

Exploring the relationship between Google Trends data and stock price data.

Author: Dartanyon Shivers, Advisor: Chris Deotte PhD, Self Published: June 2017, UCSD

Abstract: In this work, we provide a brief description of the stock market and search engines, more specifically Google's search engine. We then suggest some efficient methods to gathering historical stock price data and google search data. Additionally, we propose using a test that we created to explore the relationship, if any, of stock prices and the popularity of google searches. Finally, we share our results from the test and discuss the possibility of using the popularity of google searches to predict future stock price movement.

1. Introduction

1.1 Stock Market

The stock market is a global network in which individuals can purchase ownership, commonly known as shares, of a public company. All companies start out as privately owned companies; they do not become public until they are listed under a stock exchange. Exchanges are organized markets in which financial instruments are bought and sold. These financial instruments can be broken up into two categories, equity-based and debt-based. Equity-based financial instruments represent ownership of an asset, and debt-based financial instruments represent a loan made by an investor to the owner of that asset. Both of these categories contain different types of instruments, however for the purpose of this paper we will focus on stocks, which are shares in ownership of a company and they are sold on stock exchanges.

The collection of stock exchanges is what makes up the stock market. The United States main exchanges are the New York Stock Exchange and National Association of Securities Dealers Automated Quotations or commonly known as NYSE and NASDAQ respectively. Not just any company can be listed under these exchanges, they must meet certain requirements such as possessing a specified number of shareholders and an evaluation above some minimum worth. All businesses are considered private until they decide to "go public" and be listed under an exchange. Business owners choose to "go public" for one reason, to raise money. This could be to raise money for themselves, however in most cases they do this to use that money to grow their business, via purchasing new equipment, developing better products, expanding operations, etc. Whatever the reason, once they make this decision, shares of their business become available for anyone in the world to purchase.

Before becoming listed, banks and other financial firms come together to evaluate the company's worth. Once these financial institutions and the company agree on the number of shares to be listed and their worth, an initial public offering, IPO, is held. During an IPO, the shares of the company are all sold for the exact price that was agreed upon. The money made from the sale goes to the company, and individuals who purchased stock in the company can now call themselves a "shareholder" of that company. Although shareholders have some sense of ownership of the company, they do not own any property of the company nor do they have much control over it. Being a shareholder provides voting privileges and rights to their percentage of the company's worth if it were to be liquidated. Shareholders of the company are unable to make decisions on the company's behalf and they are not entitled to the equipment, products, or anything else that makes up the business. The tradeoff for this is that shareholders relinquish all liability to the company itself. Under the law, a corporation is treated as a legal person, in the sense that it can borrow money, own property, be sued and file taxes. This is beneficial to the investor because if a company is to go bankrupt or be sued, that investor's personal assets outside of the company are not at risk.

So if individuals have little to no authority within the company nor are they able to take what they believe is their fair share from the company, why would anyone even consider investing? Well, stocks have proven to be the best investments over time. Investing beats placing your money into a savings account because due to inflation you are essentially losing money in the long run. Inflation is the general increase in prices and decrease of purchasing power of money. So, if you were to place \$1,000 into a savings account today and not touch it in 10 years, your \$1,000 in the future will not be able to buy you the same things you could have bought 10 years

prior. It is true that most banks pay some form of interest if individuals leave their money in their savings account, however that interest is generally around .01% per year. This is problematic because inflation rises at an average rate of 3% per year. This means that you're essentially losing 2.99% of your money's purchasing power each year. Hence investing has proven to be superior to saving, and since buying stocks have proven to be the best form of investing over time, it's no mystery to why so many individuals have embraced the stock market.

There are two ways in which investors make money from the stock market. The first is via dividends payed to shareholders. When businesses are flourishing, they will often take some of the profit and award that money to its shareholders. This form of payment is referred to as a dividend and they are generally disbursed on a quarter-annual basis. Dividends are typically not substantial, however if you own enough shares of a company your dividend payment could be appreciable. The other way investors make money from the stock market is by selling a stock for more than they had originally paid for it.

After a company's IPO, their stock is then sold for whatever price an individual is willing to sell or buy the stock. The U.S. stock market only operates from 9:30am-4pm EST, and is closed on the holidays. Once the market opens, individuals could place their orders on any stock in the market, provided they have a brokerage account. A brokerage account is an arrangement between an investor and a licensed brokerage firm that allows the investor to deposit funds with the firm and place investment orders through the brokerage. Through the brokerage, two primary orders can be made, market and limit orders. Limit orders make up a queueing system in which people can propose the price that they would like to buy or sell a stock. This queueing system is known as the bid-ask spread. A very basic example of a bid-ask spread is provided below.

TOP OF BOOK		
	SHARES	PRICE
↑ ASKS	100	90.2700
	4,200	90.2600
	200	90.2500
	100	90.2400
	201	90.2200
← BIDS	200	90.2100
	272	90.2000
	26	90.1800
	100	90.1600
	300	90.1400

A market order is a buy or sell order to be executed immediately at the current market prices. So if we placed a market order to buy 5 shares we would need at least $5 * \$90.22 = \451.10 in our brokerage account to make the purchase. Now if we wanted to sell 10 shares immediately, we would place a market order to sell 10 shares and we would receive $10 * \$90.21 = \902.10 in our brokerage account. Now, there are other factors that go into the buy and sell process such as taxes from the government and fees from whatever financial firm you chose to open a brokerage account through, however these details are encouraged for the reader to explore.

Notice that after a company's IPO, the price of a stock is essentially in the hands of the public. Generally, the price at which the stocks are being bought and sold are relatively reasonable, meaning they are selling at about a company's actual worth. However now and again, a company's stock price will greatly precede or proceed its true value. This typically happens when there is a high demand for the stock, and this demand could be to buy or sell. History has proven that investors have not always behaved in the most logical manner. Since the stock market is complex, individuals are quick to listen to anyone who calls themselves an expert in the field of finance. These experts have their opinions of what they believe will happen with companies, however no one ever really knows for sure what is to come. Businesses are constantly changing, and there are too many factors that go into a business's success or failure, most of which are unpredictable. The exponential growth of technology and constant existence of competition ensure that no business is safe, and all it takes is one brilliant

innovation to be replaced. Due to this fact, speculation is common and is probably most responsible for the volatility of the market.

Despite the unpredictable price movement of stocks, many have found much success investing in the stock market. The interesting thing is that not all who've found success came about their success the same way. There are many strategies that have had their share of victories, however no one strategy has proven to be an infallible formula for success. The market is always changing and the methods in which individuals take towards investing must change with it.

One example of the stock market changing is the introduction of Exchange Traded Funds, better known as ETFs. There are different types of ETFs however we will focus on ETFs that make up a collection of stocks. ETFs have proven to be an extraordinary investment tool for two main reasons. Firstly, ETFs remove much of the risk involved in investing in stocks. Since each share of an ETF is represented by a bunch of different companies stocks, if one company's stock is to plummet, there will be very little change in the value of your ETF stock. Secondly, ETFs have made investing more affordable. A share in an ETF generally costs less than one share of any of the companies' stocks that are a part of that ETF. It should be noted that, with less risk, you are generally trading off more of the reward.

1.2 Internet Search Engines

The internet is a massive network of networks. It connects millions of computers around the globe, forming a network in which any computer can communicate with any other computer as long as they are both connected to the internet. The World Wide Web, or simply Web, is an information-sharing model that is built on top of the internet. With all the information that is out there, filtering through it all manually would take forever.

Therefore, search engines were created to do this task for you. A search engine is a program available through the internet that searches documents and files for the keywords you provide it and returns any files containing those key words. With the use of cleverly written algorithms most search engines usually return files and documents upon their importance. The most popular and arguably best search engine available is Google Search.

Google's search engine is particularly special because they track what people are searching, and they make this information available to the public via Google Trends. Google Trends is a public web facility that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world and in various languages. Access to such information could be useful to anyone who would like to do a study related towards the public's search history via google.

1.3 Research

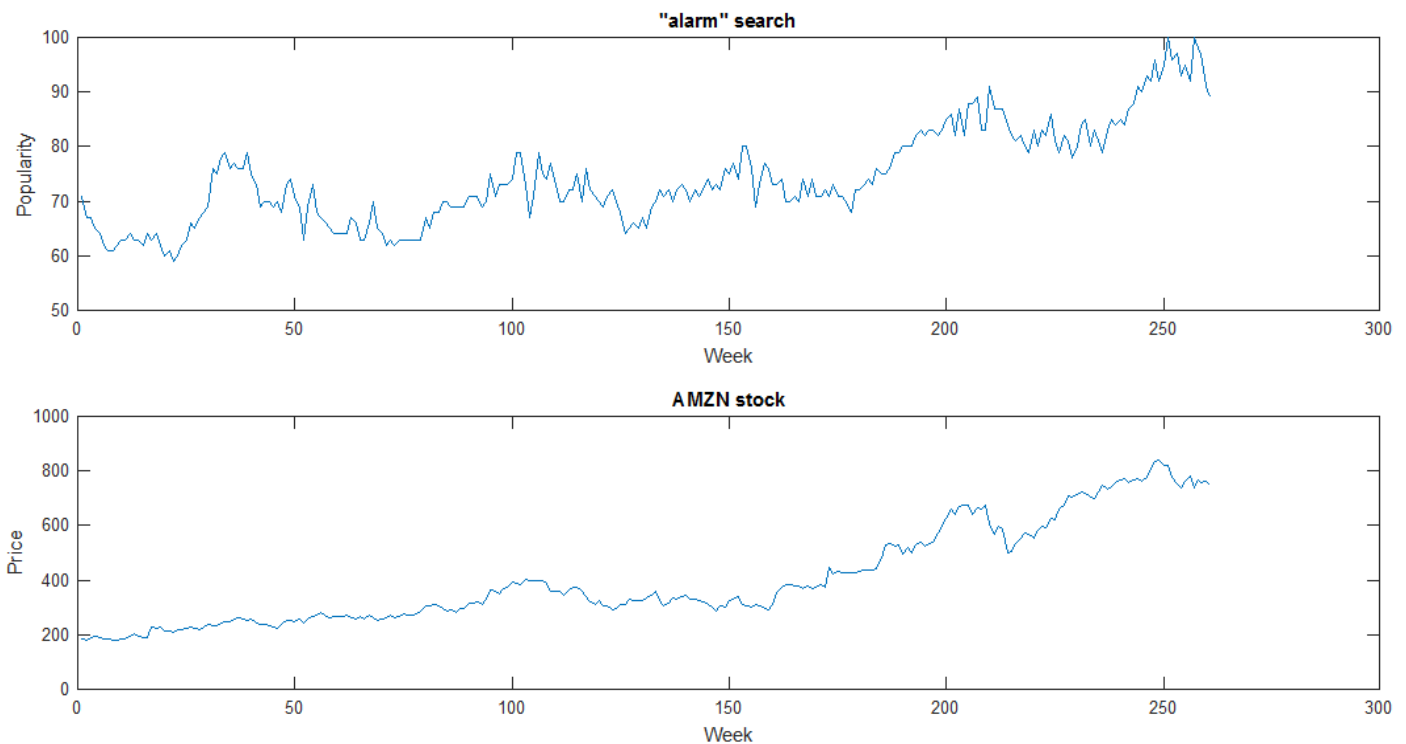
As mentioned before, many investors who have found financial success from the stock market employed different methods. Some bought and held their stocks until they needed the capital, while others purchased a stock because they believed there were short-term opportunities that they could take advantage of. In either case, all of these investors sought to sell their stock for more than they had originally bought it. Therefore having some insight into which direction a stock's price is to move would be ideal for any investor. Since stock prices have proven to be quite volatile, this is much easier said than done. Financial firms across the globe are constantly pouring resources into successfully predicting stock price movement, however most have found very little success. Less than 10% of actively managed funds actually "beat the market," meaning trying to earn an investment return greater than that of the S&P500 index, one of the most popular benchmarks of the U.S. stock market. Therefore, participating in the stock market has proven to be more of a gamble than an investment for most "experts."

Now these "experts" are generally very knowledgeable of the finance world, meaning they know the lingo and understand the theory behind finance. However, as history as shown, proficiency in business and finance has not been the answer to "intelligently" investing in the stock market. At times, the movement stock prices have been counterintuitive to the teachings of business and finance. For example, during the Dot-com bubble of the 1990s, somehow companies that had never made any revenue were pushed onto the stock exchange and were trading for extremely high values. This was when the internet first began to take off and most knew very little about it. "Internet companies" were being created left and right, most of which did not have much of a business plan. The rave about the potential of this new sector in the market attracted many investors who did not want to

miss out on “the next big thing.” Those companies made no money and soon had to file for bankruptcy, leaving investors to lose everything.

