv("STUDENTS.csv" , header = FALSE)
colnames(att)= c("account", "date", "ticker", "bs", "shares", "price", "pclose", "
p5days", "p20days", "vol", "avgvol", "cnt", "sex", "income", "age", "state", "zip"
, "homeowner", "homeyrs", "married", "mduration", "hpr", "sellvol", "year")

#Data cleaning
att$state <- NULL
att$zip <- NULL
att$homeowner <- NULL
att$homeyrs <- NULL
att$hpr <- NULL
att$married <- NULL
att$sellvol <- NULL

att$account <- as.factor(att$account)
att$vol.ratio <- mapply("/", as.numeric(as.character(att$vol)), as.numeric(as.char
acter(att$avgvol)))
```

```
## Warning in mapply("/", as.numeric(as.character(att$vol)),
## as.numeric(as.character(att$avgvol))): NAs introduced by coercion
```

```
att$vfour <- ifelse(att$vol.ratio>4, 1, 0)
att$pclose <- as.numeric(att$pclose)
att$p5days <- as.numeric(as.character(att$p5days ))
```

```
## Warning: NAs introduced by coercion
```

```
att$p20days <- as.numeric(as.character(att$p20days))
```

```
## Warning: NAs introduced by coercion
```

```
att$ret1day <- (att$pclose - att$price)/att$price
att$ret5day <- (att$p5days - att$price)/att$price
att$ret20day <- (att$p20days - att$price)/att$price
att$tradevalue <- att$price*att$shares
att$mduration <- as.numeric(as.character(att$mduration))
```

```
## Warning: NAs introduced by coercion
```

#Mean number of trades

```
att$account <- as.factor(att$account)

number.of.accounts <- length(levels(att$account))
number.of.t.trades <- nrow(att)

mean.trades <- number.of.t.trades/number.of.accounts

print(mean.trades)
```

```
## [1] 49.71179
```

#calculating invested balance

```
y = 4174
x = 0
for(i in 1:207497)
  {
  n = att$account[i]
  if(n == y)
    {
    x = x + att$tradevalue[i]
    if(x < 0)
      {
      x = 0}
      }
    else
      {
      x = 0
      y =att$account[i]
    }
  att$value[i] <- x
  }
```

#Creating new data frame

```
uaccounts <- as.data.frame (names(table(att$account)))
colnames(uaccounts) = c("account")

for (i in 1:nrow(uaccounts))
  {
  df1 <- subset.data.frame(att, account == uaccounts$account[i], select = c(value,
vol.ratio, vfour, ticker, mduration, year))
  uaccounts$max[i] <- max(df1$value, na.rm = TRUE)
  uaccounts$vr[i] <- mean(df1$vol.ratio, na.rm = TRUE)
  uaccounts$vr4[i] <- sum(df1$vfour, na.rm = TRUE)
  uaccounts$trann[i] <- nrow(df1)
  df1$ticker <- factor(df1$ticker)
  uaccounts$divr[i] <- nlevels(df1$ticker)
  uaccounts$holdmean[i] <- mean(df1$mduration, na.rm = TRUE)
  uaccounts$holdmed[i] <- median(df1$mduration, na.rm = TRUE)
  uaccounts$tperiod[i] <- max(df1$year) - min(df1$year)
  }

for (i in 1:nrow(uaccounts))
  {
   df1 <- subset.data.frame(att, account == uaccounts$account[i], select = c(bs, r
et1day, ret5day, ret20day))
   df2 <- subset.data.frame(df1, bs == " B ")
    uaccounts$r20db[i] <- mean(df2$ret20day, na.rm = TRUE)
    uaccounts$r5db[i] <- mean(df2$ret5day, na.rm = TRUE)
    uaccounts$r1db[i]<- mean(df2$ret1day, na.rm = TRUE)
  }

for (i in 1:nrow(uaccounts))
  {
   df1 <- subset.data.frame(att, account == uaccounts$account[i], select = c(bs, r
et1day, ret5day, ret20day))
   df2 <- df1 %>% filter(bs == " S ")
    uaccounts$r20ds[i] <- mean(df2$ret20day, na.rm = TRUE)
    uaccounts$r5ds[i] <- mean(df2$ret5day, na.rm = TRUE)
    uaccounts$r1ds[i]<- mean(df2$ret1day, na.rm = TRUE)
}
#Assign quantiles to accounts based on max and 20day returns for buys
uaccounts$decile1 <- ntile(uaccounts$max, 10)
uaccounts$decile2 <- ntile(uaccounts$r20db, 10)
```

#Disposition Effect

