Stock Market Analysis

Gauging Fear using Internet Search Volume

An Interactive Qualifying Project
Submitted to the Faculty of the Worcester Polytechnic Institute
In partial fulfillment of the requirements for the
Degree of Bachelor of Science by:

Chad DellaPorta | MGE - IOE | 17’
Brian Murtagh | MGE - IOE | 17’
Jason Lamb | MGE – IOE | 17’
Steven Thulin | CE | 17’

Project Advisor: Professor Koutmos

Table of Contents

Abstract

In this Interactive Qualifying Project (IQP), our team conducted a 15-week regression analysis in which we observed whether or not Google Search Volume (GSV) can serve as an explanatory variable for stock market returns, in particular, for the New York Stock Exchange (NYSE). The study attempts to provide quantitative evidence that the Efficient Market Hypothesis, which states that stock market returns cannot be explained with factual evidence, is false. The group researched the fundamentals behind stock market swings, as well as the cause for various movements in Google Trends. Through running Microsoft Excel regressions on these two sets of data, we can conclude that there is a positive correlation between search terms on GSV and the NYSE. That is, one has the potential to predict various stock index results based on GSV trends. The project gave team members valuable stock market and regression experience, which in term has a built a strong foundation for potential investment in the future.

1.0 Introduction

1.1 Goal

In today's ever-increasing 'big data' society, there are plenty of software's available that can allow one to identify patterns and trends in data. Financial markets and individual search interests are prime targets for a quantitative investigation. The objective of this IQP is to prove the idea that an individual may predict various movements in the stock market based upon publicly-viewable GSV data. It was imperative for our team to familiarize ourselves with the terminology and tools required to provide us interpretation of the market and Google Trends. Additionally, in achieving our goal, our team dedicated a significant amount of time to the retrieval and analyzation of our data sets. After comparing selective macroeconomic key words with the S&P 500, our team ran regressions to determine whether or not a key word was significant or not. After the completion of running our regression analyses, our goal is that there will be a vast number of search terms that are significant with the S&P 500.

1.2 History of the S&P 500

Dating back to over a hundred years ago, the early beginnings of the stock market can be seen as simple forms of debt trading where lenders would trade amongst themselves. The history of the Stock Market shows that around 1531 Belgium and Amsterdam were some of the first to establish a system of trading bonds and other commodities. Around this time, trading in America was established in Boston where traders were buying, selling, and trading bonds, hides, molasses, contracts, and etc. Around 1792, Wall Street was established in America as a result of organizing a small formal stock and

Apple Inc.	$710.00 billion
Google Inc.	$448.84 billion
Microsoft Corp.	$368.76 billion
Exxon Mobil Corp.	$334.24 billion
Wells Fargo & Co.	$297.29 billion
Johnson & Johnson	$274.95 billion
Facebook Inc.	$272.26 billion
General Electric Co.	$259.46 billion
JPMorgan Chase & Co.	$254.84 billion
Amazon.com Inc.	$247.77 billion
Wal-Mart Stores Inc.	$230.53 billion
Procter & Gamble Co.	$217.83 billion
Pfizer Inc.	$210.96 billion
Walt Disney Co.	$201.76 billion
Bank of America Corp. $187.45 billion	

Figure 1: Top Companies in the S&P 500

bond trading system. According to Tom DeGrace in his article titled *Stock Market History Chart and a Detailed Look at the Markets*, during this time there was even as much as 100 shares of stock being traded in one transaction.[1] Approximately 24 financials leaders constructed rules, regulations, and fees that would govern day-to-day trading from thereon out. Beginning at noon every day, traders would trade in a building located on 22 Wall Street.

The S&P 500 was first known as the "Composite Index" when it was introduced to the stock index in 1923. The index expanded from three stocks to 90 stocks in 1926, and expanded to its current 500 stocks in 1957. The index was introduced in 1957 by Standard & Poor's as a method to trade the value of 500 large corporations – all of which are listed on the New York Stock Exchange (NYSE) and the National Association of Securities Dealer Automated Quotation system (NASDAQ) composite. The S&P 500 is considered to be a leading indicator for the economy, and has also performed remarkably since perception, outpacing other major asset classes such as bonds or commodities. The indexes price appreciation has tracked U.S. economic growth by reflecting turbulent economic periods in history.[2] The S&P 500 only selects U.S. companies with a market cap of $5.3B or greater.[3]

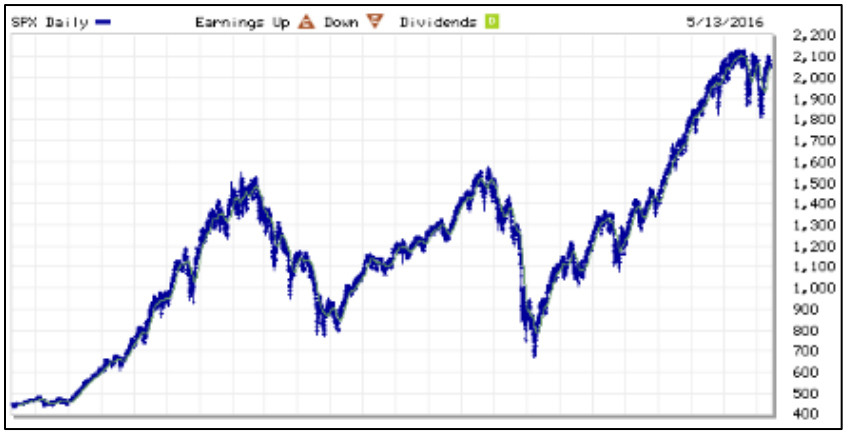

Figure 2: Historical Closing Prices of S&P 500

In attempt to provide a visual of some of the largest S&P 500 corporations, please refer to Figure 1, which displays all S&P 500 that last closed above $200/ share (as of April 20th, 2016). Additionally, please refer to Figure 2 to view a chart from Morningstar showing the historical closing prices of the S&P 500, which dates from January 1st, 1950 to April 20th, 2016.

[1] "Stock Market History Chart and a Detailed Look at the Markets." *Stock Picks System*. N.p., n.d. Web. 20 Apr. 2016.
[2] "What Is the History of the S&P 500? | Investopedia." *Investopedia*. N.p., 10 Apr. 2015. Web. 20 Apr. 2016.
[3] "S&P 500," 2015. [Online]. Available: file:///C:/Users/zwang5/Downloads/fs-sp-500.pdf.

1.3 Interdisciplinary (Global) Importance

Our project has an extremely significant global importance, as being able to predict stock market returns based off of regression data is a newly popular study in today's society. This type of study allows for the connection of ideas and concepts across different interdisciplinary boundaries. In the case of our project, we are able to connect the depths of Google data (specifically Google Trends) and stock market data. With all of this information being publically available and accessible, it is extremely convenient and easy for our team to create new discoveries in the various academic boundaries.

The rise in the recent number of these types of studies is indicative of a change in some economic viewpoints. Economists began researching the types of ways that society's behavior can affect the way the market moves. Within the past decade technology has become an increasingly large factor in explaining how the stock market fluctuates. Economists have taken full advantage of this technology, as seen by the recent study finished by the Warwick Business School (London, U.K.) and the Boston University department of physics. The study, which was titled "Quantifying Trading Behavior in Financial Markets using Google Trends," uses a group of search terms to and the Dow Jones Industrial Average in attempt to see if there is any correlation.[4] This specific project found 98 search terms, with only sixteen of them not correlating. This case was just one example of the types of cases in which economists are utilizing Google Search Trends.

A recent academic research paper complete by researchers Hyunyoung Choi and Hal Varian used Google Trends to predict the present. These two University of California at Berkeley professors used this analytical tool to create more timely forecasts for different economic industries, which are normally released every few weeks by the government. Choi and Varian referenced that the release of the economic activity in various sectors of the market tend to be delayed a couple weeks due to the government analyzing it. However they used Google Trends to document and analyze this data at a much faster pace, essentially creating a real-time forecast of how that industry was performing. They looked at sectors such as automobile sales, unemployment claims, and consumer confidence. These professors

[4] *Nature.com*. Nature Publishing Group, n.d. Web. 06 Oct. 2016.

concluded that they can create a more accurate, and faster forecast of the data using Google Trends than the government does without it.

