VINEET KAPOOR Follow
Jun 8, 2019 · 8 min read · Listen

Data science and Predictive Modelling on Cryptocurrency — Part -1

Hello Everyone!

This is my first blog in Medium. I am really excited to explain you the Data science techniques that I have applied on **Cryptocurrency** historical prices, tweets from twitter and data from news articles, blogs. This blog is Part -1 of my whole analysis. There will be more parts further. Please stay tuned as I will upload the **tutorial of applying statistics on cryptocurrencies and machine learning on twitter tweets in my next part**.

The code used for all the visuals used in this blog is given in — https://github.com/vin725k/sentiment-analysis-of-tweets-using-machine-learning

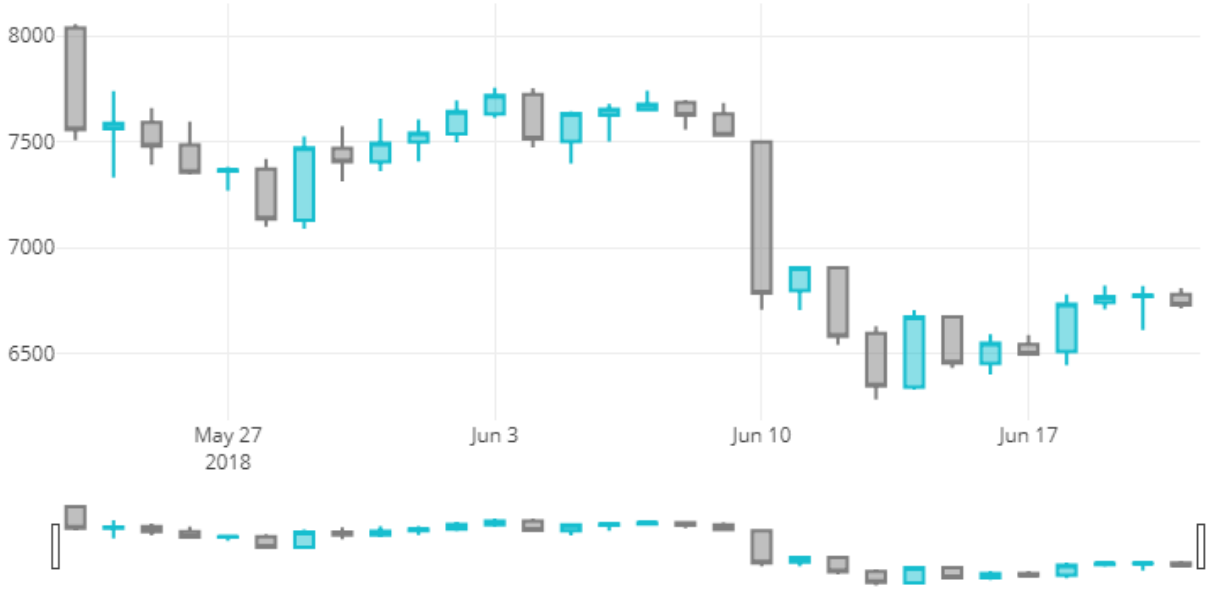
The Data science techniques that I will be explaining in this blog are:

Data science techniques applied:

1. Data Visualization and exploratory Data Analysis.

my first Candlestick Chart:





Close Price of Bitcoin for 1 month.

2. Text mining of twitter tweets.

* corpus means a collection of written texts, especially the entire works of a particular author or a body of writing on a particular subject.

Introduction:

Cryptocurrencies are distributed digital assets , in which currency is not held in digital form. It allows for speedy transactions between different parties. Cryptocurrency is not a crypto coin , it is a name given to cryptographic accounting unit. Financial transactions in these units are secured by cryptography. The most famous cryptocurrency — Bitcoin emerged after 2008 global monetary crisis. Many cryptocurrencies have been launched after bitcoin and few countries have already legalised the cryptocurrencies such as

The sentiments related to cryptocurrencies need to be reviewed from twitter tweets and articles can be used to analyse the future growth and important factors in the growth of cryptocurrencies.

Top 4 cryptocurrencies used for analysis

Methodology & Tools:

The Data has been taken from twitter using **twitter live streaming** API and various news articles, blogs. The data has also been taken from **coinmarketcap**.com. Historical prices of top 4 currencies have been scraped. **BeautifulSoup** and **Selenium** is used in python and twitter package was used in R programming for taking live streaming tweets and rest tweets. Tableau has also been used to visualise the forecast and trend of top 4 cryptocurrencies. For hypothesis testing and sentiment classifier model from articles, blogs, Python has been used. For sentiment analysis on twitter data and classifier model, R has been used.

Tools used:

Modelling Tools — Naïve Bayes and Max Entropy model.

Libraries used —

in R: **twitter, SnowballC, syuzhet, tm, ROAuth, dplyr, magrittr, ggplot2, wordcloud, stringr, udpipe, textrank, igraph, ggraph, qdap, tidytext, tidyverse, sentiment, RColorBrewer.**

In Python: **nltk, numpy, pandas, tweepy, matplotlib, textblob, random, selenium webdriver, string, seaborn, datetime, scipy, beautifulsoup, requests**

Questions & Hypothesis:

The questions/hypothesis that I would like to test using the data collected are:

1. Perform **ANOVA Hypothesis testing** between top four cryptocurrencies to determine if the average daily returns are equal or not.

2. To analyse the cryptocurrency, which is best to invest with minimum risk.

3. Check for correlation between different cryptocurrencies.

4. Check for factors responsible for predicting cryptocurrencies growth.

5. To analyse growth of cryptocurrency worldwide and regional wise.

6. To build a **Naive Bayes classifier model** to predict sentiments of documents from news articles, blogs by investors, banks, regulators.

7. Analyse the sentiments of investors, regulators, financial institutions and banks on cryptocurrency.

Data collection:

Twitter, news articles, blogs and coinmarketcap.com were predominantly used as data source for this study.

Tweets containing hashtags of “bitcoin”, “ethereum”, “blockchain” and

Cryptocurrency views data from news articles was scraped. From ….

The historical prices of bitcoin, Ethereum, bitcoincash and ripple were scraped from coinmarketcap website

The cryptocurrencies last 30days data was downloaded from coinmarketcap website.

Except the tweets from twitter and data from articles, the rest of the data was cleaned.

Summary:

Blockchain is the factor responsible for Cryptocurrencies growth and decline.

Ethereum is the currency, which can be used for investment, as it's average daily returns are better than other three cryptocurrencies — Bitcoin, Bitcoin — cash, ripple.

The pattern of twitter sentiments is similar in US and India, the pattern is like overall sentiment. Japan has a different pattern of sentiments.

Bitcoin is the currency used for most of the transactions.

There is high correlation between bitcoin and Ethereum close price.

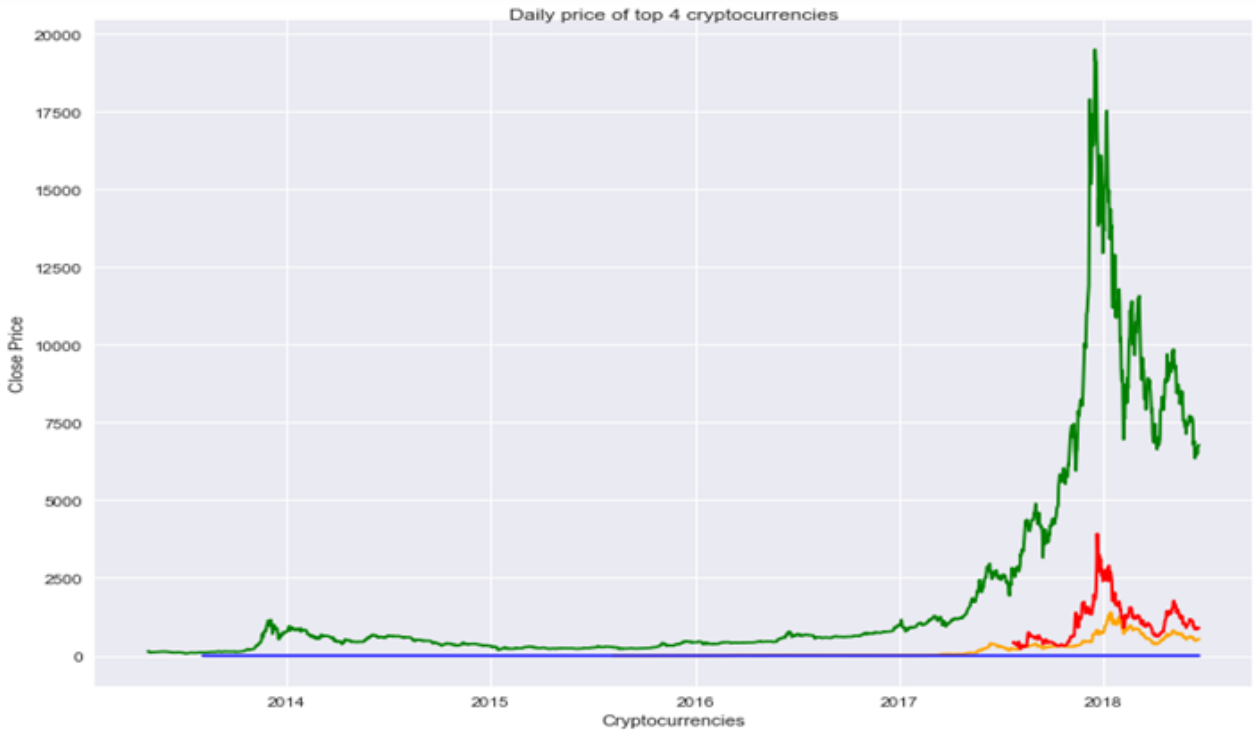
The forecast of close price of Bitcoin and Ripple shows that the prices will decline.

