Springboard Data Science Intensive Program – Capstone Project #1

Assessing Bitcoin's Place in Traditional Finance (working title)

Garrick Chu- February 2018

"Bull markets are born on pessimism, grow on skepticism, mature on optimism and die on euphoria."

- Sir John Templeton

Table of Contents

1. Introduction:

Bitcoin was created in 2009 out of a white-paper published under a nom-de-plume, Satoshi Nakamoto. Though the actual identity of the creator of Bitcoin and Satoshi are still unknown to this date, the recent media attention and meteoric rise of Bitcoin are very real. Bitcoin and the underlying Blockchain technology have been both praised as a revolutionary decentralized ledger system and criticized as a scam. However, there is no denying that Bitcoin has begun to gain the attention of the general public, media outlets and Wall Street. Stories of rags-to-riches, calls for regulation and public stances of prominent Wall Street figures are ever more present in our news feed and headlines.

For Wall Street (and to a lesser degree, Main Street), this begs the question: Does Bitcoin have a purpose in traditional finance? Do we define Bitcoin as a medium of exchange, a globally accepted currency, or conduit for fraud and shady transactions? Is Bitcoin another asset bubble that will turn out badly as we have seen in recent past? And can we formally identify driving factors in the Bitcoin market and make predictions?

Though an obvious question, the answers have proved to be much more elusive. The search for answers has become more urgent as Wall Street and individual investors fear missing their golden opportunity but are equally cautious of being the "greater fool."

I intend to take a data-driven approach using Data Science tools to uncover insights on these questions.

2. Data Acquisition

We acquired datasets from 3 main sources: Quandl (Quandl.com) for Bitcoin Market Data and market data on other cryptocurrencies, Yahoo! Finance for Stock Market Data, Chicago Board of Exchange (for Volatility Index data), the Federal Bank of St. Louis for S&P/Case-Shiller Home Price Index data.

Quandl's Bitcoin specific data also contains technical characteristics such as Number of Users (on the My Wallet service), Market Price and Transaction data. This data set spans from 2009 to December 2017.

I also obtained market data on other Cryptocurrencies including Ripple, Litecoin, Ether and Iota from Quandl and spans from 2017 through January 2018. I chose these sets because they are the next biggest coins after Bitcoin.

For my analysis on whether Bitcoin is similar to previous bubbles, I used: S&P/Case-Shiller Home Price Index data from 1997 to 2006 (known as the US Housing Bubble), Nasdaq Composite Index prices from 1997 to 2001 (also known as the Dot-Com Bubble) and a smaller data set on the Tulip Price Index spanning from 1636 to 1642 (Tulip Mania). I choose these sets as they are the most ubiquitous asset bubbles in modern finance.

3. Data Wrangling, Cleaning and Exploration

3a) Bitcoin Market and Technical Data

The Bitcoin dataset is an amalgamation of CSV files from Quandl with each variable having its own file. Here, I used a 'glob' function along with the Pandas library to read in the CSV files and combine into a single data set sharing the same time-series index. The result is a Pandas Dataframe containing 10 columns and 3259 rows:

Date	Value_AVBLS	Value_CPTRA	Value_MKPRU	Value_MWNTD	Value_MWNUS
2017-12-27	1.05	161.68	15999.04	45559.0	21272882.0
2017-12-26	1.05	146.59	14119.02	42867.0	21249422.0
2017-12-25	1.06	138.78	13949.17	49434.0	21204476.0
2017-12-24	1.06	139.69	15360.26	40957.0	21165559.0
2017-12-23	1.06	137.02	15190.94	43222.0	21100453.0

Date	Value_NADDU	Value_NTRAN	Value_NTRAT	Value_NTRBL	Value_TRVOU
2017-12-27	605853.0	247440.0	286214316.0	1742.53	1.745062e+09
2017-12-26	565074.0	228926.0	285966876.0	1695.74	1.134113e+09
2017-12-25	652209.0	279523.0	285737950.0	1838.96	1.798168e+09
2017-12-24	729637.0	308211.0	285458427.0	2217.34	1.967537e+09
2017-12-23	890731.0	380648.0	285150216.0	2455.79	5.352016e+09

We begin but taking an initial view at the raw time-series data in Fig. 1:

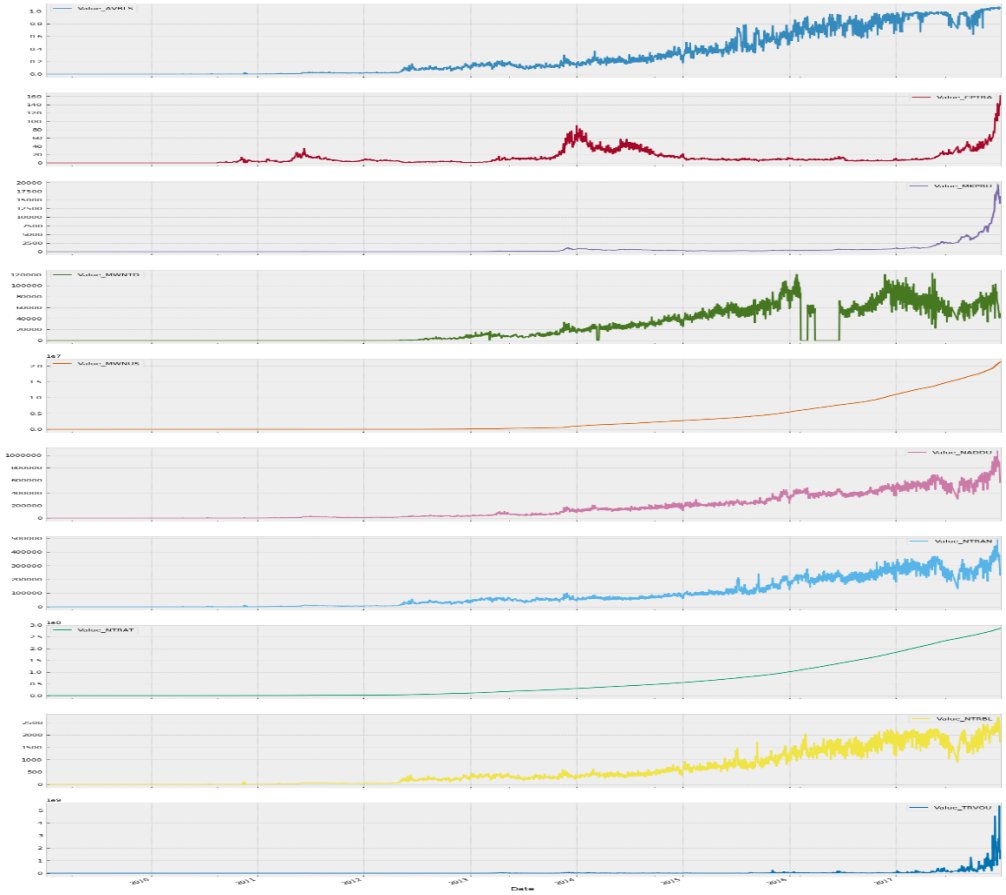


Fig. 1 – Time Series Plot of Columns in Bitcoin Data Set from 2009-2017

Note: The 4th subplot from the top indicates we have 0 values for the data. However, as I progressed in my analysis, we are able to discard this data and instead use the data in the following subplot instead.

At this stage, I wanted to figure out Bitcoin was increasingly being used as a medium of exchange. I set out to compute a new time-series set as "Average Number of Bitcoin per Transaction." I accomplished this by taking the Daily Transaction Volume in USD, divided this by Daily Closing Price in USD to get Daily Transactions. I