Recently PricewaterhouseCoopers (PwC) released a study they were conducting which also utilized the Google Trends tool.[5] This study was based on the premise that this tool may be a better indication of a company's top-line sales performance than the actual companies' numbers could be. In this study PwC compared whether Google Trends data changes from one quarter to that same quarter the following year could better explain the sales growth during this period than the actual growth of the same time period a year previous. The findings of this case were significant as the historical data only explained 6% of the current year sales growth while the Google Trends data achieved a 52% explanation. PricewaterhouseCoopers also found that in 75% of their testing scenarios, the Google Trend data was more accurate than the actual sales growth. Now PwC was not saying that companies should abandon their own top-line sales figures, as that would be foolish. The case stands to show that this analytical tool can be used to accurately represent sales figures, along with a company's already in place sales figures.
It's clear that there is a strong interest from economists, and other market researchers, on how Google Trends as well as search volume is affecting the market. They believe that this type of analytics shows that trends in the stock market can be strongly affected by the behavior of consumers. The evidence supporting behavioral economics is certainly available, and economists have begun to use it extensively.

1.4 Multidisciplinary Importance

As the use of technology as an analyzing tool has continued to grow, economists are not the only professional discipline that have found use for it. Google Trend data is used by many different professions for a multitude of different research purposes. Google Trends has given analytics an increased ability to find ways that search tools can affect different aspects of society. Whether it be in politics, sports, or public health the extent of this research has touched many aspects of society.

[5] Bauer, Scott. Klein, Ron. Coakley, Matthew. Song, Danwen. "Using Google Trends to Predict Retail Sales" 2015

One analytics company, Ugam, used Google trends in order to find a relationship between recent political primary races and Google search volume. This company wanted to see if this information could help to try and predict the winner of the primary election. Ugam traced the search volume of each of the candidates' names, and things associated with their beliefs. They ultimately found that the information was indicative of whether a candidate would trend upwards or downwards, but could was not indicative of the direction. This is due to the fact that the volume query can rise whether a candidate becomes very popular or unpopular. So even though the system can show who is being searched, it doesn't always accurately represent who is winning the political race. As you can in Figure 3, the two graphs show Bernie Sanders in blue and Hillary Clinton in red. Both of these candidates show up an upward trend, which was during the time that Ted Cruz won Iowa over Hillary Clinton (even though it was initially called a tie).[6]

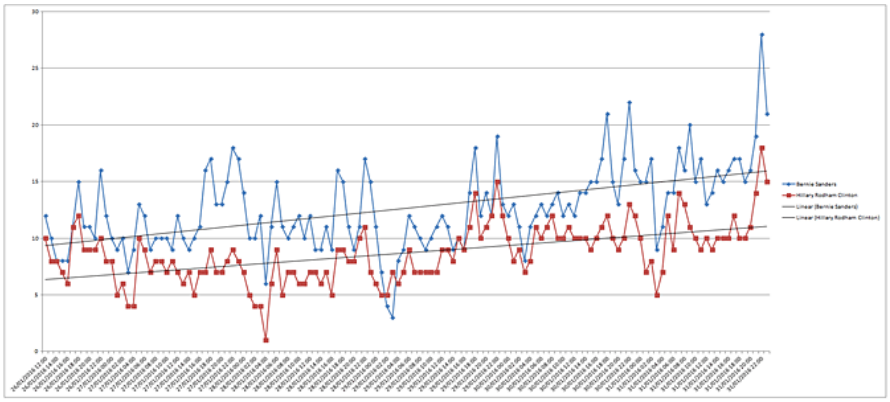

Figure 3: Google Searches for Politicians during 2016 Election

In something such as the sports world we see the relevance of this trend data as well. An ESPN, named FiveThirtyEight written by Nate Silver and Ritchie King, analyzed Google Search Trends in order to find out more about team popularity and their fan bases. These two analytics used this database to find out more about which teams tend to be the most popular, and whether or not it has to do with any external factors. These analytics used the search volume query to see what teams had the most searches, and where these searches were coming from. They ultimately wanted to see if there was any specific reason for some of the teams being the most popular, even if they weren't a successful team.

Finding cures or solutions to health problems is something that civilizations have searched to find since the beginning of time. Researchers at San Diego State University recently

[6] Can Google Trends Help Us Predict the Primary Election Results? N.p., 11 Feb. 2016. Web. 06 Oct. 2016.

conducted a study to see if there was a relation between mental illnesses, and seasonal weather.[7] These researchers used Google Trends to examine when the most highest volume of Google searches are for words such as " eating disorders," "schizophrenia," and "Bipolar disorder." The findings in this study showed that, in some cases, search volume for these words was up close to 40% in the winter compared to summer in the U.S. and Australia. This study was very conclusive for these researches as it showed that there is a positive correlation to prove that seasonal weather can affect mental health disorders. This is yet another example of how Google Trends can positively affect various aspects of the world.

It's clear that Google Trends can be used across many different disciplines, and is effective in allowing users to find trends. The multidisciplinary nature of these studies can infer that Google Trends are likely to be used increasingly in studies, and eventually by companies throughout the world. This successful analytical tool is something that will affect various market segments for years to come.

1.5 Existing Research

While observing existing research relating to stock market returns, there is an extremely popular theory in the investing world known as the 'Efficient Market Hypothesis', or EMH. The EMH is an investment theory which states that it is completely impossible to beat the market as an

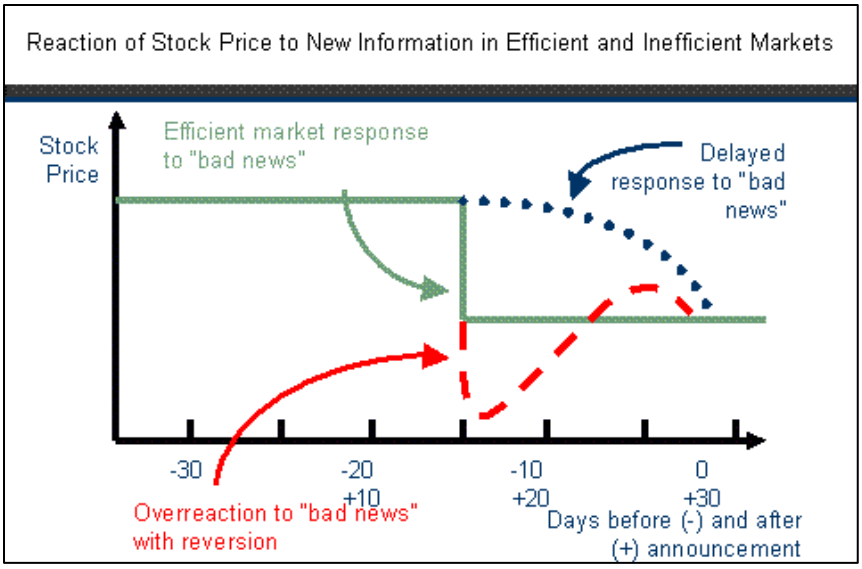

Figure 4: Efficient Market Hypothesis

[7] Numbers, By The. "News Center | SDSU | Can a Google Search Predict Mental Health?" Can a Google Search Predict Mental Health? N.p., n.d. Web. 06 Oct. 2016.

investor due to the fact that stock market efficiency causes current share prices to be reflected based upon all relevant information. In other words, the EMH states that because stocks are being traded at their fair value, investors are unable to purchase undervalued stocks or sell overvalued stocks.[8] Additionally, the EMH states that it is impossible to outperform the market through expert stock selection or selective market timing.

There are three main versions of the EMH: weak, semi-strong, and strong. The weak EMH claims that asset prices already reflect all past information that is publically available. The semi-strong EMH states that while assets reflect past information, they also constantly adjust to reflect new publically-available information. The last EMH, the strong form, claims that asset prices instantly reflect new information (and even insider information). Figure 4 displays the reaction of stock prices to new information in efficient and inefficient markets.

