CRYPTOCURRENCY ANALYSIS

Group Members:

Smrithy Dharma Sankaran,

David Marquardt,

Beau Muster,

Laura Yuan

I. Introduction

There has been an increase in the public's interest in cryptocurrencies over the past few years and is a topic that has been discussed often on social media and on financial news sites. Many young individuals and adults today are interested in investing in cryptocurrencies, with some even believing that cryptocurrencies are a worthier investment than stocks. Little is known and researched about cryptocurrencies and how their prices fluctuate.

A few months ago, Elon Musk declared that Tesla would take Bitcoin as a form of currency to purchase Tesla cars which caused Bitcoin prices to surge significantly. However, in the past week, Elon tweeted that they would no longer take Bitcoin as payment which caused Bitcoin prices to drop. Thus, we believe that it would be interesting to dive deeper into an analysis of the price of cryptocurrencies. Answering the questions we have posed will help cryptocurrency investors and traders better understand how and why cryptocurrency valuations change over time.

For this project, we will focus on cryptocurrency prices from Jan 2018 to February 2021 for the following six cryptocurrencies: Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), Cardano (ADA), XRP (XRP), and BinanceCoin (BNB). We will use daily price information from 2018 to 2021 for each cryptocurrency and measure the valuation of each in comparison to the sentiment of each via Twitter analysis. We will also compare the historical value of Bitcoin (the most valuable cryptocurrency) with the historical value of the S&P 500 and the value of gold. In addition, we will construct linear and log regression models and perform correlation analysis to determine how in sync the valuation of cryptocurrencies move together.

II. Data

We have obtained tweets containing the keyword 'cryptocurrency' through an open dataset source, Kaggle. The dataset we obtained included information regarding tweets related to cryptocurrency and already includes a sentiment analysis in the dataset. Sentiment analysis is a form of Natural Language Processing and machine learning technique that involves extracting sentiment from a given text. Using the nltk and tqdm Python packages as well as the Harvard Inquirer Dictionary dataset, we were able to extract the number of positive, neutral, and negative words for each tweet. We will then use these values in order to assess their relationship with movements in cryptocurrency prices.

Tweets Data (Kaggle)

	date	tweet	pos	neg
2	2013-04-14	, a "cryptocurrency", went on a tear last week...	0.333333	0.666667
3	2013-04-18	I'm going to make my OWN crypto-currency and e...	0.500000	0.500000
4	2013-05-09	Your momma's cryptocurrency is so virtual, she...	1.000000	0.000000
5	2013-05-09	Your momma's cryptocurrency is so virtual, whe...	0.333333	0.666667
6	2013-05-14	Venture capital firm sends a signal to Bitcoin...	0.666667	0.333333

Above is a screenshot of our tweets dataset after we have performed the sentiment analysis. The dataset details the date of the tweet, the tweet itself, and the percentage of positive and negative words. This is an example of how the percentages are calculated.

$$Positive\ words\ (pos)\ = \frac{Positive\ words\ (pos)}{Positive\ words\ (pos) +\ Negative\ words\ (neg)}$$

Another dataset we have used was also obtained from Kaggle. There, we were able to find related CSV files for all price-related data of each cryptocurrency. Lastly, we were also able to find the S&P and Gold datasets through MarketWatch to add to our analysis. We then used Python to merge the datasets together and imported them as a CSV to continue our work in R.

III. Summary Statistics

Descriptive Statistics of Cryptocurrencies (Average)

Name	High	Low	Open	Close	Volume
Binance	19.715108	17.866523	18.714476	18.979276	2.968918e+08
Bitcoin	4892.842241	4600.863853	4749.021436	4768.826575	9.201904e+09
Cardano	0.143449	0.126698	0.135877	0.136248	4.631058e+08
Ethereum	265.482446	244.386440	255.723179	256.010335	5.588557e+09
Litecoin	42.921996	39.247936	41.230862	41.200325	1.116044e+09
XRP	0.207515	0.186994	0.197485	0.197636	9.540319e+08

In the figure above, we have provided descriptive statistics regarding the various price aspects of each cryptocurrency selected. This table details the average high, low, open, close, and trading volume for each cryptocurrency. Based on this table, we can identify that on average, Bitcoin prices are relatively higher than other cryptocurrencies, and on average, Cardano and XRP prices are relatively lower than other cryptocurrencies. In terms of daily volume, it seems that on average, Bitcoin seems to be the most traded cryptocurrency followed by Ethereum, then Litecoin.