Proportion of neutral and positive sentiments from news articles and blogs are almost similar and greater than negative sentiments.

Analysis:

1. The below chart shows comparison of Close price. The green line in the chart corresponds to Bitcoin, red line shows the bitcoin cash, yellow line shows Ethereum. Blue line shows ripple. This chart shows that Bitcoin close price is very high relative to other currencies. At the end of 2017, it gained a lot.





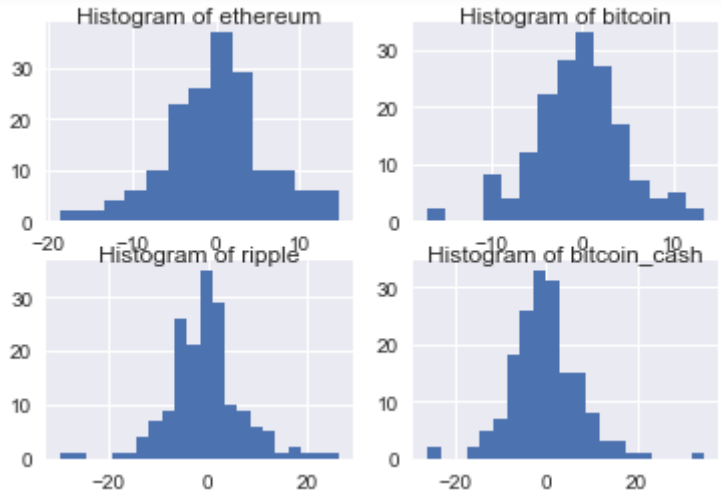
Line chart of Top 4 cryptocurrencies

2. Correlation between top 4 cryptocurrencies. There is high correlation between bitcoin and ethereum.

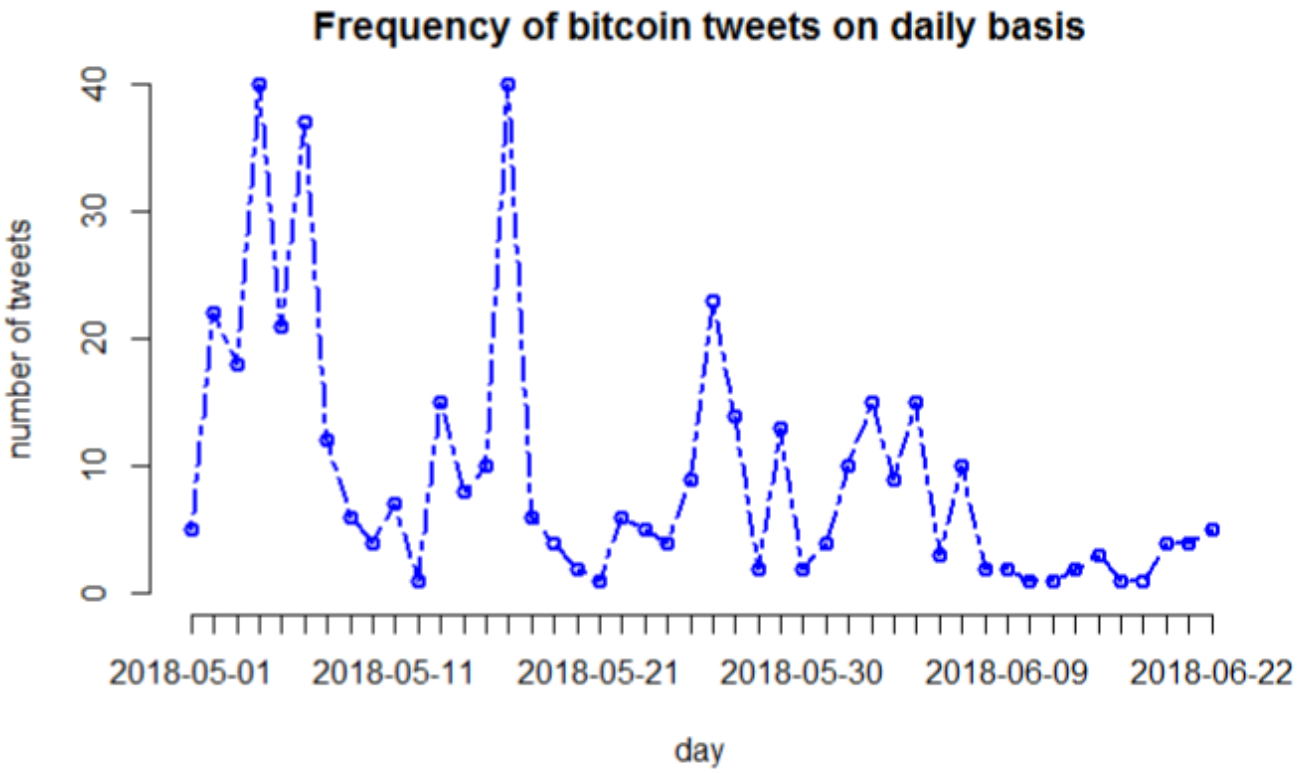
	bitcoin	ripple	ethereum	bitcoin_cash
bitcoin	1.000000	0.695727	0.813341	0.793987
ripple	0.695727	1.000000	0.728876	0.637083
ethereum	0.813341	0.728876	1.000000	0.748032
bitcoin_cash	0.793987	0.637083	0.748032	1.000000

Correlation between top 4 cryptocurrencies

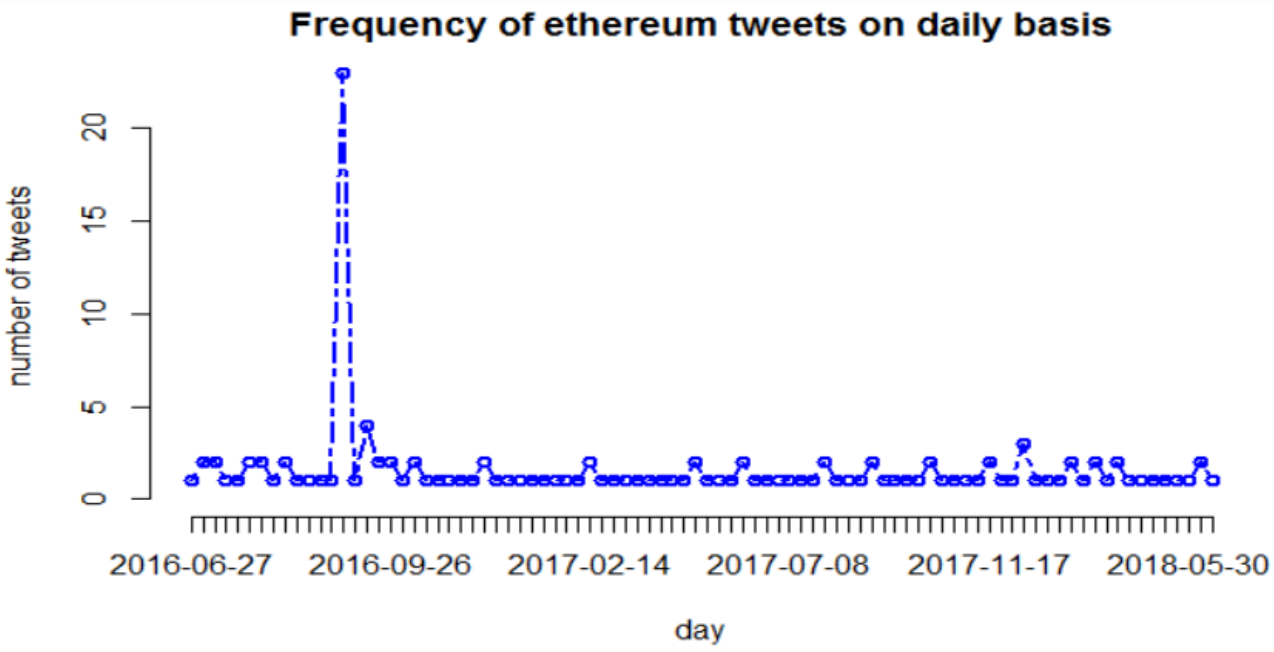
3. Histogram of close prices of all the four cryptocurrencies. All of them are following a standard normal curve.