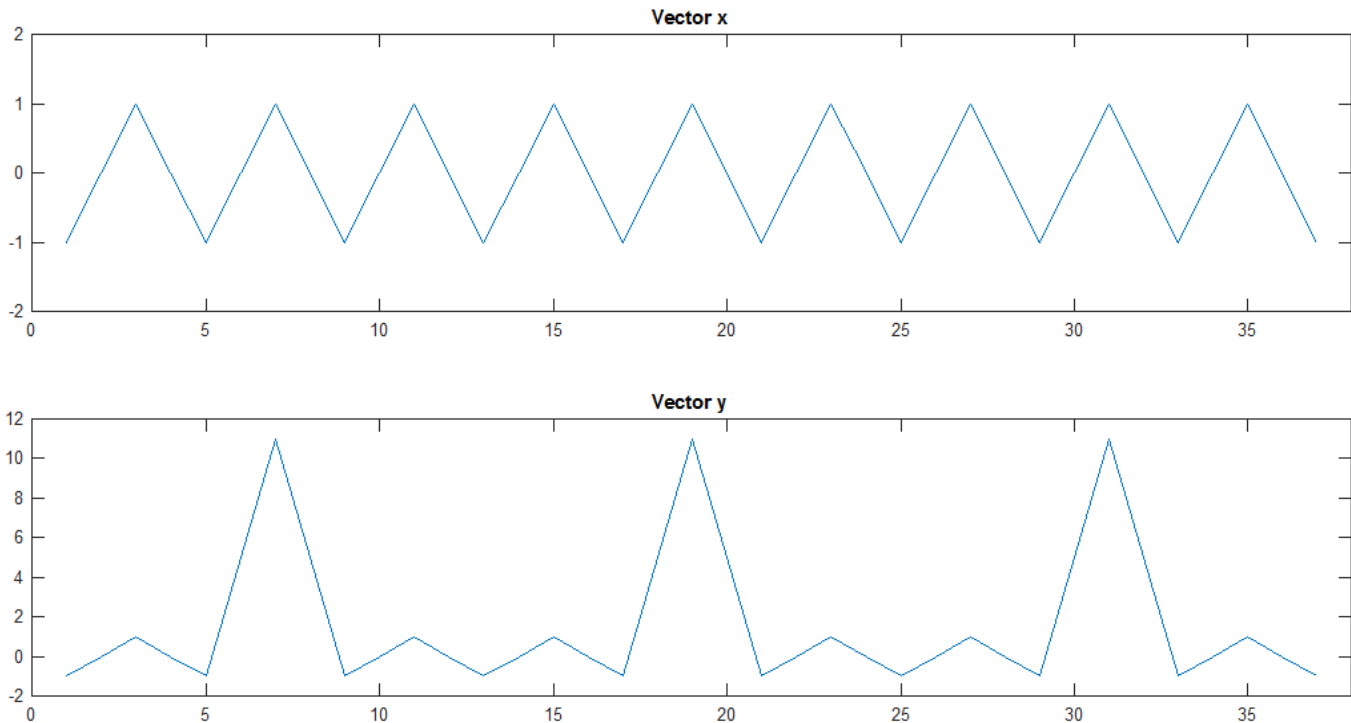
There are other examples of situations like this that go to show that competency in business and finance does not imply you can always predict which direction a stock’s price will move. This erratic behavior in the stock market has intrigued the world, and many individuals have come up with some fascinating experiments to try and predict the movement of stock prices. One interesting idea was to use news articles to predict stock price movements. An algorithm was created to search for specific words within articles about a company and provide a score on the positive or negative tone the articles had towards that company. That score was then analyzed with the company’s stock prices to determine if the articles could have possibly had any effect on its movement. After hearing about this experiment we thought about how individuals were likely finding these articles. The probable answer to this question is via a search engine. Soon after we found out that the most popular search engine, Google Search, tracks and stores its users searches through their program called Google Trends. This then motivated us to explore the relationship, if any, between Google Trends data and stock price data. If we could find such a relationship, could we then use Google Trends to inform us when to buy and sell stocks for a profit?

Once having gathered the data, we needed to use some test to measure if stock prices and popularity of searches correlated in any way. Our main goal was to find stocks and searches that when their data was plotted, the two graphs would be similar. Initially we tried to use a built in MatLab function called `corrcoef()`. This function takes in two vectors of equal dimension and spits out the Pearson correlation coefficient, say c , which is a value between -1 and 1. The closer c is to 1 or -1 the more linearly dependent the two vectors are said to be. We hoped that if we could find “good” c values, the graphs of the data would be similar. However, this is not how it turned out. Below are the graphs of the popularity of the search “alarm” and the stock prices of Amazon.com Inc, with ticker symbol “AMZN”, over a five year time period. The two had a correlation coefficient value of .8813.



As you can see, these two graphs don’t rise and fall together month to month. They only have a common overall trend. We preferred a test that would detect month to month changes, but this correlation test removes time and does not recognize localized ups and downs. Even if we removed the overall linear trends, `corrcoef()` would still be heavily influenced by long scale trends. For example, if both data sets started low, rose in the middle and ended low, `corrcoef()` would report a strong correlation regardless of month-to-month correlation. In addition

the test only considered the graphs to be similar if the overall magnitude of the peaks and valleys were proportional. The following image is an example of this.



This pair of vectors inputted into `corrcoef()` returned a value of `.5570`, which is an average result. However, we considered graphs like these to be perfect because we wanted graphs that had local extrema at the same points in time so we could use that information as buying and selling signals. Since our stock price data is in a way unbounded above, the prices could, in theory, go to infinity, however the popularity of the searches was bounded above by 100. This means that it is likely that the peaks and valleys would not be proportional as a whole, which is what the Pearson correlation coefficient is essentially calculating. These issues urged us to seek out another method in which we could test the data sets for a correlation. In our search we failed to find something that would be promising, so we decided to create our own test that would aid us in tracking down stocks and searches that would have similar graphs.

The rest of the paper is structured as follows: in Section 2, we reveal the sources in which we obtained our data and how we went about gathering it; Section 3, we discuss our correlation test; Section 4, we demonstrate the test's effectiveness.; Section 5, we present our test results comparing Google Trends with stock prices; and finally in Section 6, we disclose our conclusion on our hypothesis and close with a few final remarks.

2. Data

2.1 Stock Price Data

Yahoo! Finance is a media property that is part of Yahoo!'s network. This site provides financial news, data and commentary including stock quotes, press releases, and financial reports. One service Yahoo! Finance offers is the ability to download any company's stock price history over any length of time beyond 1962. Below is an image of the Yahoo! Finance page after searching for Coca-Cola Co stock, using its ticker symbol KO. A ticker symbol is an abbreviation used to uniquely identify publicly traded shares of a particular stock on a particular stock market.

The screenshot shows the Yahoo Finance website for KO (The Coca-Cola Company). The search bar at the top contains 'KO' and a red arrow labeled '1' points to the search button. Below the search bar, a red arrow labeled '2' points to the 'Historical Data' tab. In the 'Historical Data' section, four green arrows point to the 'Time Period' dropdown (labeled '2'), the 'Show' dropdown (labeled '1'), the 'Frequency' dropdown (labeled '3'), and the 'Apply' button (labeled '4'). A blue arrow labeled '1' points to the 'Download Data' link. A context menu is open over the 'Download Data' link, with a blue arrow labeled '2' pointing to the 'Copy link address' option.

Date	Open	High	Low	Close	Adj Close*	Volume
Dec 25, 2016	41.56	41.84	41.35	41.46	41.46	34,332,700

The red arrows refer to the steps in which you would take to get to this page:

1. Type in the company's ticker symbol and click "search."
2. Click on the "Historical Data" tab.

Once having followed these steps your screen should look fairly similar to the above picture. The green arrows point to the parameter options you can set for your data:

1. "Show" is a list of data options that you can choose from; Historical Prices, Dividend Payments, and Stock Splits.
2. "Time Period" is a parameter that allows you to gather data dating anywhere from January 2, 1962 to the present day.
3. "Frequency" is a parameter in which the data will be gathered on a daily, weekly, or monthly basis.

After choosing the data you would like to retrieve with specific parameters you would click on the "apply" button that the fourth green arrow is pointing to. The page reloads and all the data is listed by date in descending order. You can then download this data to a csv file by clicking the "download data" link that the first blue arrow points to. A csv file would then be downloaded and look like the image below.

KO (1) - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

Clipboard Font Alignment Number

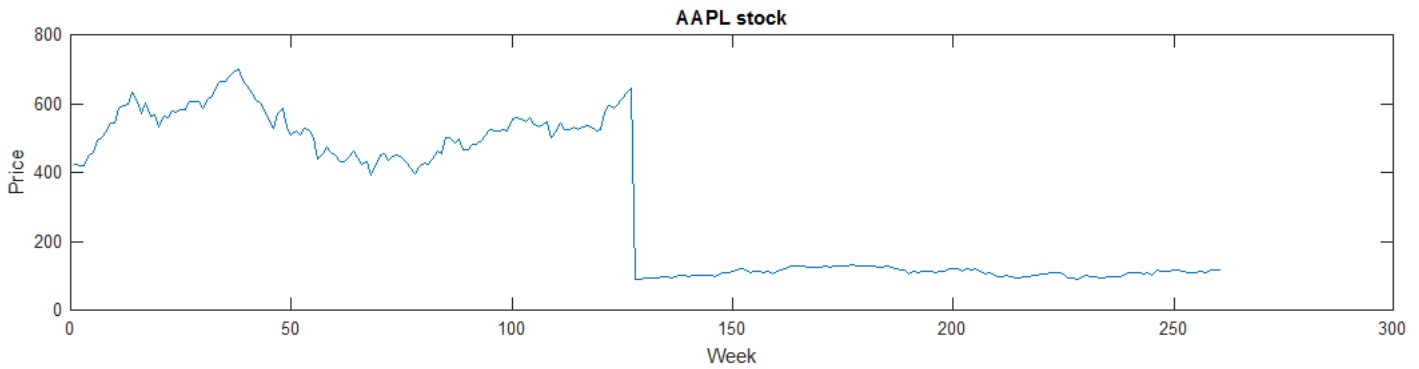
A1 : X ✓ fx Date

	A	B	C	D	E	F	G	H	I	J
1	Date	Open	High	Low	Close	Adj Close	Volume			
2	1/2/2012	35.075	35.355	34.31	68.93	34.465	57813000			
3	1/9/2012	34.5	34.76	33.285	66.99	33.495	82144600			
4	1/16/2012	33.87	34.06	33.59	68.09	34.045	61562600			
5	1/23/2012	34	34.31	33.615	67.44	33.72	74576400			
6	1/30/2012	33.645	34.245	33.515	68.08	34.04	68755800			
7	2/6/2012	33.92	34.725	33.71	67.94	33.97	82590400			
8	2/13/2012	34.21	34.625	34.045	69.05	34.525	65057600			
9	2/20/2012	34.44	34.75	34.25	69	34.5	50796600			
10	2/27/2012	34.345	34.99	34.26	69.18	34.59	71889600			
11	3/5/2012	34.51	34.87	34.25	69.51	34.755	1.05E+08			
12	3/12/2012	34.74	35.31	34.67	70.16	35.08	90019800			
13	3/19/2012	35.105	35.8	34.965	71.49	35.745	71163000			
14	3/26/2012	35.87	37.195	35.795	74.01	37.005	1.17E+08			
15	4/2/2012	36.915	37.1	36.535	73.47	36.735	51587000			
16	4/9/2012	36.48	36.555	35.91	71.94	35.97	60290200			
17	4/16/2012	36.11	37.24	36.07	74.13	37.065	72734000			
18	4/23/2012	36.99	38.91	36.665	76.63	38.315	81241200			
19	4/30/2012	38.21	38.82	37.965	77	38.5	85690000			
20	5/7/2012	38.485	38.87	38.245	77.47	38.735	56861000			

KO (1) +

READY

The yellow column represents the closing price of the stock for that day; this was the only data we needed from this table. When we first began testing our data and examining the graphs we had noticed that some of the graphs for stocks had some drastic pitfalls. An example of this is provided below.



After a brief investigation we had realized that we weren't taking stock splits into consideration. A stock split is a decision made by a company to increase its total number of shares by issuing more shares to its current shareholders. One reason for doing this is that the company's stock price has increased to levels that are either too high or beyond the price levels of similar companies in their sector. When a company decides to split its stock the shareholders are not negatively affected, they retain the same portion of ownership and the amount of money invested in the company also remains the same. We readjusted our data manually by looking up the company's stock split history and multiplied the data from the stock split day forward by the split ratio. The images below demonstrates this process.

1

	Date	Closing Price	Closing Price Adjusted
126	5/25/2014	633	
127	6/1/2014	645.57	
128	6/8/2014	638.96	
129	6/15/2014	90.91	636.37
130	6/22/2014	91.98	643.86
131	6/29/2014	94.03	658.21
132	7/6/2014	95.22	666.54
133	7/13/2014	94.43	661.01
134	7/20/2014	97.67	683.69
135	7/27/2014	96.13	672.91
136	8/3/2014	94.74	663.18

2

AAPL Split History Table	
Date	Ratio
06/16/1987	2 for 1
06/21/2000	2 for 1
02/28/2005	2 for 1
06/09/2014	7 for 1

3

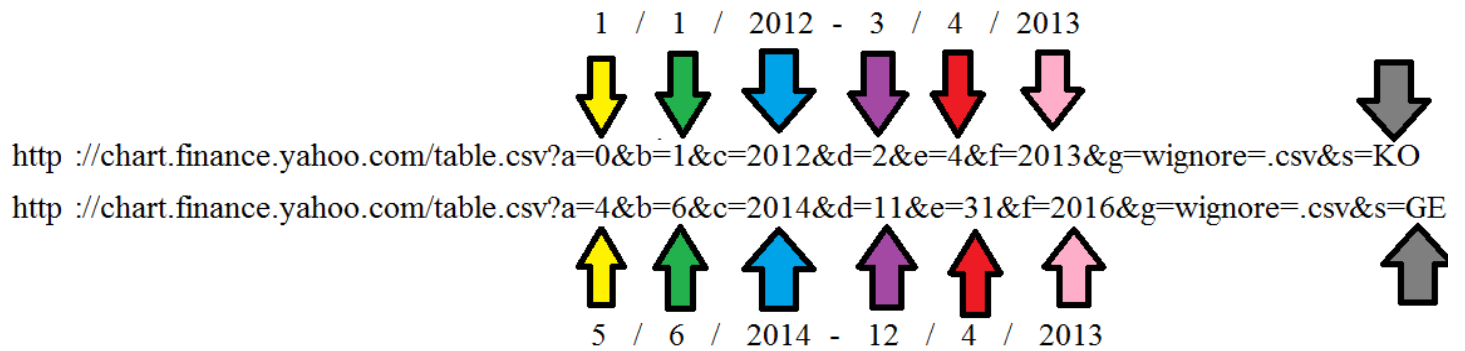
Date	Open	High	Low	Close
Jun 09, 2014				7/1 Stock Split

Image (1) displays the stock price history of AAPL around the time of its 7:1 stock split in 2014. We originally used the data tables, image (2), from stocksplithistory.com to identify stock splits of each company during our five year time period. Since the stock split was 7 for 1 we multiplied all data beyond the stock split data, 6/9/2014, by 7. Later we realized that Yahoo! Finance also provides stock split history, image (3), which we could utilize in the future to automate this process.

An alternative to this whole process is to take the "Adj Close" column from the csv file that is highlighted in blue from our previous csv file image. We learned after putting in all this work that the "Adj Close" column takes into account stock splits, however the prices become relative to the latest stock split in the data. This means that if I took the stock price data of the last two weeks, and there was a 2:1 stock split last week, the prices of the week prior would be divided by two as well.