Descriptive Statistics of S&P and Gold Prices

	High	Low	Open	Close
Name				
Gold	128.691684	127.582482	128.153078	128.140767
S&P	2428.350716	2403.588387	2416.530762	2439.198211

The figure above details descriptive statistics for the S&P index and gold. When these statistics are compared with the cryptocurrency prices above, it is observed that on average, Bitcoin prices remain the highest, then followed by the S&P index. In terms of gold, based on these statistics it seems as though gold prices on average remain higher than other cryptocurrencies but still lower than Bitcoin and Ethereum.

Descriptive Statistics for Tweets Dataset (Sentiment Statistics)

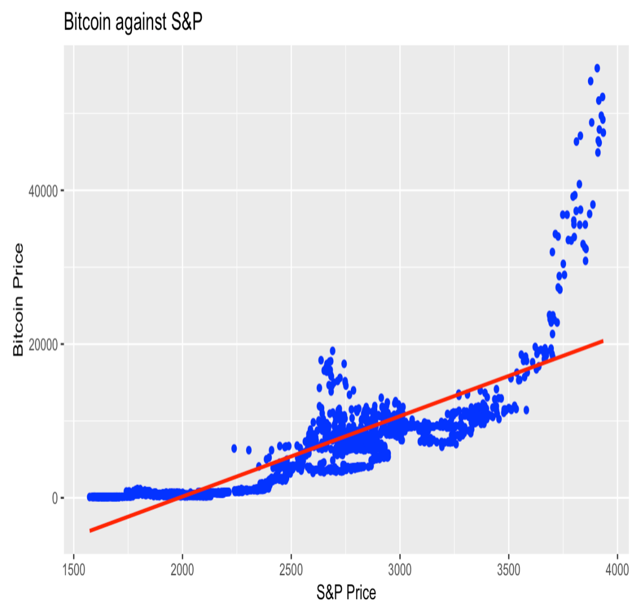
	pos	neg
count	721617.000000	721617.000000
mean	0.685677	0.314323
std	0.313603	0.313603
min	0.000000	0.000000
25%	0.500000	0.000000
50%	0.727273	0.272727
75%	1.000000	0.500000
max	1.000000	1.000000

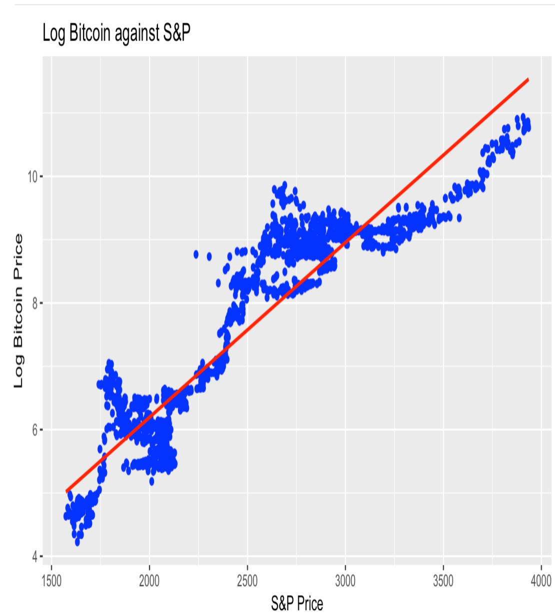
We have also provided the descriptive statistics for the tweets dataset above, which details the count, mean, min, max, standard deviation, etc. for the percent of positive and negative words per tweet. From this table, we can infer that on average, tweets involving cryptocurrencies are generally more positive (0.686) than negative (0.314).