```
#Variable Definition
att$bs <- as.integer(att$bs)
```

```
## Warning: NAs introduced by coercion
```

```
att$Realized <- (ifelse(att$mduration <= 20, 1, 0))
att$tdayg <- (ifelse(att$p20days> att$price, 1, 0))
att$paperg <- (ifelse(att$Realized ==0 & att$tdayg == 1 ,1 ,0))
att$realg <- (ifelse(att$Realized + att$tdayg == 2 ,1 ,0))
att$tdayl <- (ifelse(att$price > att$p20days, 1, 0))
att$paperl <- (ifelse(att$Realized == 0 & att$tdayl == 1, 1, 0))
att$reall <- (ifelse(att$Realized == 1 & att$tdayl ==1, 1, 0 ))

attd <- subset(att, shares > 0)
for  (i in 1:nrow(uaccounts)) {
  df1 <- subset.data.frame(attd, account == uaccounts$account[i], select = c(accou
nt, realg, reall, paperl, paperg, tradevalue, ret20day))
  rg <- sum(df1$realg, na.rm = TRUE)
  rga <- sum(subset.data.frame(df1, realg == 1)$tradevalue*(1+subset.data.frame(df
1, realg == 1)$ret20day))
  pg <- sum(df1$paperg, na.rm = TRUE)
  pga <- sum(subset.data.frame(df1, paperg == 1)$tradevalue*(1+subset.data.frame(d
f1, paperg == 1)$ret20day))
  uaccounts$PGR[i] <- rg/(pg+rg)
  uaccounts$PGRa[i] <- rga/(pga+rga)
  rl <- sum(df1$reall, na.rm = TRUE)
  rla <- sum(subset.data.frame(df1, reall == 1)$tradevalue*(1+subset.data.frame(df
1, reall == 1)$ret20day))
  pl <- sum(df1$paperl, na.rm = TRUE)
  pla <- sum(subset.data.frame(df1, paperl == 1)$tradevalue*(1+subset.data.frame(d
f1, paperl == 1)$ret20day))
  uaccounts$PLR[i] <- rl / (pl + rl)
  uaccounts$PLRa[i] <- rla / (pla + rla)
  }
```

#Correlation Matrix

