MASTER OF SCIENCE IN QUANTITATIVE FINANCE

TEAM 7 GROUP PROJECT

Research Methods for Quantitative Professionals

Author:
Padey Tan
Jonathan Wee
Le Thu Trang
Song Lu
Teo Wan Chun Jotham

Supervisor:
Dr. Scott Christopher TRELOAR

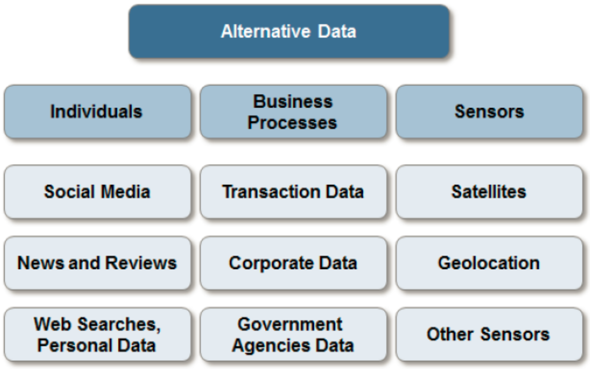
May 1, 2018

Contents

1 Introduction

Bitcoin is the first system of money not controlled by any entity and completely decentralized. It is the genesis of the blockchain movement started by the elusive Satoshi Nakamoto. Bitcoin provides a system of decentralized trust for value transfer and its popularity has garnered much attention since its inception in 2011. Traders and investors alike, have been incorporating bitcoin into their portfolios. Even large banks have acquired bitcoin exchanges and started cryptocurrency trading desks. However, unlike the usual financial assets or instruments where traditional data (e.g. quarterly earnings calls, analyst reports and financial news) is available through financial institutions, there has been a dearth of such resources on Bitcoin.

Today, the growing availability of alternative data has provided investors deeper insights of companies beyond information culled from traditional sources. The spectrum of alternative data ranges from social media, news, geo-locations, sensor data to satellite imagery. Social media, in particular, has been a rich source for new information for investors and particularly for Bitcoin.



Alternative Data

Individuals	Business Processes	Sensors
Social Media	Transaction Data	Satellites
News and Reviews	Corporate Data	Geolocation
Web Searches, Personal Data	Government Agencies Data	Other Sensors

Source: J.P.Morgan Macro QDS.

In Bitcoin's context of a decentralized system where the masses have access to, individuals are major producers of alternative data through their social media activity, forum content and web searches. Twitter is one such example. With a strong following of 330 million monthly active users, it has been a popular source of bitcoin news and official updates from famous influencers in the cryptocurrency space. These tweets have been providing a window into the public's thoughts and impressions, which are not represented in daily news coverage.

Google trend and exchange page views are other sources of alternative data that might be useful in deciphering the masses' opinion on Bitcoin. There have been earlier literature that showed positive correlation between Bitcoin prices and Google trend searches on Bitcoin using data in 2015. Other authors have explored on twitter sentiments with stocks and we intend to do likewise to Bitcoin. Bitcoin's decentralized system has also provided other network data online. Unique Bitcoin address used per day is one of such data. A Bitcoin address is an alphanumeric string, required to receive any form of payment in Bitcoins. Each user may have more than one Bitcoin address and the activity of the address is readily available online. In our research, we are keen in exploring the usage of unique Bitcoin's address per day in relation to Bitcoin value.

Our project seeks to validate if these correlations (Google trends and Twitter sentiments) still holds true today using recent data from 2017 and 2018. We are also interested to unveil if these alternative data are "causes" and effectively, forecasters of Bitcoin value by carrying out causality tests. We also expand beyond social media and web traffic data to transaction data like the use of Bitcoin addresses.

In short, our research will take on a quantitative approach to:

1. Explore relationships between alternative data/factors (Google trend interest level, Tweet sentiments, Exchange page views and BTC addresses used per day) and Bitcoin price or returns.
2. Unveil if these factors are causes or effects of the rise or fall of Bitcoin's value.
3. Transform our findings into a potential predictor of Bitcoin's price

2 Literature Review

2.1 Literature Review I

Twitter mood predicts the stock market:

- **Objective**

  The objective of this paper is to examine whether measurement of social mood as derived from twitter tweets has any correlation or predictive signal on Dow Jones Industrial Average (DJIA) over time. Using historical data of DJIA values and public mood time series, the author will quantify its relationship through various tests and modelling.

- **Methodology**

  The initial phrase is to collect mood-based information, which includes tweet identifier, posting time, and text meaning after filtering all spam messages and other information-based tweets. Using sentimental tools such as Opinion Finder and Google-Profile of Mood States [GPOMS], quantify the tweets. Opinion Finder will separate tweets into positive and negative mood, and with another tool GPOMS to capture 6 different mood: Calm, Happy, Sure, Kind, Vital and Alert. Using these time-series together with DJIA to run various model to understand their relationship. Modelling such as multiple regression, bivariate granger causality and non-linear model, Self-organizing Fuzzy Neural Network model, for forecasting were conducted. The multiple regression is conducted to quantitatively determine the relationship between GPOMS's mood dimensions and the Opinion Finder mood trends based on the statistical results. And for the bivariate granger causality analysis to testing whether any of the time series of the moods has any predictive information of the DJIA time series. To overcome the shortcomings of the linear model, a non-linear approach is to supplement the previous results.

- **Findings**

  The author mentions 3 key worthy results after inputting its twitter dataset into different modelling tests. For multiple linear regression results show that moods such as Sure, Vital and Happy are significantly correlated whereas the remaining moods are not. Based the causality analysis results, the null hypothesis that the mood time series do not predict DJIA values with a high level of confidence can be rejected. The mood with the highest predictive value with regards to the DJIA is the calm mood from the GPOMS tool. The non-linear modelling indicates, although individual moods may not be able to predict DJIA movements but by combining the different moods or factors together, it can still show significant level of prediction.

- **Discussion**

Some limitation is the data set may not fully represent the public opinion on DJIA prices as the tools completely ignores implicit sentiment statements. To widen the scope of detecting implicit sentiments, data from Google trends or public forums could be studied further. Perhaps a longer window timeframe could be studied as well. Longer timeframes might reveal insights of tweets showing consistency in predicting the DJIA. Additionally, the author assumes that the general public is currently as strongly invested in the DJIA as financial experts, and their opinion or moods will directly affect their investment decisions. Another indicator or factor can be added into the modelling, a recommend would be to use Google trend or financial news online to further improve its significance.

2.2 Literature Review II

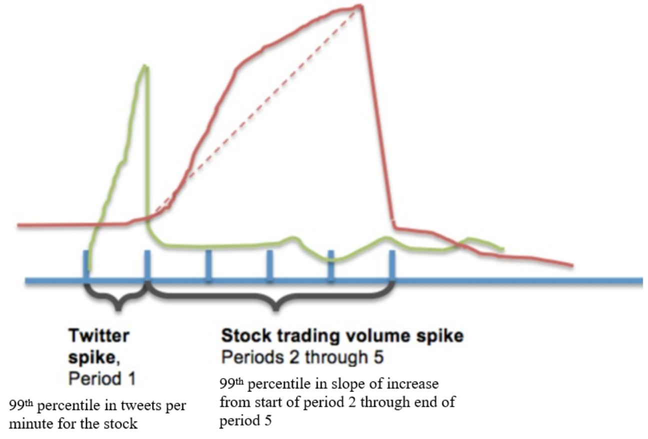
Real-Time Diffusion of Information on Twitter and the Financial Markets :

- **Objective**

In this article, the authors' objective was to examine real-time relationship between spikes in Twitter posts and stock trading activities. In contrast to the Literature review 1, the volume of stocks traded and the quantity of social chatter that revolves around the firm were used instead of sentimental analysis on just the index. The article took further steps to quantify on the relationship between Twitter news and intra-day movement of certain stock prices and to improve predictive power of models via the use social feeds at the firm level.