Existing research also shows a difference in opinion between the views of Investors on Wall Street and those of Main Street. As seen through precedent trading strategies and performances in the market, those on Wall Street are extremely bullish as they are confident in the stock market's ability to increase over time. On the other hand, Main street investors are more modern, realistic investors who play the market bearish where they believe stocks will constantly decline. Wall Street embodies global recognition and brand-name investment firms that have deep pockets and impressive resources. This may be attested to the reason investors are so confident in their investing

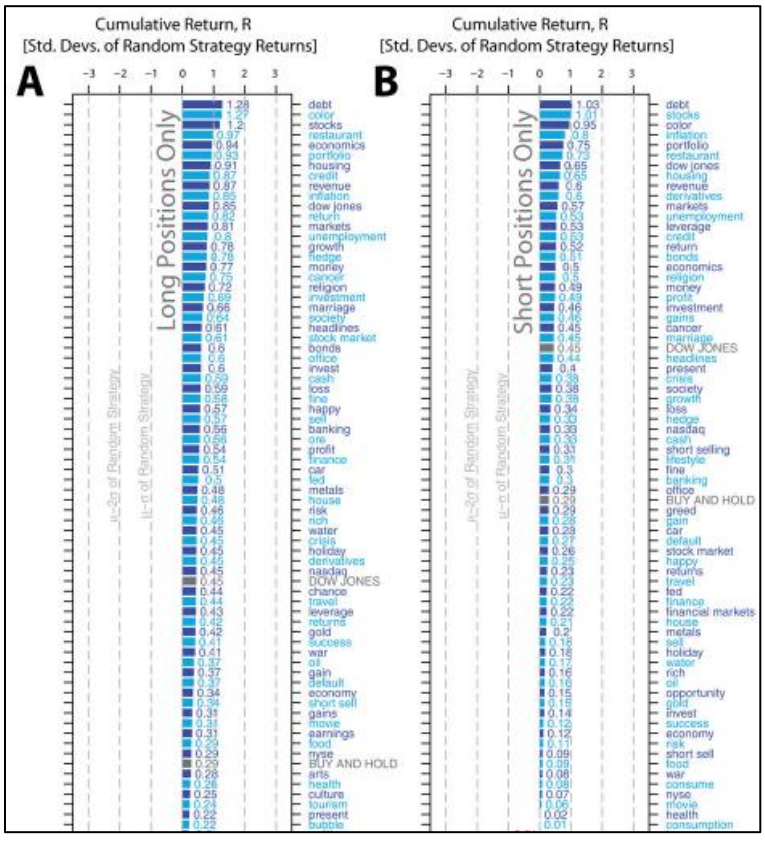

Figure 5: Standard Deviations of Random Strategy Returns

[8] Desai, Sameer (27 March 2011). "Efficient Market Hypothesis". Retrieved 2 June 2011.

strategies and engage in risky investing. Many Main Street investors got their start from Wall Street and are independent advisors. There is a trend for these investors to not engage in risky investing, as they do not have these big budgets and may even have a small number of resources to use.

As this regression analysis is new and innovative, many have begun to conduct in-depth research into the subject. Our team was able to use some of this previous research as secondary data, which could be applied to our project in various ways. While our project connects two different branches of academia: Finance/Investing and Data Analysis, research can be seen conducted by two types of individuals: those that are in the field, or practitioner, and those researching for academic purposes (like ourselves).

1.5.1 Practitioner

Prior research pertaining to our project subject has been done by many in the field of stock market investing. From contributors to Forbes magazine to investors on the floor of Wall Street, practitioners are extremely interested in this topic. We can only assume that the main reason for this is that successful discoveries in this subject matter can either make someone financial gain, or save them the agony of losing money in the stock market.

Our team was able to observe these studies and refer to them when deciding on a strategy that we recommend using in the market. For example, in examining other studies done, we observed a study done in 2013 by three respected individuals, whose article ended up being viewed over 140,000 times on Nature.com.[9] The group developed a strategy which outlined cumulative returns (standard deviations of random strategy returns) in the market.

1.5.2 Academic

In the academic realm, there is an ever-increasing effort for individuals to quantify trading behaviors in financial markets using Goggle Trends. This can be related to the fact that in academics, studies are always trying to prove a theory or wrong – or even develop a new theory. In the case of predicting stock market returns, there is a drive from the academic

[9] Nature.com. Nature Publishing Group, n.d. Web. 06 Oct. 2016.

community to want to prove EMH false. Our team found that the majority of previous academic research on the topic stems from individuals or groups involved in economic-related studies.

2.0 Recent Interest Informational Demand

2.1 Background

From the early days of the stock market economists have studied many different strategies that attempt to "beat" the market. "Beating" the stock market has proven to be no easy task however, as economists to this day still research ways to do so. The Google Trend data used in this project has given economists of all nature another tool to use in attempting to defeat the market. As there has been more research done by universities (Boston University/Warwick Business School) and big corporations (ESPN) using this trend data, data analytics have begun to wonder why. One reason to explain the recent spike in interest of this analytical data is due to the ease it takes to study and present this information. The data is very tangible, and is particularly useful in finding correlations. This allows for those looking at the data to find out whether there comparison of data is actually relevant or not. Another reason for the recent demand for this information is the desire for economists to continually seek new ways to defeat the market. This is something that will never change, as new ways to analyze data and find trends will continue to be the goals of successful economists.

2.2 Literature Review

2.2.1 Big Data; Using Google Searches to Predict Stock Market Falls

By: Tim Worstall, Forbes Contributor, August, 2014

Summary:

Tim Worstall, a contributor to Forbes, wrote an article which provides his insight into the topic of using Google search volume to predict stock market declines. In the article, Worstall puts a lot of emphasis on the fact that the majority of people really want to be able to apply the findings from this study to predict stock market moves for individual profit. The article also quotes academic studies on the topic relating to how these search results can be applied to financial gain in the markets, "In other words, what the researchers believe they've

found is a connection between what people looking at the stock market are interested in before they've made a buy-sell decision – and that the technique could be useful in other applications.[10] We suggest that extensions of these analyses could offer insight into large-scale information flow before a range of real-world events (Chirgwin, 2014).

2.2.2 'Big Data' Predicts Stock Movements, Boffins Claim

By: Richard Chirgwin. August 2014

Summary:

Richard Chirgwin, writing for The Register, discusses the usefulness of Google search volume. Specifically, how the supported power of using Google as a predictive tool has been extremely popular in academia ever since the Chocoloar Factory unveiled Google Flu Trends in 2008. In his article, Chirgwin talks about how University of Warwick researches believe they have found a connection between what people looking at the stock market are interested in before they've made buy-sell decisions. As he is quoted, "…for complex events such as financial market movements valuable information may be contained in search engine data for keywords with less-obvious semantic connections". Chirgwin's article can be applied to our project as we are able to see the buy/hold strategies from others. Additionally, our team is able to view other academic studies in a similar research area, all of which provides support to our conclusions.

2.2.3 Google can Help you Time the Market

By: Alanna Petroff, CNN Money, April 2013

Summary:

In her 2013 article for CNN Money, Alanna Petroff summarizes how the journal Scientific Report proves that you can use Google Trends to track the search volume of important financial terms, all of which can indicate whether markets are set to rise or fall. Petroff makes claims that over the course of seven years, U.S. and U.K. researchers have proven that when search volume terms such as "debt", "money", etc. rose, markets generally fell (and vice versa).[11] The idea behind this theory is that when people are Googling such financial terms, they are getting concerned about the markets and are likely to start selling. Petroff states that the reports of co-

[10] "Big Data; Using Google Searches to Predict Stock Market Falls." Forbes. Forbes Magazine, n.d. Web. 06 Oct. 2016.
[11] "Google Searches Can Predict Market Moves." CNNMoney. Cable News Network, n.d. Web. 06 Oct. 2016.

author Helen Susannah Moat, a social scientist from the University College London, looked at nearly 100 different search terms and found that “debt” best predicted market moves between 2004 and 2011. Additionally, the article states that words like “debt” may no longer be useful as the strategy needs continuous updating as new keywords become more relevant to the markets. This article provides us with insight into buy/sell strategies using our results, while also lending information into how this strategy needs to be innovated in future research.