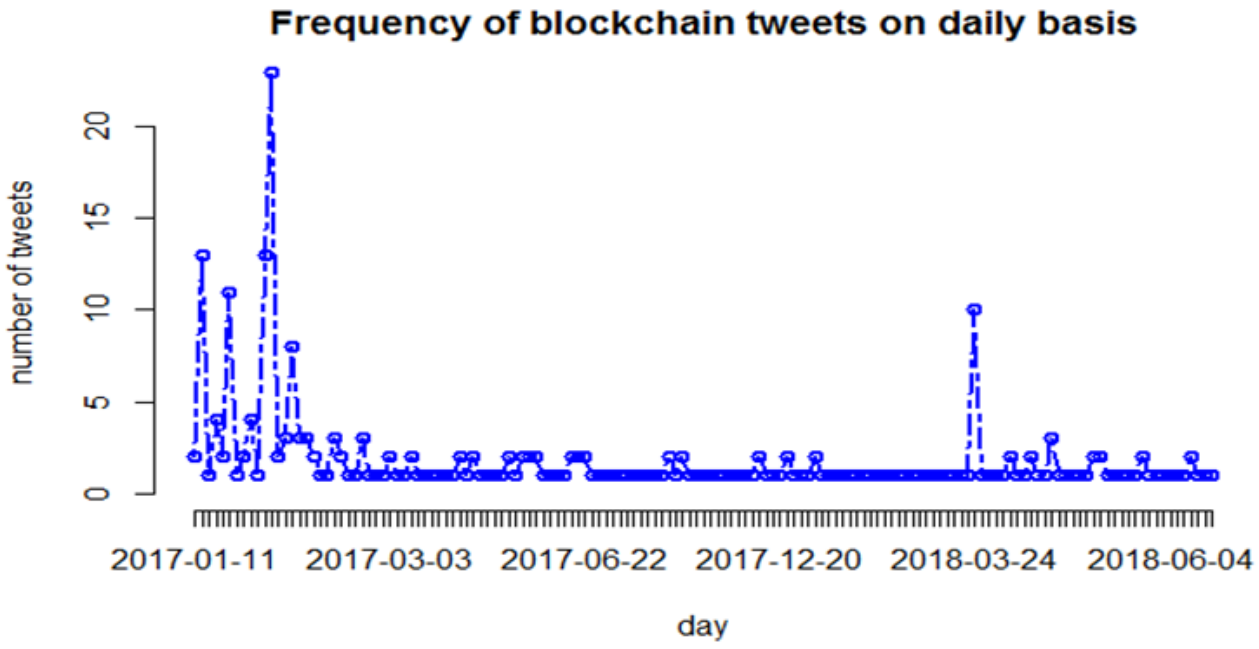
4. For Bitcoin, frequency of tweets is relatively higher in May month'2018 as compared to June month'2018. **9.3260** is the average daily number of tweets.



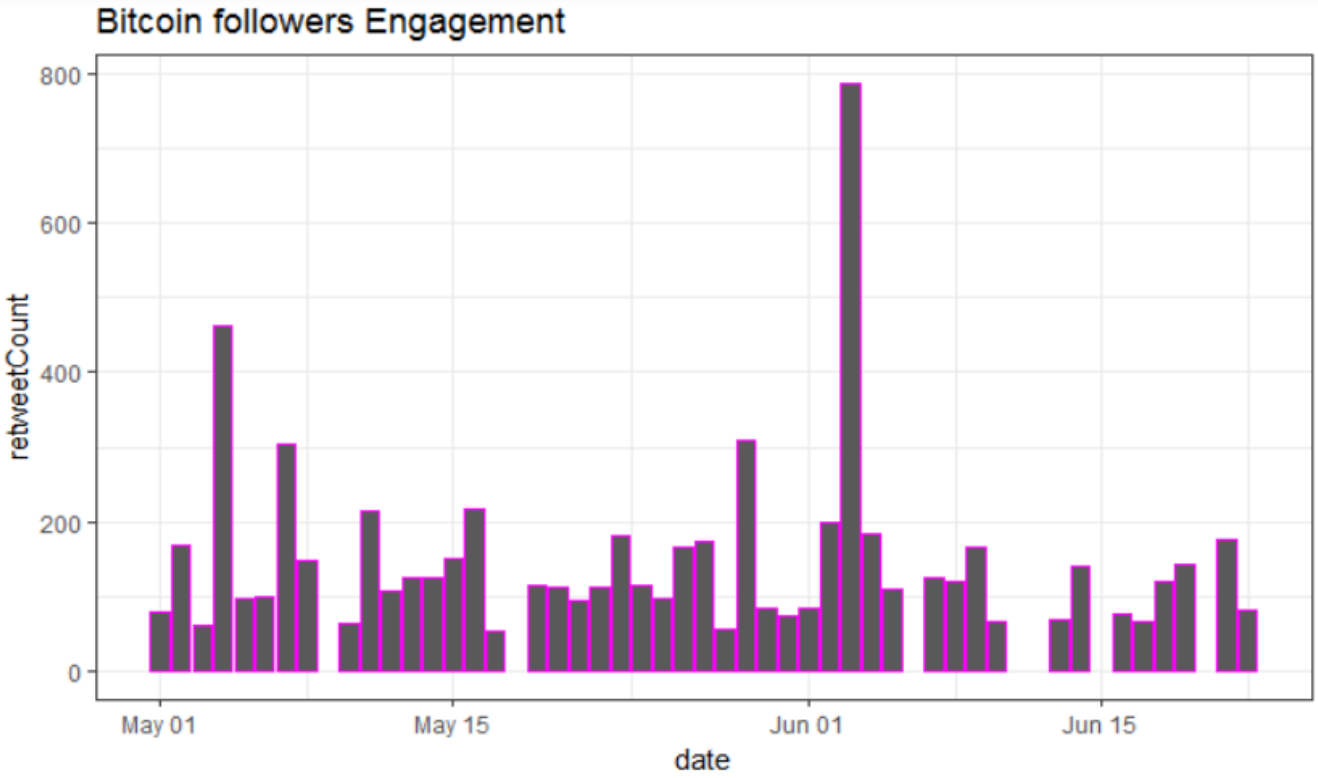
5. Only for one day, the frequency of Ethereum tweets saw an **outlier**, except that, the average frequency is 1.53 is the average daily number of tweets.



6. The mean of frequency of blockchain tweets is **1.8**. There are few outliers in January and February month.

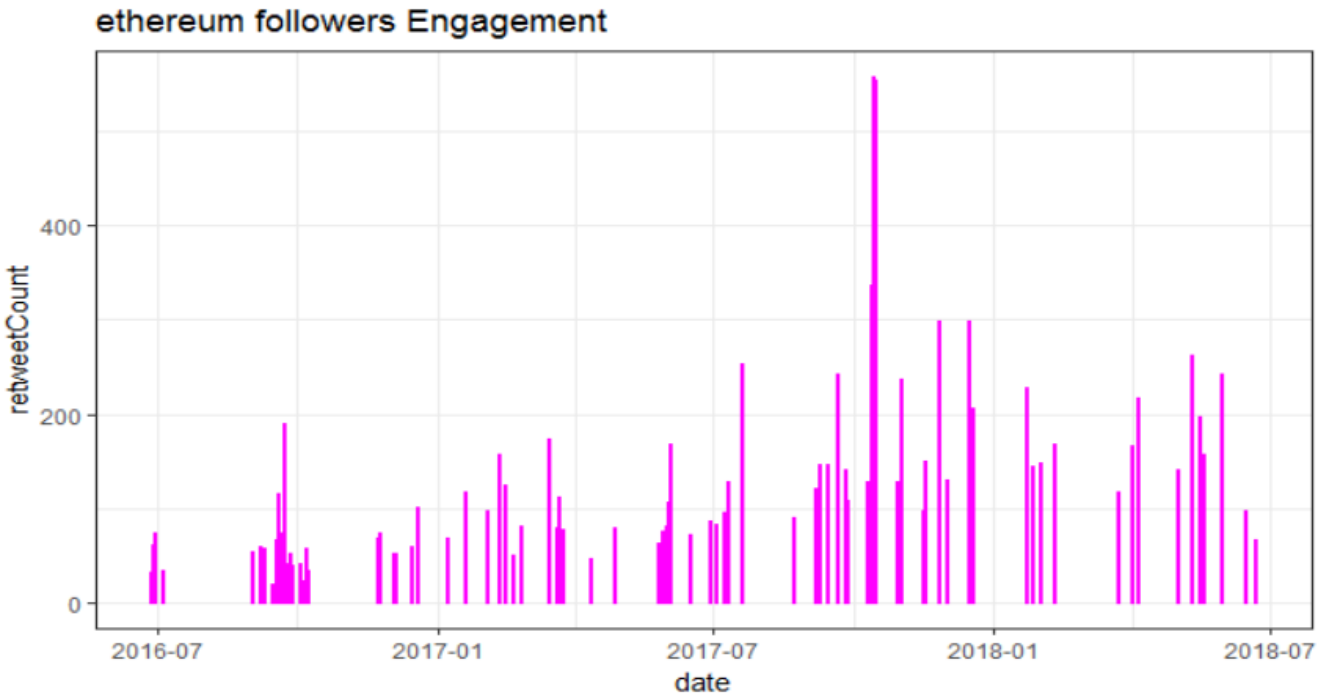


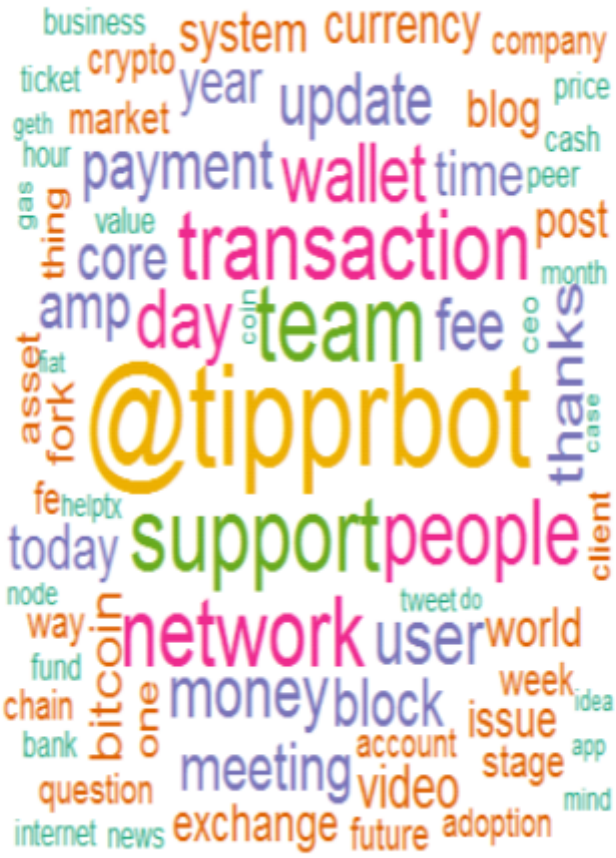
7. Followers engagement with bitcoin is decreasing after first first week of June month. In may month, the number of followers were more than 200 many times as compared to June month.