```
library("Hmisc")
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

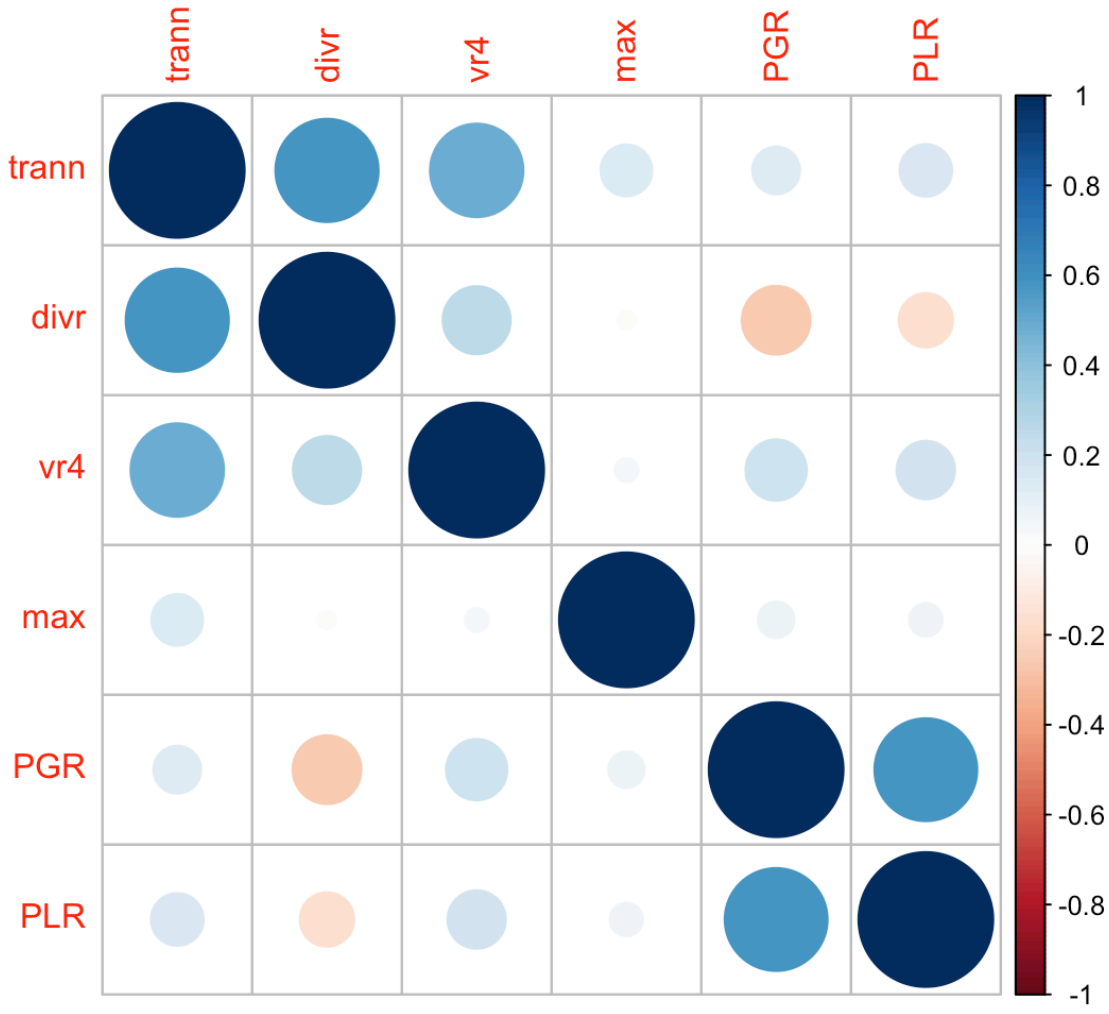
```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
mtx <- subset.data.frame(uaccounts, select = c(trann, divr, vr4, max, PGR, PLR))
mtx <- mtx[is.finite(rowSums(mtx)),]
corrplot(cor(mtx, use= "complete.obs"))
```