then divided this value by Number of Transactions to get a rough average of Bitcoins per Transaction. I then expect this measure to be inversely related to the Price of Bitcoin. That is, as the price of Bitcoin increases, fewer Bitcoins are needed to transact on the same amount.

Instead, in Fig. 2 we see that the Bitcoins per transaction falls in a dramatic fashion through 2014 and long before Bitcoin's large rise in value. If the "medium-of-exchange" theory held true, we could see the BTC per transaction fall more dramatically in 2017 and not 3 years prior.

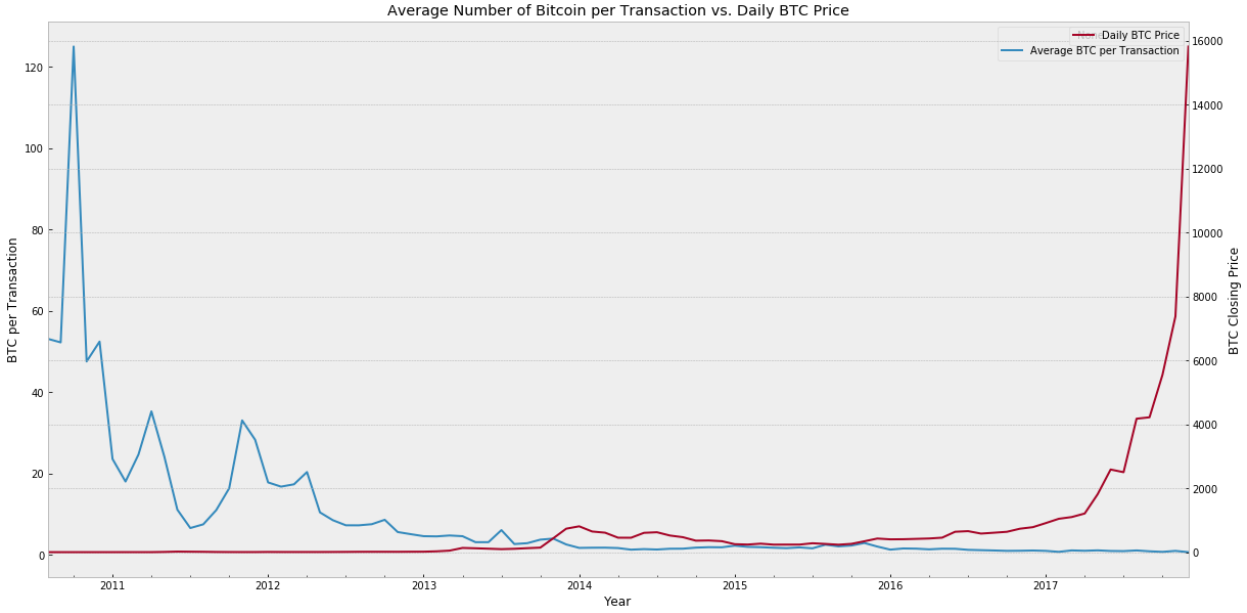
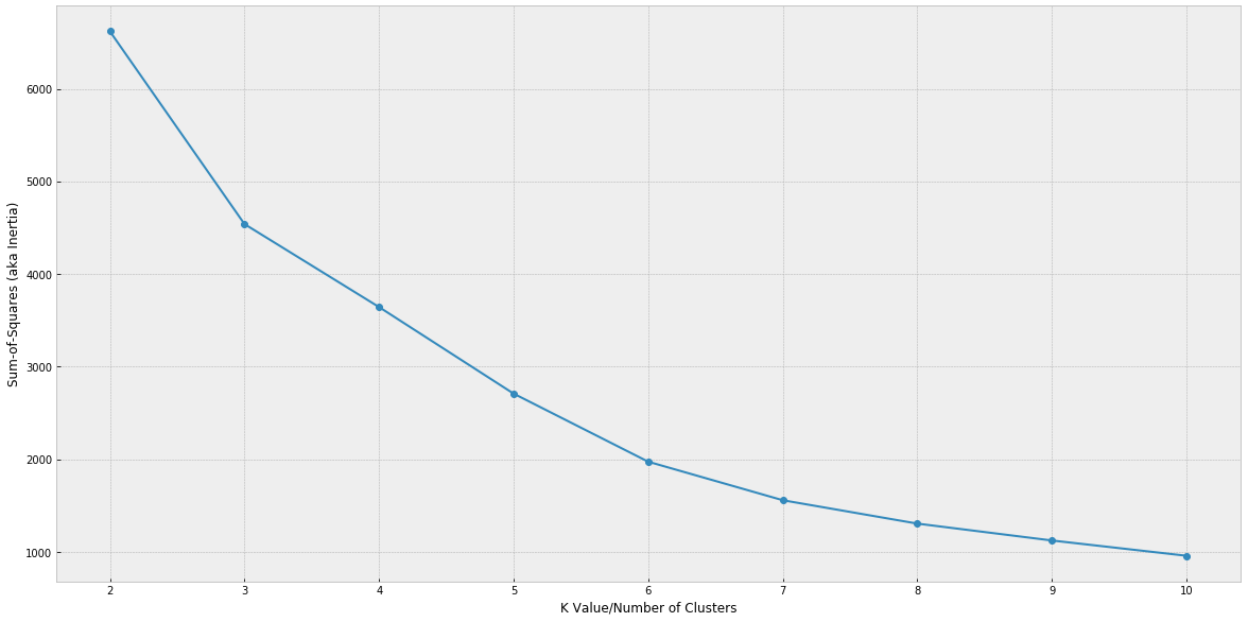