- **Methodology**

For the data collections, all Tweet that contain names of 96 common firms listed in Nasdaq 100 index were collected within a 10-minute interval. To prevent consumer interest data biases, the paper excluded the most popular household firms in Nasdaq like Microsoft, Intel, Google, Facebook and Apple. By considering only spikes in Twitter activity as unusual. This article examined statistical outlier events with the goal of understanding the relationship between such events and trading activity in a very short time. Each of them will be assigned to a random ten-minute increment matching on the firm which will input into a complex model used for event study and intra-day trading selection. Considering the issue that statistical significance might be a result of the large sample size, the authors repeated these steps with different sub-samples each time.


Twitter spike, Period 1
99th percentile in tweets per minute for the stock
Stock trading volume spike Periods 2 through 5
99th percentile in slope of increase from start of period 2 through end of period 5

Measurement of spikes in Twitter activity and trading volume for a firm.

- **Findings**

  The paper concludes that within the timeframe, Twitter volume would peak before surge of trading volumes at firm level. However, they were unable to confirm if Twitter was the sole cause. The authors hope that their results would be useful for investors to exploit the lag in the news release and the respective market reaction to the stock. On a more realistic view, the authors emphasise the difficulty in applying this idea to achieve financial benefits.

- **Discussion**

  Since this article did not consider sentiments of tweets, the cause of the surge in stock trading volume was also not explored. Further improvement on sentimental analysis can be done by adding factors or tools such as Google Trends or Opinion Finder. Although, it common practice to remove outliers from regression in order not to skew the test findings, this article only analyses extreme events that statistical outliners which may or may not be fully representative of the firm interest. Again, the same methodology could be applied with longer window size (e.g. 1 or 2 year period) and examine if conclusions still hold.

2.3 Literature Review III

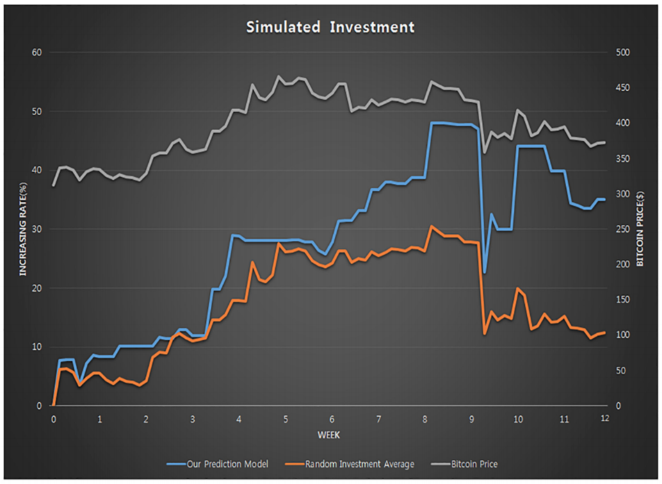
Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies:

- **Objective**

  The article objective is to predict the fluctuations in cryptocurrency's prices and transactions by analysing user comments via online cryptocurrency communities. Bitcoin, Ethereum and Ripple was chosen as the article focus given its' market size and popularity. The article's hypothesized is that the user comments in certain online cryptocurrency communities may affect fluctuations in their price and trading volume.

- **Methodology**

  The initial part the collection of user tweets, next is the data was process using an open source sentimental analysis algorithm, VADER, which classified the user sentiment into very positive, positive, neutral, negative and very negative. For data selection, the Granger causality test was adopted, where community opinions with the highest Granger causality relation (p-value < 0.05) were extracted. The Granger test is to analysis whether the time series of a community of opinions contained predictive information regarding the fluctuations in cryptocurrency prices. After extract of data, the authors utilised a form of classification machine learning technique to develop a predictive model and backtested it in a simulated investment environment.



- **Findings**

The articles have identified that opinions affecting price fluctuations varied across cryptocurrencies. For Bitcoin, positive user comments have a significant impact in the price fluctuations. Whereas, the other two currencies were significantly influenced by negative user comments and replies. The prediction model results show that user opinions proved useful to predict the fluctuations in 6 7 days. The predicted result was most precise in Bitcoin, which seems attributable to the amount of accumulated data and animated community activities, which exerted a direct effect on fluctuations in the price of the cryptocurrency. These findings suggest that the difference in community sizes may have direct effects on fluctuations in the price of cryptocurrencies.

- **Discussion**

In conclusion, the author utilised user comments and replies which have been shown to be less precise than social network data or Google search volumes at delivering results. In additional, the data used for predictive modelling was put through several stages of data selection which may not necessarily reflect the full market sentiment.

2.4 Literature Review IV

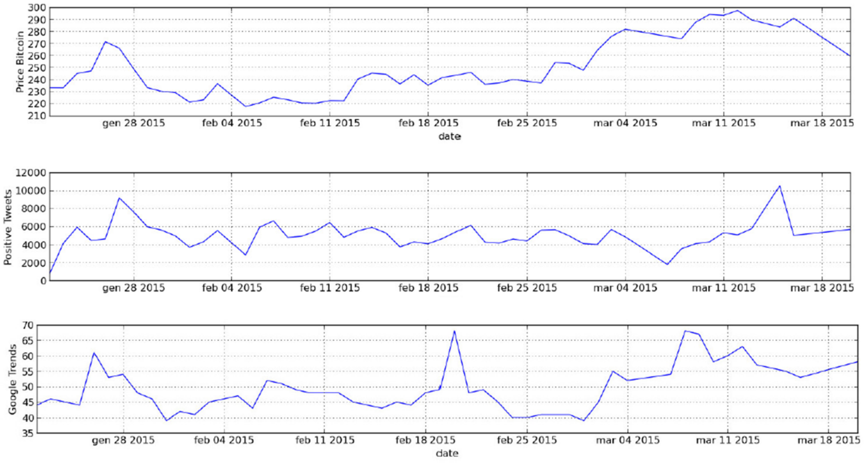
Bitcoin Spread Prediction Using Social And Web Search Media:

- **Objective**

In this article, an investigation on whether the spread of the Bitcoin's price is related to the volumes of tweets or Web Search media results. A comparison analysis was done between the trends of Bitcoin's price with Google Trends data, volume of tweets and particularly with those that express a positive sentiment.

- **Methodology**

Application of automated Sentiment Analysis on tweets scrapped online to evaluate whether public sentiment could be used to predict Bitcoin's market. Another factor used was the utilization of Google Trends to analyse Bitcoin's popularity under the perspective of Web searches volume. For a period of 60 days, variation of Bitcoin price was compared with that of tweets volume, that of positive-mood tweets volume and that of Google Trends data. The only test done on the dataset was cross-correlation.



