Can Google Trends Index data be used to predict stock returns?

November 10, 2018

Abstract

The main purpose of this paper is to study the relationship between Google Trends Index (GTI) and the stock returns in the subsequent period. Using weekly GTI data from 2013 to 2018 for 30 Dow Jones Industrial (DJI) companies, I find that a high GTI is associated with negative returns in the next period. These results are in line with Bijl et al. (2016), who conduct a similar study using data from 2008-2013. Additionally, my findings indicate that a high GTI for a particular DJI stock reduces its probability of performing better than the DJI index in the following period. A one standard deviation higher GTI decreases the probability of stock beating the DJI index in the following week by 5 percentage points or equivalently by 10 %, with mean probability of beating DJI index equal to 0.5.

Keywords: Google Search Trends, Stock Returns, Efficient Markets.

JEL Codes: G1, G4, D8

1 Introduction

According to the efficient market hypothesis price movements in the stock market are comparable to a random walk. (Malkiel and Fama, 1970). In other words, at any given time period all available information is already priced into the current stock prices and therefore any information available in the current period will not be correlated to returns on stocks in the next period or any future period. However, there is an extensive academic research that tries to find potential indicators or pattern that help predict or forecast future stock prices. Please read Malkiel (2003) for extensive literature review on this broad topic. For the purpose of my paper I will focus on how Google search volume, measured in the form of GTI, for a particular stock in the current week affects the return of that stock in the following week. My study is basically a replication of Bijl et al. (2016), however I use a more recent dataset, i.e. they use stock price data and GTI data from 2008 to 2013 and I use data from 2013 to 2018. Studying the correlation between GTI and stock return with a more recent dataset is not just a replication exercise, but rather an important issue to study since it will enhance the understanding of the efficient market hypothesis. This is arguably true because if markets are efficient and they price in such correlation in the long run, then the correlation found by Bijl et al. (2016), should go not exist in the data that I am studying. However, I find correlation between GTI and stock return very similar to Bijl et al. (2016) despite the fact that market were more volatile during 2008-2013 compared to 2013-2018 as shown in figure 4 and figure 5 in the figures section.

Insert figure 4 and figure 5 here

Secondly, comparing my study to Bijl et al. (2016), they use data from all 500 Standard and Poor's companies, while I focus only on DJI 30 companies. Lastly, in addition to estimating how GTI affect the excess stock return compared to a market index, I also estimate how GTI affects the probability of a stock beating a market index. i.e. the probability of the stock performing better than the DJI index in the subsequent week. I find that a high GTI is associated with a decrease in the probability of the stock performing better than the DJI index.

Several previous studies have used internet data to predict stock volume, returns and volatility and have found contradicting results. Takeda and Wakao (2014) found that Google search volume for a stock was positively but weakly associated with stock prices and also positively associated with trading volume. The estimated the effect using data from the Japanese stock market. Similarly, Aouadi et al. (2013) found that Google search volume is positively related with trading volume using french stock market data. On the other hand many researchers have also studied how Google trends data can be used to design a profitable investment strategy. Kristoufek (2013) argue that Google search data can be used to design a systematic invest strategy that minimizes risk or the sharpe ratio. This shows that studying stock behaviour and GTI is not a novel idea, however to the best of my knowledge, previous studies have not focused on estimating the probability of a stock beating a market index based on GTI, which is the main contribution of this paper; since estimating this probability can be used to design a dynamic profitable investment strategy.

2 Data

2.1 Stock Return Data

As mentioned in the introduction I use weekly stock returns as my dependent variable to measure the effect of GTI on stock returns. I calculate weekly stock return by comparing the stock opening price on Monday to the closing price on Friday of the same week. I use the below formula to calculate the weekly return $r_{i,t}$.

$$r_{i,t} = \frac{p_{i,t}^{M} - p_{i,t}^{F}}{p_{i,t}^{M}} \tag{1}$$

- Where, $p_{i,t}^{M}$ is the Monday opening price of stock i in time period t and $p_{i,t}^{F}$ is the Friday closing price of stock i in time period t.

- $i \in \{1, 2.., 30\}$,is a placeholder for all the stocks in (DJI). A complete list of stocks used in this paper are given in table A1 of the appendix section.

- $t \in \{1, 2.., 229\}$, indicates each week in this study. A complete summary of time period and number of weeks can be found in table A2 of the appendix section. I exclude weeks in which markets were closed on the Monday or the Friday.