2.3 Existing Gaps

2.3.1 Using the VIX and Google Search Volume to Make Stock Market Predictions

Over the years there have been many studies that have explored how Google search volumes and queries can affect volatility in the market. The VIX is the ticker symbol for the Chicago Board Operations Exchange that shows the market’s expectation of 30-day volatility. Volatility is defined as the uncertainty or risk within a given market. Researcher and Professor Michal Dzielinski of the University of Zurich in Switzerland, studied whether or not Google Search terms would correlate with market volatility, which in turn would show an ability for Google Trend terms to predict market changes. Dzielinski came to the conclusion that volatility cannot be predicted using Google Search Volume, as GSV will only reflect market uncertainty as it has happened. This case is impactful in comparison to our research as Dzielinski concluded that GSV could not predict volatility in the stock market.[12] Our research will attempt to show that GSV can predict changes in the New York Stock Exchange, not the VIX. Instead of our research predicting the risk within a given market, we will try to use GSV to predict rises and falls of the market. Our team will eventually collect data from Google Trends to analyze it versus the S&P 500, so it is important that we consider the implications of the research conducted by Dzielinski. The biggest takeaway from these findings is that the data explicitly showed that investors react to uncertainty in the market by selling their stocks. This makes finding any sort of predicative solution an impossible task. Our team understands that volatility

[12] Dzielinski, Michal. "Measuring Economic Uncertainty and Its Impact on the Stock Market." Measuring Economic Uncertainty and Its Impact on the Stock Market. Elsevier Inc., 3 Oct. 2011. Web.

and Google Search Trends may not correlate, however this is precisely why we will attempt to use the S&P 500.

2.3.2 Using Google Search Volume to Predict Market Volatility using the DOW Jones

Other researchers have also studied market trends in an attempt to predict market volatility. Thomas Dimpfl and Stephan Jank, of the University of Tubingen in Germany and the Frankfurt School of Finance and Management and Centre of Financial Research respectively, have conducted this type of research. These two professors were hoping to use Google Search Volume data to find a correlation to the volatility of the DOW Jones. Ultimately, these two researchers wanted to discover if there is a positive correlation between Google Search Volume of the term "DOW Jones," and volatility in that market. They came to some impactful conclusions after they culminated, and analyzed all of their data. The results of their research conclude that search queries of the term "DOW Jones" can effectively predict volatility within the DOW Jones.[13]

Using data from Google Search Volume as a way to predict volatility within a market has proven to be a powerful tool that researchers can use. However this is not what our research will be tailored around. By understanding the type of research that has been previously conducted with GSV and the stock market, our teams hope to find a correlation between the NYSE and GSV. The type of studies done one in this manner is not very common, and we hope to close this gap through our research.

3.0 Sample Data

3.1 Data Used

For the purpose of this project we used the Federal Reserve Economic Database (FRED) in order to obtain information pertaining to the S&P 500. Essentially FRED is a database sustained by the Federal Reserve Bank of St. Louis' research division.[14] The research division has more than 381,000 economic time series from 81 sources and allows the user to view the data

[13] Dimpfl, Thomas, and Stephan Jank. "Can Internet Search Queries Help to Predict Stock Market Volatility?" Social Science Research Network. European Financial Management, Mar. 2016. Web.
[14] Federal Reserve Economic Data. Accessed January 18, 2016.

in several different ways. The foremost that we used was graphical text that can be easily downloaded on to a spreadsheet, in our case we utilized Excel. It is important to note that the time series are accumulated by the Federal Reserve and attained from the U.S Census and Bureau of Labor Statistics. The ease of use of converting FRED S&P 500 data to an Excel sheet allowed us to run regressions which was crucial to the success of this project.

Another source of data essential to the project's success was Google Trends. This utility provided by Google Inc. feeds off of Google Search and quantifies how many times a particular term or word was searched. It monitors this volume across all regions of the world, or even specified regions. The search volume was also standardized on a 0-100 scale which made it easy to compare and contrast the volumes of different search terms.

3.1.1 Google Trend Data

Example of downloaded data from Google Trends and FRED Data:

	A	B	C	D	E	F
2	United States 2004 - present					
3						
4	Interest over time					
5	Week	asset				
6	2006-03-19 - 2006-03-25	80		#N/A		
7	2006-03-26 - 2006-04-01	82	2.5	1302.39	#N/A	
8	2006-04-02 - 2006-04-08	78	-4.87805	1298.56	-0.29407	
9	2006-04-09 - 2006-04-15	80	2.564103	1303.97	0.416615	
10	2006-04-16 - 2006-04-22	81	1.25	1290.10	-1.06367	
11	2006-04-23 - 2006-04-29	79	-2.46914	1305.13	1.165026	
12	2006-04-30 - 2006-05-06	81	2.531646	1307.12	0.152475	
13	2006-05-07 - 2006-05-13	80	-1.23457	1312.85	0.438368	
14	2006-05-14 - 2006-05-20	76	-5	1313.96	0.084549	
15	2006-05-21 - 2006-05-27	79	3.947368	1277.15	-2.80146	
16	2006-05-28 - 2006-06-03	71	-10.1266	1266.05	-0.86912	
17	2006-06-04 - 2006-06-10	77	8.450704	1275.97	0.783539	
18	2006-06-11 - 2006-06-17	86	11.68831	1259.10	-1.32213	
19	2006-06-18 - 2006-06-24	82	-4.65116	1239.57	-1.55111	
20	2006-06-25 - 2006-07-01	79	-3.65854	1244.51	0.398525	
21	2006-07-02 - 2006-07-08	63	-20.2532	1255.77	0.904774	
22	2006-07-09 - 2006-07-15	76	20.63492	1272.67	1.345788	
23	2006-07-16 - 2006-07-22	77	1.315789	1255.39	-1.35778	
24	2006-07-23 - 2006-07-29	76	-1.2987	1244.12	-0.89773	
25	2006-07-30 - 2006-08-05	78	2.631579	1267.99	1.918625	
26	2006-08-06 - 2006-08-12	75	-3.84615	1277.15	0.722403	
27	2006-08-13 - 2006-08-19	77	2.666667	1270.35	-0.53244	
28	2006-08-20 - 2006-08-26	76	-1.2987	1289.80	1.531074	
29	2006-08-27 - 2006-09-02	70	-7.89474	1296.10	0.488448	

Figure 6: Example of Google Trends Downloaded Data

We then took our data and observed the columns that we needed to fill in order to align the data. As you can see in Figure 6 above, column C displays the percent change from week to week for the key word taken from Google. We then downloaded data from FRED data regarding the stock market price at any given week and aligned the columns based on weekly date. After taking the percent change of the stock data, we were able to compare the two percent changes. There were many benefits to comparing percent changes rather than just eyeballing the results in graph form; percent change is more accurate (as you can see our data reaches the 100,000th decimal place), you cannot be visibly accurate without understanding the breakdown of week-to-week statistical data, etc.

Google Trend words used to run regressions:

Given the fact that our group applied a large amount of Google keywords to the stock market data, we needed to be accurate with words successfully applied in our regression analysis. As you can see in the chart below, there were many keywords that proved to be significant. The blocks that are highlighted green indicate the successful keywords in the group.