IV. Data Visualization

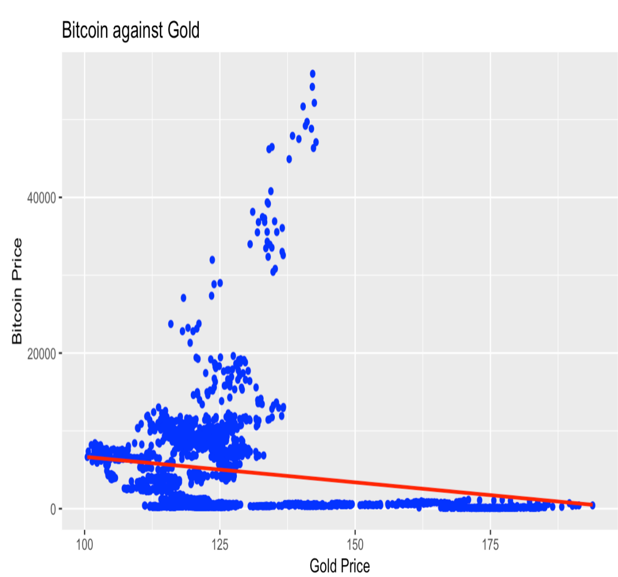
We have performed the following data visualization analyses to visualize the relationships of various variables before we have performed regression analyses.

S&P vs. Bitcoin Price





Gold Regression Data:



V. Analysis

A major part of our analysis was to measure the relationships between the different variables. More specifically, we wanted to know whether we can explain the price movement of Bitcoin through other markets or currencies. Regressing Bitcoin against S&P and Gold using the lm() function in R was a more basic first step in our analysis. For the purpose of this analysis, we have determined to focus on the Close price of the cryptocurrencies and stocks.

We also implemented a lag plot using ccf() that measures the correlation of the different cryptocurrencies with Bitcoin. The goal was to see whether Ethereum, Litecoin, Cardano, XRP, and Binance had a potentially lagged relationship with Bitcoin. There is already an expected positive correlation between Bitcoin and these currencies. The lagged relationship would show whether this relationship is instantaneous or lagged. This is also not just a one day lag, but extends up to 28 days before and 28 days after. The 28 days before is more like saying whether the movement in each of these coins impacts the future value of Bitcoin. This case might be more likely given how Bitcoin has higher trading volume and is a good market indicator for the value of cryptocurrencies in general. If lagged, that could help us have a higher likelihood of predicting the future values of these other currencies - something that could be potentially useful for investors.

VI. Results

We can see that the strongest results came from the linear regression of Bitcoin prices against the S&P. This regression goes back to the Data Visualization chart where it is clearly depicted that Log values of bitcoin is a better indicator. Bitcoin does not grow linearly as compared to S&P and Exponential growth is better covered with log, hence, the chart uses Log values for better visualization and to cope with Bitcoin's parabolic advances over time. We can also see that Bitcoin has a better relationship with S&P than Gold, and the reason could be because Gold is not as volatile and is a stable asset as compared to Equity.

From the below results we can conclude that $1 increase in bitcoin will result in a $10 increase in s&p prices. We can see that R2 value gets better with the Log function, hence the better fit.

Linear Model: Bitcoin against S&P

	Dependent variable:
	Close Bitcoin price
S&P Close	10.463 (0.357)
Constant	-20,764.140 (802.709)
Observations	1,973
R2	0.672
Adjusted R2	0.672
Residual Std. Error	3,948.585
F Statistic	4,046.757
Note:	NA

Linear Model: Log Bitcoin against S&P

	Dependent variable:
	log_bitcoin Log Bitcoin price
S&P Close	0.003 (0.00003)
Constant	0.676 (0.065)
Observations	1,973
R2	0.854
Adjusted R2	0.854
Residual Std. Error	0.617
F Statistic	11,517.040
Note:	NA

Linear Model: Bitcoin against Gold

	Dependent variable:
	Close Bitcoin price
Gold Close	-65.674 (4.325)
Constant	13,234.990 (521.916)
Observations	1,973
R2	0.032
Adjusted R2	0.031
Residual Std. Error	6,788.617
F Statistic	64.893
Note:	NA