Further, as I am interested in how a stock performs in comparison to a composite index I calculate excess return to the DJI index, $er_{i,t}$ using the below formula.

$$er_{i,t} = r_{i,t} - r_{DJI,t} \tag{2}$$

Where $r_{DJI,t}$ is the weekly return for DJI index in period t.

I downloaded the stock price data using an open source yahoo finance api, which can be found at this link. https://github.com/lukaszbanasiak/yahoo-finance

2.2 GTI Data

GTI measures the relative search interest in a given period compared to other periods. GTI is a number between 1 and 100, where a number close to 1 indicates low search interest for the term in the given period and number close to 100 implies a high level of interest. I do use the raw GTI as an independent variable in my regression but a more appropriate way to measure how variations in GTI affects stock price is to standardize the GTI index in order to make it comparable across stocks. I use the same method for standardizing GTI implemented by Bijl et al. (2016) and below is the formula to standardize GTI.

$$Standardized\_GTI_{i,t} = \frac{GTI_{i,t} - \widehat{GTI_i}}{\sigma_{GTI}} \tag{3}$$

- Where, $\widehat{GTI_i}$ is average GTI for stock i and σ_{GTI} is the standard deviation of GTI for the entire sample. Table A2 in the appendix section contains the mean and standard deviation of GTI.

Figure 1 and 2 in the figures section show how standardizing GTI makes the index comparable across stocks. Figure 1 shows the raw GTI over time for Apple, Verizon and Nike stocks. It is clear from the graph that Nike has relatively low overall search index, however it does vary over time. Therefore, as shown in figure 2 which plots standardized GTI over time, standardizing GTI makes variation in GTI for Nike comparable to variation in GTI for Apple and Verizon.

Insert figure 1 and figure 2 here

I downloaded the GTI data using an open source pseudo api for Google trends, which can be found at this link. https://github.com/GeneralMills/pytrends. I use the stock symbol as the GTI search term for each stock that has a stock symbol of length greater than 2. However, for stock symbols that have a length of 2 or less I use the stock symbol plus "Stock" as the GTI search term. For example with regards to Verizon, whose stock symbol is "VZ", the GTI search term I assign to Verizon is "VZ Stock". A complete list of search term associated with each stock can be found in Table 1A of the appendix section.

3 Results

I use the below fixed effect econometrics model to estimate the effect of GTI in excess return to DJI index.

$$er_{i,t} = \beta_0 + \sum_{l=1}^{L} \beta_l GTI_{i,t-l} + \delta_i + \theta_t + e_{it} \tag{4}$$

Where,

- $er_{i,t}$, is the excess return to DJI index for stock i in week t, calculated using the formula mentioned above.

- $GTI_{i,t-l}$ is the standardized or non standardized lagged GTI. Where l indicates the lag period. Regression result table indicates whether a standardized or non standardized GTI is used to estimate the model.

- L is the number of lag periods included in the model. Again the each regression table indicates how many lagged periods are included in the model.

- δ_i is the stock fixed effects. Individual dummy variable for each stock.

- θ_t is the week fixed effects. Individual dummy variable for each week.

Results from the above fixed effect model are displayed in table 1 of the Tables section. The results indicate that a high GTI is associated with a negative excess returns to DJI index in the very next week. The negative relation between excess return and 1 period lagged GTI is statistically significant in all four of the models. Focusing on model 3, in table 1 which includes 4 lagged standardized GTI, shows that a one standard deviation higher GTI from the mean yields a return which is 0.3% lower than the DJI index return; a substantial lower return considering that the average excess return for individual stock is -0.057%. To put this in terms of a relatable scenario, if I pick a random stock on Monday at opening bell in any given week from the dataset, my expected return on Friday at closing bell would be approximately equal to the return of the DJI index. However, if I picked the stock on Monday at opening bell which had a GTI one standard deviation higher than its average GTI, my expected return by Friday closing bell would be 0.3% less than the DJI index. The results also indicate that GTI from two periods ago have a positive impact on excess return, however it is smaller in magnitude and only statistically significant at the 90% level. The GTI from more than 2 periods do not have an impact of the stock returns. These results are very similar to the findings of Bijl et al. (2016). The fact that I find similar correlation between GTI and excess stock returns puts a doubt on the efficient market hypothesis, since efficient market hypothesis suggests that any information from a previous period cannot be used to beat the market for extended period of time (Malkiel and Fama, 1970).