Keyword	Success	Keyword	Success
Absolute Advantage	No	GDP	Yes
Asset	Yes	Inflation	Yes
Bankruptcy	No	Insurance	Yes
Benefits	Yes	Interest Rates	No
Budget	Yes	Liquidity	No
Business Cycle	No	Market	No
Capital	No	Monopoly	Yes
Charity	Yes	Oil Prices	Yes
Collusion	No	Phillips Curve	Yes
Consumer Price index	No	Recession	No
Credit Card Debt	No	Substitutes	No
Default	No	Social Security	Yes
Deflation	No	Stock Market	No
Elasticity	No	Tariff	No
Employment	Yes	Utility	Yes
Exchange Rate	No	Welfare	Yes
Total:	14/32	43.75% Success	

Figure 7: Successful Keywords

3.1.2 Graphs Used

Example of a graph for a keyword searched on Google Trends:

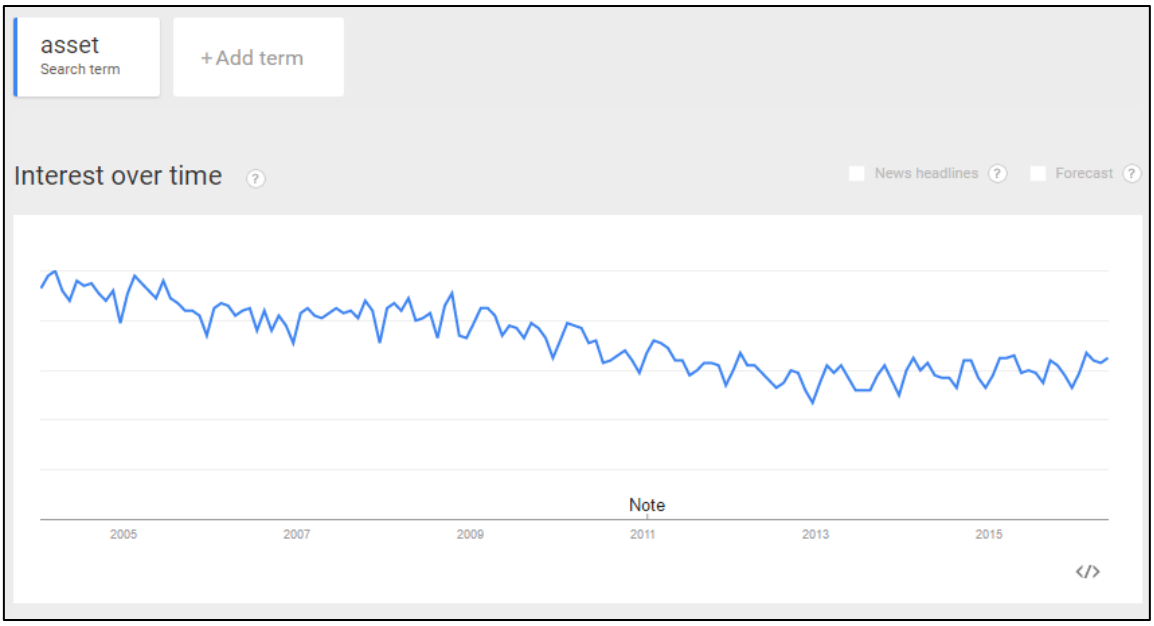

Figure 8: Example of a Graphed Google Trend Keyword

Figure 8 represents what we would receive as a graph from a Google Trend search term. What differs between the graph Google provides and the data you can publically download is in how the search volume is subjective. In the graph, the horizontal axis represents the date, while the vertical axis marks how often a search term is searched for (relative to the total number of searches in the United States).

The predictability of search trends was published through research done by Yossi Matias in 2009.[15] The New York Times even wrote about how Seth Stephens-Davidowitz was using Google Trends to measure a wide variety of terms, including comparisons between racist terms and the levels of racism in different parts of the US. He then made a cross-comparison of how

[15] On the predictability of Search Trends, Yossi Matias, Niv Efron, and Yair Shimshoni, Insights Search, Google Research blog, August 17, 2009.

this correlated with Obama's 2008 vote share where he concluded that four percentage points were lost due to racial animus.[16]

3.1.3 S&P 500 Data (FRED Data)

Example of a graph given from FRED Data (Late 2006 – Early 2016) for the S&P 500:

Figure 9: Example of Graphed FRED Data

Similar to S&P 500 data, the economic data published on Federal Reserve Economic Data (FRED) data is widely available to the public and is constantly reported in the media. FRED is a database that is maintained by the Research Division of the Federal Reserve Bank of St. Louis. These time series are then compiled by the Federal Reserve, which is collected from government agencies like the U.S. Census and the Bureau of Labor Statistics. Business Insider even published an article titled "The Most Amazing Economics Website in the World", where they quote Paul Krugman (Professor of Economics at the Graduate Center of the City University of New York) of saying "I think just about everyone doing short-order research… has become a FRED fanatic".[17] The graphs provided by FRED displays dates on the horizontal axis with the index of closing price on the vertical axis.

[16] Casey Johnston (April 6, 2012). "Google Trends reveals clues about the mentality of richer nations". Ars Technica. Retrieved April 9, 2012.

[17] "The Most Amazing Economics Website in the World", Joe Weisenthal, Business Insider, March 23, 2012.

4.0 Empirical Framework (Methodology)

4.1 Regressions

Microsoft Excel uses built-in software to help users run linear regressions. Our team used Microsoft Excel due to the fact that we had reason to believe that a linear relationship existed between two variables.

A linear regression is a modeling approach between a scalar dependent variable "y" and one or more explanatory "x" variables. In order to determine whether or not a relationship existed between Google Trend terms and the S&P 500 we had to run various linear regressions. These regressions were essential to model the relationship between the two and decide if there was any type of correlation. Microsoft Excel uses built-in software to help users run linear regressions. Our team used Microsoft Excel due to the fact that we had reason to believe that a linear relationship existed between two variables.

4.1.1 Regressions vs. S&P 50

In running regressions between search volume and S&P 500 data, our team first had to collect data. This was done through finding publically-available data online. Many financial investing platforms offer the ability to download historical stock data. Our team chose FRED data to download our S&P 500 data, as it allows for easy access into custom-entered dates. Additionally, our team used Google Trends to download our economic search terms. One of the key steps in this process was the aligning of weekly dates between both sets of data. Our data aligns at the starting date, March 19th, 2006. Using Excel, we than turned the given data into percent changes (all of which will be shown in Chapter 4.2). We than ran regressions using Microsoft Excel and observed whether or not the keyword proved to be significant.

4.2 Equations- Statistical Equations

4.2.1 Percent Change

It was critical for our team to run regressions on both sets of data using their percent of change as it allows for similar comparisons. The equation to run percent of change on week-to-week data sets is:

$$Percent\ Change = \frac{Week\ (New) - Week\ (Old)}{Week\ (Old)} \text{ X } 100$$

To provide a visual, here is our percent change data used for our keyword “Benefits”. The blue arrows represent where the percent change formula is being applied. For example, if we wanted to find the percent change of box B8, our equation would be:

$$Percent\ Change\ (B8) = \frac{B8 - B7}{B7} \text{ X } 100$$

	A	B	C	D	E
1	Web Search interest: benefits				
2	United States 2004 - present				
3					
4	Interest over time				
5	Week	Benefits			
6	2006-03-19 - 2006-03-25	47		#N/	
7	2006-03-26 - 2006-04-01	46	-2.12766	1302.39	#N/A
8	2006-04-02 - 2006-04-08	47	2.173913	1298.56	-0.29407
9	2006-04-09 - 2006-04-15	47	0	1303.97	0.416615
10	2006-04-16 - 2006-04-22	47	0	1290.10	-1.06367
11	2006-04-23 - 2006-04-29	48	2.12766	1305.13	1.165026
12	2006-04-30 - 2006-05-06	48	0	1307.12	0.152475
13	2006-05-07 - 2006-05-13	44	-8.33333	1312.85	0.438368
14	2006-05-14 - 2006-05-20	43	-2.27273	1313.96	0.084549
15	2006-05-21 - 2006-05-27	40	-6.97674	1277.15	-2.80146
16	2006-05-28 - 2006-06-03	41	2.5	1266.05	-0.86912

Figure 10: Percent Change in Stock Market Data

4.2.2 Regression Equations

Running an Excel regression provided us with the following summary output:

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.093486056							
5	R Square	0.008739643							
6	Adjusted R Square	0.006814865							
7	Standard Error	2.010724911							
8	Observations	517							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	18.35770896	18.3577	4.5406	0.033573053			
13	Residual	515	2082.152553	4.04301					
14	Total	516	2100.510262						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	0.085182242	0.088556281	0.9619	0.33655	-0.088793745	0.25916	-0.08879	0.25916
18	X Variable 1	0.02481081	0.011643522	2.13087	0.03357	0.001936168	0.04769	0.00194	0.04769

Figure 11: Regression Summary Output

Our team was mainly concerned with the t-statistic and the p-value provided in the summary output to prove if a keyword was significant or not. The t-statistic is used to compare two different sets of values, where it uses means and standard deviations of two samples to make a comparison. The formula for finding the t-statistic is as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Figure 12: t-Statistic Formula

- X1 = Mean of first set of values
- X2 = Mean of second set of values
- S1 = Standard deviation of first set of values
- S2 = Standard deviation of second set of values
- n1 = Total number of values in first set
- n2 = Total number of values in second set

For a t-Statistic to be significant for us, the value would have to be equal or greater than 2. In this example, there is indeed a significant t value.

