USC Viterbi
School of Engineering

Exploratory Analysis of Cryptocurrencies and Stock Market Indices

Eric Crouse
DSCI 510 - Introduction to Programming for Informatics
Prof. Gleb Satyukov

Fifteen years ago, there was no such thing as a Bitcoin. Today, a single Bitcoin is worth over $40,000. What happened? When it was first launched in 2009, Bitcoin was obscure and had virtually no real value. It remained this way for almost the first decade of its existence before undergoing a massive 1,338% growth in 2017. This explosive growth was followed by a less sudden crash, steady growth over a two year span, and then another massive boom in 2020-2021. Other currencies modeled after Bitcoin began to pop up left and right, and pretty soon everyone was talking about cryptocurrency. Given the nature of cryptocurrency's identity as an increasingly popular alternative to centralized fiat currency, the question naturally arises: is there a relationship between the two? If so, what kind?

To explore answers to these questions, data would need to be collected about cryptocurrencies' and the stock market's performances over time. Four indicators were chosen to represent the stock market: the DOW Jones Industrial Average, S&P 500, NASDAQ Composite, and Russell 2000. These indicators were a natural choice, considering they are widely used to gauge the overall health and performance of different segments of the stock market. Selecting cryptocurrencies was less straightforward, since several of them are not well established or indicative of major market trends. For this reason, only two cryptocurrencies were selected: Bitcoin (BTC) and Ethereum (ETH), which have the highest market volume, greatest acceptance, and most stability by a wide margin. All data collected was processed to ensure accuracy, appropriate data types, and consistent formatting, as well as to detect and resolve null values and potential duplicates. Target data was selected based on relevant information and initial data was filtered accordingly.

Historical cryptocurrency data featuring valuations in USD across time was sourced from Kaggle for BTC and ETH in the form of individual .csv files. This data was then manually compared to values sourced from Binance to ensure accuracy. The BTC dataset ranged from April 2013 to July 2021, but the data for Ethereum ranged only from August 2015 to July 2021 since Ethereum did not launch until the summer of 2015.

These datasets each contained a 'Date' column, which I formatted to match and used as a key to perform an outer join–ensuring that the resulting dataset was populated with NaNs for Ethereum values for all rows prior to Etherum's launch. To ensure that NaNs from dates prior to Ethereum's launch would not interfere with results, the original Ethereum dataset was used for calculations involving Ethereum as necessary (note that the .corr() method in Python pandas, which was used to compute a correlation matrix, automatically handles NaN values by excluding them pairwise for each pair of columns being correlated).

Historical stock market data from that same time period was fetched from Yahoo Finance API through Python's yfinance library. yfinance automatically sets 'Date' as the index, so the index had to be reset and the resulting 'Date' column was formatted to match that of the merged cryptocurrency dataset. There was another caveat when it came to this resulting financial dataset:

it was significantly smaller than the cryptocurrency dataset. Even though the dates spanned the same range, the stock indices dataset was missing several records. This is due to the fact that cryptocurrencies are traded 24/7/365, while the US stock market only has 252 trading days in a year. To resolve this, the datasets were merged using outer join on 'Date' and null values for stocks were forward filled to reflect the closing price of the most recent trading day. This preserves the consistency of the dataset and the integrity of the data.