- **Findings**

  Three-time series were produced: tweets volume, tweets with positive mood volume and Google Trends data. From results of a cross correlation analysis between these time series, the paper affirmed that positive-mood tweets may contribute to predict the movement of Bitcoin's price in 3-4 days and that Google Trends could be seen as a kind of predictor, because of its high cross correlation value with a zero lag. From these results of a cross correlation analysis between these time series, the authors affirm that positive tweets may contribute to predict the movement of Bitcoin's price in a few days. Whereas, Google Trends could be seen as a kind of predictor, because of its high cross correlation value with a zero lag.

- **Discussion**

  The methodology in this article takes a rather simplistic approach to the sentiment analysis and sentiment strengths are not quantified either. While this is useful for establishing correlation, sentiment is often subjective, given that certain people may have greater influence, and the results may not be precise. Additionally, the research is only conducted in a 60-day period which may not be a significant representation of the general market behaviour.

3 Data Collection

3.1 Bitcoin Market Data

Similar to traditional stock exchanges, bitcoins are traded on a bitcoin exchange in multiple fiat currencies or alternative cryptocurrencies. We have chosen to use exchange rates on Bitfinex (Bitcoin versus USD) as it is one of the largest cryptocurrency exchanges in the world. The data was obtained from the database Quandl for ease of collection and returns transformation using their API. Bitcoin prices are transformed into log returns and tested for stationarity using the Dickey-Fuller test to accommodate the required statistical properties for more accurate test results (e.g. Granger-Causality Analysis).

3.2 Google Search Interests

One source of alternative data would be in the form of web searches. Google is the most-used search engine on the World Wide Web and handles 3 billion searches each day. A free service called Google

Trends displays how often a particular search-term is searched for relative to the total number of searches for their search engine. Our group selected the keyword “Bitcoin” and retrieved the weekly search interest data from the Google Trends API.

3.3 Twitter Sentiment Data

Twitter is a social media network which has grown rapidly as an important tool for businesses and individuals to share comments and information including financial news and investment decisions. Prior research by Bollen, Mao and Zeng (2011) has shown that the emotional content of tweets can be useful in predicting market movements. As a rich source of real-time alternative data, sentiment of the information diffused through Twitter can be useful in determining the overall market perspective towards cryptocurrency and bitcoin.

3.4 Twitter historical data

For the collection of historical tweets within our specified timeframe, our team utilised a python library called “TwitterScraper” to retrieve a sample size of approximately 10,000 tweets a month based on the selected keywords “bitcoin” and “BTC”. We were able to extract a total of 152,293 tweets which were then resampled into weekly data for sentiment analysis.

3.5 Sentiment Analysis Tool – VADER

Valence Aware Dictionary and sEntiment Reasoner (VADER) is a lexicon and rule-based sentiment analysis tool suited towards analysing and quantifying social media text sentiments. The algorithm detects the polarity and intensity of sentiments in texts and generates a normalized sentiment score between -1 (most negative) and +1 (most positive). By putting the weekly resampled Twitter historical data through the sentiment analysis tool, we were able to derive the overall weekly Twitter sentiment based on the sample data retrieved.

3.6 Pageviews

Web traffic and pageviews can also be another source of alternative data. Pageviews record the total number of times a webpage is visited within a selected time period. Estimations of pageviews for Coinbase.com, a highly reputable bitcoin exchange, during the investigation period were downloaded using Alexa which is an analysis software developed by Amazon.com. The number of pageviews would an indicator for the general trend of investors’ interests in bitcoin prices during the period.

3.7 Bitcoin Unique Address Used

To receive bitcoin payments, users are required to create Bitcoin addresses to direct payments from their counterparties. The number of bitcoin unique addresses used for each period can then be utilised as a measure of individual transaction data for that period. This information is provided by Blockchain.info which deals in Bitcoin statistics and market information, and retrieved using Quandl’s API.

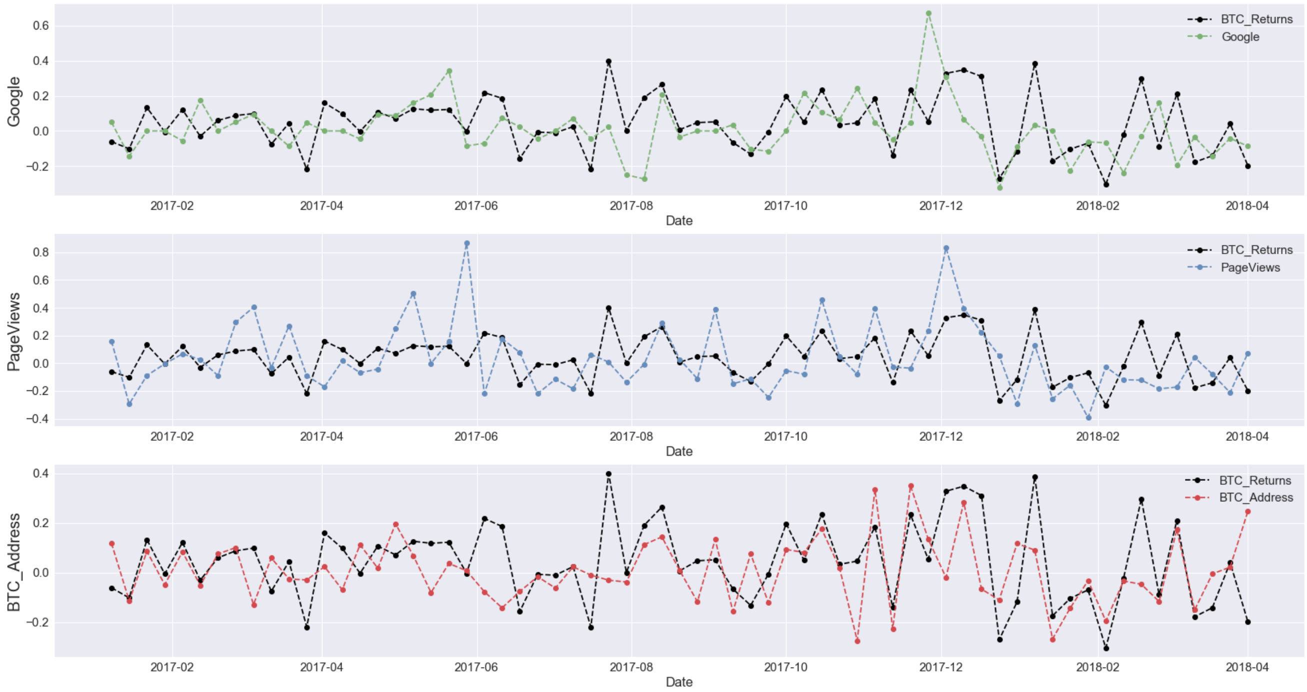
4 Methodology and Results

Despite earlier literature in 2015 reporting positive correlations between Google trend data and tweets' sentiment with Bitcoin price, our research challenges if these findings still carry water with recent data and a longer time frame of study (1 January 2017 to 31 March 2018). On a more important note, what distinguishes our research from other paper is the fact that we use Bitcoin returns instead of price.