For last 30 days. in 2018

8. Followers engagement of Ethereum currency are is more at the end of the year 2017. The retweets have increased in 2018 relative to 2017.

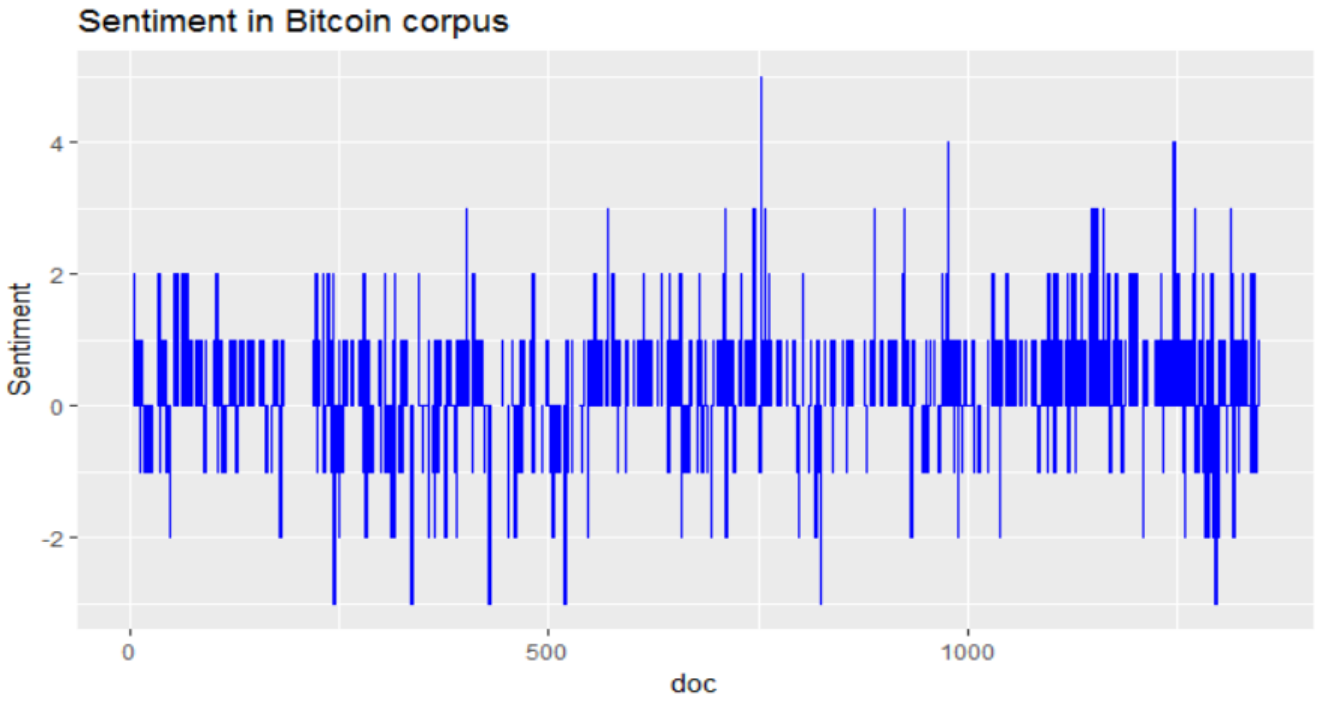




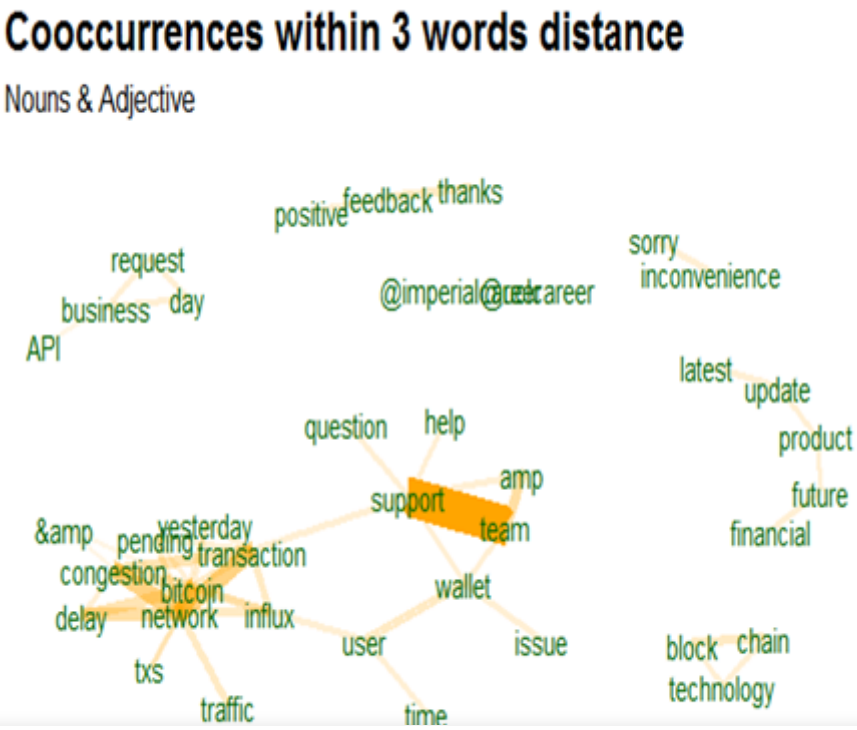
Wordcloud of ethereum, bitcoin, ripple corpus.

10. Wordcloud of common tokens in Cryptocurrency user time line corpus. High frequency occurring words are bitcoin, Ethereum, network, cash.

11. The sentiments are almost neutral. There is not clear pattern of negative and positive sentiments using bing Dictionary.



12. Cooccurrences between words for nouns and adjectives in Blockchain corpus.



13. Most positive sentiment line in cryptocurrencies corpus

[1] "@DewiPutri Hi Dewi! Please contact our support team so we can advise & assist on this issue: aVTaYqufM Thank you."

14. Most negative sentiment line in cryptocurrencies corpus.

```
[1] "@Rastachildd Hi there! When the network is congested, this can unfortunately
result in delayed transactions."
```

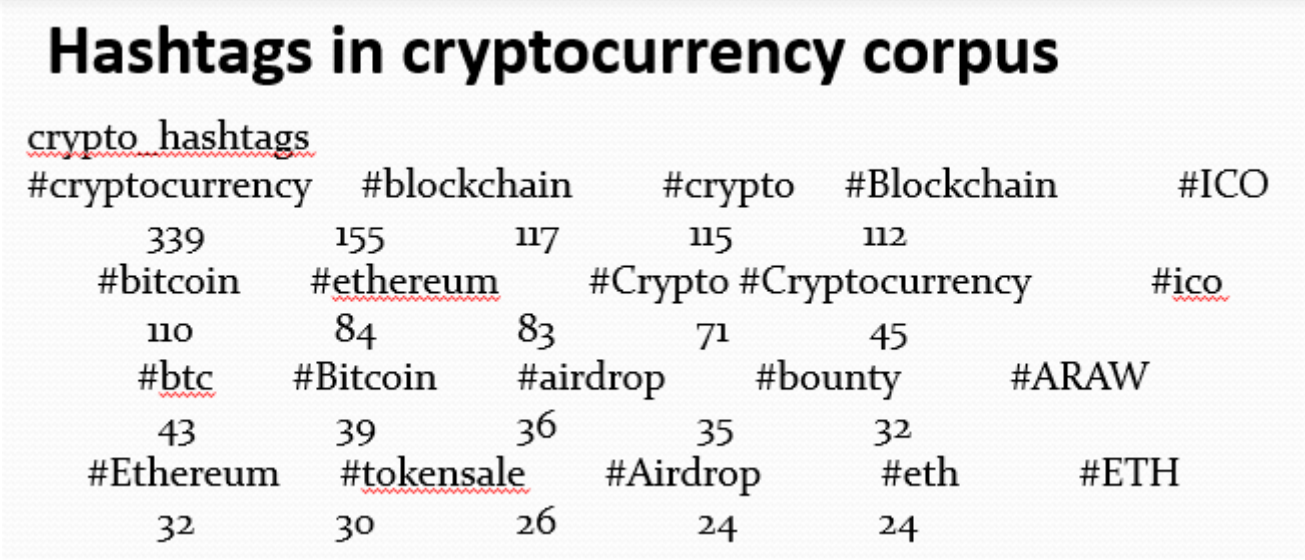
15. Positive words in the Cryptocurrency corpus.

```
[1] "secure"       "worth"        "like"         "glad"         "cool"         [6]
"improving"    "magic"        "work"         "awesome"      "works"        [11] "trust"
"fans"         "available"    "winners"      "cashback"     [16] "nice"         "interesting"
"valuable"     "free"         "reward"       [21] "right"        "innovation"   "amazing"
"encourage"    "excellent"    [26] "successfully" "upgraded"     "popular"      "balanced"
"good"         [31] "useful"       "benefactor"   "recovery"     "sensitive"    "easy"
[36] "better"       "win"          "top"          "benefits"     "great"        [41] "holy"
"worked"       "superior"     "reliably"     "support"      [46] "wow"          "excited"
"enough"       "smart"        "perfectly"
```