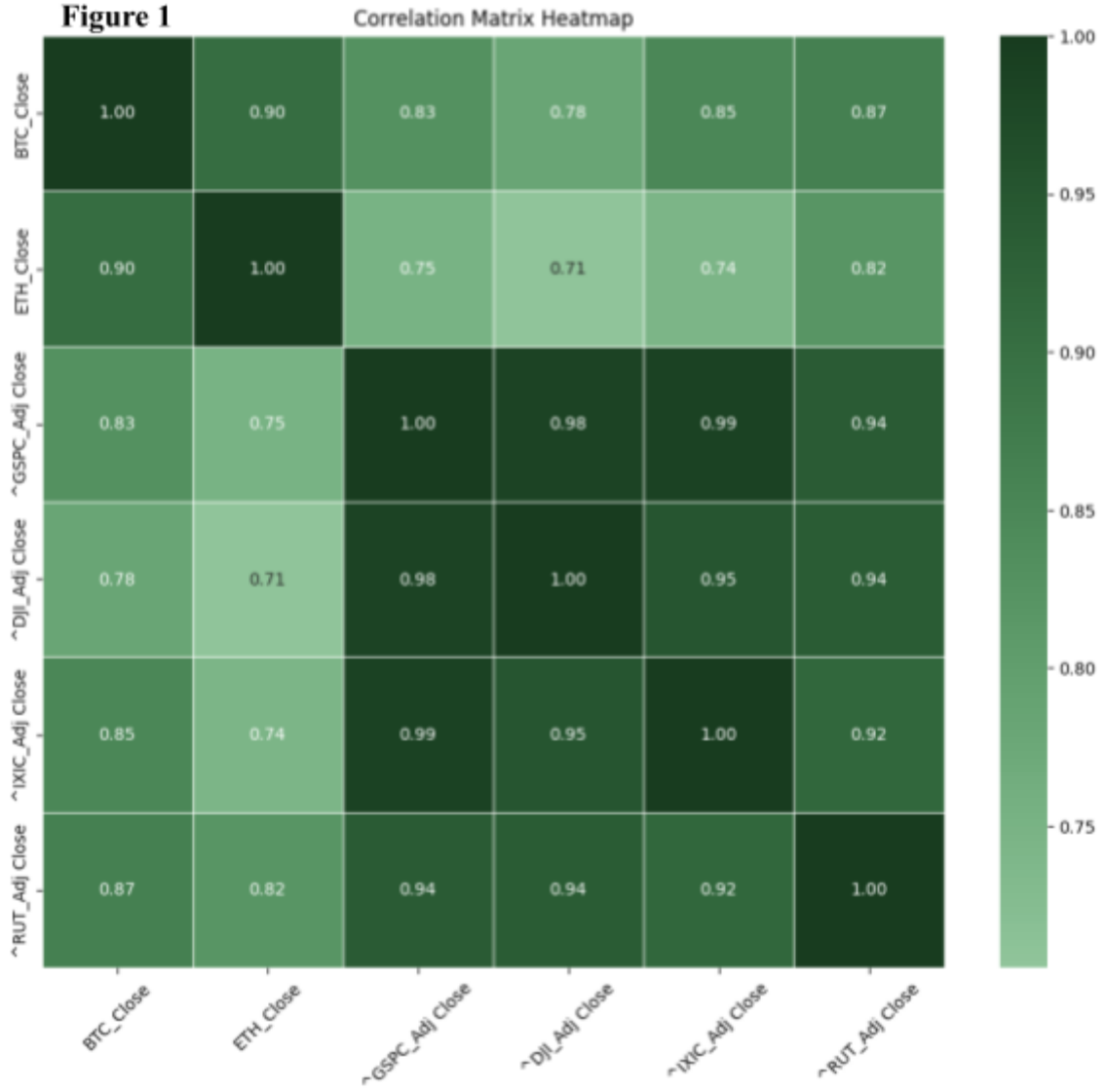
These datasets also included several attributes, such as Name, Symbol, Open, High, Low, Close, Adjusted Close, Volume, and Market Cap. Since most of these attributes were not necessary for time-series analysis of overall performance over the eight-year span, the merged dataset was filtered to select target data. Adjusted Close was chosen over Close for stock market indicators, as domain knowledge asserts that it is a more accurate reflection of the market's performance.

I had originally intended to provide live visuals using continuously updated dynamic visualizations generated from yfinance and CoinAPI, but abandoned these efforts primarily because the insight offered / information gain of such visualizations was deemed not enough to justify the amount of work required to generate them. Additionally, I had originally collected data for and planned to analyze several cryptocurrencies but later decided that since Bitcoin and Ethereum had the highest market cap and were the most popular by a wide margin, they would be more indicative of the state of cryptocurrency as a whole and other coins may uncharacteristically skew the results.

Also, I had originally hoped to incorporate a sentiment analysis aspect using Twitter/X to gauge public sentiment towards the stock & cryptocurrency markets and compare it to their respective performances. This, however, turned out to be beyond the scope of this semester-project. Since both of these ideas deviate from the rest of the project and require further efforts to integrate into a single cohesive endeavor, I opted to push forward with the data already collected.

The processed data frame had attributes for the closing prices of all four indicators and both cryptocurrencies across approximately eight years (six years for ETH), organized by date. It was first used to generate a correlation matrix in order to get an understanding of how closely related the individual components were to each other. The correlation discovered between cryptocurrencies and stock market indicators was higher than anticipated. See below for the resulting matrix.

Figure 1



The correlation matrix offers a few interesting results. Across the board, correlation coefficients are very high. The lowest value to be found anywhere is 0.71 between ETH and the Dow. Both cryptocurrencies' correlation coefficients with market indicators were significant, indicating a positive linear relationship, and their correlation with each other indicates a very strong positive relationship. It is interesting to note that the Russell 2000 had the strongest relationship with both cryptocurrencies of any of the stock market indicators. This is likely due to the facts that the Russell 2000 represents the smallest publicly traded companies in our analysis and smaller companies tend to have more volatility in their share prices.

