Exploration of factors impacting bitcoin

Project motivation

The genesis of this project was an interest in the volatility of the cryptocurrency market. A **cryptocurrency** is a digital asset designed to work as a medium of exchange that uses cryptography to secure its transactions, to control the creation of additional units, and to verify the transfer of assets. Cryptocurrencies use decentralized control as opposed to centralized electronic money and central banking systems.

The most topical example would be that of the Bitcoin. In the fall and winter of 2017, the price of “Bitcoin”  a period of time that saw the price of bitcoin climb from below $1,000 to nearly $20,000 on the CoinDesk Bitcoin Price Index (BPI).1. The momentum of this increase was slowed in early 2018 but still fluctuated wildly, as of April 25th the price $8,977.19 per bitcoin.2.

The lack of prescient for such a rapid increase in the value of a cryptocurrency, motivated me to choose the topic as a project, using the skills and techniques developed in INFO 550 Software Engineering to explore and gain insight into the unprecedented nature of this trend.

Intent

To gain some kind of insight into what

History

The project had a genesis mostly focusing on the cryptocurrencies as a whole. While the topic provides a lot of material to explore, I

Hypotheses

Hypothesis One: Are the local peaks of these top cryptocurrencies similar in any way?

Hypothesis Two: are mentions of bitcoin in news media on twitter and price of bitcoin related?

Hypothesis Three: are mentions of bitcoin by the public related to market value of bitcoin?

Hypothesis Four: Is there a time of year that bitcoin value is generally high?

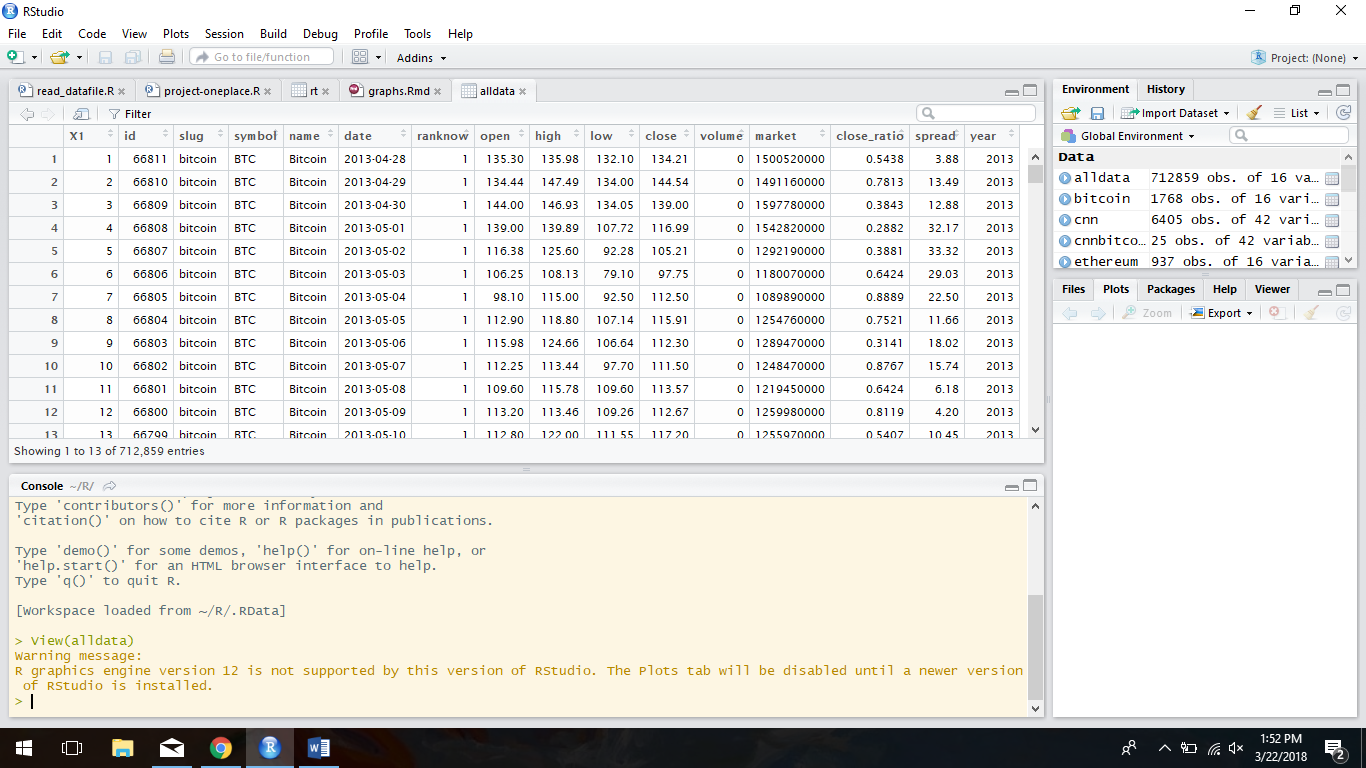
Hypothesis Five: Is there a relationship between number of bitcoin transactions and its market value?