The P-value is the area under the null-distribution curve which is in disagreement with the null hypothesis. If the P-value of a data set is below .05, than we can reject our null hypothesis; this means that we rule out the hypotheses that our experiment variables have no meaningful effect on the results. The following figure shows the corresponding conclusions associated with the test statistic.

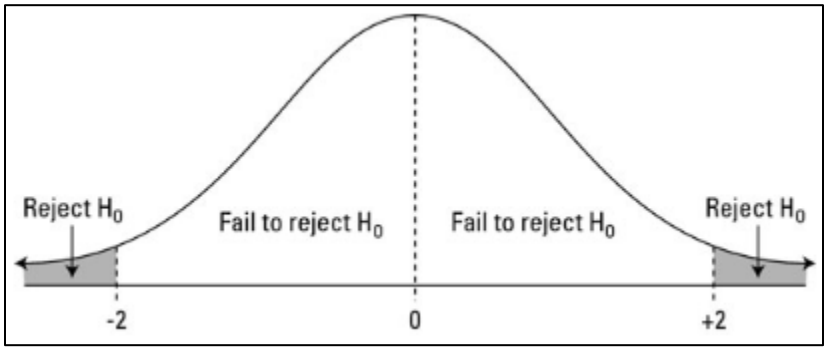

Figure 13: P-Value Null Distribution

5.0 Major Findings & Socioeconomic Implications

The search volume data for a ten year span was downloaded in an Excel file for each of the keywords. We also downloaded the S&P 500 returns over the past ten years off of FRED Data. Once we had both data sets we were ready to compare the two. The formulas used to calculate the change in the Google Trends keywords and S&P 500's returns are stated earlier. With each having its own column for change we were able to run a regression analysis for each word and determine which keywords correlated best with the S&P 500.

Our team ran regressions on a total of 32 keywords words from Google Trends. These words were chosen carefully based off the research we did on other projects similar in nature and also through our knowledge of the stock market. We tried to pick words which were thought to have relation to the stock market as a whole. Out of the 32 words searched on Google Trends, 14 of them had a satisfactory correlation to the S&P 500. Of those 14, 3 had a P-value under 10% and a t Stat varying around 2-2.5. The other 11 words all had a strong positive correlation with the P-value being under 5% and the t Stat being generally around 2.5.

5.1 Significant Search Terms

Once we ran all of our keyword regressions, our team removed the keywords that were irrelevant to our study. As you can see in Figure 14, there are 14 keywords that have significant

P-values. The top three significant P-values are highlighted in yellow: GDP, Utility, and Welfare. Also included in the table are the t-statistic and the R-squared variable.

Key Word	T-Stat	P-Value	R Squared
Asset	2.11997	0.03449	0.00865
Benefits	2.13087	0.03357	0.00874
Budget	2.52088	0.01209	0.01219
Charity	2.26942	0.02366	0.00991
Employment	2.40007	0.01675	0.01106
GDP	1.83341	0.06731	0.00648
Inflation	2.13592	0.03316	0.00878
Insurance	2.51322	0.01227	0.01212
Monopoly	2.50339	0.01261	0.01202
Oil Prices	5.80086	1.15 x 10-8	0.06133
Phillips Curve	1.90261	0.04765	0.00698
Social Security	1.98164	0.04805	0.00758
Utility	1.68895	0.09183	0.00551
Welfare	1.82661	0.06833	0.00644

5.2 Example of Regression Analysis for Search Term (Budget)

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.110404193							
5	R Square	0.012189086							
6	Adjusted R Square	0.010271006							
7	Standard Error	2.007223345							
8	Observations	517							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	25.6032999	25.6033	6.354839	0.012007049			
13	Residual	515	2074.906962	4.028946					
14	Total	516	2100.510262						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	0.087046441	0.088336712	0.985394	0.324893	-0.086498184	0.26059107	-0.08649818	0.260591066
18	X Variable 1	0.026860471	0.010655194	2.520881	0.012007	0.00592748	0.04779346	0.00592748	0.047793462

Figure 15: Regression Analysis for Keyword "Budget"

5.2.1 Example of S&P 500 vs first three terms over 10 Years

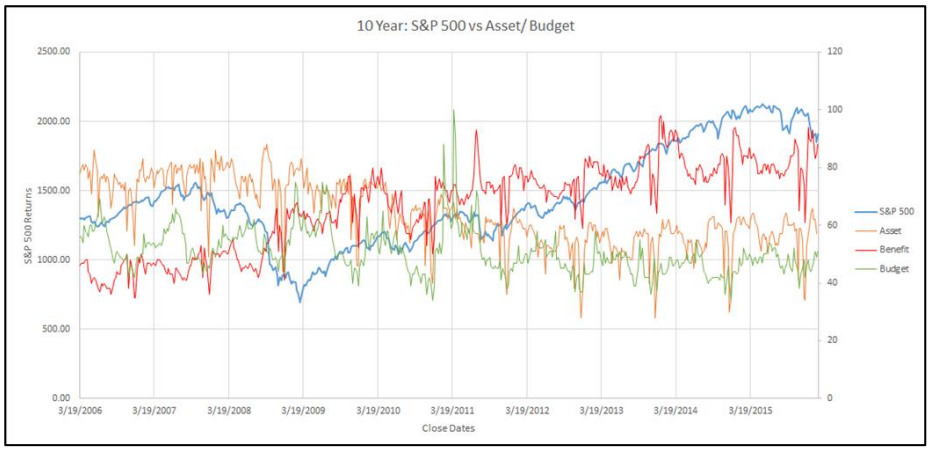

Figure 16: Ten-Year S&P 500 vs. "Asset" and "Budget"