#Regression

```
n = 1
for  (i in 1:nrow(uaccounts)){
    while(uaccounts$account[i] == attd$account[n])
    {
      if(n == nrow(attd))
      {
      attd$quantile[n] <- attd$quantile[n-1]
      break
      }
    attd$quantile1[n] <- uaccounts$decile1[i]
    attd$quantile2[n] <- uaccounts$decile2[i]
    n = n+1
    }

}
#disposirtion effect regressions by quantile, quantile 1 ->
Dispreg10 <- glm(Realized ~ ret20day , data = subset.data.frame(attd, quantile1 ==
10), family = "binomial")
summary(Dispreg10)
```

```
##
## Call:
## glm(formula = Realized ~ ret20day, family = "binomial", data = subset.data.fram
e(attd,
##     quantile1 == 10))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1689  -0.5485  -0.5204  -0.4793   2.5016
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.90436    0.02497 -76.251   <2e-16 ***
## ret20day     1.50750    0.15266   9.875   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11356  on 14525  degrees of freedom
## Residual deviance: 11259  on 14524  degrees of freedom
##   (29 observations deleted due to missingness)
## AIC: 11263
##
## Number of Fisher Scoring iterations: 4
```

```
Dispreg1 <- glm(Realized ~ ret20day , data = subset.data.frame(attd, quantile2 ==
1), family = "binomial")
summary(Dispreg1)
```

```
##
## Call:
## glm(formula = Realized ~ ret20day, family = "binomial", data = subset.data.fram
e(attd,
##     quantile2 == 1))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6474  -0.5262  -0.5194  -0.5075   2.1244
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9129     0.0315 -60.736   <2e-16 ***
## ret20day      0.3045     0.1635   1.862   0.0626 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7727.3  on 10179  degrees of freedom
## Residual deviance: 7723.8  on 10178  degrees of freedom
##   (57 observations deleted due to missingness)
## AIC: 7727.8
##
## Number of Fisher Scoring iterations: 4
```

```
Dispreg10 <-glm(Realized ~ ret20day , data = subset.data.frame(attd, quantile1 ==
10), family = "binomial")
summary(Dispreg10)
```

```
##
## Call:
## glm(formula = Realized ~ ret20day, family = "binomial", data = subset.data.fram
e(attd,
##     quantile1 == 10))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1689  -0.5485  -0.5204  -0.4793   2.5016
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.90436    0.02497 -76.251   <2e-16 ***
## ret20day     1.50750    0.15266   9.875   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11356  on 14525  degrees of freedom
## Residual deviance: 11259  on 14524  degrees of freedom
##   (29 observations deleted due to missingness)
## AIC: 11263
##
## Number of Fisher Scoring iterations: 4
```

```
Dispreg1 <- glm(Realized ~ ret20day , data = subset.data.frame(attd, quantile2 ==
1), family = "binomial")
summary(Dispreg1)
```

```
##
## Call:
## glm(formula = Realized ~ ret20day, family = "binomial", data = subset.data.fram
e(attd,
##     quantile2 == 1))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6474  -0.5262  -0.5194  -0.5075   2.1244
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.9129     0.0315 -60.736   <2e-16 ***
## ret20day      0.3045     0.1635   1.862   0.0626 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7727.3  on 10179  degrees of freedom
## Residual deviance: 7723.8  on 10178  degrees of freedom
##   (57 observations deleted due to missingness)
## AIC: 7727.8
##
## Number of Fisher Scoring iterations: 4
```

#Quantile Comparison

```
Dec1 <- as.data.frame (names(table(uaccounts$decile1)))
colnames(Dec1) = c("Decile")
Dec2 <- as.data.frame (names(table(uaccounts$decile2)))
colnames(Dec2) = c("Decile")

uaccounts$vr <- ifelse(uaccounts$vr == Inf, NA, uaccounts$vr)
uaccounts$r20db <- ifelse(uaccounts$r20db == Inf, NA, uaccounts$r20db)
uaccounts$r5db <- ifelse(uaccounts$r5db == Inf, NA, uaccounts$r5db)
uaccounts$r20ds <- ifelse(uaccounts$r20ds == Inf, NA, uaccounts$r20ds)
uaccounts$r5ds <- ifelse(uaccounts$r5ds == Inf, NA, uaccounts$r5ds)
uaccounts$PGRa <- as.numeric(uaccounts$PGRa)
uaccounts$PLRa <- as.numeric(uaccounts$PLRa)

min_max_norm <- function(z) {
    (z - min(z)) / (max(z) - min(z))
}

for (i in 1:10){
  df1 <- lapply(subset.data.frame(uaccounts, uaccounts$decile1 == i, select = c(ma
x, vr, vr4, trann, divr, holdmean, holdmed, tperiod, r20db, r20ds, r5db, r5ds, PGR
, PLR, PGRa, PLRa)), min_max_norm)
```

```
  Dec1$max[i] <- mean(df1$max, na.rm = TRUE)
  Dec1$vr[i] <- mean(df1$vr, na.rm = TRUE)
  Dec1$vr4[i] <- mean(df1$vr4, na.rm = TRUE)
  Dec1$trann[i] <- mean(df1$trann, na.rm = TRUE)
  Dec1$divr[i] <- mean(df1$divr, na.rm = TRUE)
  Dec1$holdmean[i] <- mean(df1$holdmean, na.rm = TRUE)
  Dec1$holdmed[i] <- mean(df1$holdmed, na.rm = TRUE)
  Dec1$tperiod[i] <- mean(df1$tperiod, na.rm = TRUE)
  Dec1$r20db[i] <- mean(df1$r20db, na.rm = TRUE)
  Dec1$r20ds[i] <- mean(df1$r20ds, na.rm = TRUE)
  Dec1$r5db[i] <- mean(df1$r5db, na.rm = TRUE)
  Dec1$r5ds[i] <- mean(df1$r5ds, na.rm = TRUE)
  Dec1$PGR[i] <- mean(df1$PGR, na.rm = TRUE)
  Dec1$PLR[i] <- mean(df1$PLR, na.rm = TRUE)
  Dec1$PGRa[i] <- mean(df1$PGRa, na.rm = TRUE)
  Dec1$PLRa[i] <- mean(df1$PLRa, na.rm = TRUE)
}
for (i in 1:10){
  df1 <- lapply(subset.data.frame(uaccounts, uaccounts$decile2 == i, select = c(ma
x, vr, vr4, trann, divr, holdmean, holdmed, tperiod, r20db, r20ds, r5db, r5ds, PGR
, PLR, PGRa, PLRa)), min_max_norm)
  Dec2$max[i] <- mean(df1$max, na.rm = TRUE)
  Dec2$vr[i] <- mean(df1$vr, na.rm = TRUE)
  Dec2$vr4[i] <- mean(df1$vr4, na.rm = TRUE)
  Dec2$trann[i] <- mean(df1$trann, na.rm = TRUE)
  Dec2$divr[i] <- mean(df1$divr, na.rm = TRUE)
  Dec2$holdmean[i] <- mean(df1$holdmean, na.rm = TRUE)
  Dec2$holdmed[i] <- mean(df1$holdmed, na.rm = TRUE)
  Dec2$tperiod[i] <- mean(df1$tperiod, na.rm = TRUE)
  Dec2$r20db[i] <- mean(df1$r20db, na.rm = TRUE)
  Dec2$r20ds[i] <- mean(df1$r20ds, na.rm = TRUE)
  Dec2$r5db[i] <- mean(df1$r5db, na.rm = TRUE)
  Dec2$r5ds[i] <- mean(df1$r5ds, na.rm = TRUE)
  Dec2$PGR[i] <- mean(df1$PGR, na.rm = TRUE)
  Dec2$PLR[i] <- mean(df1$PLR, na.rm = TRUE)
  Dec2$PGRa[i] <- mean(df1$PGRa, na.rm = TRUE)
  Dec2$PLRa[i] <- mean(df1$PLRa, na.rm = TRUE)
  }
```