Methodology

I first needed to locate the primary data source that I would use. The simplest contender was the crypto package, an r package that web scrapes the most popular index aggregator for cryptocurrency prices, “coinmarketcap.com”

Dataset: Crypto

Justification: Scrapes the most commonly used cryptocurrency price index aggregator. Used by both BBC3 and CNNMoney4. A picture of the dataset follows. The Bitcoin values were compared to two other indexes, the blockchain.info bitcoin value csv5 and bitcoin value csv obtained from coindesk6. While there were differences (which was expected due to the nature of each index being determined by multiple exchanges that bitcoin is being traded on, and can vary form indeed to index) there was negligible difference and I decided to proceed with crypto.



Intro.Rmd

I proceed to make the dataset more manageable, by sub setting the dataset generated from the crypto package. With over 1152 cryptocurrencies, and the relative ease that comes with creating a cryptocurrency, I decided that this would be the best approach going forward. I also decided to cutoff all prices that were acciocated with dates before January, 1st, 2014. The motivation for this decision is that only three of the top 15 cryptocurrencies existed on that date(Bitcoin, Ripple, Litecoin).

LanguageProcessing.Rmd

Having looked at the data, I was motivated to look for particular news events that may have defined cryptocurrency affected it’s market value and reach some insight into what causes cryptocurrency to fluctuate. The first step of this process was to take advantage of Rtimes package.

The Rtimes package is a package that acts as an R interface with the New York Times API, to retrieve headlines based on a number of parameters(Date Range, Keyword in Headline).7

Dataset: Rtimes

Justification: Used as a proxy for media coverage of cryptocurrency. The New York Times is a mainstream publication and fetching headlines from the website from it’s API would be helpful to see what kind of coverage cryptocurrency has received.

I first used the Rtimes package to generate a data frame that contains headlines from January 2010 to April 2018, search specifically for articles that contain “Bitcoin” and “Cryptocurrency” tags specifically. I returned two data frames, the “bitcoin” data table containing over 1153 observations, and “cryptocurrency” data table containing 309. The resulting list contains four tables and I pull the “data” table from the list and it is with this table that I perform my analyses. I am motivated to take the headlines that have been retrieved from the package and use language processing tools to plot rating against market value and see if there are any observable trends.

*Sentiment Analysis*

There are a number of headlines in my dataset that would not be useful to perform any kind of sentiment analysis (i.e. “Daily Update: November 7th”). Assuming the most significant cryptocurrency related headlines would get headlines of their own, I decided to filter the headlines by character string. This action reduces the number of available data to use, reducing the “cryptocurrency” table to a data table containing 2 observations. The filtered bitcoin table however, contained 309 observations.