Insert Table 1 here

Next I estimate how GTI index relates to the probability of the stock beating the DJI index in the next period, using the below fixed effect model similar to equation (4).

$$(er_{i,t} > 0) = \beta_0 + \sum_{l=1}^{L} \beta_l GTI_{i,t-l} + \delta_i + \theta_t + e_{it} \tag{5}$$

Where,

- $(er_{i,t} > 0)$, is a dummy variable equal to 1 if $er_{i,t} > 0$ for stock i in time period t

Insert Table 2 here

Results from the above model are presented in table 2 of the tables section. Similar to the results from the first model, a high GTI in the previous week reduces the probability of the stock return being higher than the return on the DJI index. These results are statistically significant at the 99 % level. Again focusing on the model with 4 lagged standardized GTI values, a one standard deviation higher than average GTI reduces the probability of that stock beating the DJI index in the following week by 5 percentage points. This is a 10 % lower probability of beating the DJI from the average probability of beating the stock market which is approximately 0.5. In other words, if I were to randomly pick a stock in any given week from the dataset, whether or not it will yield a return higher than the DJI index is similar to a coin flip, however if I were to pick a stock with one standard deviation lower GTI than the mean in the previous week, then the probability of that stock beating the DJI index is 0.55 greater probability than a coin flip.

4 Discussions

Insert figure 3 here

The results from the two regression models indicate that there is negative relation and statistically significant relation between GTI in period $t-1$ and excess stock return in period t. This phenomena can also be observed in figure 3 of the figures section which plots standardized GTI on the x axis and excess stock return on the y axis, with the line of best fit and the associated bandwidth for all 30 DJI stocks. From the figure it appears that even at the individual stock level the relation is either negative or flat and only for few stock such as IBM and McDonald's the relation is positive.

One shortcoming of this research is that despite the highly statistically significant correlation, I cannot formulate a reasonable theory or even an explanation that can justify the relation.

5 Conclusion

Whether or not markets are efficient has been a long standing question in the finance literature and my paper makes no attempt at solving the puzzle and putting this issue at rest. However, it is striking to find that information, like the GTI, that is easily available to just about anyone with an internet connection can be used as a strong predictor of the probability of an individual stock performing better than the overall market; despite the fact that this relation has already been documented in the literature.

References

Aouadi, A., Arouri, M., and Teulon, F. (2013). Investor attention and stock market activity: Evidence from france. *Economic Modelling*, 35:674–681.

Bijl, L., Kringhaug, G., Molnár, P., and Sandvik, E. (2016). Google searches and stock returns. *International Review of Financial Analysis*, 45:150–156.

Kristoufek, L. (2013). Can google trends search queries contribute to risk diversification? *Scientific reports*, 3:2713.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1):59–82.

Malkiel, B. G. and Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Takeda, F. and Wakao, T. (2014). Google search intensity and its relationship with returns and trading volume of japanese stocks. *Pacific-Basin Finance Journal*, 27:1–18.

Tables

Table 1: Fixed effect model to estimate the effect of GTI on excess stock return in the following week

	Dep. Var: Excess Return to DJI (%)			
	(Stand. GTI) 1 Lag	(Stand. GTI) 2 Lag	(Stand. GTI) 4 Lag	(Non-Stand. GTI) 4 Lag
$GoogleTrendsIndex_{t-1}$	-0.1738**	-0.2841***	-0.3324***	-0.0179***
	(0.0785)	(0.1097)	(0.1163)	(0.0063)
$GoogleTrendsIndex_{t-2}$		0.1816**	0.1306*	0.0070*
		(0.0849)	(0.0770)	(0.0041)
$GoogleTrendsIndex_{t-3}$			0.0765	0.0041
			(0.0516)	(0.0028)
$GoogleTrendsIndex_{t-4}$			0.0835	0.0045
			0.0516	(0.0030)
Constant	-0.0537***	-0.0436***	-0.0309***	0.0339
	(0.0013)	(0.0028)	(0.0041)	(0.1148)
Stock Fixed Effects	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes
N	6,524	6,494	6,434	6,434
R^2	0.0022	0.0038	0.0049	0.0049

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 2: Fixed effect model to estimate effect of GTI the probability of excess stock return in the following week greater than 0%.