Our research required us to gather historical price data on many stocks, and going through this process proved to be very timely. So we sought some way to automate this process. We noticed that the "download data" link allowed us to copy its link address (Blue arrow from Yahoo! Finance image). Here is a picture of two link

address, the first coming from searching KO stock with a time period of 1/1/2012-3/4/2013 on a weekly frequency and the other coming from searching GE stock with a time period of 5/6/2014-12/31/2016 on a weekly frequency.



Notice the two links are the same except for the values at the arrows. Using this discovery, we could exploit this pattern and use the MatLab function `urlread()` to automate retrieving this data. Below is a portion of our code that helped us accomplish this task.

```

1 function x = getstock(symbol, startdate, enddate)
2 % INPUT: symbol = stock symbol
3 % start = starting date. Example '01-01-12'
4 % Date should be a string with 2 digits for each thing
5 % end = ending date. Example '12-31-16'
6 % Date should be a string with 2 digits for each thing
7 % OUTPUT: queries internet and returns stock prices from 1/1/12 and
8 % 12/31/17 with 1 week resolution. Modify the URL below if
9 % you wish for something else
10 %
11 x=zeros(261,1);
12 if size(startdate,2)~=8 || size(enddate,2)~=8
13 disp('ERROR: Start and/or end date is formatted wrong');
14 return;
15 end
16 a = num2str(str2num(startdate(1:2))-1);
17 b = startdate(4:5);
18 c = ['20' startdate(7:8)];
19 d = num2str(str2num(enddate(1:2))-1);
20 e = enddate(4:5);
21 f = ['20' enddate(7:8)];
22 url=['http://chart.finance.yahoo.com/table.csv?a=' a '&b=' b '&c=' c '&d=' d '&e=' e '&f=' f];
23 url=[url '&g=w&ignore=.csv&s=' symbol];
24 %%% in newer versions of Matlab, you can use webread instead of urlread %%%
25 try
26 y=urlread(url);
27 catch
28 fprintf('ERROR: Yahoo finance doesnt have info on stock %s\n',symbol);
29 return;

```

Connecting to the internet and running `urlread()` via Matlab we could access the Yahoo! Finance site using the following URL:

<http://chart.finance.yahoo.com/table.csv?a=0&b=1&c=12&d=11&e=31&f=16&g=w&ignore=.csv&s=SYMBOL>

Where `s` is the stock symbol, `g` is frequency, `a/b/c` month/day/year (example 00/01/12) is the start date with 'a' equal to the month minus 1, `d/e/f` is the end date with 'd' equal to month minus 1. This URL returns a comma separated value file and then we would extract the closing stock price column.

Using this function we created a program that would prompt the user to type a company's ticker symbol and would then retrieve the stock price history of that company over a five year time period, 1/1/12 to 12/31/16, on a weekly basis. The program would continue to ask the user for ticker symbols until a blank string was entered, which would then end the program. As the ticker symbols were entered, our program would merge all stocks with their data into a single excel document.

In our research we gathered data from over 200 stocks; 48 stocks from Vanguard's ETF selection, 75 from Standard and Poor's S&P400 Midcap index, and 103 Standard and Poor's S&P500 index. Below is a list of all the stocks ticker symbols from each category.

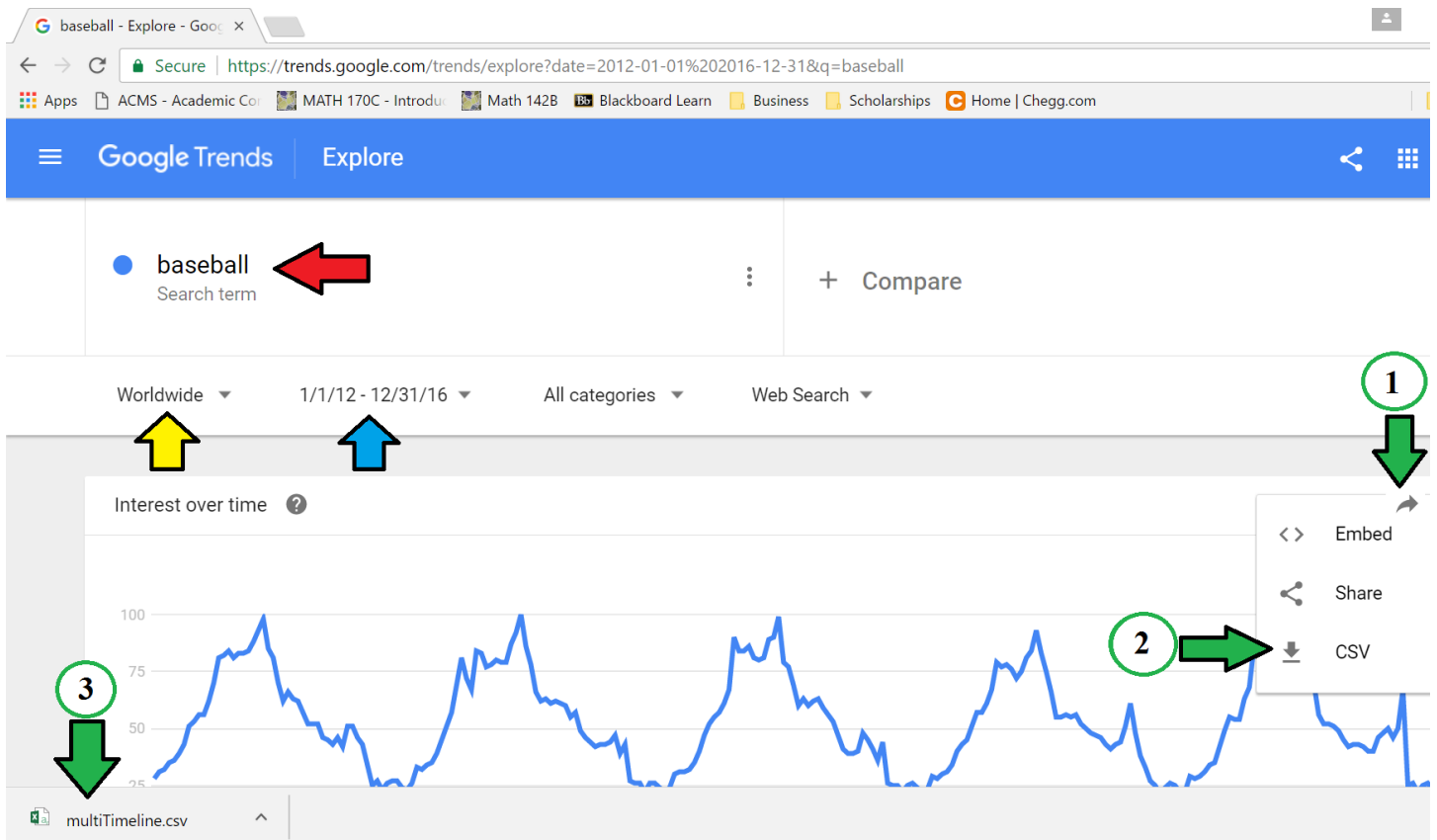
VANGUARD									
EDV	BIV	VGIT	BLV	VGLT	VMBS	BSV	VGSH	BND	VCIT
VCLT	VCSH	VIG	VUG	VYM	VV	MGC	MGV	VOO	VTI
VTV	VXF	VO	VOT	VOE	VB	VBK	VBR	VEU	VSS
VEA	VWO	VGK	VPL	VNQI	VXUS	VT	VCR	VDC	VDE
VFH	VHT	VIS	VGT	VAW	VNQ	VOX	VPU		

S&P400 Midcap									
AAN	ABMD	AEO	AFG	AKRX	AMCX	ARW	ASH	BEAV	BIG
BKH	BOH	BWLD	CAA	CAKE	CASY	CBT	CEB	CHDN	CHS
CLGX	CNK	CR	CRI	CRS	CRUS	CSC	CTB	CXW	DDD
DECK	DEI	DF	DNKN	DO	DPZ	DV	EAT	EDR	FAF
FICO	FII	FSLR	GEO	HII	HSNI	IDTI	INGR	INT	ISCA
ITT	JACK	JBLU	JCP	KBH	KEX	LPNT	LPX	LSI	MCY
MD	MSA	MSCC	NE	NEU	NJR	ODP	OGE	PAY	PNRA
SM	TDS	TR	WEN	ZBRA					

S&P500									
AAPL	MSFT	AMZN	XOM	JNJ	JPM	GE	T	GOOGL	WFC
GOOG	BAC	PG	CVX	PFE	VZ	HD	CMCSA	PM	MRK
INTC	CSCO	V	C	DIS	KO	PEP	UNH	IBM	MO
ORCL	AMGN	MMM	MDT	SLB	WMT	MCD	MA	BA	CELG
HON	BMJ	AVGO	GILD	PCLN	UNP	SBUX	GS	UTX	QCOM
USB	LLY	TXN	CVS	AGN	ABT	TWX	NKE	ACN	COST
UPS	LOW	CHTR	WBA	DOW	NFLX	MS	FDX	GM	BLK
F	YHOO	EBAY	DAL	FOXA	GIS	TGT	MAR	INTU	ROST
SY	CBS	DLTR	K	TSN	RCL	HSY	EXPE	FOX	DISH
BBY	ALK	HAS	TIF	FL	ETFC	WFM	M	NDAQ	JEC
SPLS	HRB	UAA							

2.2 Google Trends Data

As previously mentioned, Google Trends is a database that shows how often a particular search-term is entered relative to the total search-volume across various regions of the world. Google Trends data is a random sample of Google search data. Therefore, it is not a direct reflection of what is being searched, however it is a pretty good approximation. Data that is excluded are searches made by very few people and repeated searches from the same person over a short period of time. Google Trends also filters out apostrophes and other special characters in a search. The samplings are taken from a specified geographical region over a designated period of time determined by the user. The following is an image of the Google Trends site after searching "baseball" over the five year time period 1/1/12-12/31/16.



The red arrow points to the search term. The yellow and blue arrows point to parameters in which you could adjust to observe that search terms popularity over a specified region and time. The graph just under those arrows is a plot of the popularity of the search over the given time period. Notice that the max for the data is 100. This is because each data point is divided by the total searches of the geography and time range it represents and those resulting numbers are scaled on a range of 0 to 100 based on the proportionality of its popularity over that time period. The data could then be collected by following the green arrows:

1. Click the gray arrow in the right direction and the box with the following three options will appear.
2. Click the “download CSV” option.
3. A csv file with the data will be downloaded to the default download folder for your web browser.

Unlike the data from Yahoo! Finance, Google Trends makes it difficult to automate the download of data with scripts. Therefore, we downloaded the search data’s CSV files into a targeted folder, and then ran a program that read all the files, extracted the data, and merged that data into a single CSV file. Below is a portion of our program that combined the data into one file.

```

1  function gettrends(dirname)
2  % INPUT: dirname = name of folder containing CSV from Google Trends
3  %       Download CSV files from following URL into a folder
4  %       https://trends.google.com/trends/explore?date=2012-01-01%202016-12-31&geo=US&q=WORD
5  %       Change WORD at the end of the URL to the word you're interested in
6  % OUTPUT: Reads all the CSV files and writes all info into one file
7  %         named trends.csv
8  %
9  -   addpath('mFiles');
10 -   F=dir(dirname);
11 -   n=size(F,1);
12 -   if n==0
13 -       fprintf('ERROR: cannot find folder %s\n',dirname);
14 -       return;
15 -   end
16 -   c=0;
17 -   for i=1:n
18 -       m=max(size(F(i).name));
19 -       if m>=46&&strcmp('.csv',F(i).name(m-3:m))
20 -           c=c+1;
21 -           fil=fopen([dirname '/' F(i).name] , 'r');
22 -           str=fscanf(fil,'%100c');
23 -           fclose(fil);
24 -           j=1;

```

For our research we gathered over 800 searches. Some of the searches were words or phrases used in the financial world such as “bankruptcy” and “stock split.” Others were random words in which we used a random word generator from the web. The last set of searches were the ticker symbols of the S&P500 stocks. All Google Trends data was from the 5 year period, 1/1/12-1/31/16, on a weekly basis.


2.3 Remarks about gathering the data

The following subsections provide some last minute remarks about our data gathering methods. We hope this information could be useful to the reader in their own data gathering processes.

2.3.1 Stock Price Data


Once our research was all said and done we noticed that Yahoo! Finance had updated its URL link for downloading the data. Our program no longer worked with the URL we were using. Below is an image of the old URL compared to the new URL using our KO example from before.