In order to visualize these relationships across different points in time, a dynamic time-series graph was created using Python's plotly.express showing valuations in USD of each variable

over the chosen interval. A static version of this plot is shown below:

Figure 2

On this plot we see the explosive growth of BTC coupled with what appears to be general upward trends of the other variables. To get a better understanding of each variable's true performance in this context, the data was then normalized using a logarithmic transformation with base e, and the results were plotted in the same manner (see below).

Figure 3

Looking at the normalized data, we can get a much clearer picture. First, it becomes readily apparent that ETH's performance resembles BTC much more strongly than it resembles performances of market indicators, which was not at all clear prior to the transformation. Since the natural logarithmic transformation models these variables' growth rates over time, it is also much easier after the transformation to see the consistent trend in cryptocurrency valuations of steady overall growth despite volatility.

The next analysis conducted was an Ordinary Least Squares (OLS) Regression, which was initially performed using the four stock market indices as independent variables and Bitcoin

valuation as the dependent variable. This resulted in a high R^2 value of 0.830 (meaning the model predicts that 83% of the variance in BTC can be explained by stock market indicators), at which point I realized that using all four market indicators as independent variables gave rise to the issue of strong multicollinearity. To avoid this, I repeated the regression for each stock market indicator separately and found R^2 values ranging from 0.749 to 0.612, which better aligned with intuition.

The OLS summaries were generated using Python statsmodels.api and included Durbin-Watson scores which were extremely low, suggesting high autocorrelation. High autocorrelation would indicate that the variables display patterns over time and are subject to momentum, which has been anecdotally observed in cryptocurrencies. To explore these patterns, I selected the most relevant variable from each category–which I deemed to be BTC and the S&P 500, since the S&P 500 offers a broad representation of the stock market–and began to plot them with and against each other. Consider the three plots shown below.