A number of functions are created to assist in my analysis

*Sentimentoutputb*i - Generates data frame of NYT headlines and binary sentiment analysis results (positive/negative) from the following dictionaries: General Inquirer(Harvard-IV), Loughran-McDonald financial dictionary, and QDAP (polarity words from qdap package). Also adds a variable called "score" where the results of the dictionaries are converted to a numeric factor (1="positive" or -1="negative") and added together to form an aggregate score

*Freqnews* - Function to count number of headlines a day and filter headlines by character string

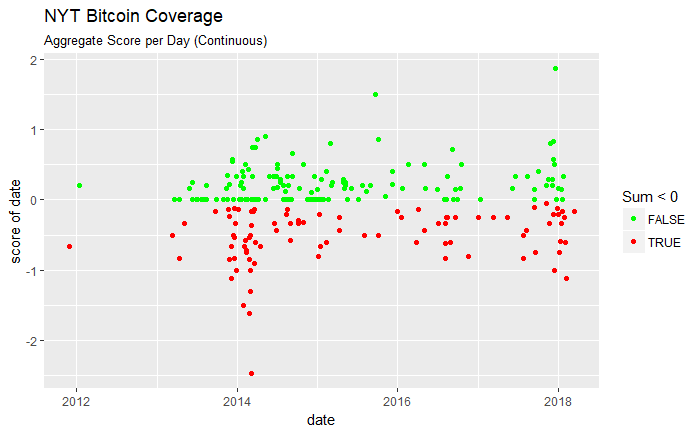
*Aggregatescore*- Function to aggregate the scores of all headlines from a particular date into one aggregate score per date

*Topheadline* - Function to pull filters table to include headlines from dates with that have a score within a specified range

*Sentimentoutputunique*- Identical to previous sentimentoutput function except only returns unique headlines

*Sentimentoutputcont*- Identical to *sentimentoutputunique*, but returns a continuous sentiment score, allowing for a more neutrality/nuance in sentiment score.

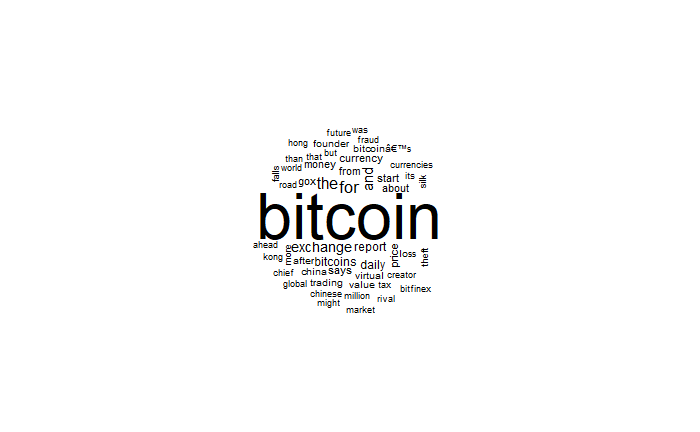
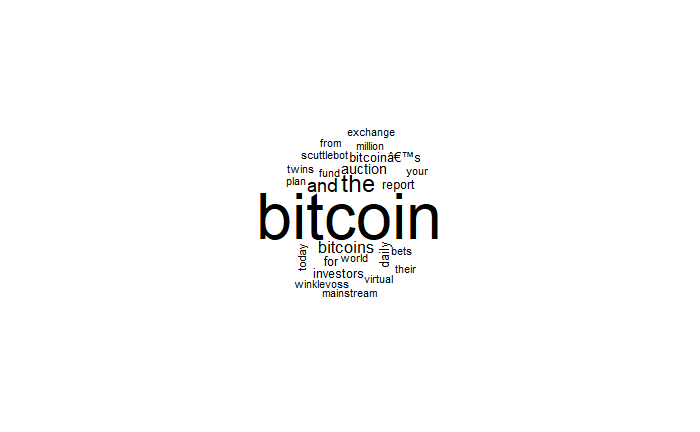
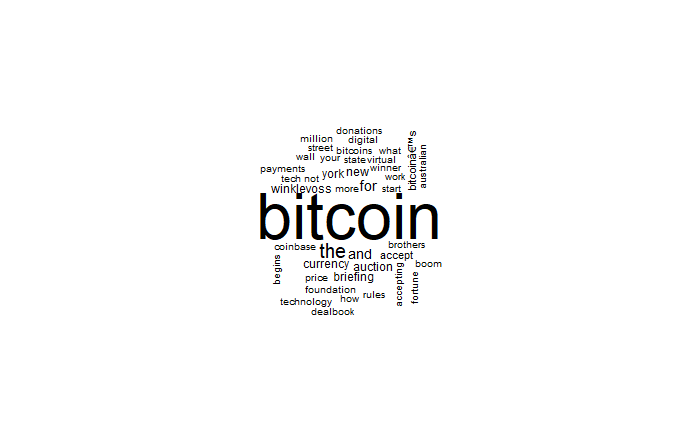
We see that NYT’s bitcoin’s coverage peaks at two points, early 2014 and late 2017/early 2018. Bitcoin’s unprecedented market value increase in late 2017 would likely be the reason coverage increased in the latter period, however the increase in 2014 is not apparent at this point and more exploration of the data is necessary. According to the sentiment analysis performed, it seems that media coverage of bitcoin in the NYT is split evenly among covering events that are favorable to bitcoin and negative.