OLD URL




<http://chart.finance.yahoo.com/table.csv?a=0&b=1&c=2012&d=2&e=4&f=2013&g=wignore=.csv&s=K̂O>

NEW URL



<https://query1.finance.yahoo.com/v7/finance/download/KO?period1=1325404800&period2=1362384000&interval=1wk&events=history&crumb=I868n4zuWrc>



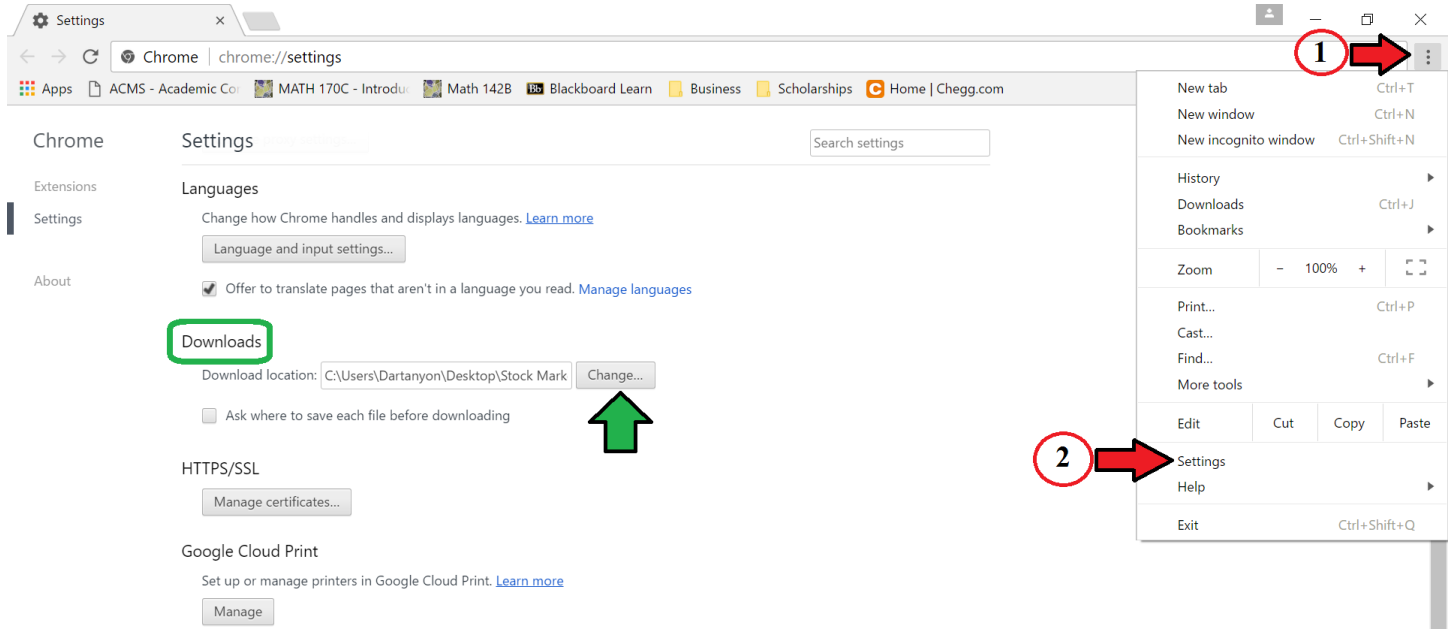
<https://query1.finance.yahoo.com/v7/finance/download/GE?period1=1399359600&period2=1483171200&interval=1wk&events=history&crumb=I868n4zuWrc>

As you can see the old URL and new URL are very different. However, the new URLs seem to be fairly similar with differing only at the blue underlined section. This section refers to the time period parameter you can set,

however, it is difficult to identify how they have coded the dates. After doing some research, we discovered that this was not the first time that Yahoo! Finance has updated its URL links. It is likely that Yahoo! Finance does not want computers gathering massive amounts of this data at once because they need actual users to visit their site for the advertisements on the page. In spite of the update impeding our program from running with the original URL, we were able to make a small tweak to the old URL that would then work. Notice the red arrows in the above image; the new URLs have an “s” at the end of “http” while the old URL does not. By adding that “s” to our old URL, our program worked just as it did before. We are unsure of how long the URL will work, however we can say that for now we have a sort of “back door” for gathering this data off of Yahoo! Finance. Although this “back door” exists for now, we don’t expect it to be there much longer. Therefore, data automation via this site may still be possible, however it will likely require more cleverness in constructing a URL that will actually work.

2.3.2 Google Search Data

As mentioned before, when downloading the Google Trends data, it would automatically be downloaded to whatever the default download location was for your browser. Instead of trying to access this default location through a series of directory commands in MatLab, we changed our web browser’s download location to a folder located in the same folder as our gettrends() program. The web browser we used was Google Chrome and below is an image with directions on how to achieve this task.



1. Click the button of the first red arrow
2. Select the “Settings” option (Second Red Arrow)
3. Scroll down until you see a blue link titled “Advanced settings” and click on it
4. Scroll down until you see “Downloads” (Green Box)
5. Click the “Change” button (Green Arrow)
6. Select your new download location

3. Correlation Test

Since both sets of data are a series of points indexed in time order, we will refer to, from this point forward, each subset of data, i.e. a single stock’s price history or a search’s “popularity” over time, as a time series.

3.1 Description

When we originally began testing our data, we used the MatLab function `corrcoef()`. This test was not very effective in providing us with time series pairs that would likely have matching graphs because it essentially did

not recognize ups and downs. Therefore we created our own test that simply counted matching ups and downs of the two time series. For example, when one time series increases in value, does the other increase? When one decreases does the other decrease? Our hope was that time series with high percentages of matching ups and downs would imply matching graphs.

3.1.1 Mathematical Description of our test:

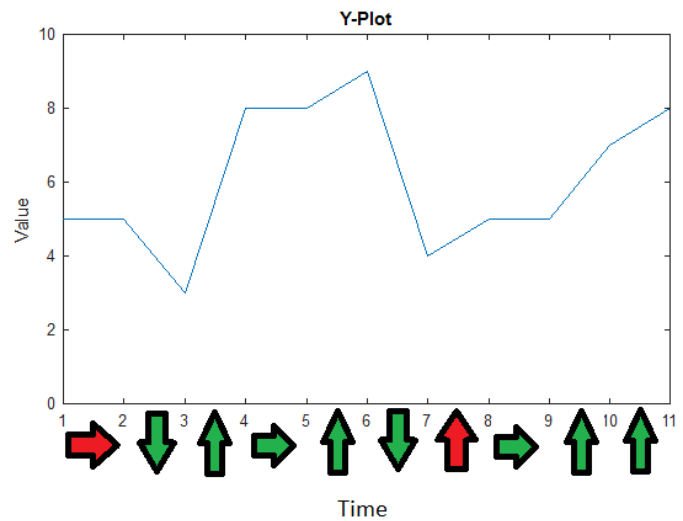
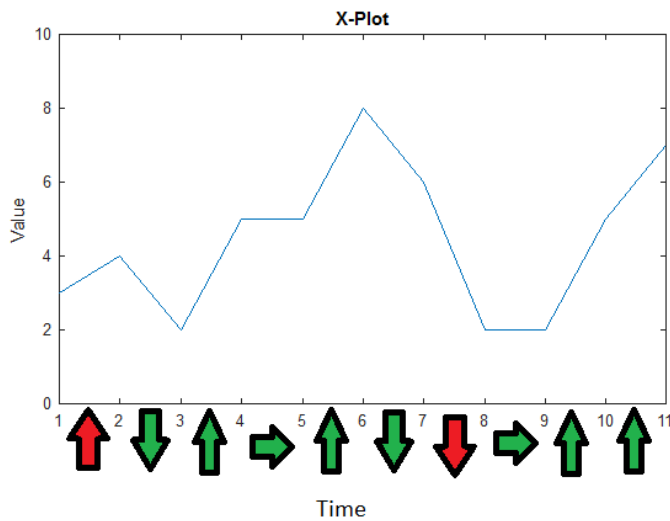
Let $\{x_k^{(j)}\}$ be our original time series for $1 \leq k \leq n$ and $j = 1, 2$. The number k in $1 \leq k \leq n$, denotes 5 years of weekly data with $n = 261$ and the number $j = 1, 2$ denotes the two sequences we are comparing. Our new

time series is $s_k^{(j)}(x^{(j)}) = \begin{cases} 1 & \text{if } x_{k+1}^{(j)} > x_k^{(j)} \\ 0 & \text{if } x_{k+1}^{(j)} = x_k^{(j)} \\ -1 & \text{if } x_{k+1}^{(j)} < x_k^{(j)} \end{cases}$ for $1 \leq k \leq n - 1$. Our test denoted by T returns a number

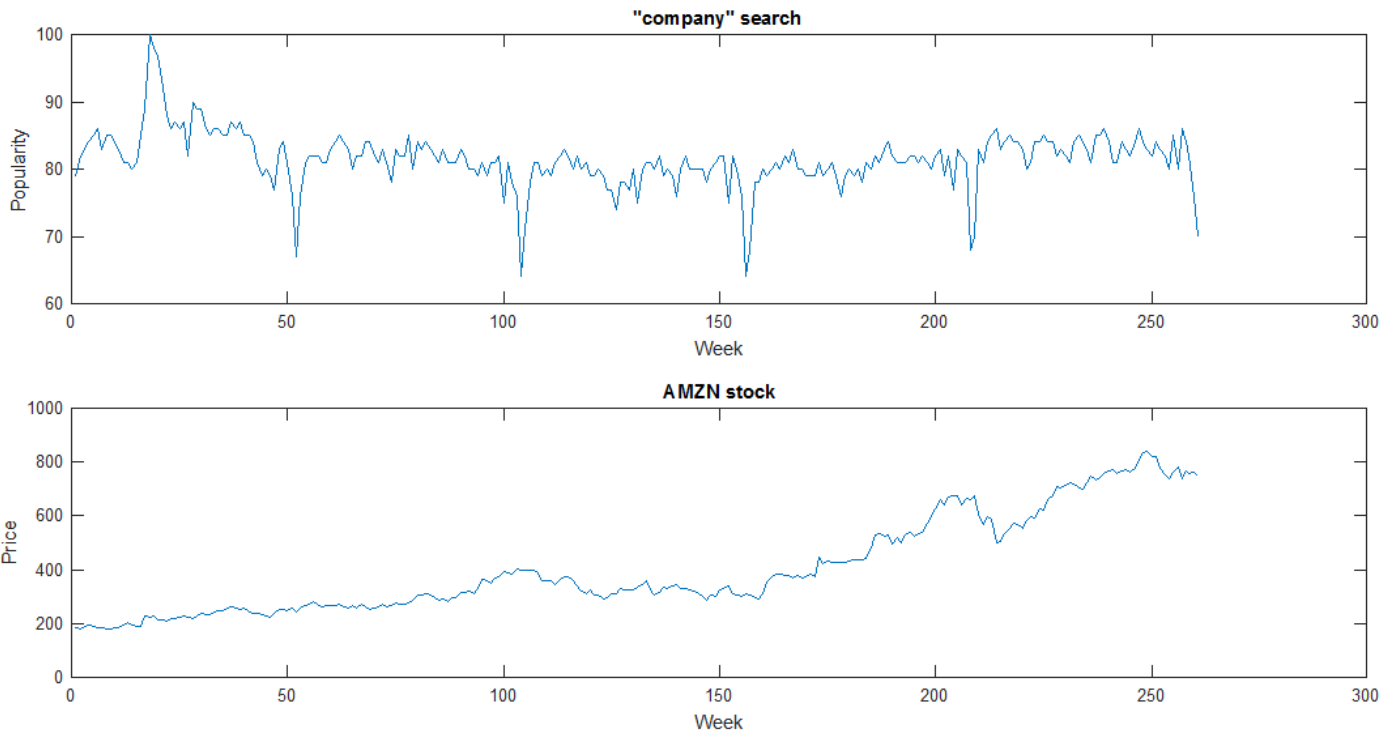
between 0 and 1 inclusive, $T(x^{(1)}, x^{(2)}, d) \in [0, 1]$. The number d is a time shift. The test converts the sequences $\{x^{(1)}\}$ and $\{x^{(2)}\}$ into sequences $\{s^{(1)}\}$ and $\{s^{(2)}\}$ as describe above and then compares them. When comparing the two sequences, we compare element $s_k^{(1)}$ with element $s_{k+d}^{(2)}$ for $1 \leq k \leq n - 1 - d$. The test T reports what percentage of elements from sequence $\{s^{(1)}\}$ are the same as $\{s^{(2)}\}$ with time shift d . If $d < 0$, we compare element $s_{k-d}^{(1)}$ with element $s_{k+d}^{(2)}$ for $1 \leq k \leq n - 1 + d$.

Below is a plot of two time series:

1. $x = [3 \ 4 \ 2 \ 5 \ 5 \ 8 \ 6 \ 2 \ 2 \ 5 \ 7]$
2. $y = [5 \ 5 \ 3 \ 8 \ 8 \ 9 \ 4 \ 5 \ 5 \ 7 \ 8]$



The green and red arrows represent matching and mismatching directions of the graph respectively. Since there are 11 data points there are 10 comparisons, and 8 of those 10 comparisons matched which implies an 80% similarity according to our test. The two graphs somewhat look similar, however the jaggedness of the graphs makes it more difficult to see this. With our data, this issue was even more so the case as you can see below.

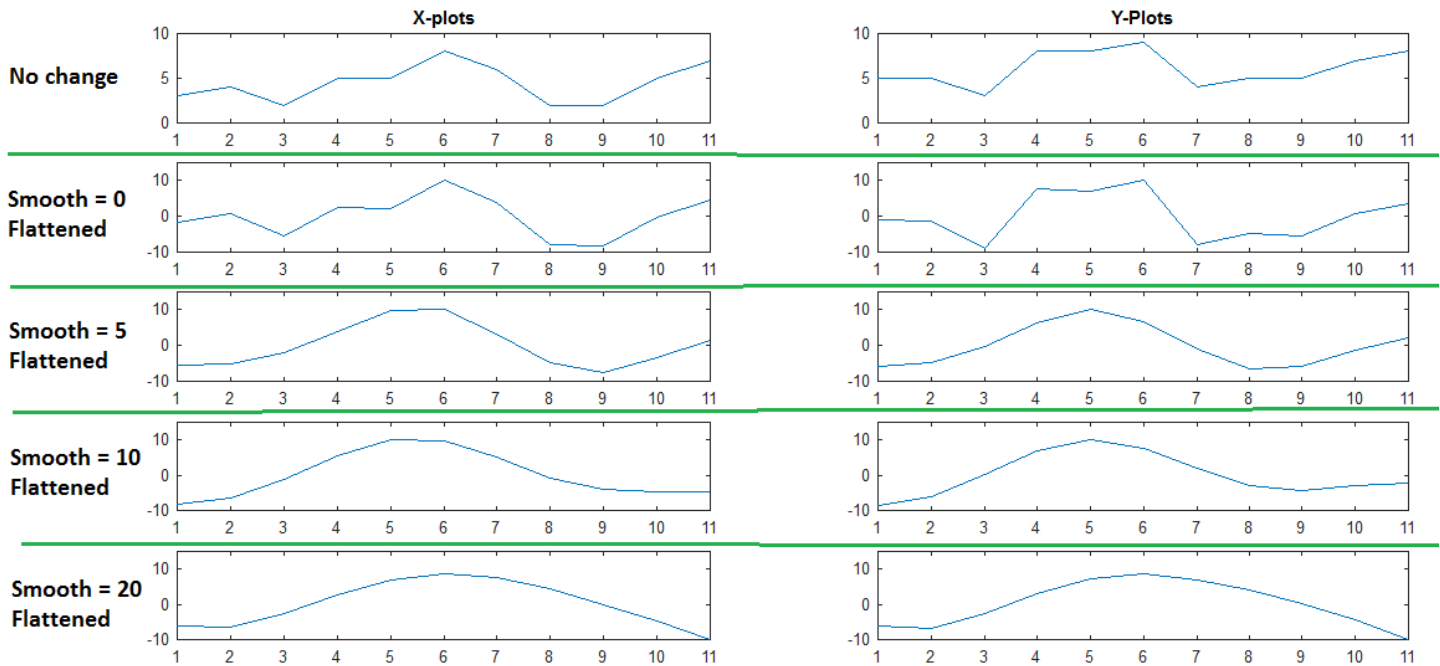


To solve our problem with the jaggedness of the graphs, we averaged the data points with their neighbors. This removed the “noise” and allowed the overall patterns to stand out. In addition to our jaggedness problem, the stocks data generally had an upward trend. This trend is likely due to inflation; with the overall cost of goods rising, a company’s value would, in effect, rise as well. We tackled this issue by removing the linear trend of both data sets. Although the search data did not have much of a trend, we applied this technique to that data as well to stay consistent.