Linear Model: Log Bitcoin against Gold

	Dependent variable:
	log_bitcoin Log Bitcoin price
Gold Close	-0.041 (0.001)
Constant	12.702 (0.152)
Observations	1,973
R2	0.229
Adjusted R2	0.229
Residual Std. Error	1.418
F Statistic	585.473
Note:	NA

Below you will find the lag plots to see the relationship of Bitcoin with other cryptos both instantaneously and with several days of lag. In other words, the +26 is basically saying that for each of these different smaller crypto exchanges (let's say in this first case with Binance), Binance is lagging Bitcoin by 26 days; then what is the correlation at that point in time. The -26 would be that Binance leads Bitcoin by 26 days and then measures the correlation at that time. Either way, the results show that the relationship with all the crypto exchanges are best at time 0. Therefore, the strongest relationships happen to be instantaneous. Additionally, we can see that the strongest correlation happens to be with Ethereum (86% correlation at time 0). At the very bottom, we also tested the relationship with the S&P. The results were consistent with the strongest relationship and correlation seems to be at time 0, so that relationship as well seems to be instantaneous.

Binance Lag Plot:

```
Autocorrelations of series 'X', by lag

  -26   -25   -24   -23   -22   -21   -20   -19   -18   -17   -16   -15   -14   -13   -12   -11   -10    -9    -8    -7    -6    -5    -4    -3
0.211 0.213 0.216 0.218 0.221 0.223 0.226 0.228 0.233 0.238 0.243 0.249 0.258 0.269 0.287 0.311 0.333 0.358 0.383 0.416 0.454 0.523 0.577 0.625
   -2    -1     0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20    21
0.677 0.728 0.774 0.768 0.761 0.747 0.732 0.711 0.686 0.666 0.644 0.618 0.597 0.577 0.557 0.536 0.518 0.509 0.500 0.494 0.490 0.487 0.484 0.486
   22    23    24    25    26
0.492 0.493 0.494 0.490 0.487
```

Cardano Lag Plot:

```
Autocorrelations of series 'X', by lag

  -26   -25   -24   -23   -22   -21   -20   -19   -18   -17   -16   -15   -14   -13   -12   -11   -10    -9    -8    -7
0.123 0.130 0.137 0.145 0.153 0.159 0.167 0.175 0.186 0.198 0.211 0.223 0.240 0.260 0.281 0.310 0.339 0.369 0.398 0.429
   -6    -5    -4    -3    -2    -1     0     1     2     3     4     5     6     7     8     9    10    11    12    13
0.460 0.490 0.526 0.556 0.590 0.623 0.661 0.657 0.653 0.645 0.637 0.624 0.608 0.595 0.581 0.570 0.558 0.545 0.535 0.522
   14    15    16    17    18    19    20    21    22    23    24    25    26
0.508 0.503 0.499 0.492 0.486 0.481 0.474 0.468 0.466 0.460 0.455 0.447 0.440
```

Ethereum Lag Plot:

```
Autocorrelations of series 'X', by lag

  -28   -27   -26   -25   -24   -23   -22   -21   -20   -19   -18   -17   -16   -15   -14   -13   -12   -11   -10    -9    -8    -7    -6    -5
0.496 0.507 0.519 0.528 0.538 0.548 0.559 0.567 0.577 0.587 0.598 0.611 0.625 0.639 0.654 0.670 0.686 0.702 0.718 0.734 0.750 0.767 0.784 0.802
   -4    -3    -2    -1     0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19
0.818 0.831 0.843 0.854 0.864 0.855 0.845 0.834 0.822 0.808 0.795 0.782 0.768 0.756 0.745 0.734 0.724 0.714 0.703 0.695 0.688 0.681 0.675 0.668
   20    21    22    23    24    25    26    27    28
0.661 0.655 0.649 0.641 0.634 0.625 0.617 0.607 0.598
```