In order to efficiently look at headlines from specific dates, I create the following functions:

*Wordc* - Function to generate word cloud from data frame generated from aforementioned sentimentoutput functions.

*Wordcnofilter*- Function to generate word cloud from data frame generated from aforementioned sentimentoutput functions.



Good neutral bad

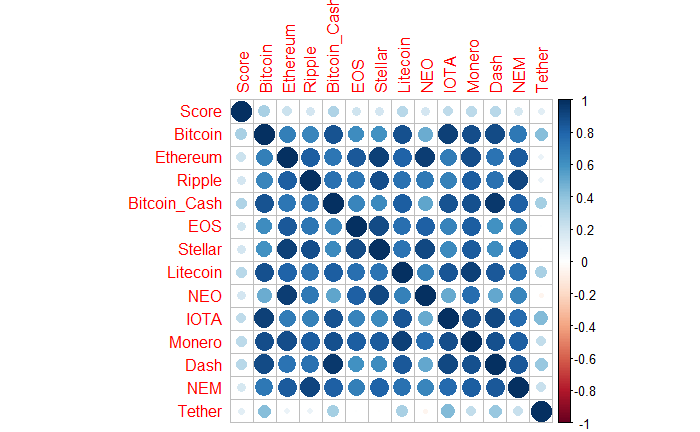
ForumWebScrapper.Rmd

TwitterAPI.Rmd

I moved to utilizing the Rtweet package, to generate more data relating to coverage of cryptocurrencies and bitcoin. I selected six well known new organizations that cover cryptocurrencies on a daily basis to scrape from. These are Coindesk, Bitcoinnews, Cointelegraph, The Merkle, CCN (CryptoCurrencyNews). The twitter API only allows 3200 tweets per account, and tweets older than the 3200th oldest tweet are not accessible from the API. After scrapping the date range of my tweets were from 2017-09-07 to 2018-04-30.

Pivot: I decided to subset my analysis, taking advantage of the amount of tweets I have for a short time period in comparison to the NYT articles which were more spread out.

Using modified versions of the previous functions, I perform sentiment analysis on the tweets retrieved and generate scores that are an aggregate of the sentiment analysis scores on a particular date. To not let dates with a larger normal of tweets stand out I will average the score by the amount of tweets that contribute to that score.



Sources

1. From $900 to $20,000: Bitcoin's Historic 2017 Price Run Revisited. (2017, December 30). Retrieved from <https://www.coindesk.com/900-20000-bitcoins-historic-2017-price-run-revisited/>
2. C. (n.d.). Bitcoin Price Index - Real-time Bitcoin Price Charts. Retrieved from <https://www.coindesk.com/price/>
3. Weeks, I. R. (n.d.). What is ripple? Here's why everyone's talking about the cryptocurrency. Retrieved from <http://money.cnn.com/2018/01/02/technology/what-is-ripple-cryptocurrency/index.html>
4. Bitcoin braces for worst week since 2013. (2018, February 02). Retrieved from <http://www.bbc.com/news/business-42915437>
5. Market Price (USD). (n.d.). Retrieved from <https://blockchain.info/charts/market-price?timespan=all>
6. C. (n.d.). Bitcoin Price Index - Real-time Bitcoin Price Charts. Retrieved from <https://www.coindesk.com/price/>
7. Chamberlain, S. (n.d.). Client for New York Times 'APIs' [R package rtimes version 0.5.0]. Retrieved from https://cran.r-project.org/web/packages/rtimes/index.html