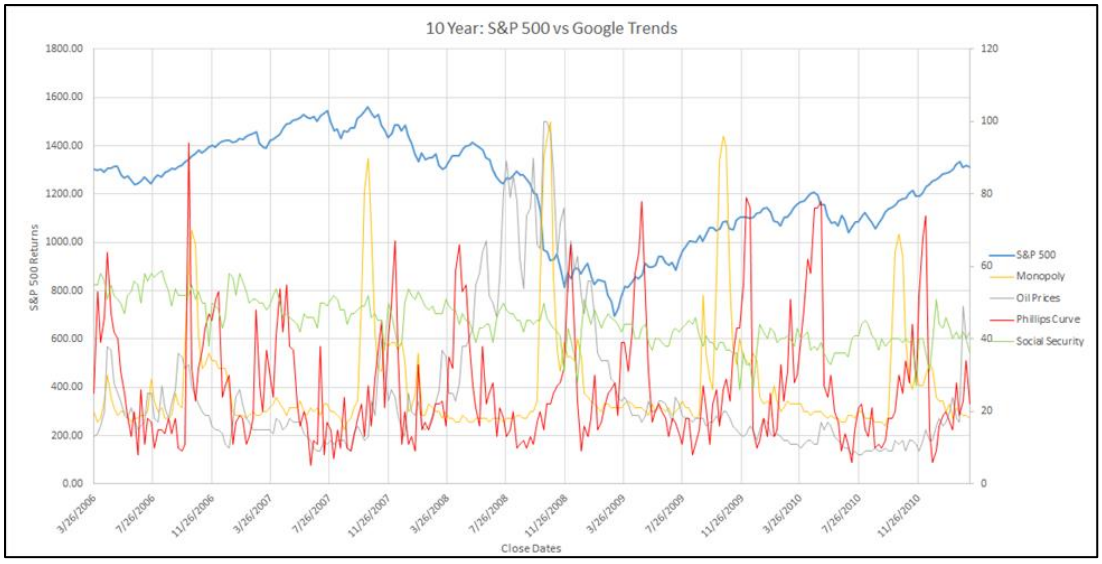

Figure 17: Ten-Year S&P 500 vs. Google Trends

5.2.2 Example of Change in S&P500 vs. Benefits

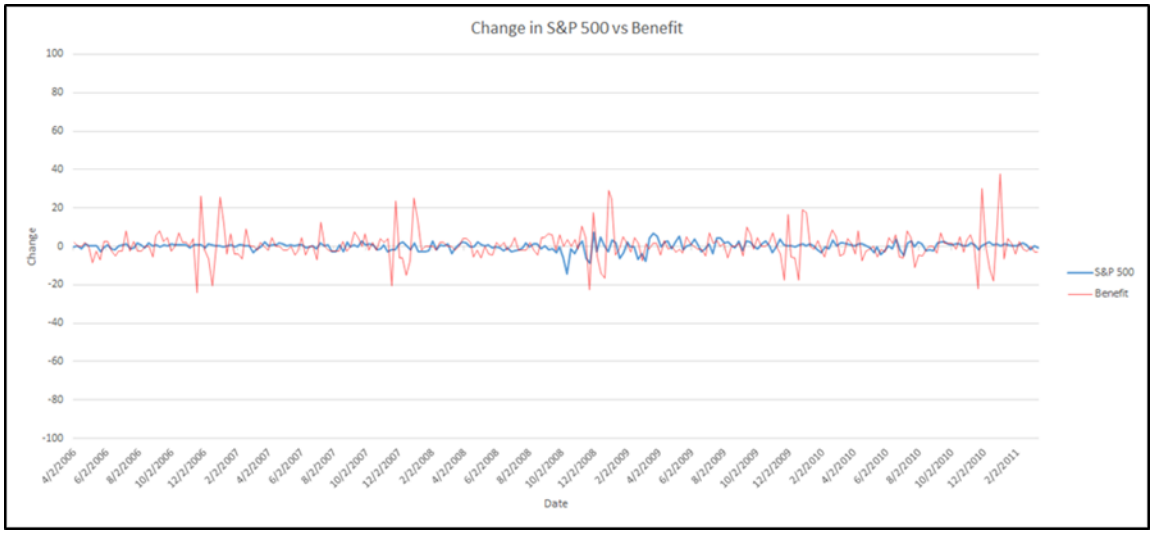

Figure 18: Ten-Year Change in S&P 500 vs. "Benefit"

5.3 Buy and Sell Strategies

After aggregating all of the data and determining which terms correlated best with the S&P 500, we focused on creating a buy and hold strategy. We wanted to create the simplest strategy possible and therefore concentrated on using only one term, "debt." Comparing the google trend and S&P 500 data side by side, we found that the higher the spike in google searches for "debt" the sharper the drop in the S&P 500 and vice versa. This analysis was made on a weekly basis for both the google trend term "debt" and change in the S&P 500. We found that the most drastic rises and drops in the S&P 500 occurred when the google trend search volume for "debt" climbed or sunk outside of the 18-40 range. From this range we developed a nonaggressive buy and hold strategy that was meant to make small returns steadily over longer periods of time. If the google search volume for "debt" ever moved above 40 then you would sell, because it meant that a steep drop was coming in the S&P 500. Inversely if the google search volume for "debt" dropped below 18 you would buy because it meant a rise in the S&P 500 was coming. Lastly if the search volume was inside of the

Figure 19: Example of Buy/Hold Strategy

18-40 range you would simply hold until it went out of that range. Examining this strategy psychologically helps to understand why it works. If people are searching "debt" more on google it most likely means people are pessimistic about the market, which in turn would result in a bearish trending market. Inversely if people are searching "debt" on google less then generally they are more optimistic concerning the market, which would show an overall bullish trend. The Efficient Market Theory states that it is impossible to "beat" the market, as the market shows and reflects all relevant information. This means is that if someone invests in the market they shouldn't be able to sell a stock when it's overachieving to obtain a higher return, nor should an investor be able to make gains if they buy a stock when it is underperforming. Economists have battled back and forth on this theory for hundreds of year, and will most likely continue to for as long as the stock market exists. The information that is represented in this project can be used to challenge this hypothesis even more. The findings from the Google Trend data portrayed in this project attempt to show there is a correlation between certain search terms, and how the market reacts. By analyzing the above data it's clear that the data shows certain words can be indicators for whether or not the market is going to change. The buy and hold strategies portrayed show that certain words can give indicators to investors when they should or should not invest in a stock, as well as showing when they should sell a share. This data shows investors a strategy in which they can invest in the market when it is "underachieving," and then exit the market when it is "overachieving." This in and of itself gives economists a strong case to challenge the efficient market hypothesis. The findings of this project have certainly allowed for more

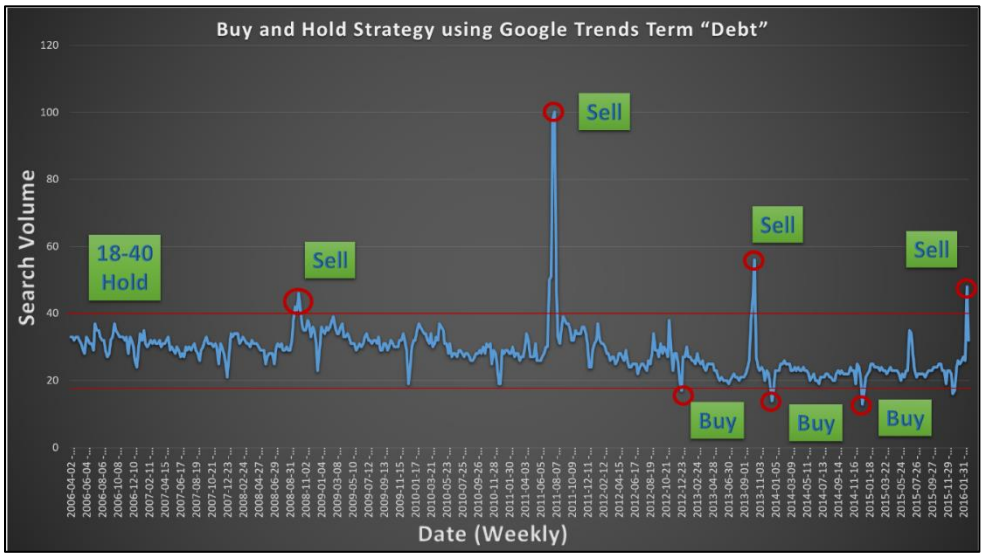

Figure 20: Buy/Hold Strategy using "Debt"

speculation into whether or not the efficient market hypothesis can be proven accurate.

Examining one buy to sell period in the previous graph it is clear that the rate of return would be positive thus proving this method to be profitable.

Using the previously established buy and hold strategy the first drop below 18 for search volume of "debt" occurred December 23, 2012. This indicated a time to buy stock in the S&P 500, which at that time was approximately 1430 points. The search volume did not go outside the 18-40 threshold until October 6, 2013. At that point it jumped to "45" which would indicate a time to sell, and the S&P 500 Index was approximately 1700 points.