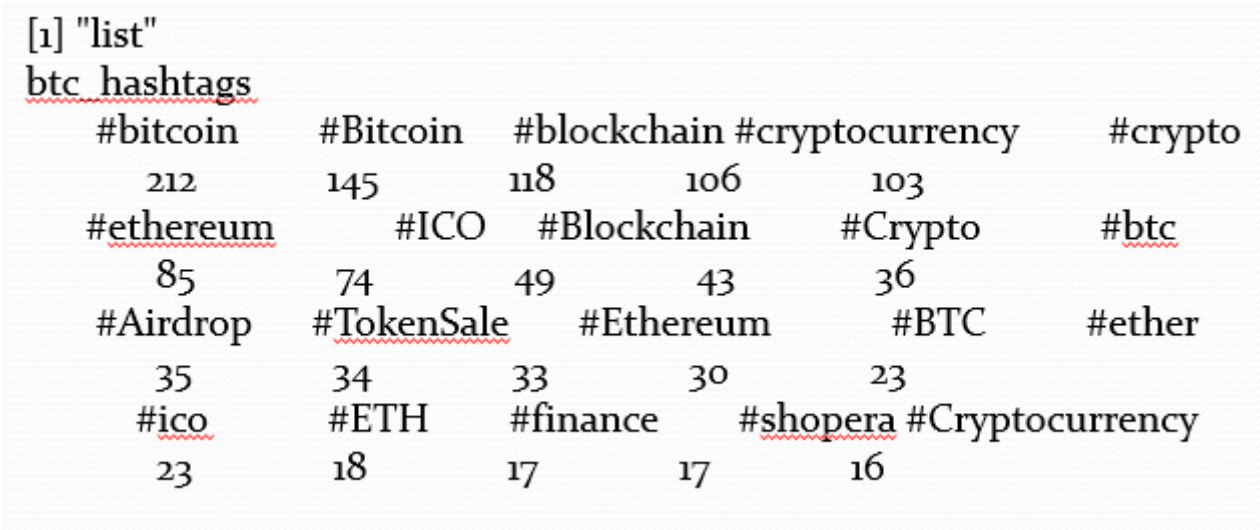
16. Negative words in cryptocurrency corpus.

```
[1] "warning"       "failure"       "useless"       "worse"          [5]
"corrupt"       "crash"         "criminal"      "set up"         [9]
"concerns"      "lost"          "miss"          "inadequate"    [13]
"volatile"      "limit"         "irony"         "confuse"       [17] "accuse"
"hedge"         "lie"           "hard"         [21] "dangerous"
"incorrectly"   "delusions"     "drawbacks"    [25] "funny"         "lack"
"cripple"       "choppy"       [29] "slower"        "cheap"
"overwhelming"  "self interest"[33] "lose"          "harass"
"fanatics"      "punish"       [37] "disagree"      "steal"         "crazy"
```

17. Hashtags in cryptocurrency corpus.

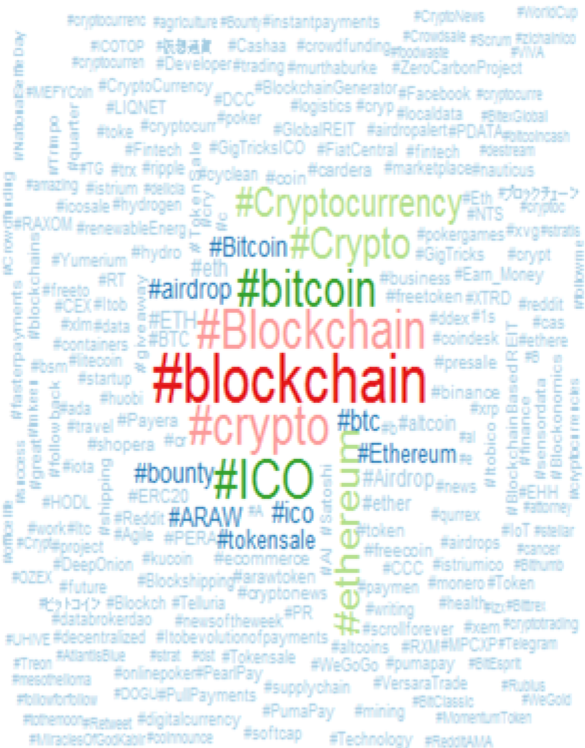


18. Hashtags in Bitcoin corpus.

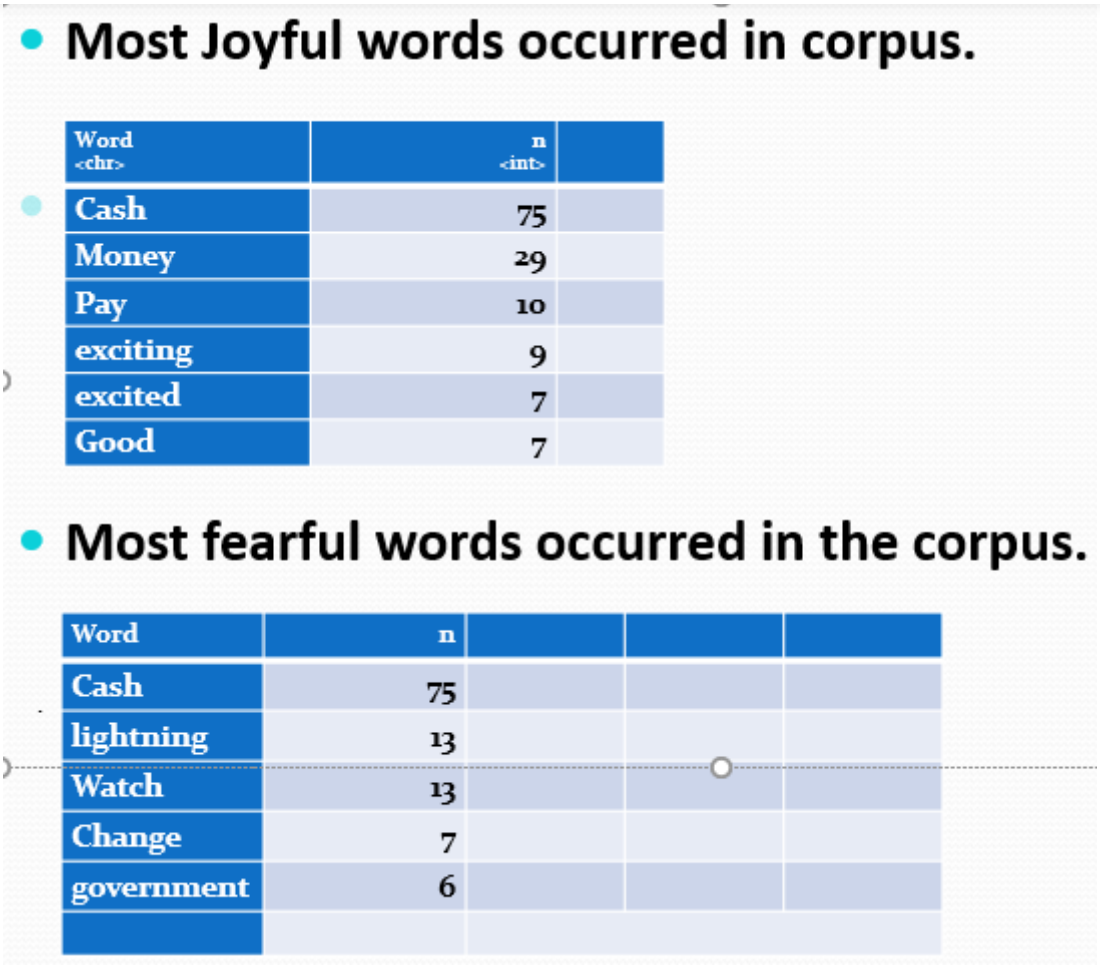


19. Comparison of live tweets of Cryptocurrency and Bitcoin


#bitcoin
#Bitcoin
#blockchain
#crypto
#cryptocurrency
#ethereum
#ICO
#Crypto
#Blockchain
#Ethereum
#TokenSale
#Airdrop
#btc
#ether
#BTC
#ico
#cardera
#ETH
#shopera
#finance
#crowdfunding
#Cryptocurrency
#tokensale


#blockchain
#Blockchain
#bitcoin
#crypto
#ICO
#Cryptocurrency
#Crypto
#Bitcoin
#airdrop
#ethereum
#btc
#Ethereum
#bounty
#ARAW
#ico
#tokensale
#ETH
#BTC
#eth
#Airdrop

20.