#ANOVA Testing

```
# Compute the analysis of variance
max.aov <- aov(max ~ decile1, data = uaccounts)
summary(max.aov)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## decile1        1 2.945e+13 2.945e+13   919.9 <2e-16 ***
## Residuals   4172 1.336e+14 3.202e+10
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vr.aov <- aov(vr ~ decile1, data = uaccounts)
summary(vr.aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## decile1        1     22  21.608    3.34 0.0677 .
## Residuals   4165  26945   6.469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 7 observations deleted due to missingness
```

```
vr4.aov <- aov(vr4 ~ decile1, data = uaccounts)
summary(vr4.aov)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## decile1        1    590   589.6   35.16 3.28e-09 ***
## Residuals   4172  69951    16.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
trann.aov <- aov( trann~ decile1, data = uaccounts)
summary(trann.aov)
```

```
##               Df  Sum Sq Mean Sq F value Pr(>F)
## decile1        1  121622  121622   392.2 <2e-16 ***
## Residuals   4172 1293577     310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
divr.aov <- aov( divr~ decile1, data = uaccounts)
summary(divr.aov)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## decile1        1    765   764.6    9.51 0.00206 **
## Residuals   4172 335439    80.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
holdmean.aov <- aov( holdmean~ decile1, data = uaccounts)
summary(holdmean.aov)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## decile1        1 6.847e+08 684676994   172.1 <2e-16 ***
## Residuals   4167 1.658e+10   3978555
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

```
holdmed.aov <- aov(holdmed ~ decile1, data = uaccounts)
summary(holdmed.aov)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## decile1        1 2.089e+09 2.089e+09   108.6 <2e-16 ***
## Residuals   4167 8.012e+10 1.923e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 5 observations deleted due to missingness
```

```
r20db.aov <- aov( r20db~ decile1, data = uaccounts)
summary(r20db.aov)
```

```
##               Df Sum Sq   Mean Sq F value Pr(>F)
## decile1        1  0.000 0.0000732   0.043  0.835
## Residuals   4165  7.069 0.0016973
## 7 observations deleted due to missingness
```