Fig. 2 – Bitcoin Per Transaction vs. Bitcoin Prices from 2010-2017

The second objective with this dataset was to explore if the measures in the dataset can classify distinct "phases" of Bitcoin's life. That is, I want to explore whether it is possible to classify "eras" of Bitcoin and to classify future Bitcoin characteristics as reminiscent as prior periods.

To accomplish this, I employed a K-Means clustering model and StandardScaler to normalize the data (both from the Sci-kit Learn library). Using the Elbow-method in Fig 2, I determined that the appropriate number of Clusters to use for the model is 6.



To visualize this, I created a scatter plot of Bitcoin Price to Bitcoin Transaction Volume with the data points color-coded with their respective Cluster/Grouping (Fig. 3).

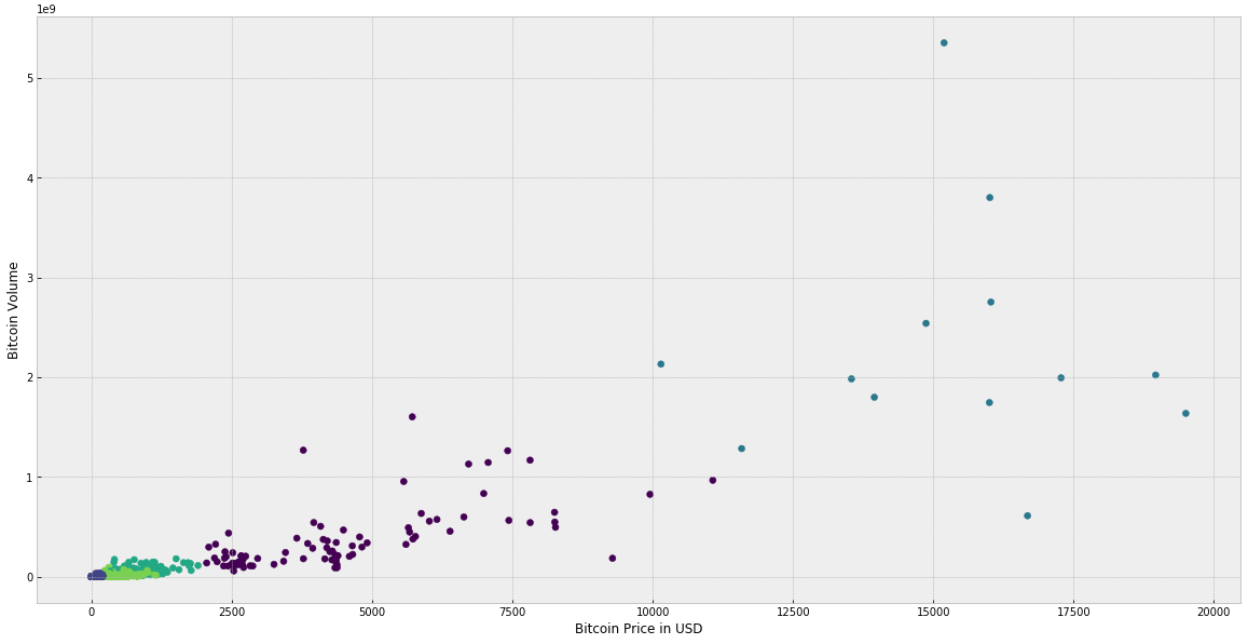
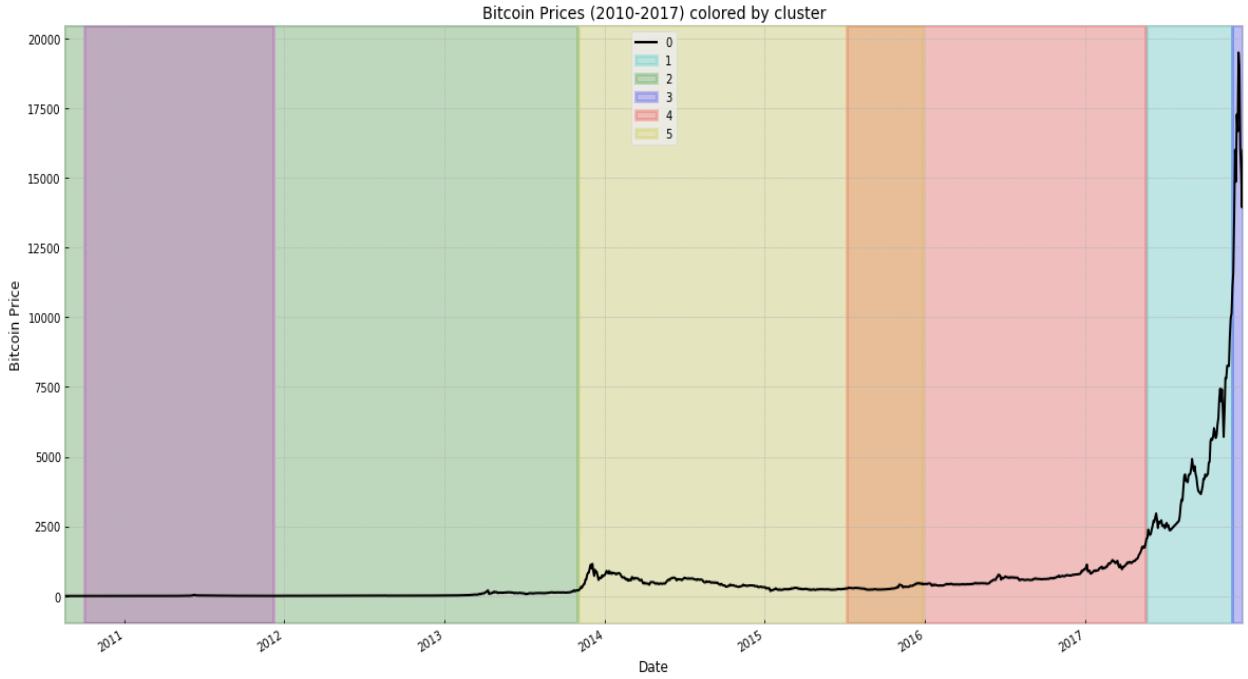


Fig. 3 – Bitcoin Price in USD and Bitcoin Transaction Volume (color-coded by respective cluster).

We see that the clusters are distinctive, even determined, by Bitcoin Prices and Volumes. Near the origin, a cluster may be defined as Early Bitcoin when volume and prices were low. On the other end, we see a cluster defined by high prices and high volumes.

On a timeline basis, we see the Bitcoin has gone through “life-stages” with respective contexts which I will explain in further detail (Fig. 4).



- The Green period is Bitcoin’s early growth stage where the dynamics of Bitcoin remained largely uniform. The exception is the Purple Section is overlaid on green which represents when alternative coins began to be introduced.
- The yellow section is marked by the shutdown for Silk Road, retailers such as Overstock, Newegg and Dell began accepting Bitcoin as payment.
- The Red section is when academic research around Bitcoin begins to increase, indicating Bitcoin finds its “normal” growth trajectory.
- The Cyan section is the pre-cursor to the huge December 2017 rally and when Bitcoin gains more legitimacy among lawmakers and legacy financial institutions (I call this the “rumor phase”).
- The last Blue section is solely December 2017 and that it is a standalone cluster may be indicative that in this era of time, Bitcoin was/is in an asset bubble (and is fueled by the “greater fool theory”).

Even with the recent correction and prices around $7000 USD, my K-Means model predicts that Bitcoin remains in the Blue cluster. In other words, correction or not, we are still in an over-valued/fever pitch phase.

3b) Bitcoin as a Portfolio Management Tool and Asset Bubble Analysis

As with the Bitcoin data, I acquired time-series data from different sources. I utilized Pandas to load in the various CSV files and combined into appropriate dataframes for analysis.

With the analysis of Bitcoin as a source of diversification/portfolio management, I needed to take static values of Bitcoin and traditional assets and find their respective returns. The main tenant of diversification is to create a portfolio of assets with uncorrelated returns. To get this, I created new columns of an asset's Percentage Return, indexed to the first point of available data.