3.1.2 Mathematical Description of smooth and flatten technique:

Let $\{x^{(1)}\}$ and $\{x^{(2)}\}$ be the original sequences. The preconditioning function A determines the least squares regression line of $\{x^{(j)}\}$ and subtracts it from the sequence. The preconditioning function B smooths the sequence by averaging each element with its neighbors, $\bar{x}_k^{(j)} = .25 \cdot x_{k-1}^{(j)} + .5 \cdot x_k^{(j)} + .25 \cdot x_{k+1}^{(j)}$. Then $\bar{\bar{x}}^{(j,n)} = B(A(x^{(j)}), n)$ where n is the number of smoothing steps. We then perform the test on the preconditioned sequences $\{\bar{\bar{x}}^{(1,n)}\}$ and $\{\bar{\bar{x}}^{(2,n)}\}$ as in $T(\bar{\bar{x}}^{(1,n)}, \bar{\bar{x}}^{(2,n)}, d) \in [0,1]$. If we wish to test for negative correlation, then we perform the test $T(-1 \cdot \bar{\bar{x}}^{(1,n)}, \bar{\bar{x}}^{(2,n)}, d) \in [0,1]$.

Below is a visual of how the smoothing and flattening process works using our previous example with the vectors x and y .



As you can see the graphs started looking much more similar after five iterations of smoothing. Applying these two conditioning functions to the data allowed us to capture the clearer patterns of the data and aided our efforts in identifying sets that are behaving similarly.

The following two images are portions of our smoothing and flattening functions respectively. In both cases we rescaled the y-axis to values between 10 and -10 to make it easier to compare the graphs of the two data sets.

```

1  function z = smooth2(x,k)
2  % INPUT: a vector x and number of smoothing iterations k
3  % OUTPUT: a vector z that fluculates than input x.
4  %         and then scaled so max(abs(z))=10
5  %
6  if k<=0
7      z=x;
8      return;
9  end
10 m=size(x);
11 n=m(1);
12 if (m(1)==1)
13     x=x';
14     n=m(2);
15 end
16 D = toeplitz([0.5 0.25 zeros(1, n-2)]);
17 D(1,1)=0.75;
18 D(n,n)=0.75;
19 z=D*x;
20 for i=2:k
21     z=D*z;
22 end
23 d=max(abs(z));
24 z=10*z/d;
25 end
26

```

```

1 function z = flatten2(y,s)
2 % INPUT: y is a vector of any length
3 %       s=1 for rescale to max(abs)=10. s=0 don't rescale
4 % OUTPUT: a vector z equal to y minus the linear regression line
5 %         and then optionally scaled so max(abs(z))=10
6 %
7     m=size(y);
8     n=m(1);
9     if (m(1)==1)
10        y=y';
11        n=m(2);
12    end
13    x=1:n;
14    X=[x' ones(n,1)];
15    c=X\y;
16    z=zeros(n,1);
17    for i=1:n
18        z(i)=y(i)-c(1)*i-c(2);
19    end
20    if (s==1)
21        d=max(abs(z));
22        z=10*z/d;
23    end
24 end
25

```

3.2 Statistical Significance

Statistical significance is the likelihood that a relationship between two or more variables is caused by something other than random chance. When developing this test we had to consider the possibility of a random search and stock having a high correlation value according to our test. The similarity of our two sequences is a binomial distribution. A binomial distribution is the probability of a specific number of successes or failures in an experiment that is surveyed multiple times. Since we were simply answering yes or no to whether the two data sets were similar in sign for the difference of their points on an interval our probability value is .5. Our test is essentially equivalent to flipping two coins n times and measuring the number of times that they land on the same side. Using the binomial distribution model we could calculate the probability of our results randomly occurring. The less probable a result is, the more statistically significant it is.

3.2.1 Mathematical explanation of statistical significance of our test (without smoothing):

If the sequences $\{x^{(1)}\}$ and $\{x^{(2)}\}$ contain random data with $x_{k+1}^{(j)} \neq x_k^{(j)}$ then the similarity of sequences $\{s^{(1)}\}$ and $\{s^{(2)}\}$ is a binomial distribution with $p = .5$. Therefore, we expect $T(x^{(1)}, x^{(2)}, d) = .5$ with standard deviation $\sigma = \frac{1}{n-1-|d|} \sqrt{.25(n-1-|d|)}$. For the case of $n = 261$ and $d = 0$, the standard deviation is $\sigma = .031$. Therefore given two random sequences $\{x^{(1)}\}$ and $\{x^{(2)}\}$ the probability of $T(x^{(1)}, x^{(2)}, 0) \geq .5 + z\sigma$ is $p = 0.05, 0.025, 0.01, 0.005, 0.001, \text{ and } 0.0005$ for $z = 1.645, 1.960, 2.327, 2.576, 3.091, 3.291$ respectively. For example, there is only a 1% chance that $T(x^{(1)}, x^{(2)}, 0) \geq .5 + 2.327\sigma = 0.5721$.

We confirmed all of these values by generating 1,000 random fake stock price data series and empirically calculating these values. Additionally we confirmed that actual stock prices appear to be random data because

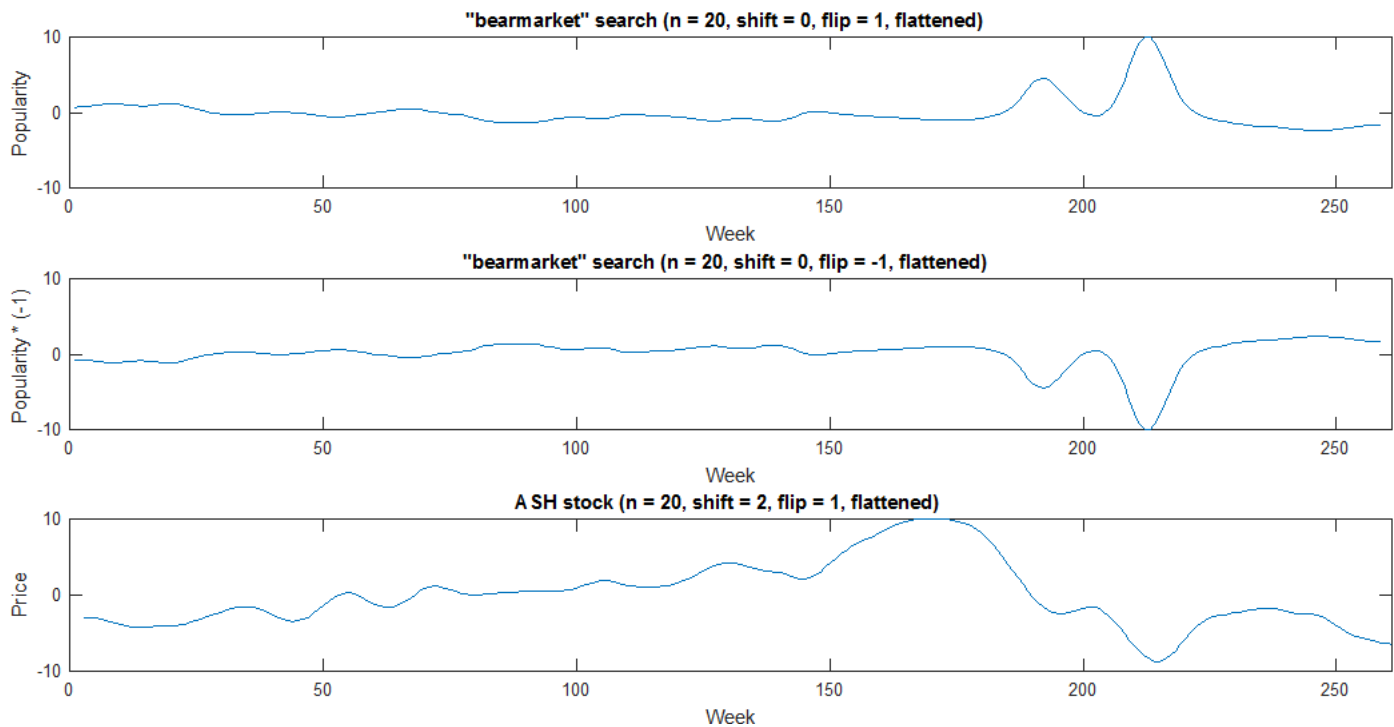
we compared 100 real stocks against themselves and found the mean of $T = 0.5$ and the standard deviation = 0.03.

3.2.2 Mathematical explanation of statistical significance of our test (with smoothing):

In order to determine the statistical significance when using preconditioned sequences $\{\bar{x}^{(1,n)}\}$ and $\{\bar{x}^{(2,n)}\}$, we calculated this empirically using 1,000 random fake stock price data series. In this case, we found that for each n , the mean of $T(\bar{x}^{(1,n)}, \bar{x}^{(2,n)}, d) = .5$ and for $n = 5, 10, 20$ the standard deviation is $\sigma = .05170, .06000, .07083$ respectively. Therefore for the case of $n = 5, 10, 20$, there is only a 1% chance that $T(\bar{x}^{(1,n)}, \bar{x}^{(2,n)}, 0)$ is greater than .6203, .6396, .6648 respectively. When the value d is changed, $-4 \leq d \leq 4$, the above values are nearly the same.

3.3 “Visual Test”

After running our test we noticed that pairs returning correlation values less than .75, meaning the two data sets only match 75% of the time in ups and downs, had graphs that were nowhere near similar. Also, some values above .75 had graphs that weren’t as similar as we would have thought. The following is an example of this, with the search word = “bearmarket”, stock = ASH (Ashland Global Holdings), with the number of smoothing iterations = 20, flip = -1, and shift = 2.



Since we were smoothing the data, we had to be cautious of results like this. Although our test calculated these two to have a correlation value of .7364 and it’s likely that the local extrema from weeks 3 to approximately 180 are the same, we would not accept these results to be sufficient enough in possibly predicting future stock price movement. Results like this occur because after enough smoothing iterations, we will eventually get a line. Hence in theory, after an infinite amount of smoothing and a removal of the linear trend of the data sets, they will be exactly the same, horizontal lines. It’s important to clarify that although we originally stated that we weren’t worried about the overall proportionality of the peaks and valleys, we still wanted these peaks and valleys to be perceivable.

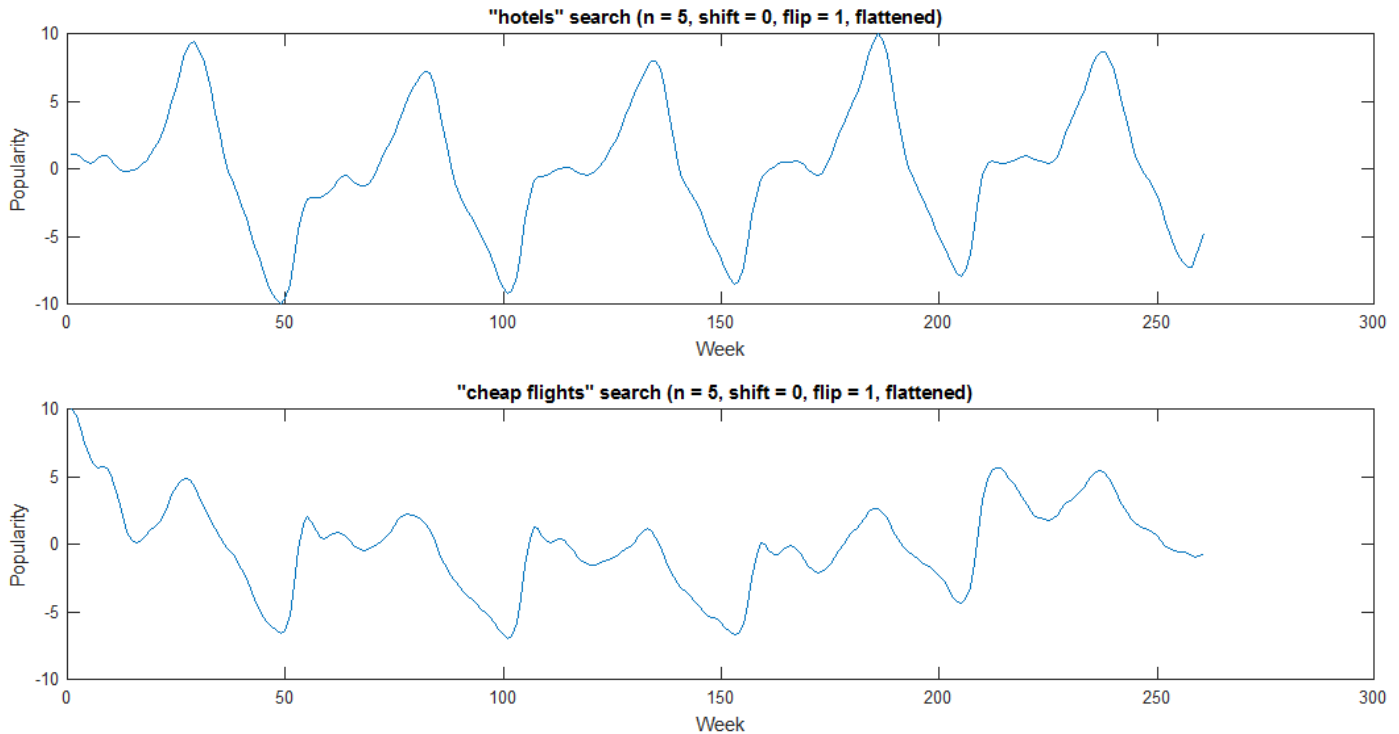
The final step in our research process was to use the results given to us to plot the correlating data sets and determine if the graphs were good enough to consider the possibility of using the search data to predict future stock price movement. We referred to this process as our “visual test.”