```
r5db.aov <- aov( r5db~ decile1, data = uaccounts)
summary(r5db.aov)
```

```
##               Df Sum Sq  Mean Sq F value Pr(>F)
## decile1        1 0.0035 0.003476   6.283 0.0122 *
## Residuals   4165 2.3040 0.000553
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 7 observations deleted due to missingness
```

```
r20ds.aov <- aov( r20ds~ decile1, data = uaccounts)
summary(r20ds.aov)
```

```
##               Df Sum Sq  Mean Sq F value Pr(>F)
## decile1        1  0.004 0.003739   0.534  0.465
## Residuals   4167 29.151 0.006996
## 5 observations deleted due to missingness
```

```
r5ds.aov <- aov( r5ds~ decile1, data = uaccounts)
summary(r5ds.aov)
```

```
##               Df Sum Sq  Mean Sq F value Pr(>F)
## decile1        1  0.004 0.003679   0.953  0.329
## Residuals   4167 16.079 0.003859
## 5 observations deleted due to missingness
```

```
PGR.aov <- aov(PGR ~ decile1, data = uaccounts)
summary(PGR.aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## decile1        1   2.94  2.9372     103 <2e-16 ***
## Residuals   4164 118.75  0.0285
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 8 observations deleted due to missingness
```

```
PLR.aov <- aov( PLR~ decile1, data = uaccounts)
summary(PLR.aov)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## decile1        1   1.00  0.9970   61.81 4.78e-15 ***
## Residuals   4162  67.13  0.0161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 10 observations deleted due to missingness
```

```
PGRa.aov <- aov(PGRa ~ decile1, data = uaccounts)
summary(PGRa.aov)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## decile1        1   3.93   3.929   102.2 <2e-16 ***
## Residuals   4164 160.04   0.038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 8 observations deleted due to missingness
```

```
PLRa.aov <- aov( PLRa~ decile1, data = uaccounts)
summary(PLRa.aov)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## decile1        1   1.45  1.4522   65.73 6.73e-16 ***
## Residuals   4162  91.95  0.0221
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 10 observations deleted due to missingness
```

#Compute returns for attention grabbing stocks

```
ags <- subset.data.frame(attd, attd$vfour == "1", select = c(ret20day))
nags <- subset.data.frame(attd, attd$vfour == "0", select = c(ret20day))
retags <- mean(ags$ret20day, na.rm = TRUE)
retnags <- mean(nags$ret20day, na.rm = TRUE)
retags <- (1+retags)^(365/20)-1
retnags <- (1+retnags)^(365/20)-1
print(retags)
```