Figure 4

Figure 5

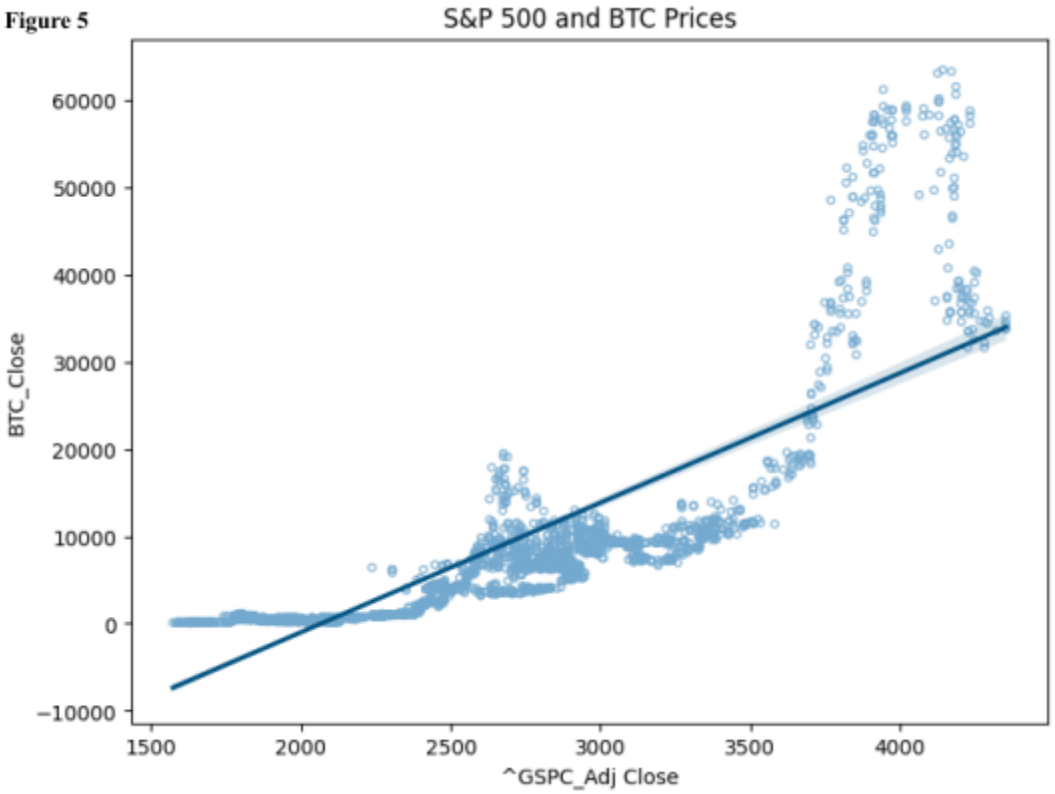
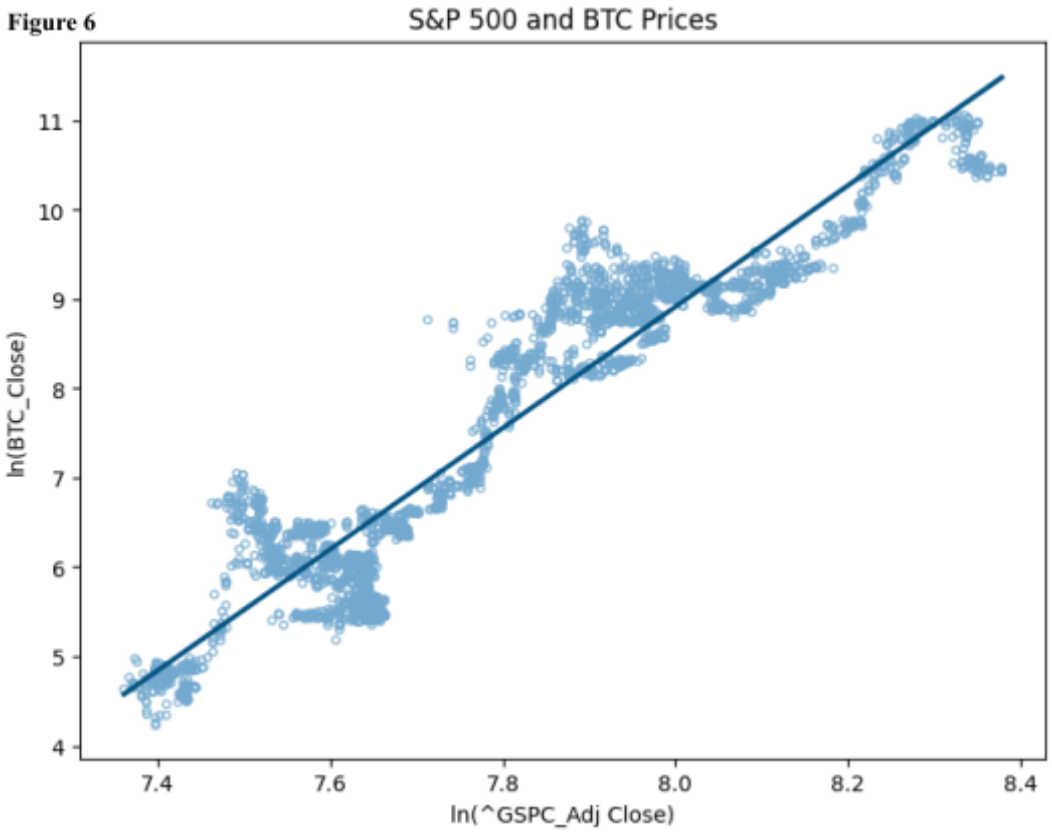


Figure 6



First we have valuations in USD of BTC and the S&P 500 over time. GSPC increases steadily while BTC is subject to high variation, yet increases massively overall.

The next graph (Figure 5) plots the relationship between GSPC and BTC valuations. Since the S&P 500 has increased steadily over time, increased values of the GSPC on the X-axis are generally equivalent to later points in time and the graph is mostly stable. This is why the trend of BTC values on this plot is strikingly similar to the trend shown in Figure 4.

In the third graph (Figure 6), we see something interesting. Plotting the natural logarithm values of GSPC and BTC against each other reveals a helix formation (easily visualized as spiraling around the regression line). Since values of ln(x) represent the growth rate of x, we know that increasing values of ln(x) represent increasing rate of growth of x, and decreasing values of ln(x) represent decreasing rate of growth of x. In the plot, values of ln(BTC) are increasing and decreasing on a fairly regular interval with respect to ln(GSPC).

From this information, we can interpret the following:

The helix formation on this graph is a visual representation of the volatility of Bitcoin with respect to US stock market indicators (specifically the S&P 500).

The regularity (symmetry) of the helix shows that this volatility occurs in roughly equal parts positive and negative.

The upward trend shows that while the rate of change of Bitcoin may swing dramatically between positive and negative, overall there is substantial growth not only in the valuation of Bitcoin, but in the rate at which this valuation is increasing.