4. Demonstration of the test

To demonstrate the effectiveness of our test we have provided examples of a pair of searches and a pair of stocks that our program calculated to have a strong correlation and which we would expect this to be the case.

4.1 Google Trends versus Google Trends

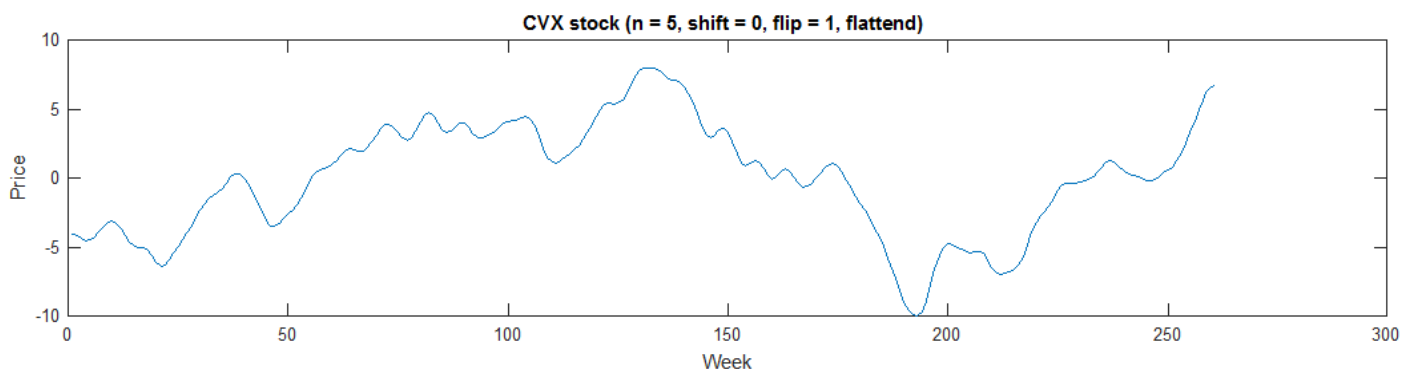
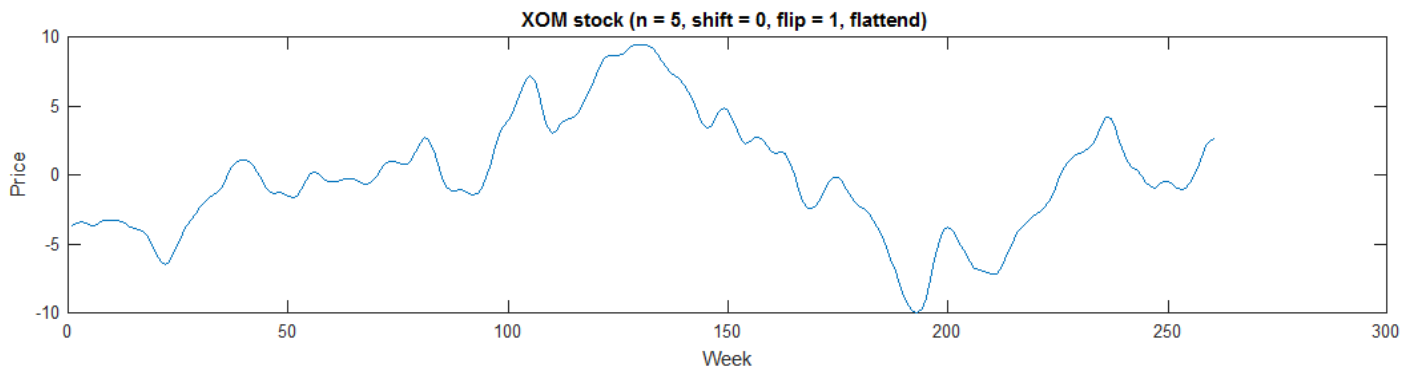
Using smoothing iterations = 5, shift = 0, and flip = 1, our test determined that the searches “hotels” and “cheap flights” correlated with a value of $T = .842$. This is statistically significant with the chances of this randomly occurring being less than one in one billion (using z-score = 6 where actual z-score = 6.61). We would expect this pair to correlate well.



Although the magnitude of the hills are not the same, it is incredible that the peaks and troughs of the graphs are all at the same points in time. This means that “hotels” and “cheap flights” are at their lowest and highest points of being searched at the same time.

4.2 Stock prices versus stock prices

Using smoothing iterations = 5, shift = 0, and flip = 1, our test determined that the stock prices of the companies Exxon Mobil Corporation and Chevron Corporation correlated with a value of $T = .815$. This is statistically significant with the chances of this randomly occurring being less than one in one billion (using z-score = 6 where actual z-score = 6.09). Both of these companies are in the oil industry, which makes sense of why these two would be correlated. The following image illustrates this high correlation. Exxon and Chevron have ticker symbols XOM and CVX respectively.



The two companies' graphs of stock prices are nearly identical. Much of their peaks and valleys share the same magnitude which means they are almost perfectly correlated. This is impressive since this data reflects a five year time span, and begs the question "Can we use one stocks price movement to predict another?" Here are tables of some other stocks that correlated well with one another according to our test.

S&P500 stocks and S&P500 stocks with correlation $\geq .8$ and shift ≥ 1							
smooth iter	shift	flip	stock predictor	stock predicted	corrValue	z-score	
5	1	1	'FOX'	'FOXA'	0.8571	6.91	
5	1	1	'JPM'	'GS'	0.8224	6.24	
5	1	1	'CVX'	'XOM'	0.8031	5.86	
10	1	1	'C'	'BAC'	0.8494	5.82	
10	2	1	'C'	'BAC'	0.8023	5.04	
10	1	1	'GS'	'C'	0.8185	5.31	
10	1	1	'FOXA'	'FOX'	0.8880	6.47	
10	1	1	'JPM'	'GS'	0.8224	5.37	
10	1	1	'C'	'GS'	0.8185	5.31	
10	1	1	'MS'	'GS'	0.8378	5.63	
10	1	1	'MMM'	'HON'	0.8108	5.18	
10	1	1	'GS'	'JPM'	0.8069	5.12	
10	1	1	'UPS'	'MMM'	0.8378	5.63	
10	1	1	'JPM'	'USB'	0.8031	5.05	
10	1	1	'WFC'	'USB'	0.8263	5.44	
10	1	1	'CVX'	'XOM'	0.8108	5.18	

S&P500 stocks and S&P500 stocks with correlation $\geq .8$ and shift ≥ 1							
smooth iter	shift	flip	stock predictor	stock predicted	corrValue	z-score	
20	1	1	'C'	'BAC'	0.8764	5.31	
20	1	1	'GS'	'BAC'	0.8147	4.44	
20	2	1	'C'	'BAC'	0.8527	4.98	
20	2	1	'GS'	'BAC'	0.8217	4.54	
20	3	1	'C'	'BAC'	0.8132	4.42	
20	3	1	'GS'	'BAC'	0.8054	4.31	
20	1	1	'JPM'	'C'	0.8263	4.61	
20	1	1	'BAC'	'C'	0.8224	4.55	
20	1	1	'GS'	'C'	0.8764	5.31	
20	2	1	'JPM'	'C'	0.8023	4.27	
20	2	1	'GS'	'C'	0.8450	4.87	
20	3	1	'GS'	'C'	0.8054	4.31	
20	1	1	'UTX'	'CBS'	0.8069	4.33	
20	1	1	'XOM'	'CVX'	0.8340	4.72	
20	2	1	'XOM'	'CVX'	0.8101	4.38	
20	1	1	'FOXA'	'FOX'	0.9189	5.91	
20	1	1	'FOX'	'FOXA'	0.9189	5.91	
20	1	1	'JPM'	'GS'	0.8494	4.93	
20	1	1	'C'	'GS'	0.8610	5.10	
20	1	1	'MS'	'GS'	0.8417	4.82	
20	2	1	'C'	'GS'	0.8101	4.38	
20	1	1	'MMM'	'HON'	0.8108	4.39	
20	1	1	'UTX'	'HON'	0.8378	4.77	
20	1	1	'C'	'JPM'	0.8340	4.72	
20	1	1	'GS'	'JPM'	0.8571	5.04	
20	1	1	'MS'	'JPM'	0.8147	4.44	
20	2	1	'GS'	'JPM'	0.8178	4.49	
20	1	1	'GS'	'MS'	0.8378	4.77	
20	1	1	'INTC'	'MSFT'	0.8031	4.28	
20	1	1	'MMM'	'UTX'	0.8031	4.28	
20	1	1	'HON'	'UTX'	0.8147	4.44	
20	2	1	'MMM'	'UTX'	0.8023	4.27	
20	1	1	'CVX'	'XOM'	0.8263	4.61	

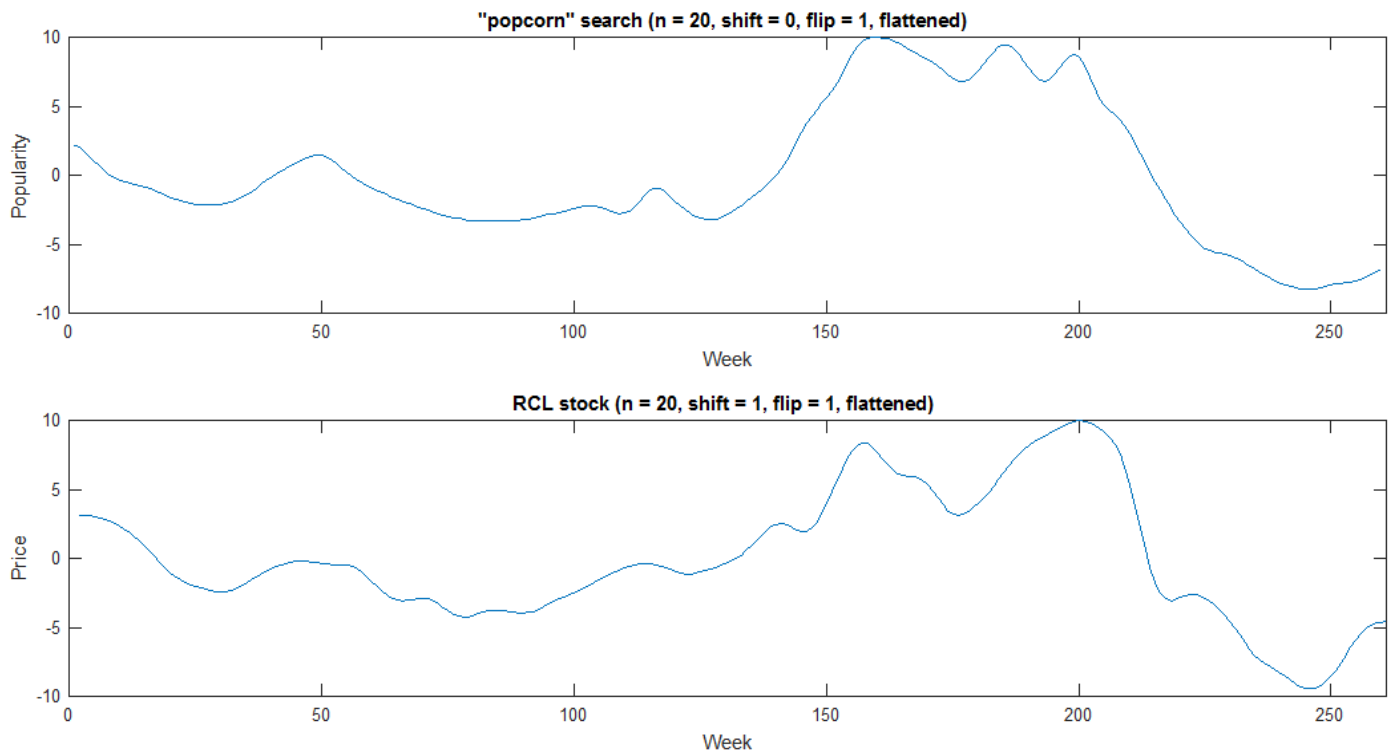
z-score Table						
z-score	1	2	3	4	5	6
*x	6	44	741	31,574	3,486,914	1,014,713,328
*1 in x chances of randomly occurring						