```
## [1] -0.1944822
```

```
print(retnags)
```

```
## [1] 0.01209304
```

Normalize and compare quantiles of trann

```
uaccounts$disp  <- uaccounts$PGRa-uaccounts$PLRa
uaccounts2 <- as.data.frame(lapply(subset.data.frame(na.omit(uaccounts), select =
c(max, vr,  trann,  r20db,  r5db,  disp)), min_max_norm))
uaccounts2$tranns<- ntile(uaccounts2$trann, 10)
top1 <- subset.data.frame( uaccounts2, uaccounts2$tranns == 1)
top10 <- subset.data.frame( uaccounts2, uaccounts2$tranns == 10)
summary(top1)
```

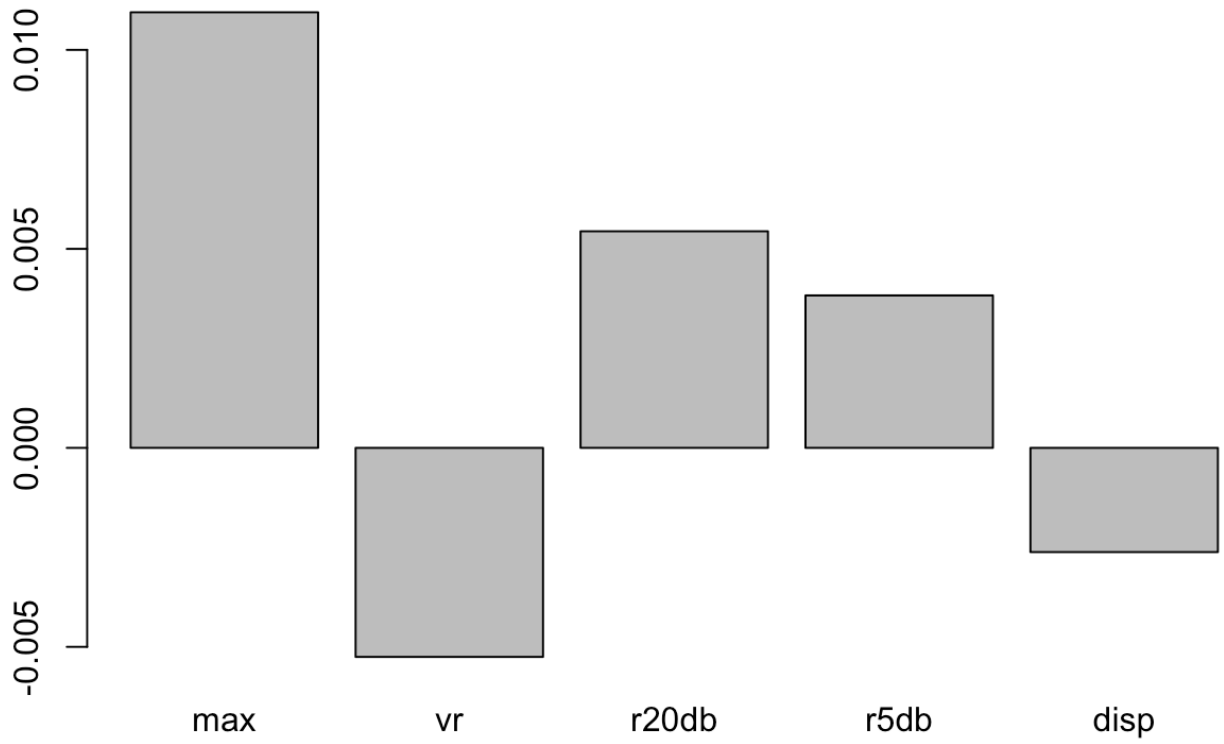
```
##       max                 vr              trann             r20db       
##  Min.   :0.000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000  
##  1st Qu.:0.001991   1st Qu.:0.01569   1st Qu.:0.000000   1st Qu.:0.3031  
##  Median :0.003558   Median :0.02413   Median :0.000000   Median :0.3539  
##  Mean   :0.005718   Mean   :0.04749   Mean   :0.005063   Mean   :0.3510  
##  3rd Qu.:0.006801   3rd Qu.:0.04649   3rd Qu.:0.010101   3rd Qu.:0.3997  
##  Max.   :0.049155   Max.   :0.92548   Max.   :0.020202   Max.   :0.6049  
##       r5db             disp            tranns
##  Min.   :0.0000   Min.   :0.04519   Min.   :1
##  1st Qu.:0.4744   1st Qu.:0.50540   1st Qu.:1
##  Median :0.5145   Median :0.50540   Median :1
##  Mean   :0.5101   Mean   :0.53831   Mean   :1
##  3rd Qu.:0.5513   3rd Qu.:0.56928   3rd Qu.:1
##  Max.   :1.0000   Max.   :0.98441   Max.   :1
```