Finally, the helix formation of BTC values indicates the presence of cyclic behavior. To further investigate this, I chose to plot the normalized values of both Bitcoin and the S&P 500 over time.

In order to do so, the ‘Date’ attribute was converted from datetime data type to integer value as ‘NumericDate.’ This revealed similar trends in the two over the same time interval, with the primary difference being higher volatility of Bitcoin. Both variables appear to display cyclical patterns, which indicate that future prices can be predicted based on current and past trends. These patterns are more noticeable with BTC due to its higher volatility.

Figure 7 Normalized BTC Performance

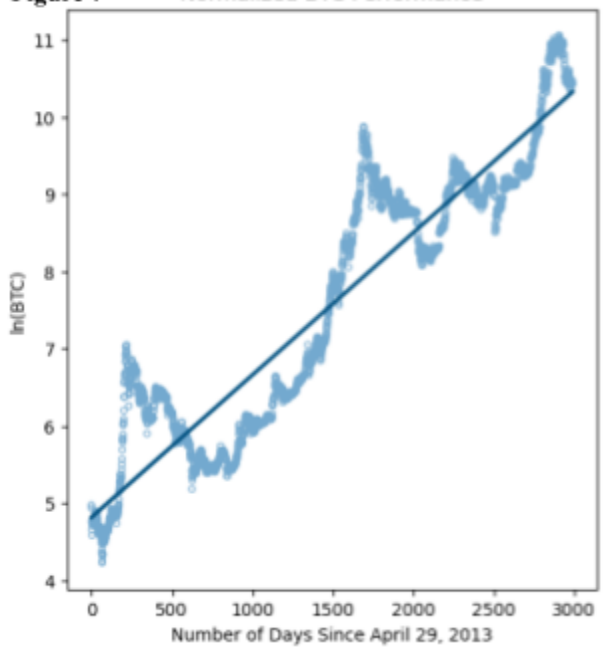
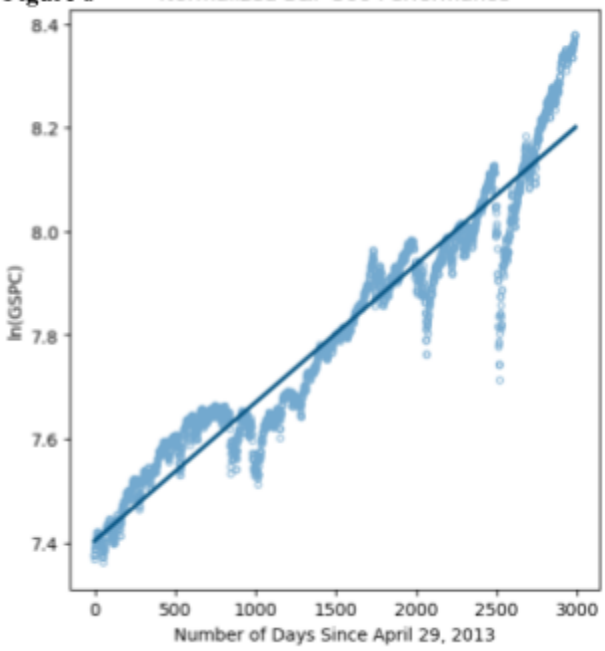


Figure 8 Normalized S&P 500 Performance



Overall, the findings of this exploratory analysis consistently demonstrate a strong positive correlation between valuations of leading cryptocurrencies and US stock market indices. Growth periods, inflection points, and overarching trends are more congruent underneath the surface than one would initially expect. Additionally, both stock market indices and cryptocurrencies appear to follow cyclical patterns that can be exploited in future predictive models.

I had initially expected to find some form of negative relationship wherein cryptocurrency booms closely followed crashes of the stock market. While my analysis did not demonstrate evidence of this, it may still be true that public fear / uncertainty during the 2020 stock market crash and associated events was one of multiple driving factors behind the cryptocurrency boom that took place that year. Experiential observations indicate that many of the same factors that drive stock market growth also drive cryptocurrency growth, so isolation of these variables' direct impact on each other may reveal a different relationship that is overshadowed by other factors.

Interesting and unexpected findings include the insights offered by logarithmic transformation of values, particularly in the patterns they were able to uncover. This would make an excellent basis for future exploration and I would like to look into these relationships further given more time.

Additionally, incorporating sentiment analysis as previously mentioned could offer another dimension to the understanding of these two categories of financial tools. If more time were available, I would be interested in integrating an API from Twitter/X or another social media platform to explore the public sentiment regarding cryptocurrency and the financial market during the same time frame.