21. Top 10 nouns in Blockchain and Cryptocurrencies corpus.

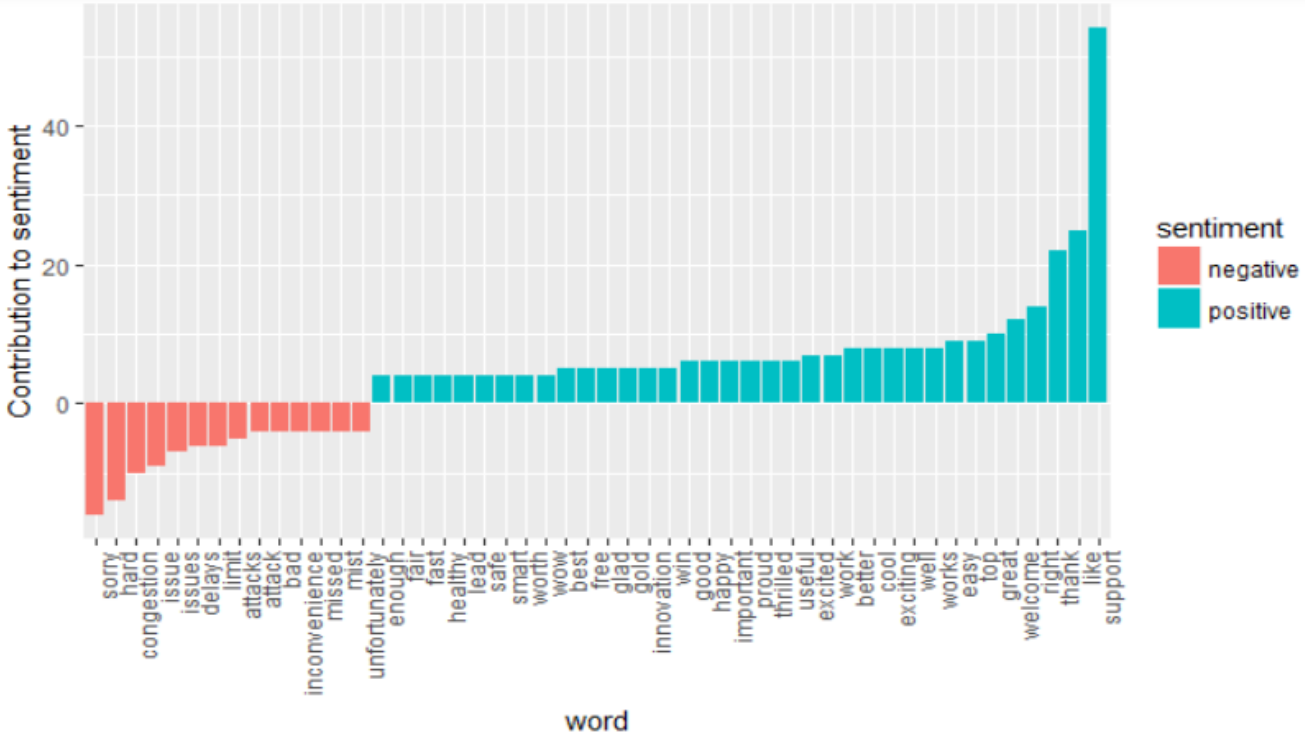
Top 10 nouns in Blockchain corpus

	key <chr>	freq <int>	freq_pct <dbl>	
1	support	32	4.134367	
2	team	28	3.617571	
3	wallet	23	2.971576	
4	network	19	2.454780	
5	transaction	18	2.325581	
6	bitcoin	16	2.067183	
7	user	16	2.067183	
8	thanks	15	1.937984	
9	amp	15	1.937984	
10	update	10	1.291990	

Top 10 nouns in Cryptocurrencies

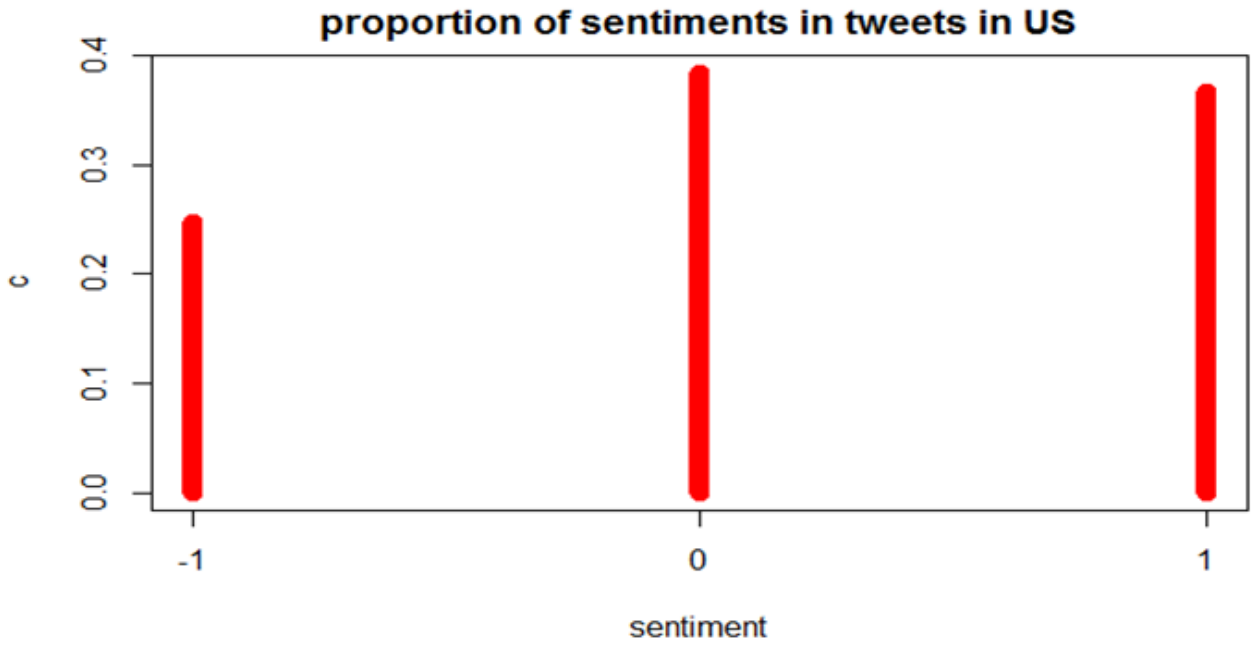
	key <chr>	freq <int>	freq_pct <dbl>	
1	@tipprbot	54	1.6922595	
2	team	38	1.1908493	
3	support	37	1.1595111	
4	transaction	36	1.1281730	
5	network	36	1.1281730	
6	people	32	1.0028204	
7	wallet	29	0.9088060	
8	day	28	0.8774679	
9	user	27	0.8461297	
10	money	26	0.8147916	

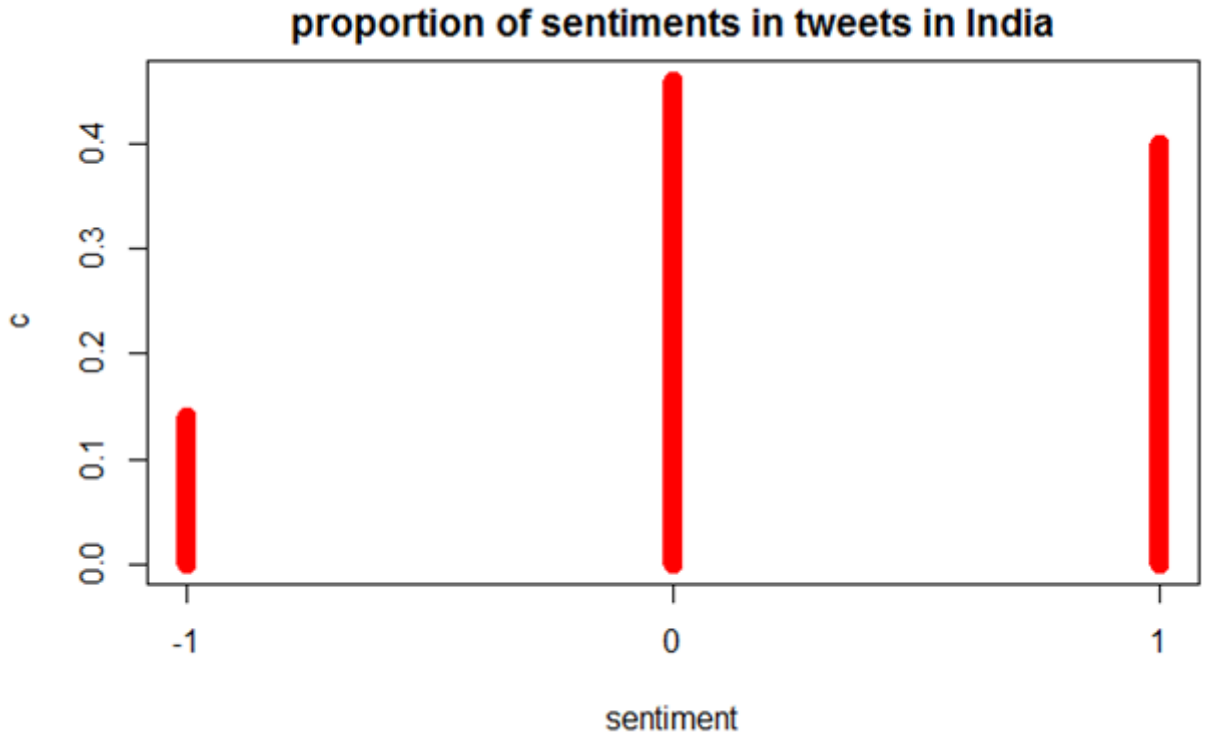
22. Words like sorry, hard, congestion, issue, attacks are occurring most in negative sentiments. Words like thank, right, welcome, cool, exciting, most of the words are sentiment words in cryptocurrency corpus.