Here I used statistical analysis to determine Pearson Coefficients and found the below:

	BTC_TR	VIX_TR	SP500_TR	Gold_Return	Oil_Return
BTC_TR	1.000000	-0.316832	0.542559	-0.334866	-0.250007
VIX_TR	-0.316832	1.000000	-0.618065	0.653232	0.205093
SP500_TR	0.542559	-0.618065	1.000000	-0.785429	-0.730141
Gold_Return	-0.334866	0.653232	-0.785429	1.000000	0.724267
Oil_Return	-0.250007	0.205093	-0.730141	0.724267	1.000000

Visualized as Fig. 5:

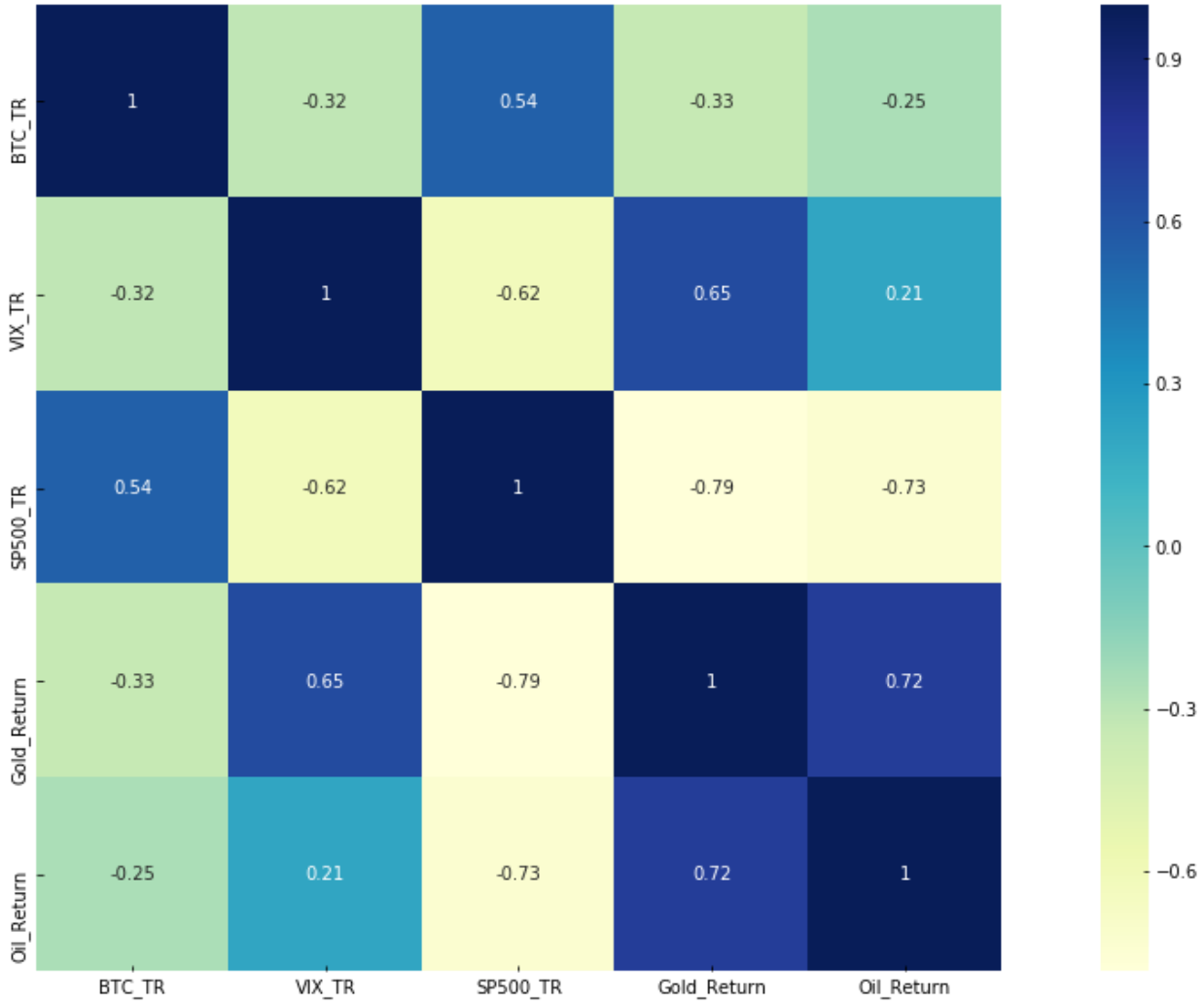
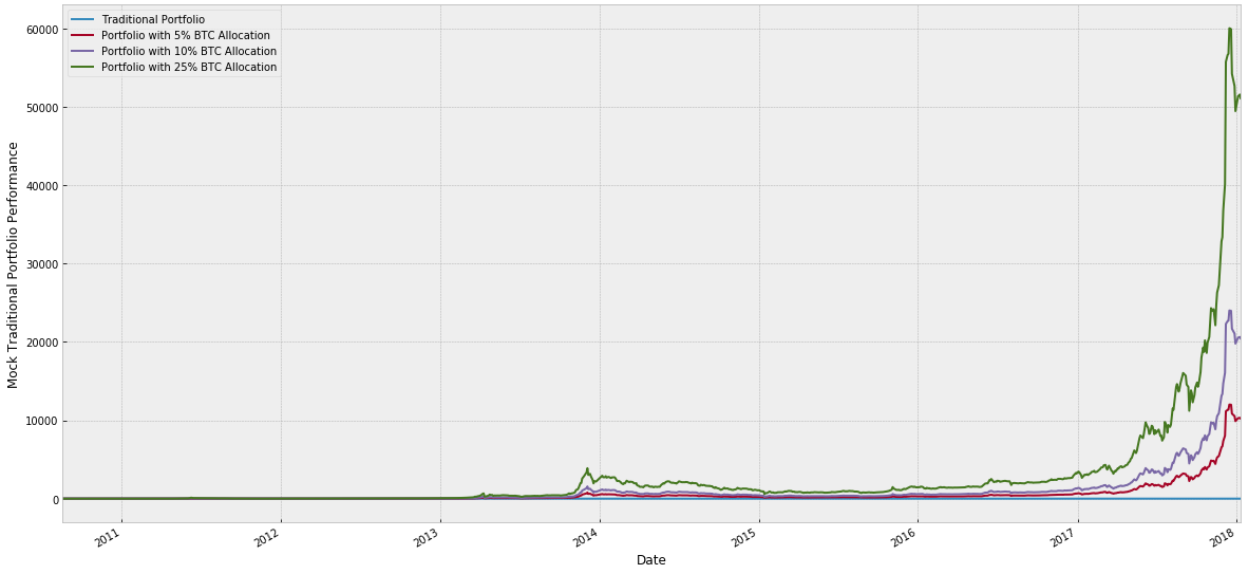


Fig. 5 – Correlation Coefficient Matrix Heatmap for the Total Return for Bitcoin, VIX, Gold, and Oil.