Prices are, unfortunately, not stationary and persistent of 'shocks' will be many in a non-stationary series. Shocks are evident in the price history of Bitcoin and as such we use Bitcoin returns that allow shocks to the system to gradually die away. This also eliminates any attribution to spurious regressions: if two variables are trending over time, a regression of one on the other could have a high R2 even if the two are unrelated. In addition, when running regressions, it is not possible to validly undertake hypothesis tests about the regression parameters if data are stationary. We then carried out Augmented Dickey-Fuller test as confirmation that Bitcoin returns follow a stationary process.

4.1 Time Series

In this part, we plotted the weekly time series overlay for simple returns as below, to provide us an glimpse of how BTC returns are related to percentage change of the 3 factors:



By intuitively running through the graph, we saw that there exists similar trends between BTC and the 3 factors. An estimation about the relative lag period of the overlays was conducted, following which certain rational assumptions was made.

- **BTC Return** vs **Google Trend percentage change**
  There exists a 2-week lag between BTC and Google Trend. And we could make a fairly reasonable assumption regarding this: If you think of yourself as a potential Bitcoin investor or speculator, before you really kick into this trading activity of cryptocurrency, you might as well search about it thoroughly in Google, in order to make a solid decision; After that, registration for certain trading platform may take a while as well; The whole process might aggregate to around 2 weeks.
- **BTC Return** vs **Page views percentage change**
  There exists a 1-week lag between BTC and PageViews. Once again, think about yourself as a keen BTC investor, one thing you normally do on a weekly basis is checking the BTC exchanges pages; BTC price increases as investors become more interested in crypto-currency trading and checked exchange pages more frequently.
- **BTC Return** vs **BTC Address percentage change** There exists a 4-week lag between BTC and BTC Address.

4.2 Dickey-Fuller

If two variables are trending over time, a regression of one on the other could have a high R^2 even if the two are totally unrelated. The stationarity or otherwise of a series can strongly influence its behavior and properties, e.g., persistence of shocks will be infinite for non-stationary series. If the variables in the regression model are not stationary, then the standard assumptions for asymptotic analysis will not be valid. In other words, the usual "t-ratios" will not follow a t-distribution, so we cannot validly undertake hypothesis tests about the regression parameters. In order to obtain a valid regression, we conducted the Augmented Dickey-Fuller test as below:

$$\Delta y_t = \psi y_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta y_{t-i} + \mu_t$$

$$test\ statistic\ = \frac{\widehat{\psi}}{SE(\psi)}$$

	BTC_Returns	Google	PageViews	BTC_Address
ADF Statistic	-8.272954	-5.796783	-6.705582	-10.247936
P-Value	-8.272954	-5.796783	-6.705582	-10.247936
Critical Values 1%	-3.536928	-3.536928	-3.536928	-3.536928
Critical Values 5%	-2.907887	-2.907887	-2.907887	-2.907887
Critical Values 10%	-2.591493	-2.591493	-2.591493	-2.591493

[Table 1: Dickey-Fuller Test]

ADF Statistics of the four series (simple returns) are all less than the 1% critical value -3.54, implying that they are all stationary time series for the 99% confidence interval.

4.3 Single Factor Linear Regression

Single Factor Simple Linear Regression:
Based on the stationarity, we conducted single factor simple linear regression in the following section to investigate the relationship between BTC price and the 3 factors. In addition, another factor,

Twitter Sentimental Analysis, was tested and produced quite an unexpected result.
The fitted linear regression model can be written as:

$$R = \alpha + \beta R_{factor} + residue$$

We implemented OLS for **BTC Return vs Google Trend percentage change** as an example for illustration:

T-static for the regression coefficients (95% confidence interval): The null hypothesis is written as,

$$H_0 : \alpha = 0 \quad \text{vs} \quad H_1 : \alpha \neq 0$$
$$H_0 : \beta = 0 \quad \text{vs} \quad H_1 : \beta \neq 0,$$

t-static for α :

$$t_\alpha = 2.028 > t_{critical} = 1.9718$$

t-static for β :

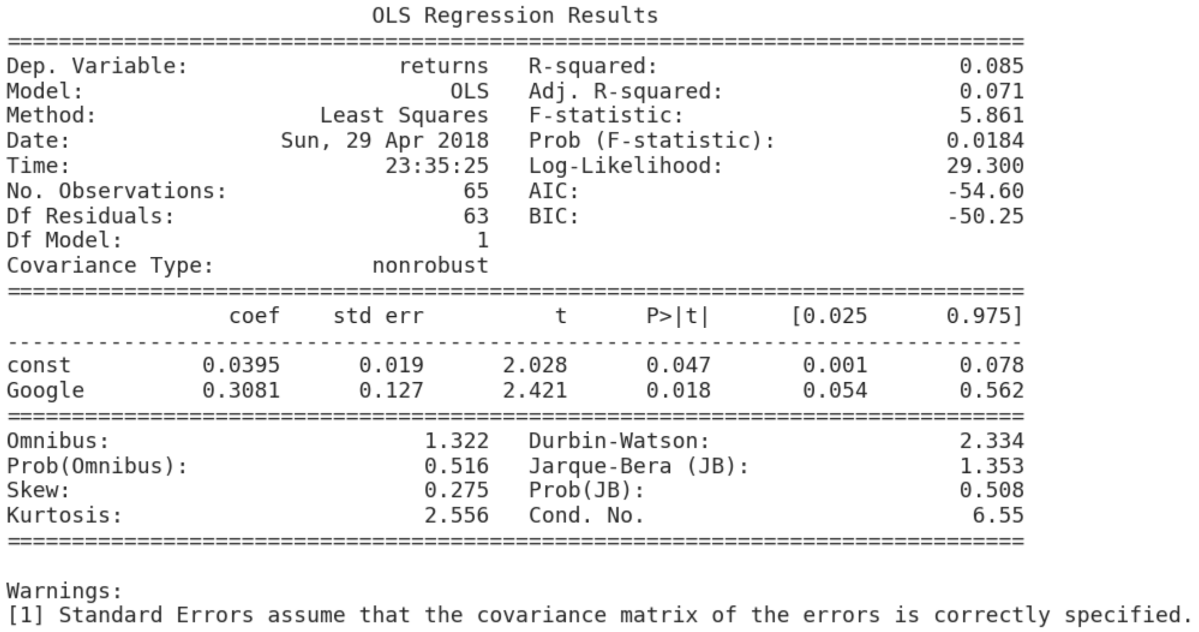
$$t_\beta = 2.421 > t_{critical} = 1.9718$$

Thus, we reject the NULL hypothesis for both α and β, which implied that there do exist positive linear correlation between BTC return and Google Trend percentage change, and the slope coefficient β is unlikely to be 0.

F-static for the regression coefficients: The NULL hypothesis for F static in regression is that all of the regression coefficients are equal to zero. In other words, the model has no predictive capability. Basically, the f-test compares our linear regression with zero predictor variables (the intercept only model), and decides whether the added slope coefficient improved the model. Since $f_{static} = 5.861$ and $p \approx 0$ in our case, the β coefficients included in our model improved the model's fit, further confirming that the slope coefficient is unlikely to be 0.