	Dep. Var: Excess Return to DJI (%) > 0			
	(Stand. GTI)	(Stand. GTI)	(Stand. GTI)	(Non-Stand. GTI)
	1 Lag	2 Lag	4 Lag	4 Lag
$GoogleTrendsIndex_{t-1}$	-0.0252***	-0.0466***	-0.0520***	-0.0028***
	(0.0096)	(0.0121)	(0.0129)	(0.0007)
$GoogleTrendsIndex_{t-2}$		0.0360**	0.0302*	0.0016 *
		(0.0154)	(0.0158)	(0.0008)
$GoogleTrendsIndex_{t-3}$			0.0074	0.0004
			(0.0130)	(0.0007)
$GoogleTrendsIndex_{t-4}$			0.0127	0.0007
			(0.0132)	(0.0007)
Constant	0.5009 ***	0.5032***	0.5062 ***	0.5088 ***
	(0.0002)	(0.0004)	(0.0006)	(0.0217)
Stock Fixed Effects	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes
N	6,524	6,494	6,434	6,434
R^2	0.0009	0.0018	0.0024	0.0020

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Figures

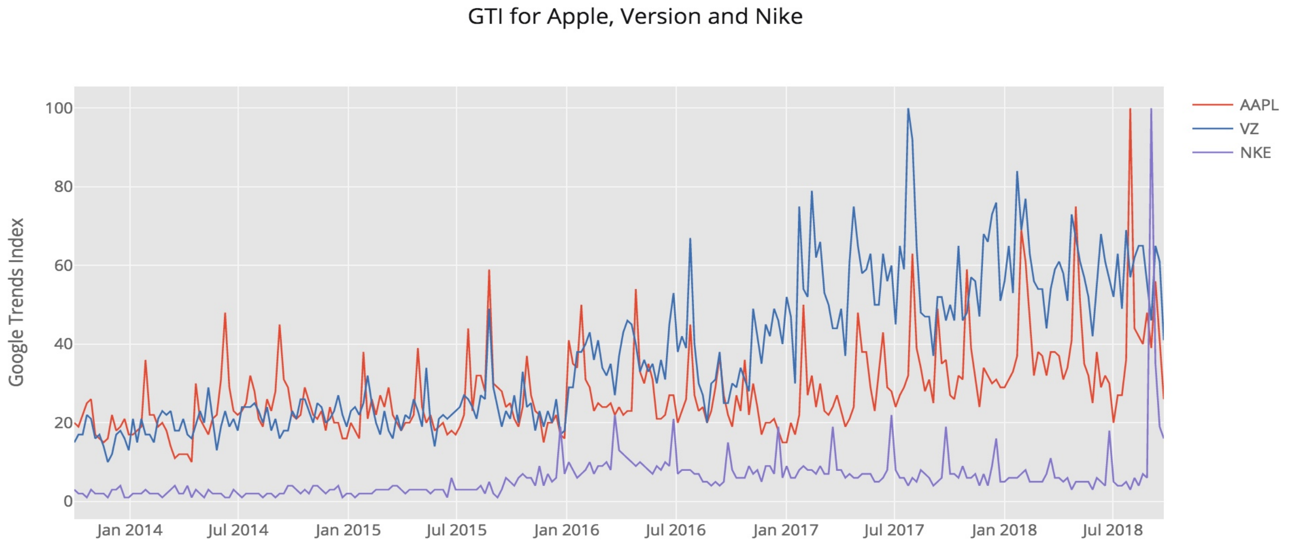

Figure 1: Non Standardized Google Trend Index

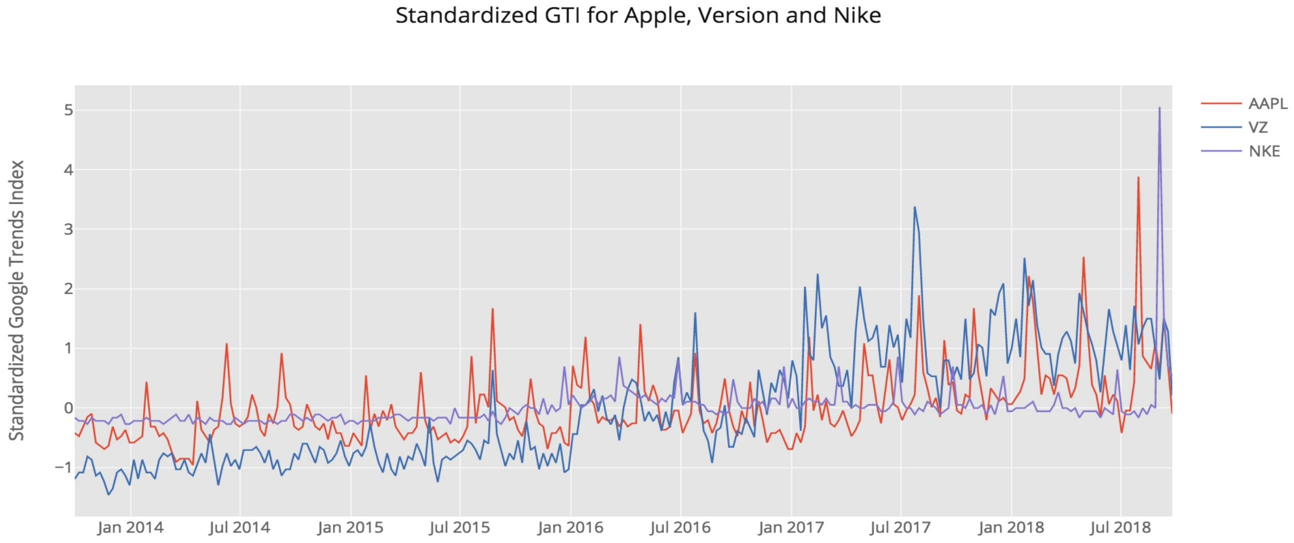

Figure 2: Standardized Google Trend Index

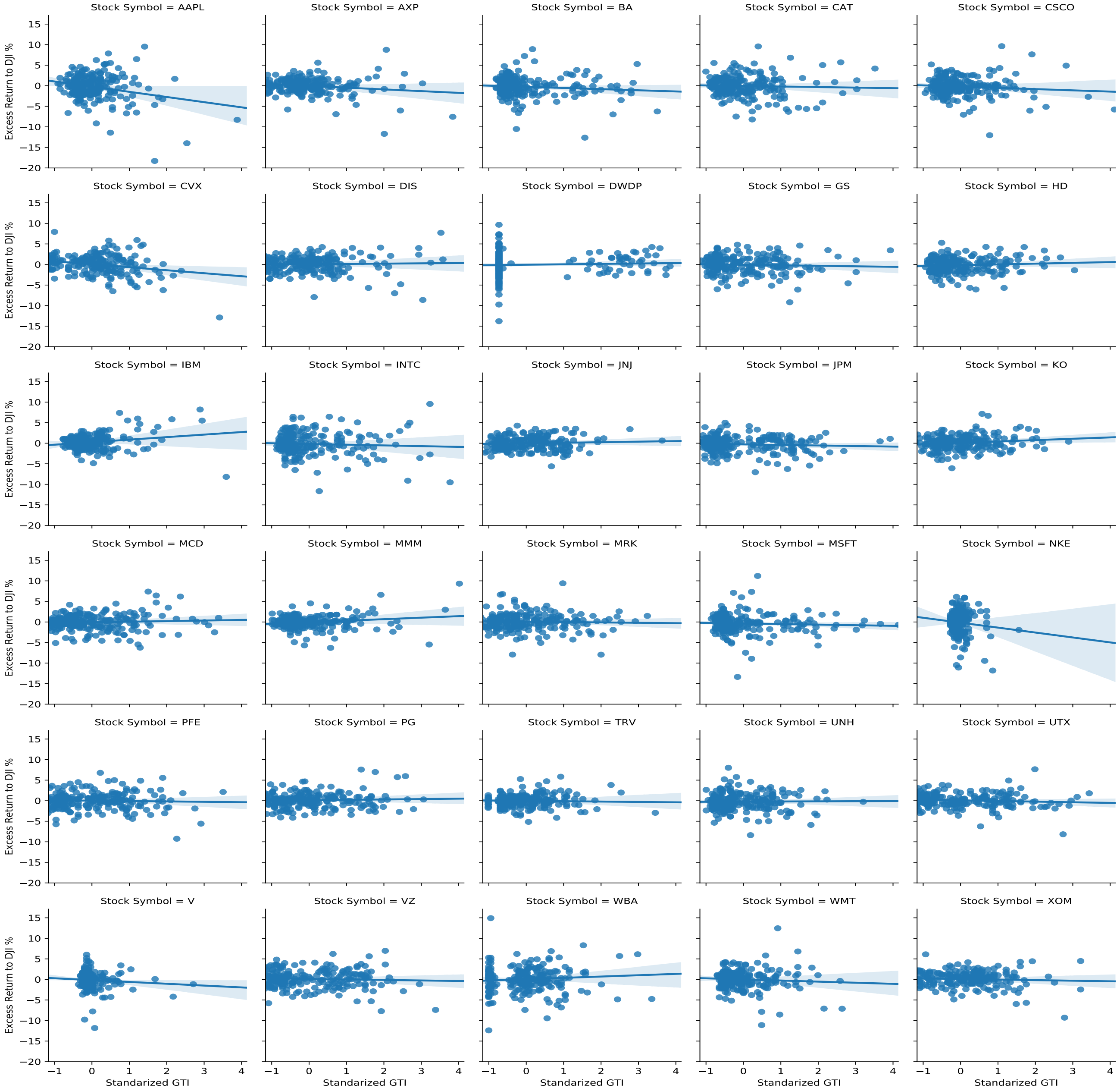

Figure 3: Individual Stock's Excess Weekly Return and Standardized GTI

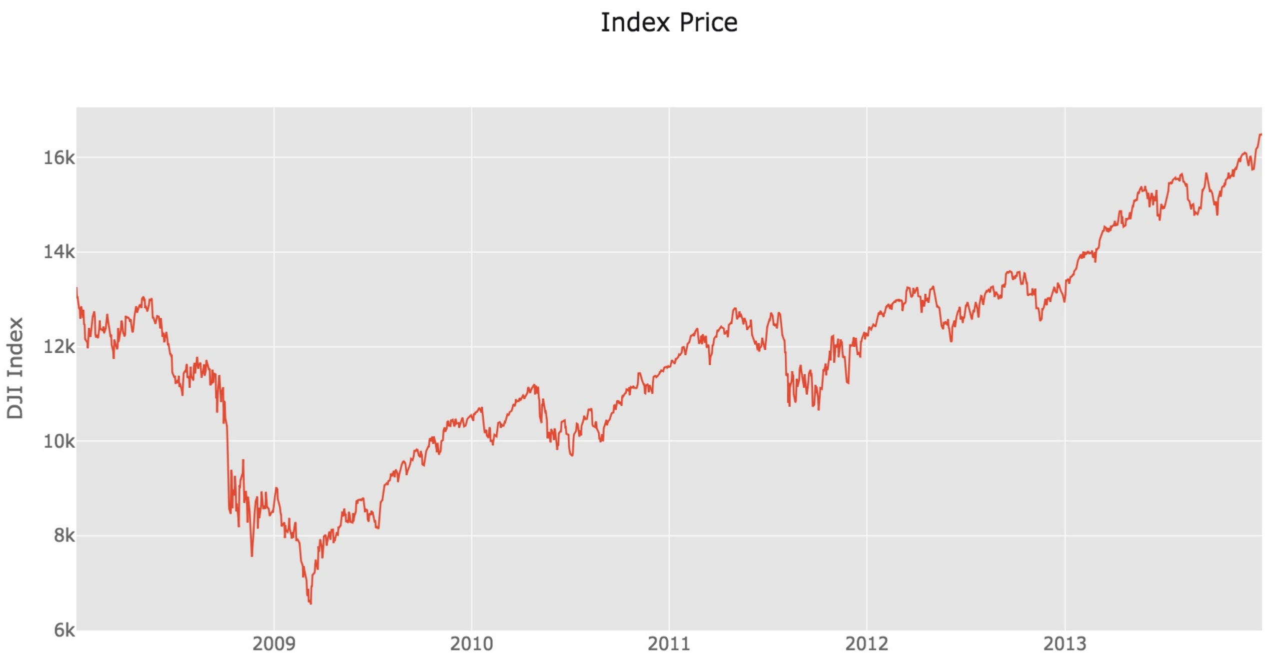

Figure 4: DJI Index 2008-2013

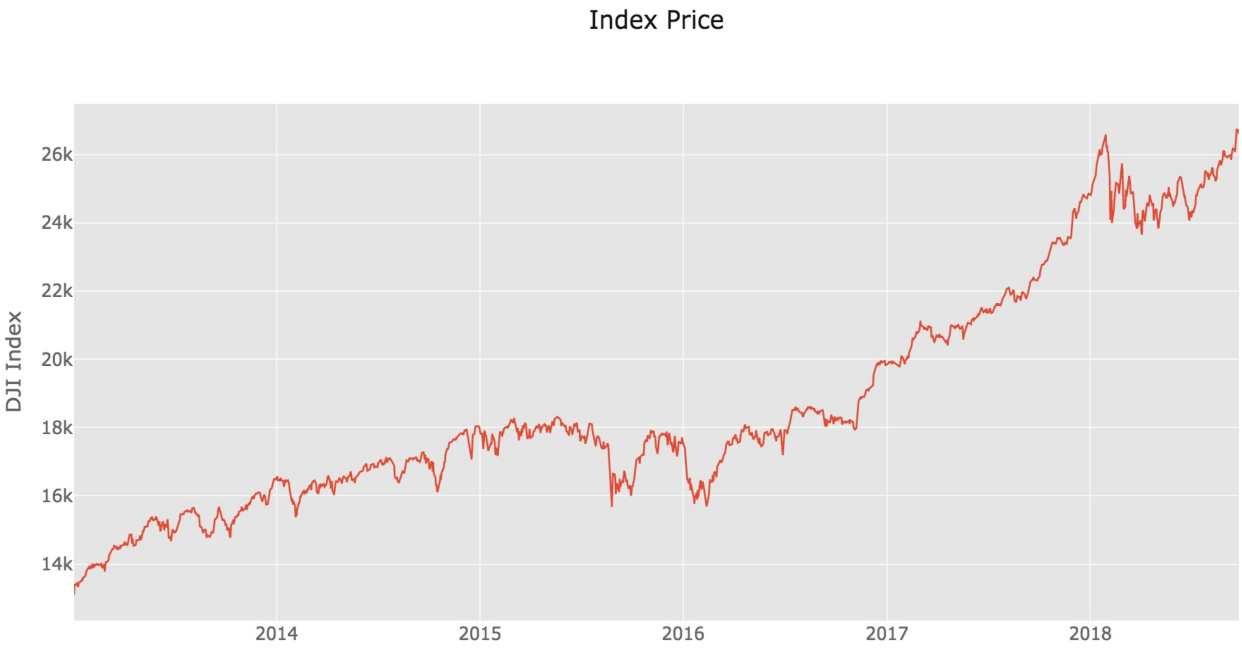

Figure 5: DJI Index 2013-2018

Appendix