We find that the S&P 500 (a proxy for the broader stock market) is most correlated with Bitcoin's performance (from 2011 to 2017 at least). As a result, Bitcoin has very little correlation with assets found in a traditional portfolio and can be integrated into a traditional portfolio to benefit from added diversification.

For additional insight, I wanted to view how differently a traditional portfolio has performed since 2011 vs. a portfolio with 5%, 10% and 25% allocated to Bitcoin.



We find that a portfolio with just 5% of total capital invested in Bitcoin would have outperformed the standard portfolio by nearly 10,000%. In the case of a 10% allocation to Bitcoin, that portfolio would have outperformed by 20,000%. Lastly with a 25% allocation, the number jumps to nearly 50,000%.

This analysis was done without on a non-risk-adjusted basis. Further analysis using a Sharpe Ratio may yield different results.

The second objective in this area was to explore the similarities, if any, between Bitcoin and the most widely recognized asset bubbles. Here I used the S&P/Case-Shiller National Home Price Index from 1996 to 2007 (widely accepted as the US Housing Bubble), Tulip Price Index Data from 1634 – 1642, and Nasdaq Composite Index data from 1997 to 2001 (the focal point of the Dot-Com Crash). For the housing and Nasdaq data, no wrangling was required. All datasets were loaded in into their own dataframes for this exploration.

To visually explore the data side-by-side, I downsampled the Dot-Com data and also only used 2017 prices of Bitcoin (Fig. 6).

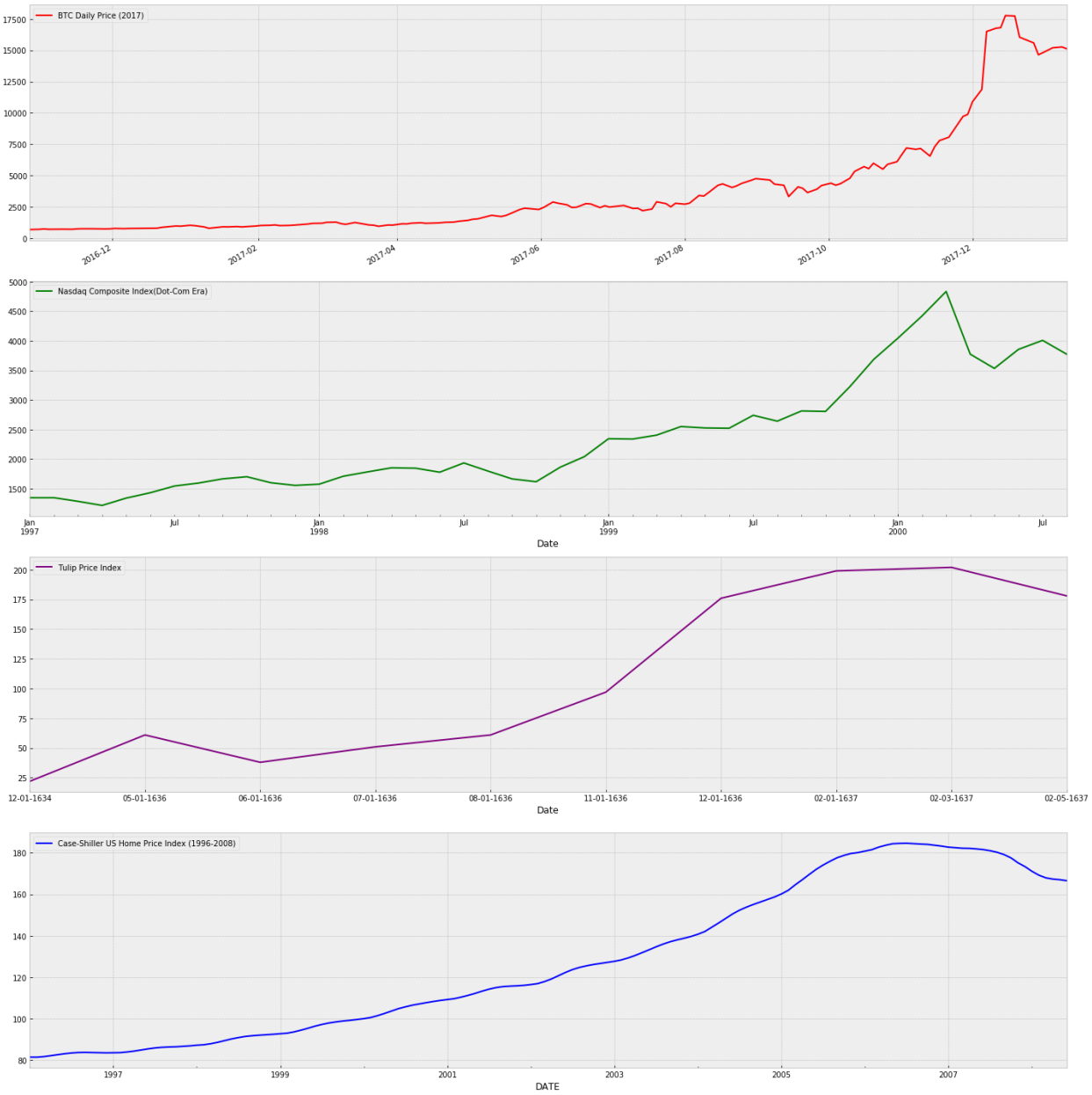


Fig. 6 – From Top to Bottom: Bitcoin Prices (2017), Nasdaq Composite Index (1997-2001), Tulip Prices in Netherlands (1634-1642), Case-Shiller Home Price Index (1997-2007).