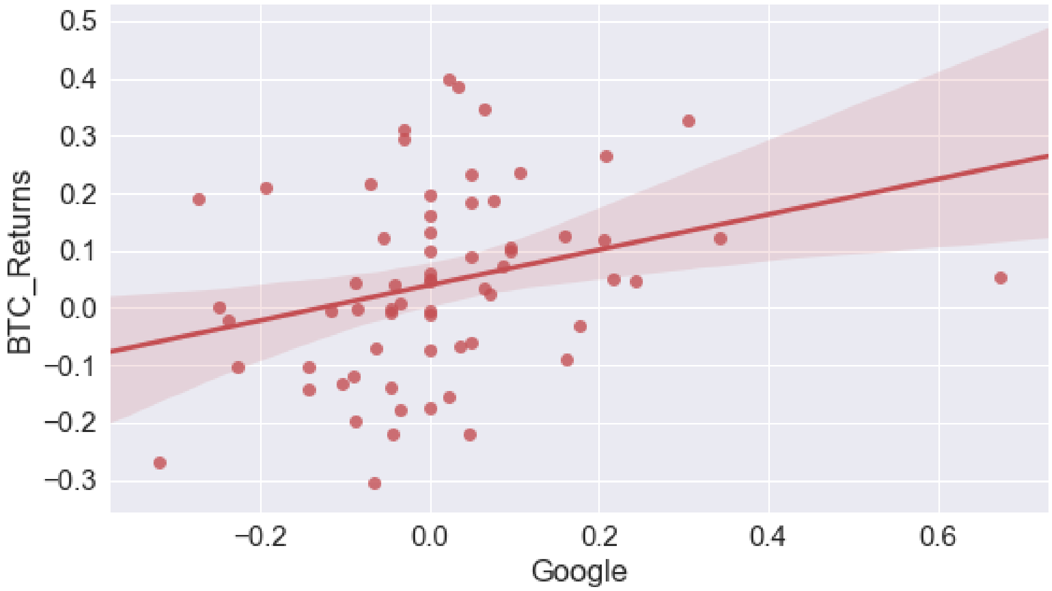
We conducted OLS for the rest two factors and obtained similar results, listed as below:

1) `BTC Return vs Google Trend percentage change`

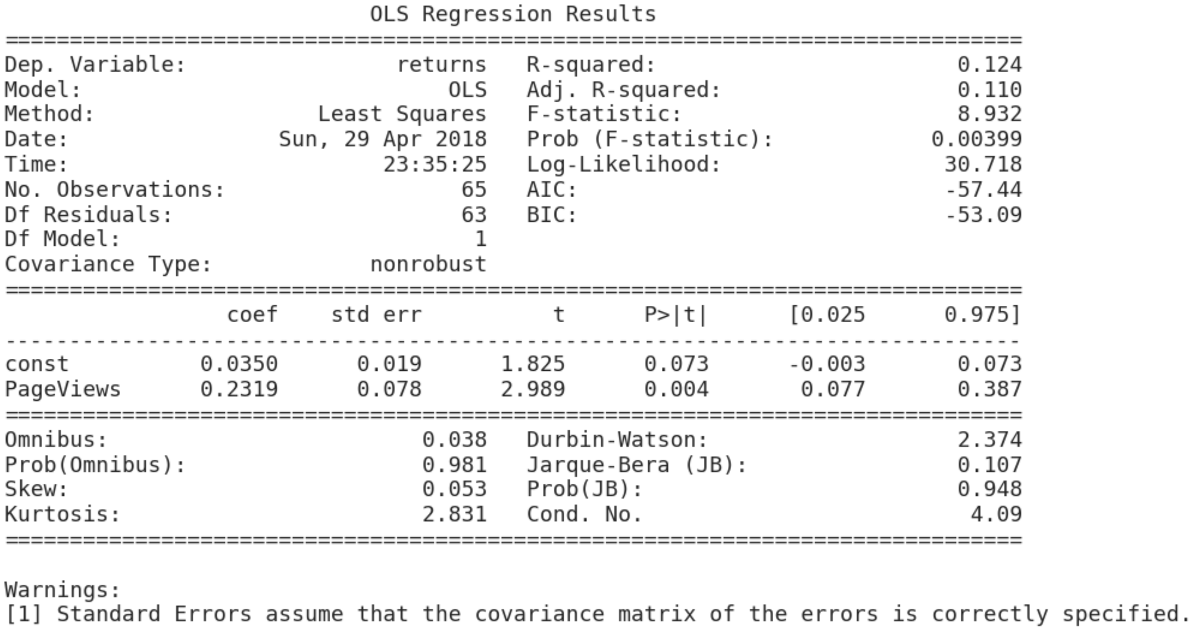


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                returns   R-squared:                       0.085
Model:                            OLS   Adj. R-squared:                  0.071
Method:                 Least Squares   F-statistic:                     5.861
Date:                Sun, 29 Apr 2018   Prob (F-statistic):             0.0184
Time:                        23:35:25   Log-Likelihood:                 29.300
No. Observations:                  65   AIC:                            -54.60
Df Residuals:                      63   BIC:                            -50.25
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0395      0.019      2.028      0.047       0.001       0.078
Google         0.3081      0.127      2.421      0.018       0.054       0.562
==============================================================================
Omnibus:                        1.322   Durbin-Watson:                   2.334
Prob(Omnibus):                  0.516   Jarque-Bera (JB):                1.353
Skew:                           0.275   Prob(JB):                        0.508
Kurtosis:                       2.556   Cond. No.                         6.55
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