Table A1: Stock name from DJI and corresponding Google trends index search term.

Company Name	GTI Search Term
3M Company	MMM
American Express Company	AXP
Apple Inc.	AAPL
Caterpillar Inc.	CAT
Chevron Corporation	CVX
Cisco Systems, Inc.	CSCO
DowDuPont Inc.	DWDP
Exxon Mobil Corporation	XOM
Intel Corporation	INTC
International Business Machines Corporation	IBM
Johnson & Johnson	JNJ
JPMorgan Chase & Co.	JPM
McDonald's Corporation	MCD
Merck & Co., Inc.	MRK
Microsoft Corporation	MSFT
NIKE, Inc.	NKE
Pfizer Inc.	PFE
The Boeing Company	BA Stock
The Coca-Cola Company	KO Stock
The Goldman Sachs Group, Inc.	GS Stock
The Home Depot, Inc.	HD Stock
The Procter & Gamble Company	PG Stock
The Travelers Companies, Inc.	TRV
The Walt Disney Company	DIS
United Technologies Corporation	UTX
UnitedHealth Group Incorporated	UNH
Verizon Communications Inc.	VZ Stock
Visa Inc.	V Stock
Walgreens Boots Alliance, Inc.	WBA
Walmart Inc.	WMT

Table A2: Summary of the Data

Number of Stocks	30 (DJI)
Time frame	09-30-2013 to 09-24-2018 (5 Years)
Number of t (Weeks)	229
Total Number of Observations	6,524
Stock Return Summary	
Mean of Excess Return	-0.057 %
Mean of Excess Return to DJI (%)> 0 (Dummy)	0.5004
GTI Summary	
Mean of GTI	28.73
Standard Deviation of GTI	18.59