5. Test results

5.1 Google Trends versus stock prices

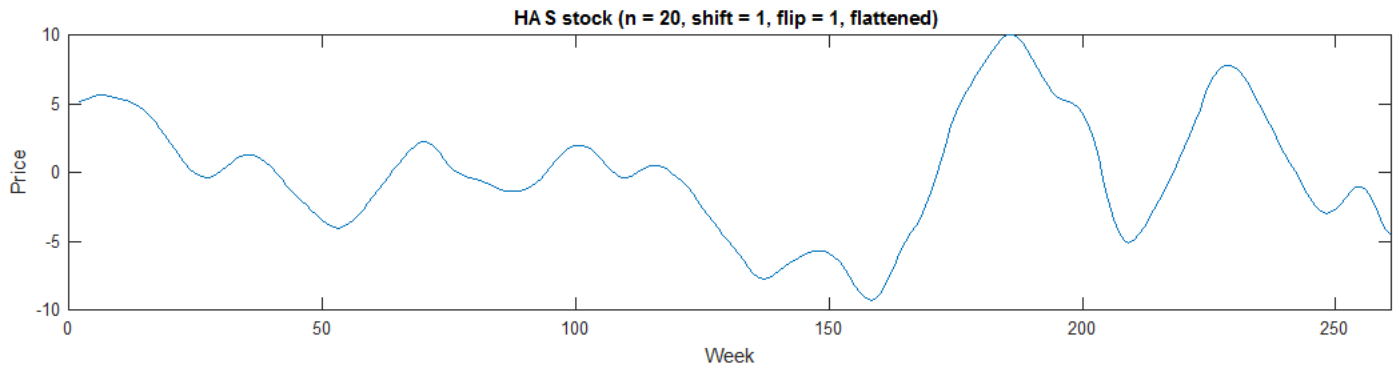
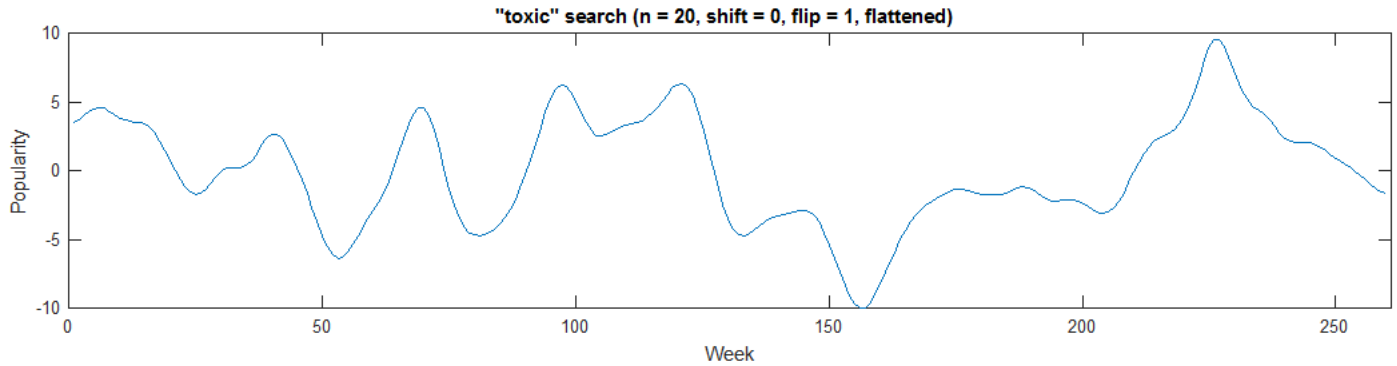
The purpose of our test was to aid us in finding searches that could possibly be used to predict future stock price movement. The stocks from Vanguard's ETF and S&P's midcap options did not correlate well enough with our searches. However, there were quite a few stocks from the S&P500 that had not only high correlation values with a search but also had very similar looking graphs. The following are three of our best and most interesting examples.

This first example is with the search word "popcorn" and Royal Caribbean Cruise stock, ticker symbol RCL. The correlation value of these two was .776 with smoothing iteration = 20, shift = 1 week, and no flip.

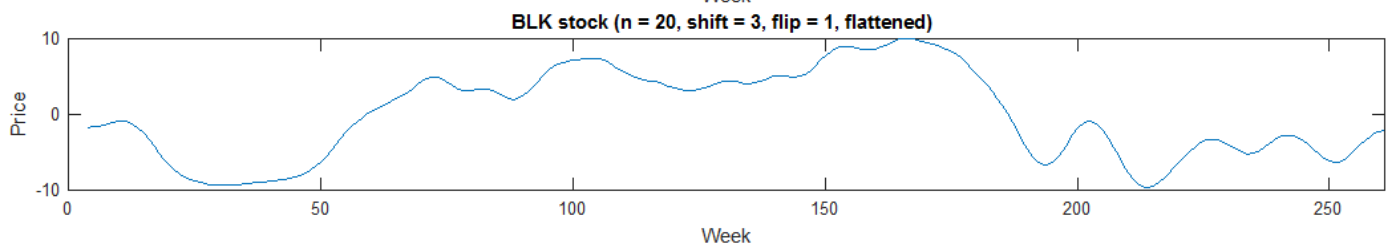
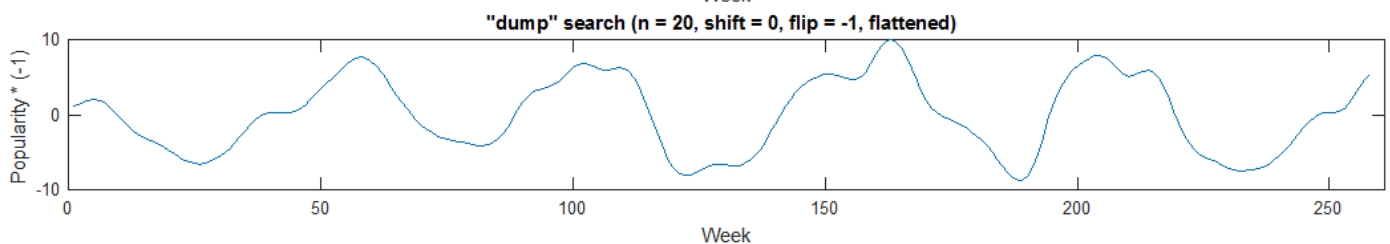
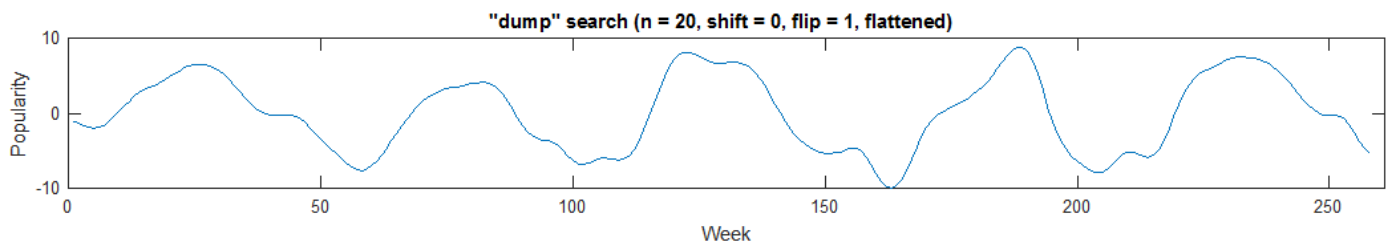


These two graphs are not exactly the same, however they look fairly similar. Since there was no flip and the search's data was shifted one week, we can say that when "popcorn" was increasingly being searched, the stock price for RCL was likely to rise. Notice that the location of many of the local max and min points of the search's data were also local max and min points for the data of RCL. With our test returning a correlation value of .776 we can say that the chances of this event being random are one in 20,510.

Our next example is with the search word "toxic" and Hasbro Inc. stock, ticker symbol HAS. Our test determined that these two also had a correlation value of .776 with smoothing iterations = 20, shift = 1 week, and no flip.



The magnitude of the rises and falls are not equal however the placement of the local maxima and minima are almost spot on. This pair was particularly interesting because Hasbro is a company that produces toys for children. Having a correlation value of .776 with the search word “toxic” being shifted forward one week implies that when the word “toxic” was being used in the search engine frequently, it was likely that the price of HAS stock would go up a week later. It would be fascinating to find out if the company’s revenue also correlated with this search word. If this were the case, it would appear that the correlation between the popularity of “toxic” and stock price of HAS would be more than just a one in 20,510 chance of being random. Our final example involves the search word “dump” and BlackRock Inc. stock, ticker symbol BLK. This pair was calculated to have a correlation value of .7588 with smoothing iterations = 20, shift = 3 week, and a flip.



Blackrock Inc. is a global investment management corporation. Although these two graphs look the least similar of our three examples, we found this to be another interesting case. Notice that much of the local minima of the bottom graphs of the illustration are the same. This means that when “dump” was being entered in search engines, it was likely that three weeks later BLK stock would drop. The correlation value being .759 meant that this would randomly occur one in 7,825 times. Note that the term “dump a stock” is frequently used in the financial world when an investor or investment firm is seeking to sell all or, perhaps, a large portion of its shares of a company they are invested in. Generally this means placing market orders to sell which results in a decline in that stocks price because a stocks “current price” is determined by the last price it was bought/sold for.

Our test provided search and pair matches that had a correlation value above some specified threshold. Using a threshold value of .70 proved to be most reasonable because this essentially meant that we could expect the graphs to look about 70% similar. The following tables exhibit other search and stock pairs that our test calculated to have a correlation value of .725 or higher.

Google Searches and Vanguard ETFs with correlation \geq .725					
smooth iter	shift	flip	search	stock	corrValue
20	4	-1	'peace'	'VB'	0.7422
20	3	-1	'peace'	'VBK'	0.7393
20	4	-1	'peace'	'VBK'	0.7617
20	1	-1	'telephone'	'VCSH'	0.7375
20	1	-1	'cook'	'VCSH'	0.7297
20	2	-1	'telephone'	'VCSH'	0.7287
20	2	-1	'bath'	'VCSH'	0.7287
20	3	-1	'cook'	'VCSH'	0.7432
20	3	-1	'bath'	'VCSH'	0.7549
20	4	-1	'guitar'	'VCSH'	0.7266
20	4	-1	'cook'	'VCSH'	0.7305
20	4	-1	'bath'	'VCSH'	0.7500
20	4	1	'fiduciary'	'VCSH'	0.7266
20	1	-1	'crib'	'VCSH'	0.7297
20	1	1	'mdt'	'VCSH'	0.7413
20	2	1	'mdt'	'VCSH'	0.7558
20	3	1	'mdt'	'VCSH'	0.7626

Google Searches and Vanguard ETFs with correlation \geq .725					
smooth iter	shift	flip	search	stock	corrValue
20	4	1	'mdt'	'VCSH'	0.7422
20	1	-1	'dow'	'VEA'	0.7297
20	1	-1	'dow'	'VEU'	0.7375
20	1	-1	'dow'	'VGK'	0.7568
20	4	-1	'donald trump'	'VHT'	0.7266
20	4	-1	'serious'	'VNQI'	0.7266
20	4	-1	'alien'	'VNQI'	0.7305
20	2	-1	'ring'	'VNQI'	0.7326
20	3	-1	'ring'	'VNQI'	0.7354
20	4	-1	'ring'	'VNQI'	0.7305
20	2	1	'pg'	'VOX'	0.7403
20	3	1	'pg'	'VOX'	0.7432
20	4	1	'pg'	'VOX'	0.7305
20	3	1	'borrow'	'VSS'	0.7276
20	1	-1	'dow'	'VT'	0.7297
20	3	-1	'peace'	'VXF'	0.7315
20	4	-1	'peace'	'VXF'	0.7617

Google searches and S&P400 stocks with correlation \geq .725					
smooth iter	shift	flip	search	stock	corrValue
20	1	-1	'football'	'AMCX'	0.7297
20	2	-1	'bear market'	'ASH'	0.7364
20	4	-1	'gamble'	'BEAV'	0.7266
20	2	1	'pm'	'BIG'	0.7326
20	3	1	'pm'	'BIG'	0.7354
20	1	-1	'pep'	'CHS'	0.7259
20	2	-1	'pep'	'CHS'	0.7558
20	3	-1	'pep'	'CHS'	0.7588
20	4	-1	'pep'	'CHS'	0.7500
20	1	-1	'mmm'	'CLGX'	0.7568
20	2	-1	'mmm'	'CLGX'	0.7636
20	3	-1	'mmm'	'CLGX'	0.7471
20	4	-1	'mmm'	'CLGX'	0.7266
20	1	-1	'gamble'	'CR'	0.7259
20	2	-1	'gamble'	'CR'	0.7364
20	3	-1	'gamble'	'CR'	0.7471
20	4	-1	'gamble'	'CR'	0.7578
20	1	1	'stock split'	'DEI'	0.7375
20	2	1	'stock split'	'DEI'	0.7364

Google searches and S&P400 stocks with correlation \geq .725					
smooth iter	shift	flip	search	stock	corrValue
20	3	1	'stock split'	'DEI'	0.7276
20	1	-1	'bankruptcy'	'DF'	0.7490
20	2	-1	'bankruptcy'	'DF'	0.7519
20	1	-1	'mmm'	'FAF'	0.7259
20	2	-1	'mmm'	'FAF'	0.7481
20	3	-1	'mmm'	'FAF'	0.7471
20	4	-1	'mmm'	'FAF'	0.7422
20	1	-1	'facebook'	'HII'	0.7259
20	2	-1	'facebook'	'HII'	0.7287
20	3	-1	'facebook'	'HII'	0.7315
20	1	-1	'china'	'INT'	0.7413
20	2	-1	'china'	'INT'	0.7481
20	3	-1	'china'	'INT'	0.7393
20	1	-1	'fall'	'INT'	0.7452
20	1	1	'gold'	'JACK'	0.7259
20	4	1	'violent'	'ODP'	0.7266
20	1	1	'silk'	'WEN'	0.7490
20	2	1	'silk'	'WEN'	0.7364