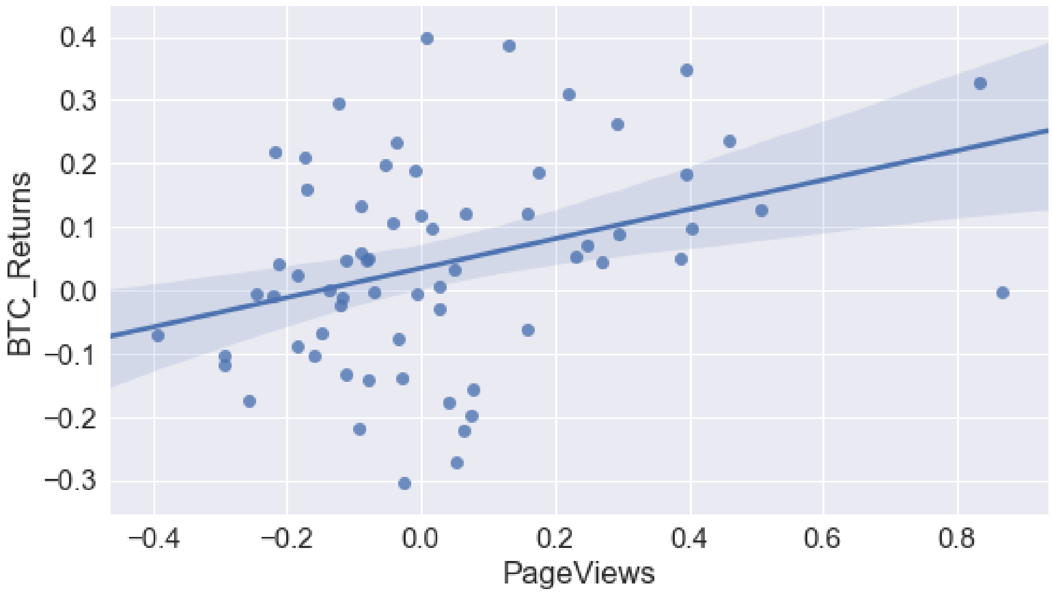


2) BTC Return vs PageViews percentage change

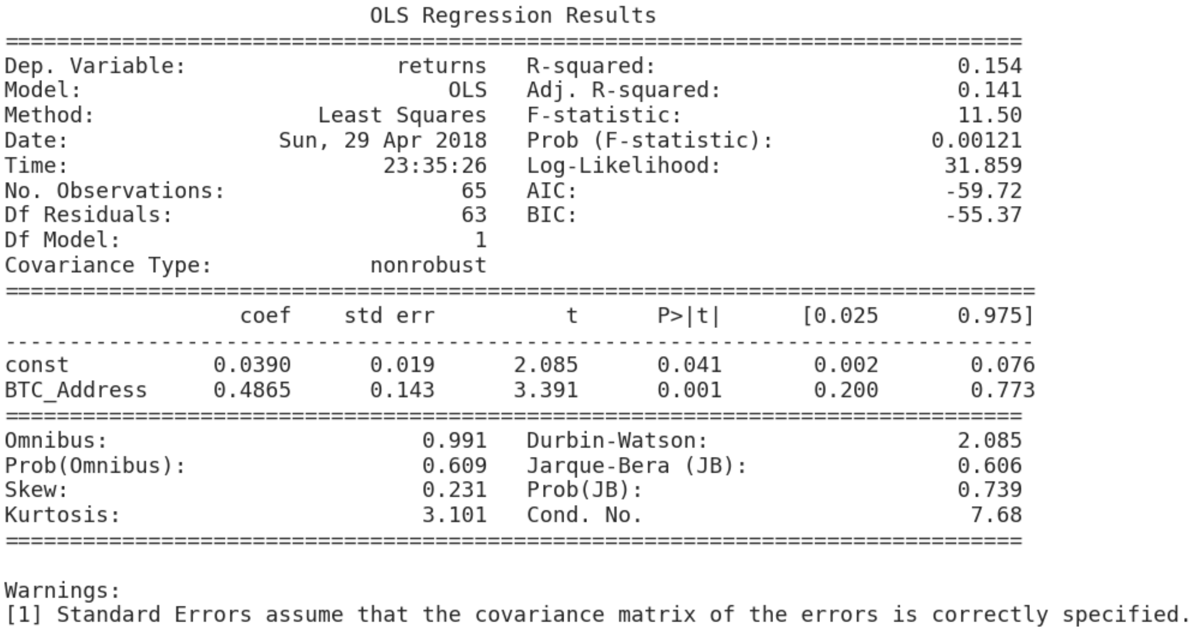


```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                returns   R-squared:                       0.124
Model:                            OLS   Adj. R-squared:                  0.110
Method:                 Least Squares   F-statistic:                     8.932
Date:                Sun, 29 Apr 2018   Prob (F-statistic):            0.00399
Time:                        23:35:25   Log-Likelihood:                 30.718
No. Observations:                  65   AIC:                            -57.44
Df Residuals:                      63   BIC:                            -53.09
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0350      0.019      1.825      0.073      -0.003       0.073
PageViews      0.2319      0.078      2.989      0.004       0.077       0.387
==============================================================================
Omnibus:                        0.038   Durbin-Watson:                   2.374
Prob(Omnibus):                  0.981   Jarque-Bera (JB):                0.107
Skew:                           0.053   Prob(JB):                        0.948
Kurtosis:                       2.831   Cond. No.                         4.09
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

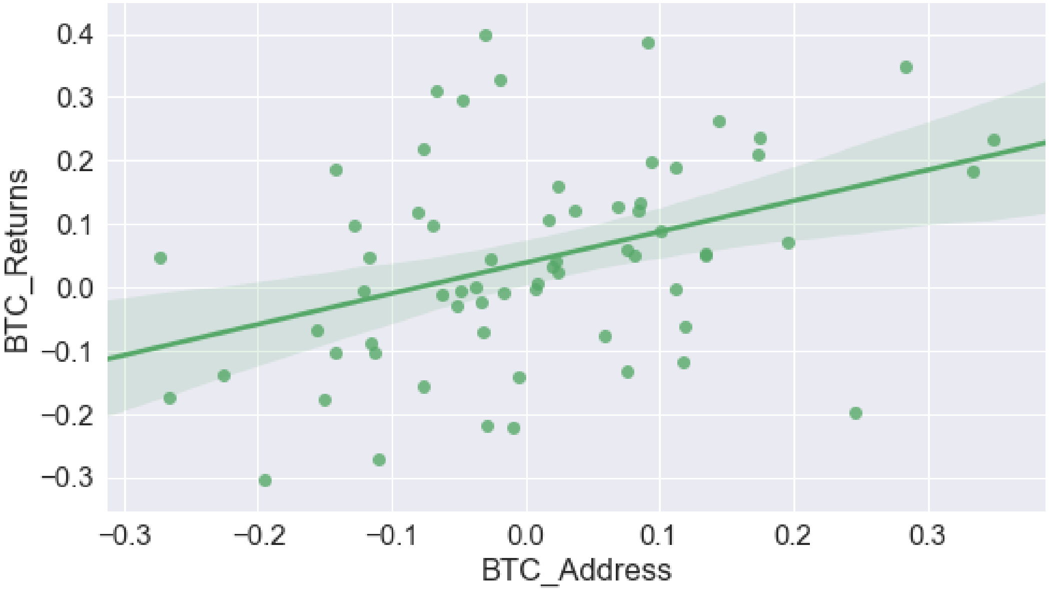


3) `BTC Return vs BTC Address percentage change`



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                returns   R-squared:                       0.154
Model:                            OLS   Adj. R-squared:                  0.141
Method:                 Least Squares   F-statistic:                     11.50
Date:                Sun, 29 Apr 2018   Prob (F-statistic):            0.00121
Time:                        23:35:26   Log-Likelihood:                 31.859
No. Observations:                  65   AIC:                            -59.72
Df Residuals:                      63   BIC:                            -55.37
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------
const           0.0390      0.019      2.085      0.041       0.002       0.076
BTC_Address     0.4865      0.143      3.391      0.001       0.200       0.773
==============================================================================
Omnibus:                        0.991   Durbin-Watson:                   2.085
Prob(Omnibus):                  0.609   Jarque-Bera (JB):                0.606
Skew:                           0.231   Prob(JB):                        0.739
Kurtosis:                       3.101   Cond. No.                         7.68
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```



Overall our linear regression produce significant results for the three factors: Google Trends, Pageviews and Bitcoin Address. Based on the above regression table, the most factor with the most significant results is BTC address. This review an insight that as Bitcoin prices increases, investors will either purchase more or sell existing Bitcoins for profit. Thus an increased in the activity of Bitcoin address has a permanent affect on Bitcoin returns. For Google and Pageviews however, it seems that public may just be interested in the acquiring knowledge of Bitcoin which may or may not lead to a transaction of purchasing or selling Bitcoin. Therefore, this concludes that Bitcoin address is statistically and intuitively better in predicting Bitcoin returns as compared to Google trends and Pageviews.

Twitter Sentimental Analysis Simple Linear Regression:

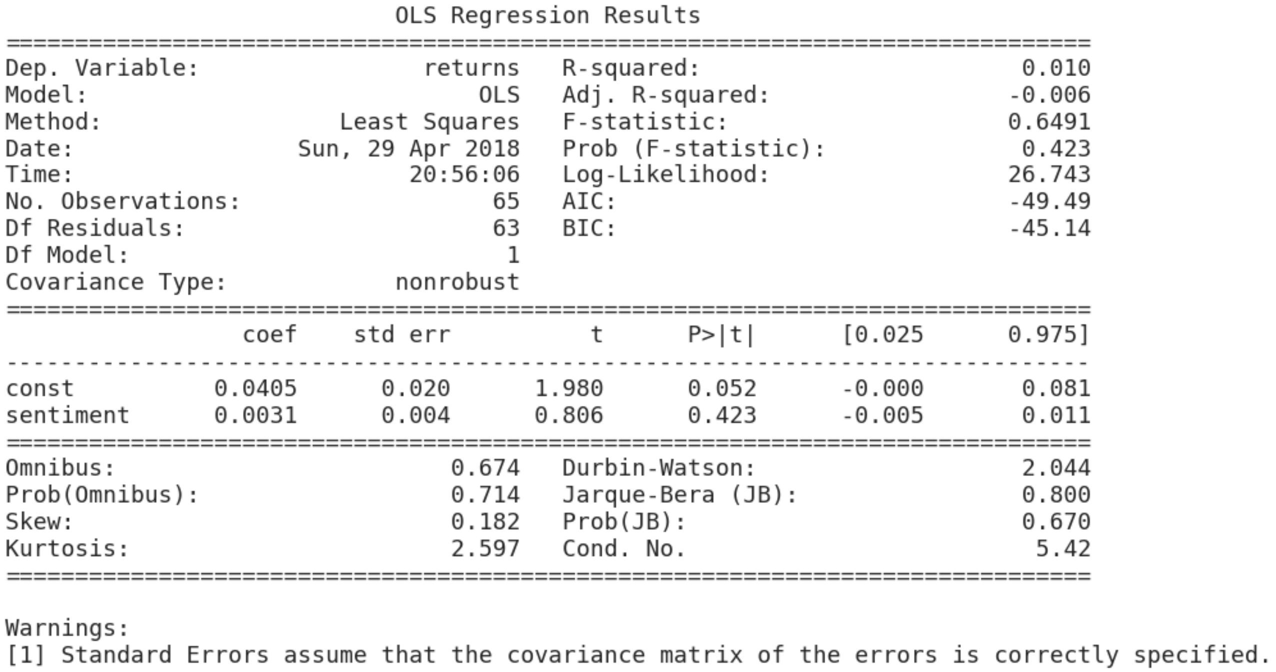
The collection Twitter Data, was based on based on the keyword "*Bitcoin*" in each twitter account. The reason for selection was due to their influence on the general public or investors who are planning to buy Bitcoin or do a Bitcoin transsaction.

Twitter Account Data is categories based on:

- Bitcoin News or Bitcoin Exchange Accounts: Cointelegraph, BitcoinNetworks, BitcoinMagazine, ForbesCryto, Coindesk, BTCTN, BTCnewsBOT, BTCNewsletter.
- Key Bitcoin Influencers Accounts: NeerajKA, VitalikButerin, aantonop, SatoshiLite, WhalePanda, NickSzabo4, gavinandresen, brianarmstrong, starkness, twobitidiot, lopp, rogerkver, Excellion, ErikVoorhees, TuurDemeester

After consolidation of data, it was input into an opensource Twitter Sentimental Analysis tools, Vader. With this results, we would be able to quantify the tweets ranging from $-1 to + 1$. Where $+1$ indicate extreme positive and -1 indicates extreme negative.

Lastly, transform this data into sentimental change against Bitcoin simple returns and run via a one factor linear regression test.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                returns   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                 -0.006
Method:                 Least Squares   F-statistic:                    0.6491
Date:                Sun, 29 Apr 2018   Prob (F-statistic):              0.423
Time:                        20:56:06   Log-Likelihood:                 26.743
No. Observations:                  65   AIC:                            -49.49