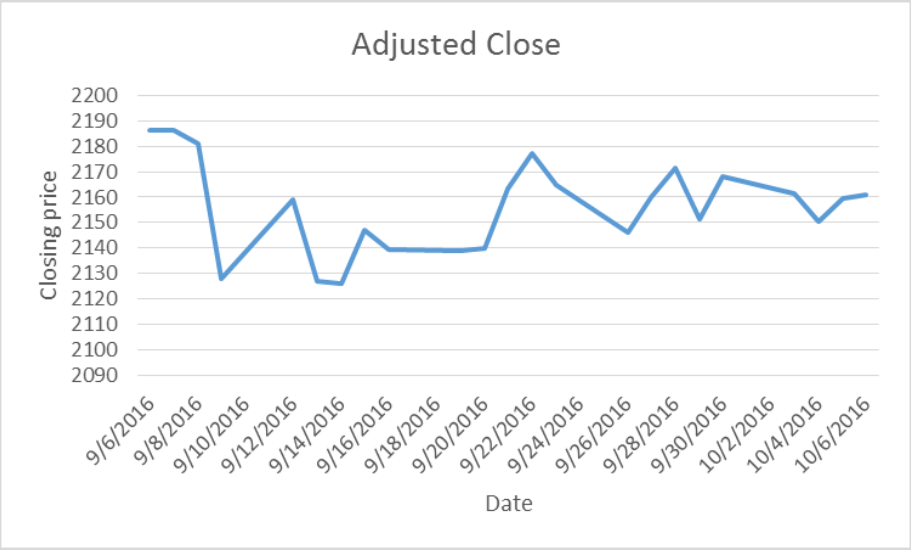

Figure 21: S&P 500 Adjusted Closes (09/06/2016 - 10/06/2016)

Looking at the increase in points of the S&P 500 between this buy and sell period it is clear that this technique works in terms of having steady returns over a longer period of time. After you sell you shares at the indicated buy point, you should weight to buy again until the search volume drops below 18. Then you simply repeat the strategy and steady returns over a long period of time are very likely. It is a simple and less risky way to make money in the S&P 500 index.

6.0 Conclusions

The main challenge of this project was to systematically challenge the efficient market hypothesis and determine whether or not it could be refuted. When first researching Google Trends we found that its success as an estimation tool was not limited to just the stock market, but to various facets of society. For one we saw that it had been used to predict

upcoming flu epidemics by analyzing a spike in search volume of flu-related symptoms. We saw its success when Ugam found the relationship between political races and search volume of different political party terms.[18] With all of its triumph in diverse areas we decided to apply Google Trends to the stock market, specifically the S&P 500. We wanted to hone this tool and tailor it in order to find correlations between how society felt and the fluctuation of the S&P 500. After running regressions of different terms and the returns of the S&P 500 we found strong correlations. From the terms that held strong correlation with the S&P 500 we developed a buy and hold strategy centered on the term "debt." It was hard to test this strategy on the market because we developed it to have small steady returns over a longer period of time.

While there is no such thing as a guaranteed positive investment, all investments are made with the intent to make positive gains in the stock market. With new technology continuing to arise, techniques are beginning to form with the intentions of helping individuals have peace of mind when investing their hard-earned monetary funds. Our method of applying Google Search Volume to trends in the stock market can help investors have a better sense of risk and returns that are associated with these risks in the market. This project aimed to help analyze how the flow of information from information providers online can help investors reduce the risks and increase the gains associated with investments in publically-traded corporations. This is an extremely beneficial study if executed correctly, as even some of the top investors in the world do not have a method of investing that can produce positive returns constantly.

During our project, our team utilized skills obtained through elements of finance, psychology, mathematics, Excel coding, and behavioral economics to be able to access, obtain, and analyze data required to present worthwhile material. In today's society, increasing methods of analyzing data are heavily available to public, and our team tried to utilize these methods for monetary gain.

We can draw many conclusions from our project, including the successful proof that Google Search Volume can in fact be used as a tool to predict risk and return in the stock

[18] Jeremy Ginsburg. "Detecting Influenza Epidemics Using Search Engine Query Data." Www.Google.com. N.p., n.d. Web.

market. Our assumptions are as such: weeks that hold large volumes of search data for words listed throughout our project signify curiosity and/or concern pertaining to specific stocks. During this time, the stock market fluctuated based upon those volumes. Rather than constantly searching for the ability to predict returns, our theory provided us with understanding risk in the market. Throughout the chapters in our project, we elaborate, with conclusive evidence, that there is in fact a relationship between levels of Google Search Volume and risk/returns in the stock market at those points in time. In other words, our results from this project confirm that the documented certainty of those entering keywords in Google has a direct impact on the variability and risk/returns in the stock market. There have been many other projects done on adjacent topics by those respected in academia, all of which lead to similar conclusions like ours.

Our team would like to take the time to thank Dr. Dimitrios Koutmos, Professor of finance, risk analysis, and asset pricing at WPI's Foisie School of Business. Without the help of Professor Koutmos and WPI's Foisie school of Business, our project would not have had the ability to be as successful as it was. Both allocated the necessary resources required for our team to learn and apply what was needed to understand and execute our study.

7.0 References

"Stock Market History Chart and a Detailed Look at the Markets." *Stock Picks System*. N.p., n.d. Web. 20 Apr. 2016.

"What Is the History of the S&P 500? | Investopedia." *Investopedia*. N.p., 10 Apr. 2015. Web. 20 Apr. 2016.

"S&P 500," 2015. [Online]. Available: file:///C:/Users/zwang5/Downloads/fs-sp-500.pdf.

Nature.com. Nature Publishing Group, n.d. Web. 06 Oct. 2016.

Bauer, Scott. Klein, Ron. Coakley, Matthew. Song, Danwen. “Using Google Trends to Predict Retail Sales” 2015

Can Google Trends Help Us Predict the Primary Election Results? N.p., 11 Feb. 2016. Web. 06 Oct. 2016.

Numbers, By The. "News Center | SDSU | Can a Google Search Predict Mental Health?" Can a Google Search Predict Mental Health? N.p., n.d. Web. 06 Oct. 2016.

Desai, Sameer (27 March 2011). "Efficient Market Hypothesis". Retrieved 2 June 2011.

Nature.com. Nature Publishing Group, n.d. Web. 06 Oct. 2016.

"Big Data; Using Google Searches to Predict Stock Market Falls." Forbes. Forbes Magazine, n.d. Web. 06 Oct. 2016.

“Google Searches Can Predict Market Moves." CNNMoney. Cable News Network, n.d. Web. 06 Oct. 2016.

Dzielinski, Michal. "Measuring Economic Uncertainty and Its Impact on the Stock Market." Measuring Economic Uncertainty and Its Impact on the Stock Market. Elsevier Inc., 3 Oct. 2011. Web.

Dimpfl, Thomas, and Stephan Jank. "Can Internet Search Queries Help to Predict Stock Market Volatility?" Social Science Research Network. European Financial Management, Mar. 2016. Web.

Federal Reserve Economic Data. Accessed January 18, 2016.

On the predictability of Search Trends, Yossi Matias, Niv Efron, and Yair Shimshoni, Insights Search, Google Research blog, August 17, 2009.

Casey Johnston (April 6, 2012). "Google Trends reveals clues about the mentality of richer nations". Ars Technica. Retrieved April 9, 2012.

"The Most Amazing Economics Website in the World", Joe Weisenthal, Business Insider, March 23, 2012.

Jeremy Ginsburg. "Detecting Influenza Epidemics Using Search Engine Query Data." Www.Google.com. N.p., n.d. Web.

Highlights. *Using Google Trends to Predict Retail Sales* (n.d.): n. pag. Web.

Person, and Nate Silver and Ritchie King. "The Distribution of Fandom in Pro Leagues." DataLab. N.p., 02 Apr. 2014. Web. 16 May 2016.

Predicting the Present with Google Trends (n.d.): n. pag. Web.

Preis, Tobias, Helen S. Moat, and Eugene H. Stanley. "Quantifying Trading Behavior in Financial Markets Using Google Trends." Nature.com. Nature Publishing Group, n.d. Web. 16 May 2016.

Rios, Felix. "Blog." Can Google Trends Help Us Predict the Primary Election Results? Market Research Technology, 11 Feb. 2016. Web. 16 May 2016.

"The Efficient Market Hypothesis - Boundless Open Textbook." Boundless. N.p., n.d. Web. 30 May 2016.

White, Daniel. "Here's What Google Results Hint about the New Hampshire Primary." TIME. N.p., n.d. Web.

"S&P 500©." - FRED. N.p., n.d. Web. 16 May 2016.

"Google Trends." Google Trends. N.p., n.d. Web. 16 May 2016.