```
summary(top10)
```

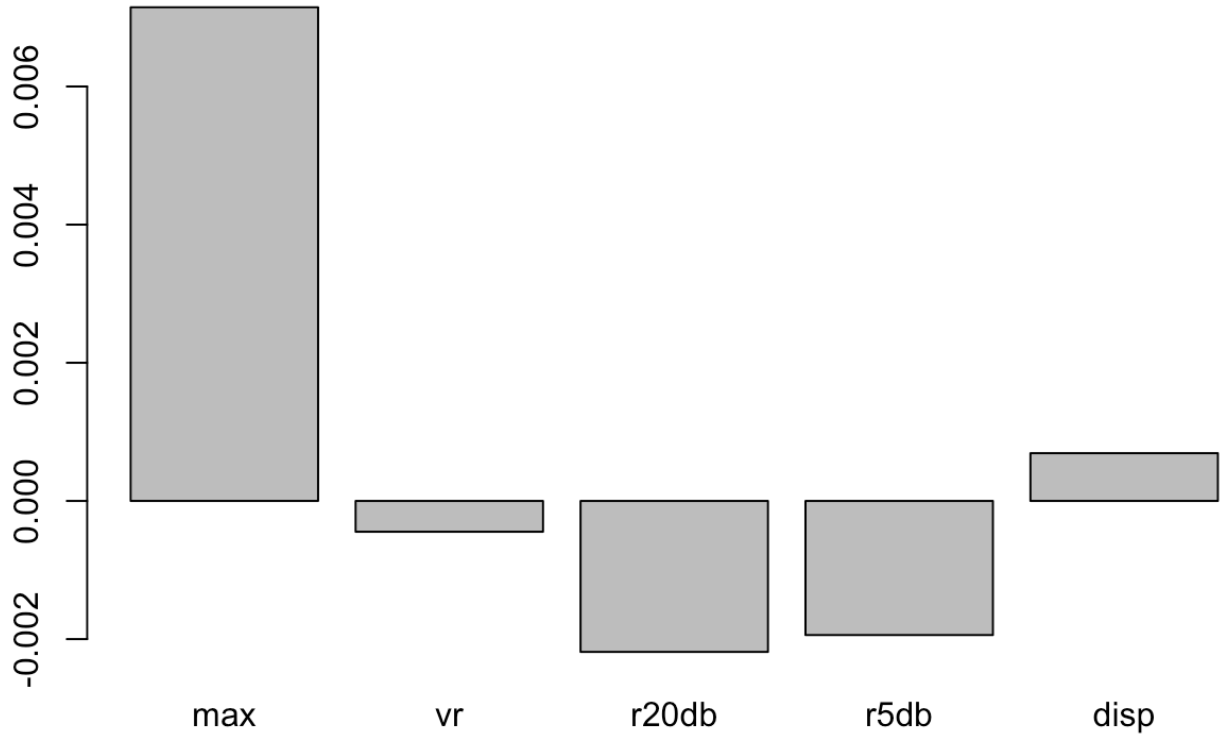
```
##       max                  vr               trann            r20db       
##  Min.   :0.0006176   Min.   :0.002429   Min.   :0.4848   Min.   :0.1307  
##  1st Qu.:0.0051641   1st Qu.:0.018942   1st Qu.:0.5354   1st Qu.:0.3281  
##  Median :0.0097919   Median :0.027940   Median :0.5859   Median :0.3614  
##  Mean   :0.0166630   Mean   :0.042236   Mean   :0.6155   Mean   :0.3565  
##  3rd Qu.:0.0180584   3rd Qu.:0.049635   3rd Qu.:0.6843   3rd Qu.:0.3849  
##  Max.   :0.3007536   Max.   :0.419683   Max.   :1.0000   Max.   :0.6646  
##       r5db             disp            tranns
##  Min.   :0.3121   Min.   :0.2729   Min.   :10
##  1st Qu.:0.4899   1st Qu.:0.4824   1st Qu.:10
##  Median :0.5157   Median :0.5058   Median :10
##  Mean   :0.5139   Mean   :0.5357   Mean   :10
##  3rd Qu.:0.5392   3rd Qu.:0.5903   3rd Qu.:10
##  Max.   :0.6273   Max.   :0.9420   Max.   :10
```

```
top1  <- subset.data.frame( top1, select = -c(tranns, trann))
top10 <- subset.data.frame( top10, select = -c(tranns, trann))
uaccounts2 <- subset.data.frame( uaccounts2, select = -c(tranns, trann))

avg <- sapply(na.omit(uaccounts2), mean)
top1 <- sapply(na.omit(top1), mean, )
top10 <- sapply(na.omit(top10), mean)
diff1 <- top10-top1
diff2 <-top10-avg
barplot(diff1)
```




```
barplot(diff2)
```



#Write CSVs

```
write.csv(uaccounts, "Uaccounts")
write.csv(Dec1, "Dec1")
write.csv(Dec2, "Dec2")
```