Df Residuals:                      63   BIC:                            -45.14
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0405      0.020      1.980      0.052      -0.000       0.081
sentiment      0.0031      0.004      0.806      0.423      -0.005       0.011
==============================================================================
Omnibus:                        0.674   Durbin-Watson:                   2.044
Prob(Omnibus):                  0.714   Jarque-Bera (JB):                0.800
Skew:                           0.182   Prob(JB):                        0.670
Kurtosis:                       2.597   Cond. No.                         5.42
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

From the above regression results, Tweets sentitmental analysis did not indicate any significant relationship with Bitcoin returns. The contribution to this results seems to be from the Vader tools used. The Vader tools has limitations such as it is not able to interpret sarcasm and also it penalise negative sentimental more heavily as compared to positive comments. Another issues with our approach, maybe due to insufficient data samples that we collected and the clean or removal of another language, such as chinese language or emoji icons. Overall, our project was not able to replicate the literature review results of Twitters tweets who had significant results.

4.4 Granger Causality

As we intuitively outlined the lag period for the return series overlays in the previous section, to further quantify the correlation and lag period between BTC price and the 3 factors, we conducted Granger causality test in the following.

	Google	PageViews	BTC_Address
statistic	[F-critical, F-test, P-value]	[F-critical, F-test, P-value]	[F-critical, F-test, P-value]
lag=1	[0.26596, 2.92538, 0.09228]	[0.26596, 4.81463, 0.03204]	[0.26596, 0.02166, 0.88347]
lag=2	[0.39976, 3.07497, 0.05379]	[0.39976, 2.56187, 0.08587]	[0.39976, 0.20928, 0.81178]
lag=3	[0.49657, 2.59186, 0.06188]	[0.49657, 1.3661, 0.26274]	[0.49657, 0.68469, 0.56522]
lag=4	[0.57371, 1.89101, 0.12593]	[0.57371, 0.87423, 0.48575]	[0.57371, 0.81295, 0.52268]

[Table 2: Granger Causality Test]

The Granger causality test is a statistical hypothesis test for determining whether one time series is useful in forecasting another. From above, we can concluded that:

- Google Trend provides statistically insignificant information about BTC return although the most optimal lag is in 2 weeks' time.

- PageViews provides statistically significant information about BTC return in 1 week's time.
- BTC Address provides statistically significant information about BTC return although the most optimal lag is in 4 weeks' time.

4.5 Pearson Correlation Test

Specifically, we want to know whether the each of the pairs are related and also how strongly they are related. We do that by calculating the Pearson's Correlation coefficient which would be the average of the simple return or percentage change.

	Google	PageViews	BTC_Address
Coeff	0.291732	0.352377	0.392903
P-value	0.018378	0.003992	0.001206

[Table 3: Pearson Correlation Test]

The Pearson Correlation Coefficient, P-value is shown in Table x. The correlation between the factors series and bitcoin return ranges from 0.29 to 0.49 which indicates some linear depen- dence between them. Google trends seem to have the least significant results whereas bitcoin address has the highest coefficient. This proves that bitcoin address carries a greater weight in the changes of bitcoin return, as high activity in bitcoin address may indicate that many investors are trading or transferring bitcoin. Furthermore between each factors, the result indicates that none of them are highly correlated with each other, thus representing a different sets of information to bitcoin returns movement.

Overall, the Pearson Correlation results was only used for the purpose of verification and to supplement the other test findings.

5 Conclusion

In conclusion, after running various statistical tools comprising of linear regression model and Pearson's correlation, our results showed that Google interest level of Bitcoin, Coinbase PageViews and unique Bitcoin address used per day showed positive correlation to Bitcoin returns - with unique Bitcoin address used per day showing the greatest R2 (0.154) and the greatest Pearson correlation coefficient (0.393). This make sense as Bitcoin prices increases, investors will either purchase more or sell existing Bitcoins for profit. In any case, Bitcoin addresses need to be accessed to send Bitcoins to exchanges for transacting.

We go on further to scrutinize if these factors are "determinants" or Granger-causes of Bitcoin returns and results showed that only Coinbase PageViews displayed greatest significance with a lag of 1 week. Surprisingly, Google trend data and unique Bitcoin address used per day may not be significant Granger-causes to Bitcoin returns after all. Coinbase PageViews being a likely Granger-cause to Bitcoin returns is likely as investors will need to access an exchange before carrying out buy-sell transactions of Bitcoin.