XRP Lag Plot:

```
Autocorrelations of series 'X', by lag

  -29   -28   -27   -26   -25   -24   -23   -22   -21   -20   -19   -18   -17   -16   -15   -14   -13   -12   -11   -10    -9    -8    -7    -6
0.419 0.418 0.418 0.418 0.418 0.417 0.417 0.418 0.418 0.419 0.421 0.424 0.427 0.430 0.434 0.438 0.441 0.445 0.449 0.454 0.460 0.465 0.471 0.477
   -5    -4    -3    -2    -1     0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18
0.485 0.492 0.498 0.504 0.508 0.512 0.511 0.511 0.510 0.509 0.508 0.506 0.505 0.505 0.505 0.505 0.504 0.502 0.499 0.497 0.495 0.493 0.489 0.485
   19    20    21    22    23    24    25    26    27    28    29
0.480 0.476 0.471 0.467 0.462 0.459 0.454 0.450 0.446 0.441 0.437
```

Litecoin Lag Plot:

```
Autocorrelations of series 'X', by lag

  -29   -28   -27   -26   -25   -24   -23   -22   -21   -20   -19   -18   -17   -16   -15   -14   -13   -12   -11   -10    -9    -8    -7    -6
0.542 0.547 0.552 0.557 0.561 0.566 0.570 0.574 0.578 0.582 0.586 0.591 0.597 0.604 0.610 0.617 0.624 0.633 0.642 0.650 0.659 0.669 0.681 0.692
   -5    -4    -3    -2    -1     0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18
0.703 0.713 0.721 0.728 0.735 0.741 0.736 0.730 0.724 0.716 0.708 0.699 0.690 0.680 0.671 0.663 0.655 0.647 0.639 0.631 0.625 0.619 0.614 0.608
   19    20    21    22    23    24    25    26    27    28    29
0.604 0.599 0.595 0.591 0.586 0.581 0.576 0.572 0.567 0.562 0.555
```

S&P 500 Lag Plot:

```
Autocorrelations of series 'X', by lag

  -29   -28   -27   -26   -25   -24   -23   -22   -21   -20   -19   -18   -17   -16   -15   -14   -13   -12   -11   -10    -9    -8    -7    -6
0.753 0.755 0.757 0.760 0.762 0.765 0.767 0.770 0.772 0.774 0.775 0.777 0.779 0.781 0.784 0.786 0.789 0.792 0.794 0.797 0.799 0.802 0.805 0.807
   -5    -4    -3    -2    -1     0     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18
0.810 0.812 0.814 0.816 0.818 0.820 0.813 0.806 0.798 0.790 0.781 0.772 0.764 0.755 0.747 0.739 0.732 0.724 0.717 0.709 0.703 0.697 0.692 0.686
   19    20    21    22    23    24    25    26    27    28    29
0.681 0.675 0.670 0.665 0.660 0.655 0.650 0.645 0.639 0.633 0.627
```

VII. **Conclusion**

In general, the use of the regression analysis and the Twitter analysis could help us paint a better picture of the causal relationship. In the majority of financial analyses between two variables, the cause and effect of price movements could never be 100% determined. It becomes especially a more difficult task if we are dealing with such highly liquid investments with high trading volume where movements are happening in nanoseconds. Bitcoin and the stock market tend to have a decently strong relationship. There could be some evidence that they are motivated by each other's movements. It is also likely that there is an overlap in the information that affects both.

As for recommendations, we believe that investment advice could not be given based on the performance of another sector or cryptocurrency. There might be an argument in favor of investing if there is some sort of lag where the increase in one market, currency or crypto could be a good enough reason to invest in another market, currency or crypto. The instantaneous relationship between each other is a good case for further evaluation rather than a reason for making an investment decision. So these results are actually important because if you understand the forces that cause each of these investments to go up, you get a better idea of the different sectors or markets. Sometimes this can be measurable as well. With the twitter sentiment analysis, we understand what people are talking about and how they feel about specific events. Knowing what drives these investments as a result of understanding why such a relationship exists are the useful tools in making a prudent investment decision.