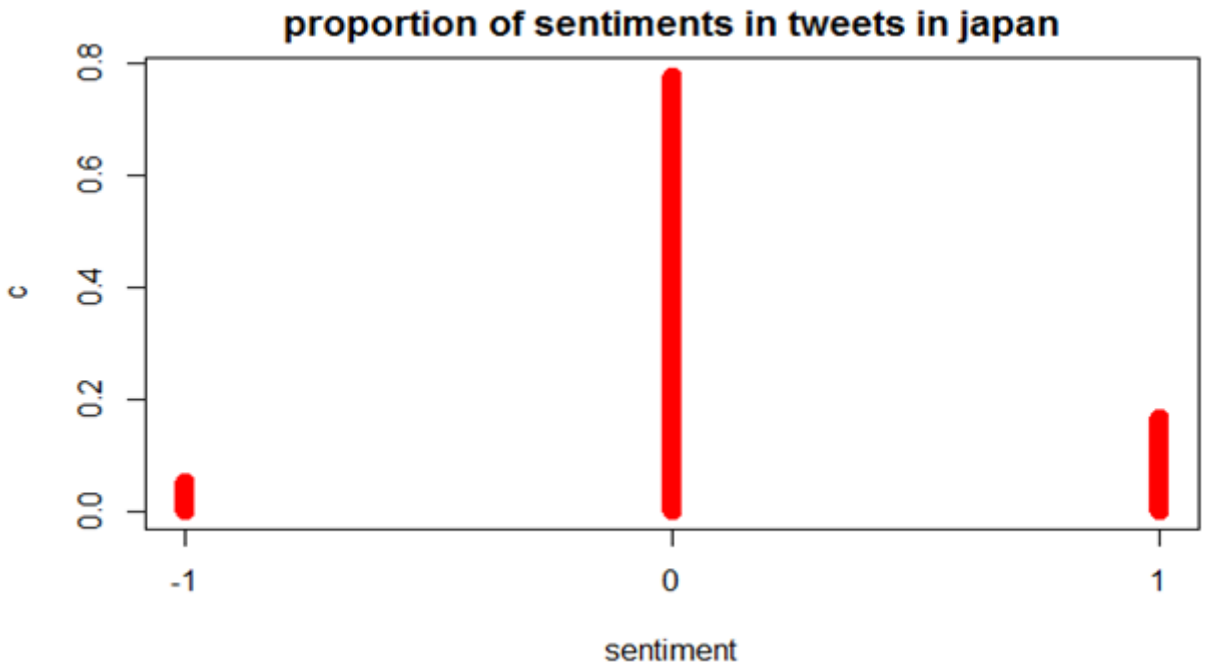
Regional wise analysis of Cryptocurrency

Proportion of sentiments of tweets of Bitcoin in US — Neutral and Positive tweets are almost equal. The negative tweets are 25% of the total tweets. The sentiment is diverse.

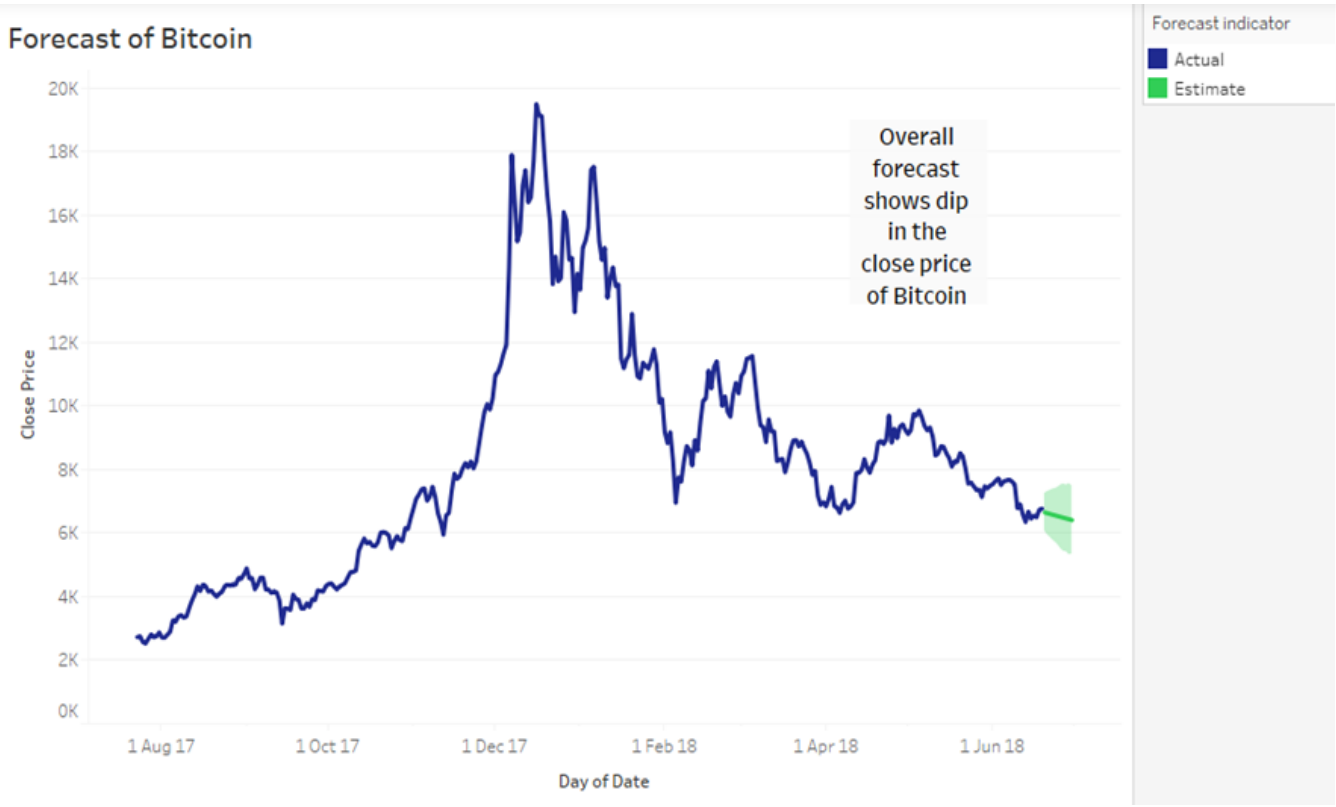




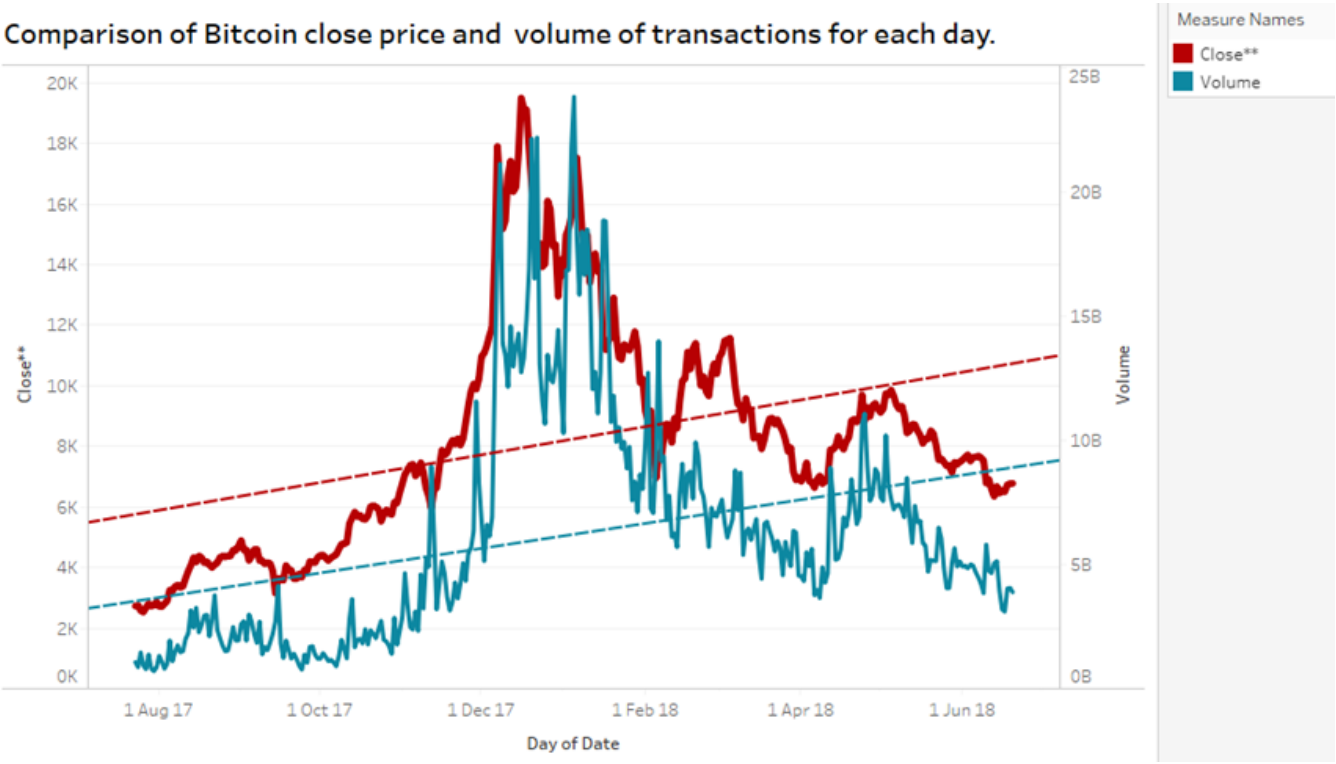
— Most of the tweets are neutral in Japan. Most the users are having neutral sentiment of Bitcoin.