On the contrary to earlier papers on positive correlations between tweet sentiments and Bitcoin, our findings showed that there is no correlation found. In examining further, most of the tweets were either irrelevant to Bitcoin (e.g. advertisements, unrelated topics like football tagged with BTC hashtags, etc.). With growing Bitcoin's massive rise, many are also keen on taking every chance to ride this wave of interest by tagging their tweets with "♯Bitcoin" or "♯BTC". That being said, we also face limitations in our sentiment analysis tool, VADER and the volume of tweets scraped. VADER has been poor in interpreting context of tweets like sarcasm. Also, we would need a better method in extracting past tweets ever since Twitter has changed its access to its API.

For future studies, we would use daily Bitcoin returns and data collected on a data basis (Note: this may not apply to Google trends as Google limits to weekly data). This may help further in the results of our Granger causality test for Coinbase PageViews. 1 week lag time may not be realistic since investors react in a matter of minutes or hours to Bitcoin price changes on exchanges. For twitter sentiments, we would be keen to use a better sentiment analysis tool for interpretation of tweets.

6 References

1. Johan Bollen, Huina Mao, Xiao-Jun Zeng, ”Twitter mood predicts the stock market”, Journal of Computational Science, March 2011, Pages 1-8

2. Ali Tafti1, Ryan Zotti, Wolfgang Jank, ”Real-Time Diffusion of Information on Twitter and the Financial Markets”, PLoS ONE 11(8): e0159226.

3. Marko Kolanovic,Rajesh T. Krishnamachari, ”Big Data and AI Strategies Machine Learning and Alternative Data Approach to Investing Quantitative”, J.P.Morgan Securities LLC

4. Chris Randle, E, Quantifying Alternative Data Alpha Opportunities, Retrieved March 4th, 2018, ModernTrader.com

5. Liu, Yanhui, Gopikrishnan, Parameswaran , Cizeau, Pierre , Meyer, Martin , Peng, Chung-Kang , Stanley, H. Eugene, “The statistical properties of the volatility of price fluctuations, Physical Review E, vol. 60, pp. 1390-1400, 1999.

6. Gilbert, E Karahalios, K. (2010) Widespread worry and the stock market.

7. Gruhl, D, Guha, R, Kumar, R, Novak, J, & Tomkins, A. (2005) The predictive power of online chatter. (ACM, New York, NY, USA), pp. 78–87

8. Young Bin Kim1, Jun Gi Kim, Wook Kim, Jae Ho Im, Tae Hyeong Kim, Shin Jin Kang,Chang Hun Kim3,Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies, August 17, 2016.

9. Martina Matta, Ilaria Lunesu, Michele Marchesi, Bitcoin Spread Prediction Using Social And Web Search Media, 09 July 2015, https://www.researchgate.net/publication/279917417

10. Nofsinger, J. (2005) Journal of Behaviour Finance. 6, 144–160.

11. Edmans, A, Garca, D, & Norli, . (2007) Journal of Finance 62, 1967– 1998.

12. Wilson, T, Hoffmann, P, Somasundaran, S, Kessler, J, Wiebe, J, Choi, Y, Cardie, C, Riloff, E, & Patwardhan, S. (2005) OpinionFinder: A system for subjectivity analysis. pp. 34–35.

13. Leng, G, Prasad, G, & McGinnity, T. M. (2004) Neural Netw. 17, 1477– 1493.

14. Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market. Journal of Computational Science. 2011;2(1):1–8.

15. Burniske, C. (2018). Cryptoassets, The Innovative Investor’s Guide to Bitcoin and Beyond. New York, NY: McGraw-Hill.

16. Verhage and Kharif. (2018, February 26). Goldman-Backed Circle Agrees to Buy Crypto Exchange Poloniex, Retrieved March 21, 2018 from https://www.bloomberg.com/news/articles/2018-02-26/goldman-backed-circle-buys-digital-exchange-poloniex

17. Son, Campbell, and Basak. (2017, December 22). Goldman Is Setting Up a Cryptocurrency Trading Desk. Retrieved March 20, 2018 from https://www.bloomberg.com/news/articles/2017-12-21/goldman-is-said-to-be-building-a-cryptocurrency-trading-desk

18. Statista 2018. (2018) Number of monthly active Twitter users worldwide from 1st quarter 2010 to 4th quarter 2017 (in millions). Retrieved March 20, 2018 from https://www.statista.com/statistics/282087/numb of-monthly-active-twitter-users/

19. What are some great sources for BitCoin and Blockchain news. Retrieved March 21, 2018 from http://www.kryptographe.com/what-are-some-great-sources-for-bitcoin-and-blockchain-news/

20. Kasanmascheff, M. (2018, March 28). World's Fifth Largest Crypto Exchange Bitfinex Wants To Move To Switzerland. Retrieved March 28, 2018 from https://cointelegraph.com/news/worlds-fifth-largest-crypto-exchange-bitfinex-wants-to-move-to-switzerland

21. Bitcoin Markets (bitfinexUSD). (2018). Retrieved March 20, 2018 from https://www.quandl.com/data/BCHAR
Bitcoin-Markets-bitfinexUSD

22. Google Search Statistics. Retrieved March 20, 2018 from http://www.internetlivestats.com/google-search-statistics/

23. Bollen, Johan & Mao, Huina & Zeng, Xiao-Jun. (2010). Twitter Mood Predicts the Stock Market. Journal of Computational Science. 2. 10.1016/j.jocs.2010.12.007.

24. Google Trends. Retrieved March 20, 2018, from https://trends.google.com/trends/explore?q=bitcoin

25. VaderSentiment. Retrieved March 21, 2018 from https://github.com/cjhutto/vaderSentiment

26. Brooks, Chris (2008). Introductory Econometrics for Finance. New York, Cambridge University Press.

27. Fuller, W. A. (1976). Introduction to Statistical Time Series. New York: John Wiley and Sons