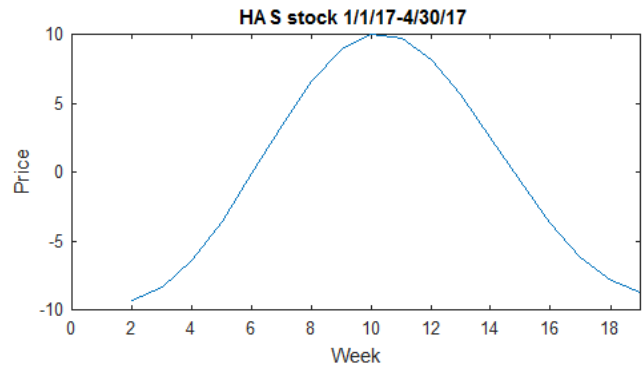
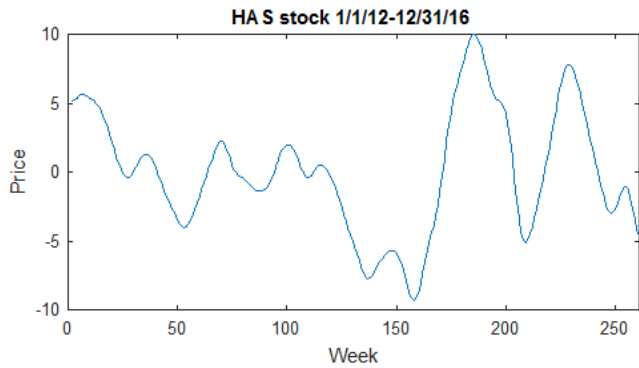
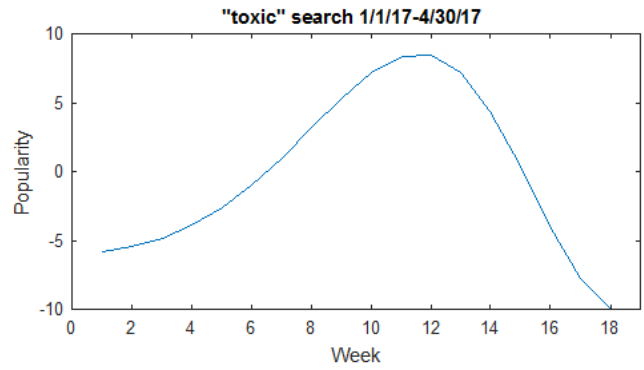
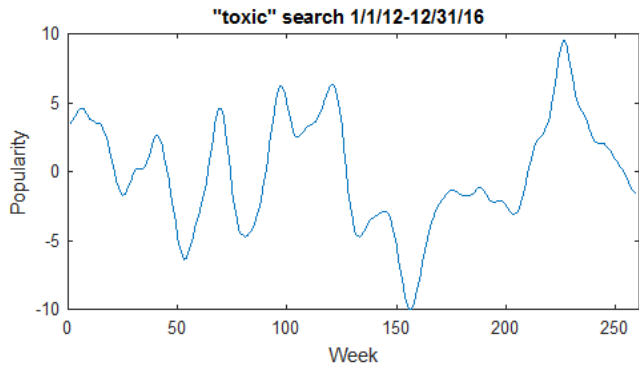
Google Searches and S&P500 stocks with correlation $\geq .725$					
smooth iter	shift	flip	search	stock	corrValue
20	1	-1	'facebook'	'BBY'	0.7259
20	2	-1	'dump'	'BLK'	0.7326
20	4	-1	'dump'	'BLK'	0.7344
20	2	-1	'border'	'BMY'	0.7558
20	3	-1	'border'	'BMY'	0.7588
20	4	-1	'border'	'BMY'	0.7383
20	4	-1	'peace'	'CHTR'	0.7305
20	1	-1	'musical'	'CHTR'	0.7375
20	4	-1	'destroy'	'CSCO'	0.7422
20	2	-1	'bear market'	'CSCO'	0.7287
20	3	-1	'bear market'	'CSCO'	0.7354
20	4	-1	'bear market'	'CSCO'	0.7266
20	2	-1	'note'	'DIS'	0.7326
20	3	-1	'note'	'DIS'	0.7315
20	1	-1	'drown'	'ETFC'	0.7375
20	2	-1	'drown'	'ETFC'	0.7597
20	1	1	'video games'	'FDX'	0.7375
20	2	-1	'gamble'	'GM'	0.7364
20	3	-1	'gamble'	'GM'	0.7665
20	4	-1	'gamble'	'GM'	0.7734
20	1	1	'butter'	'GM'	0.7259
20	1	1	'rule'	'HAS'	0.7375
20	2	1	'toxic'	'HAS'	0.7636
20	3	1	'toxic'	'HAS'	0.7549
20	1	1	'mdt'	'HAS'	0.7297
20	3	-1	'mold'	'HSY'	0.7276
20	4	-1	'mold'	'HSY'	0.7500
20	4	-1	'destroy'	'JNJ'	0.7266

Google Searches and S&P500 stocks with correlation $\geq .725$					
smooth iter	shift	flip	search	stock	corrValue
20	4	-1	'destroy'	'JNJ'	0.7266
20	1	-1	'call'	'LLY'	0.7683
20	2	-1	'call'	'LLY'	0.7674
20	3	1	'mmm'	'MCD'	0.7276
20	4	1	'mmm'	'MCD'	0.7305
20	2	1	'video games'	'MS'	0.7287
20	4	1	'video games'	'MS'	0.7266
20	1	-1	'bruise'	'RCL'	0.7259
20	1	-1	'yellow'	'RCL'	0.7375
20	1	1	'cheese'	'RCL'	0.7259
20	1	1	'fog'	'RCL'	0.7413
20	2	1	'fog'	'RCL'	0.7287
20	1	1	'christmas'	'RCL'	0.7297
20	2	1	'christmas'	'RCL'	0.7287
20	1	-1	'shorts'	'RCL'	0.7297
20	1	1	'oven'	'RCL'	0.7413
20	2	1	'oven'	'RCL'	0.7403
20	1	-1	'fork'	'RCL'	0.7297
20	1	1	'lipstick'	'RCL'	0.7452
20	2	1	'lipstick'	'RCL'	0.7326
20	4	1	'shame'	'SLB'	0.7422
20	2	1	'pm'	'SLB'	0.7326
20	3	1	'pm'	'SLB'	0.7276
20	3	1	'gis'	'VZ'	0.7393
20	4	1	'gis'	'VZ'	0.7578
20	3	1	'opera'	'YHOO'	0.7315
20	4	1	'opera'	'YHOO'	0.7305

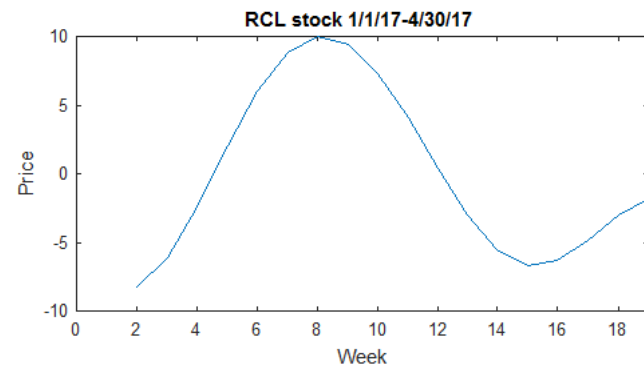
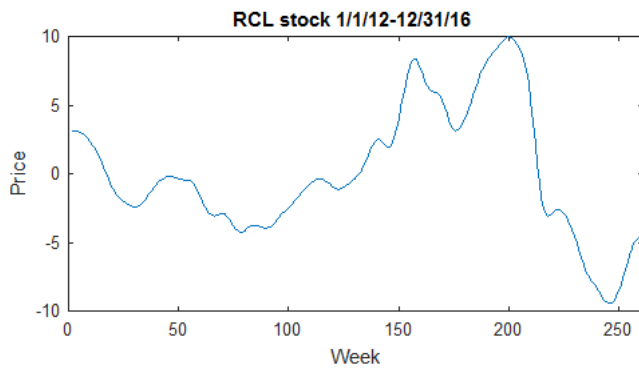
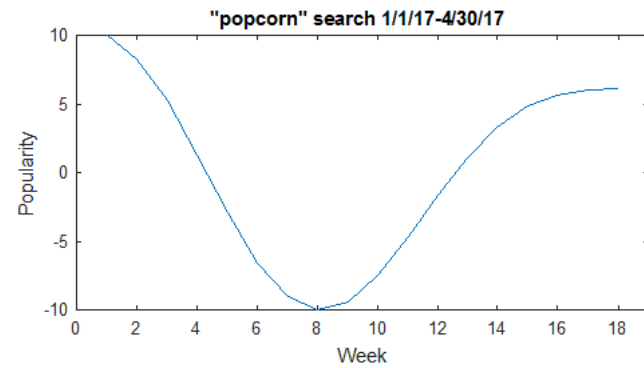
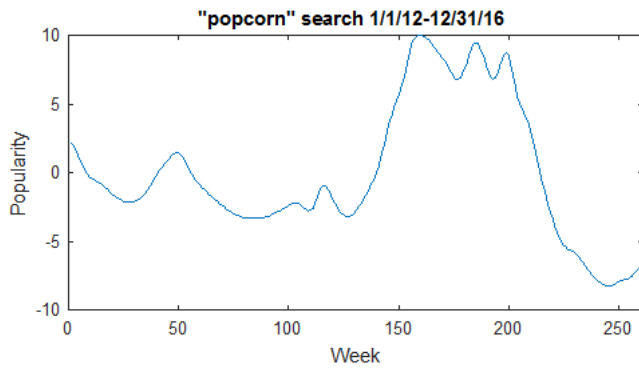
6. Conclusion

6.1 Thoughts on our Hypothesis

Although we found some searches and stocks that correlated well with one another, it would not have been wise for us to take this information and start investing solely based on our results. Therefore, we gathered data for the search and stock pairs for the four months following our five year time period, 1/1/17-4/30/17, on a weekly basis to determine whether or not we could use the results to predict future stock price movement. The following examples, “toxic” vs HAS and “popcorn” vs RCL, demonstrate the possibility of this. Note that we used the same amount of smoothing iterations = 20, shift = 1, and flip = 1 for the new four month data sets.



As you can see here, the correlation appears to continue into the next time period. Our test calculated the new data sets to have a correlation value of .6875 with the shift = 1 week. If we did not shift the data at all, i.e. set shift = 0, the pair would have a correlation value of .8824. This is impressive and suggests that it may be possible to use our results to predict future stock price movement. Our next example, however, contradicts this possibility.

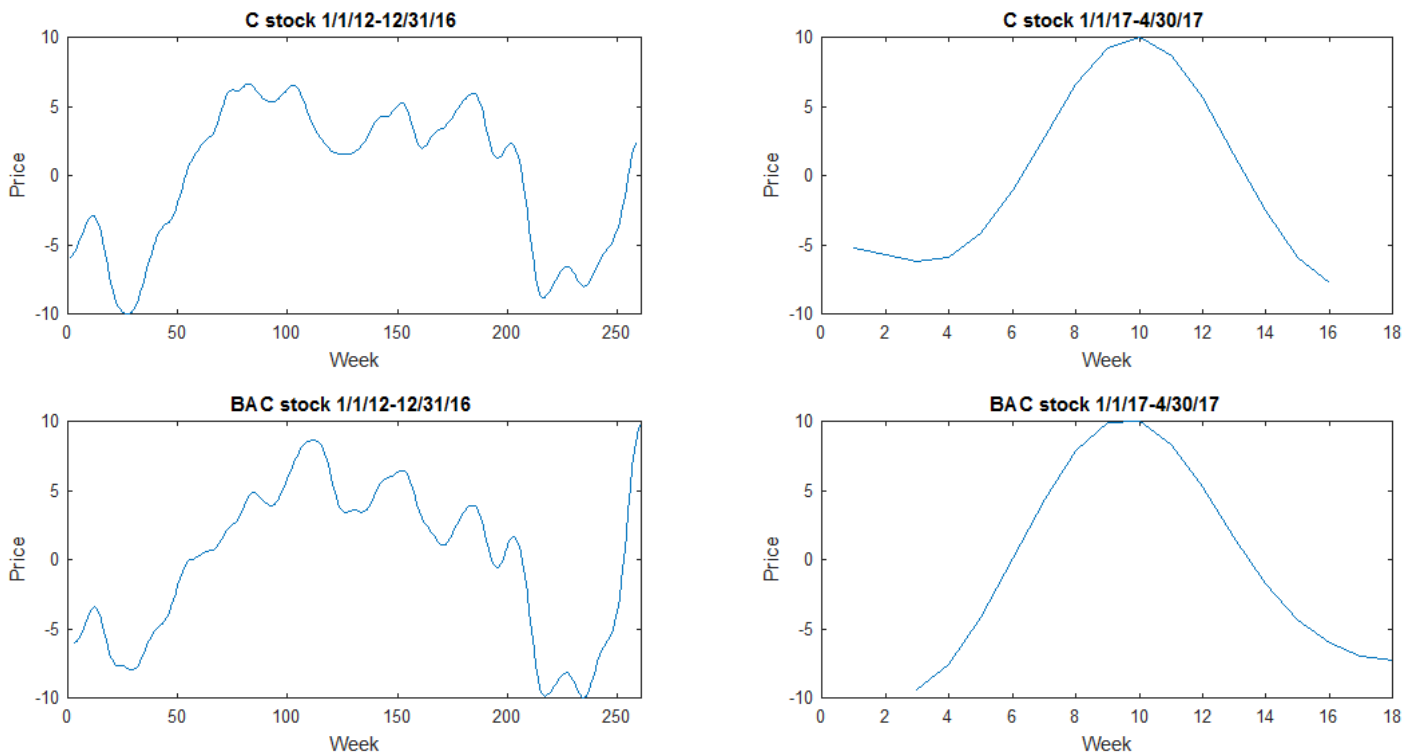


This example is quite interesting because our test determined that the new data sets of these two had a correlation value of .3125 when flip = 1. Conversely, when flip = -1, they had a correlation value of .6875. Their correlation inverted from that five year period to the four month period. If we based our investment strategy exclusively on our previous results we would most definitely lose money.

From these examples we can conclude that there exists searches whose popularity correlates with stock prices over a given period. Although the probability of this randomly occurring may be small, it does not imply that the pair's correlation will continue into the future. Therefore, it's difficult to say whether or not we could use the popularity of searches to predict future stock price movement. On the other hand, our example with the search word "toxic" and Hasbro Inc. stock did raise a few questions for us. Another search that we gathered data for is the word "hazardous," which is closely related to "toxic," however our test calculated "hazardous" and HAS to have a correlation value no higher than .6470 under all possible conditions during that five year time period. This suggests that the word "toxic" was more related to Hasbro Inc. than the word "hazardous" during those five years. So although we may not be able to use Google Trends data to predict stock price movement, could our test discover possible words or phrases that may be linked to a company? It may also be interesting to investigate whether or not those words or phrases are used in any articles written about companies that they are determined to correlate with.

6.2 Final Remarks

In spite of this hypothesis leading us to a conclusion of uncertainty, we have not been left empty handed. From this process we have gained a better understanding of the stock market and search engines. In addition, this research motivated us to seek out efficient methods for gathering large amounts of data and develop an original test for determining whether the data sets were correlated or not. Most importantly, our research has led us to even more hypotheses worth testing. One of which is to whether or not we can use stocks to predict the price movement of other stocks. Below is an example from our previous stock vs. stock table with the companies Citigroup Inc., ticker symbol C, and Bank of America Corp., ticker symbol BAC, using smoothing iterations = 20, shift = 2, and flip = 1.



The correlation value of these stocks during the five year period was .8530 and .7333 for the following four month period. We are currently exploring this hypothesis, however, this example is a demonstration of its promise.