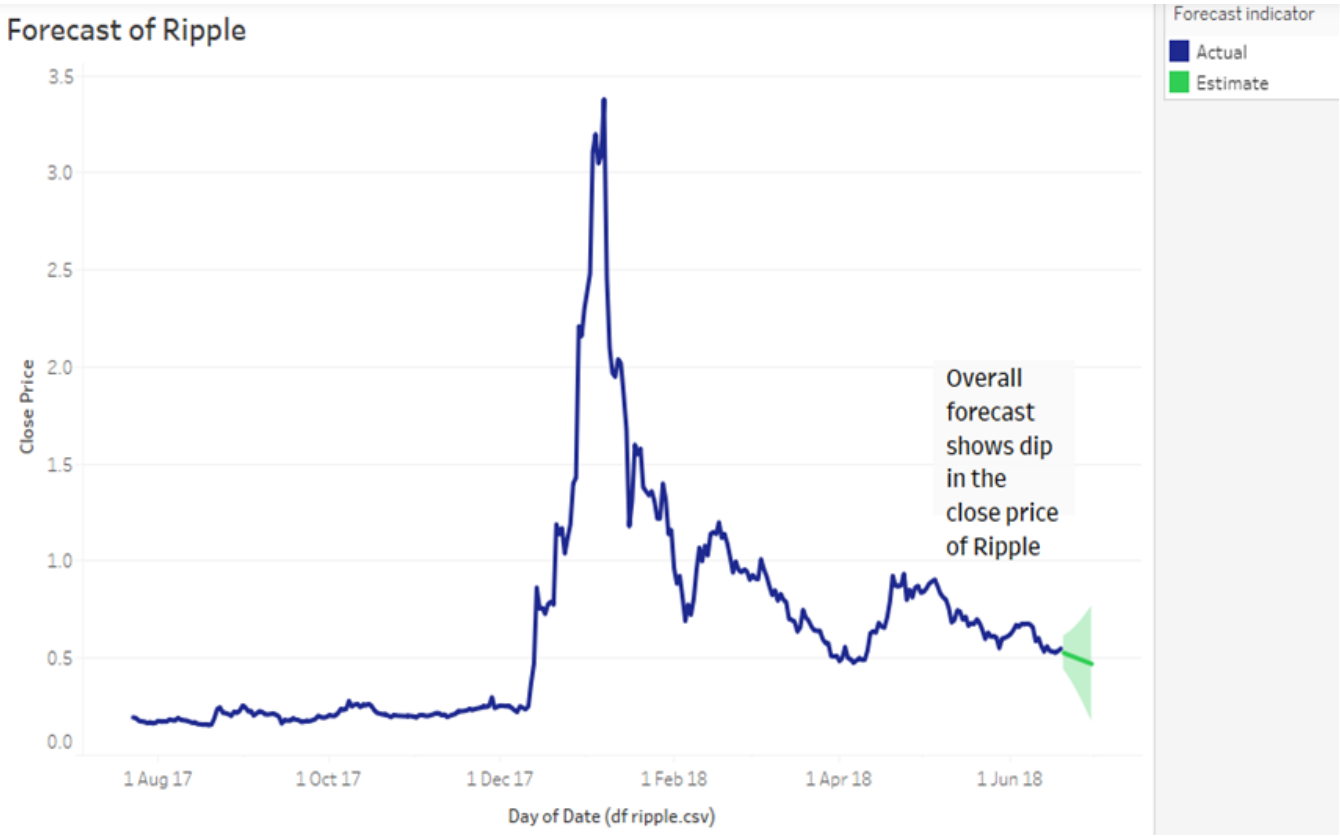


Appendix and other notable related works:

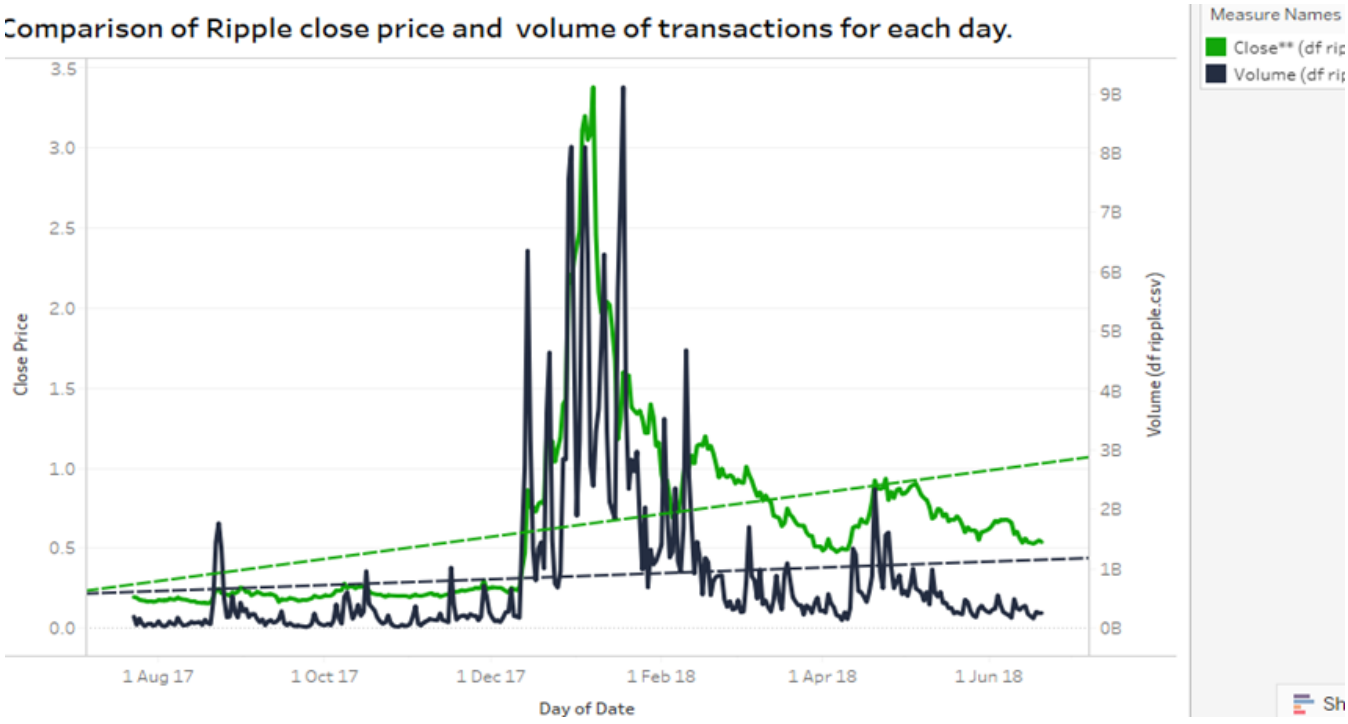


2. Trend line of Close price and trend line of volume of transactions for each day are positive for Bitcoin.





4. Trend line of Close price is positive as compared to trend line of volume of transactions for each day for ripple.



The code used for all the visuals used in this blog is given in —

References and Credits:

https://coinmarketcap.com/

https://economictimes.indiatimes.com/markets/stocks/news/cryptocurrencies-are-like-ponzi-schemes-world-bank-chief-says/articleshow/62830841.cms

https://economictimes.indiatimes.com/markets/stocks/news/how-cryptocurrencies-split-global-central-banks/articleshow/62715511.cms

https://economictimes.indiatimes.com/markets/stocks/news/anger-shock-confusion-as-rbi-bars-banks-from-cryptocurrencies/articleshow/63638799.cms

https://coinpupil.com/altcoins/advantages-disadvantages-of-cryptocurrency/

https://economictimes.indiatimes.com/wealth/invest/7-reasons-why-you-should-not-invest-in-bitcoins-cryptocurrencies/articleshow/60891341.cms

https://www.investinblockchain.com/7-signs-bad-cryptocurrency/

http://thecircular.org/cryptocurrencies-bad-sides-bitcoin/

This report is done as part of my course work for Indian School of Business (ISB) Certification in Business Analytics (CBA) program.