The price trends among all 4 plots are strikingly similar but requires further analysis to confirm. Unfortuantely, Pandas uses a 64-bit integer for its Time-Series functionality and is limited the scope to years 1677 to 2220. As my Tulip data is out of this range, I was unable to manipulate the data for further analysis.

To quantify the above visual correlations, I again utilized statistical analysis by way of Pearson Coefficients. With respect to the Nasdaq data, I resampled the data on a Weekly basis and took the means within that period. This was necessary to ensure an equal size of data comparison. I received the following results:

```
Pearson R correlation for Bitcoin (2017) and the Dot-Com Bubble:
0.7513945799997369
P-Value: 1.6451377714107817e-28
Pearson R correlation for Bitcoin (2017) and US Housing Bubble:
0.7525625687801886
P-Value: 1.217934985768466e-28
```

This suggests that Bitcoin has a strong correlation with both the Dot-Com and US Housing bubble. A very low P-Value attests to the strength validity of this correlation.

3c) Comparison of Bitcoin and Biggest Alternative Coins (Ethereum, Litecoin, Ripple and Iota)

Similar to the Bitcoin data, I acquired the data from Quandl where each file contains time-series data from 2017 through January 2018 and High, Low, Mid, and Closing prices in addition to Daily Volume for each coin. Iota was introduced in 2017 and thus the combined/merged dataframe excludes data pre-June 2017. As a result, the dataframe for this analysis had 35 columns and 201 observations (days).

For the sake of simple visualization, I plotted the Mid-Daily Price of each coin and their respective volume over this time period:

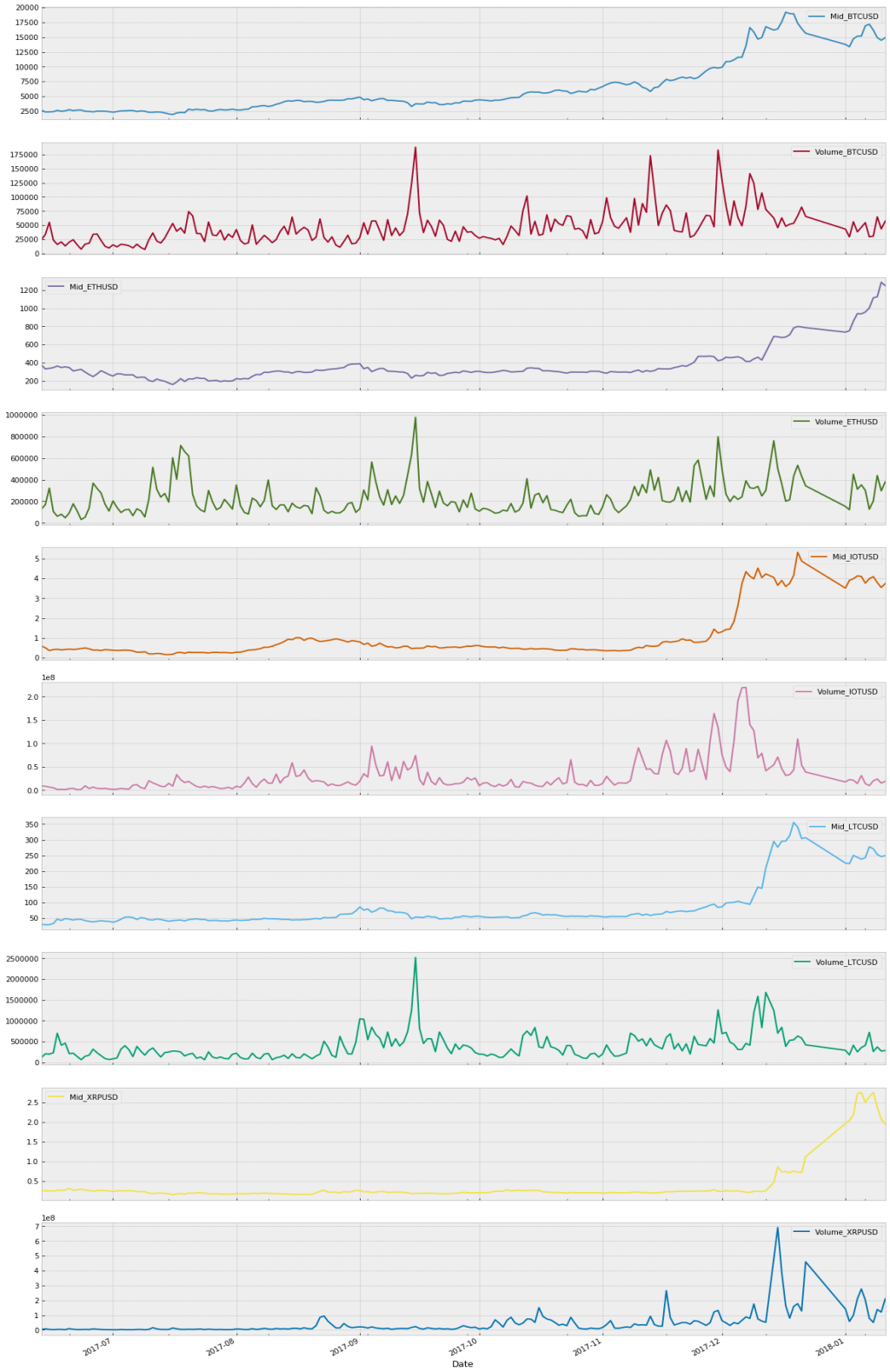
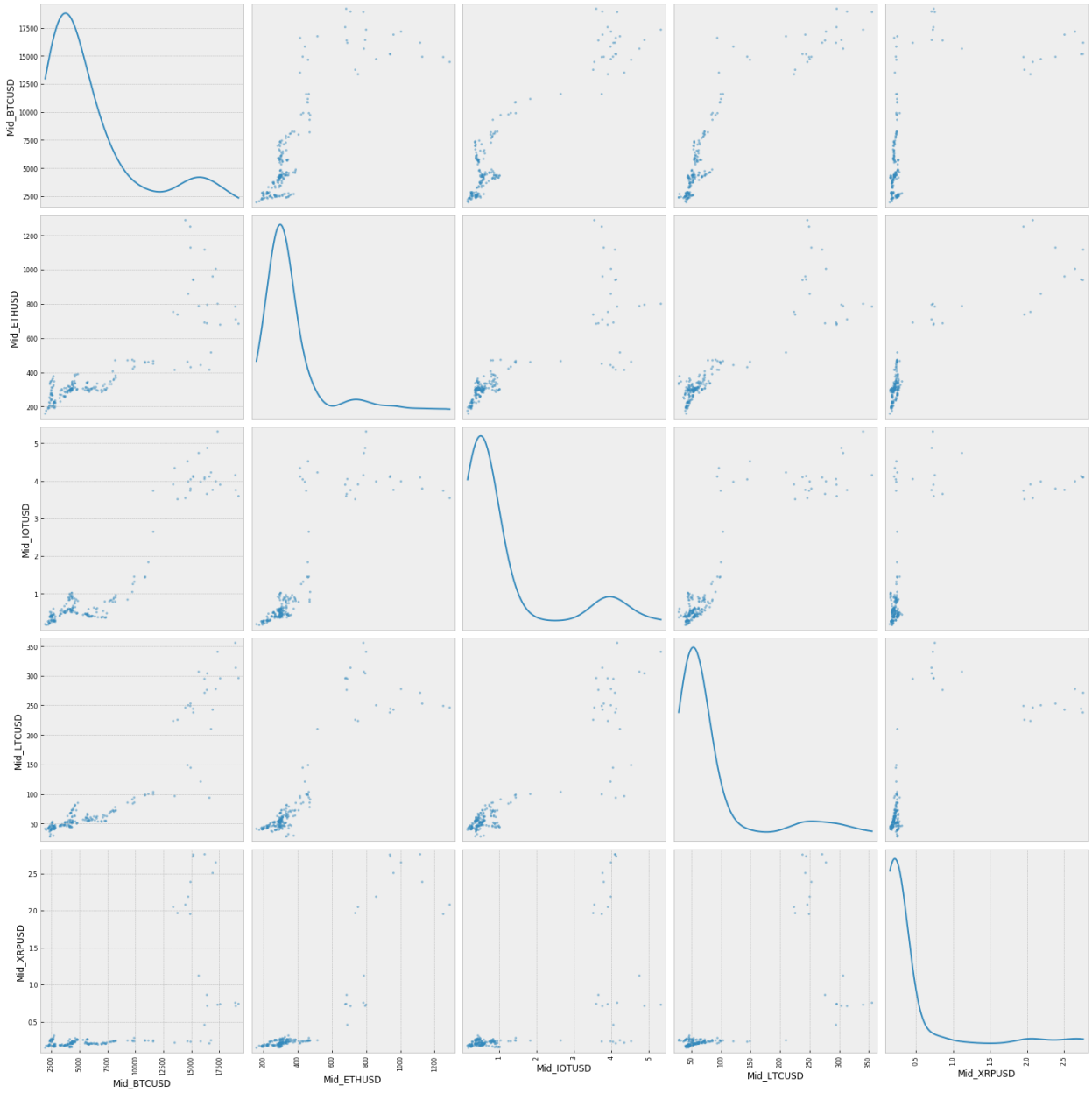


Fig. 7 – Average Daily Prices of Coins and respective volumes. From Top to Bottom: Bitcoin, Ethereum, Iota, Litecoin and Ripple.

Despite a relatively short time-series of less than a year, we see some stark similaraties in price trends in November – January. Furthermore, we see more similarities in the volumes of Bitcoin, Ethereum and Litecoin. (but not Ripple nor Iota).

I believe these correlations are more apparent using a Scatter Matrix:



While this visualization was not helpful to me. A closer look at their correlations coefficients uncover a high degree of correlation that may not be obvious visually:

	Mid_BTCUSD	Mid_ETHUSD	Mid_IOTUSD	Mid_LTCUSD	Mid_XRPUSD
Mid_BTCUSD	1.000000	0.831611	0.921704	0.899781	0.625674
Mid_ETHUSD	0.831611	1.000000	0.840506	0.891562	0.879821

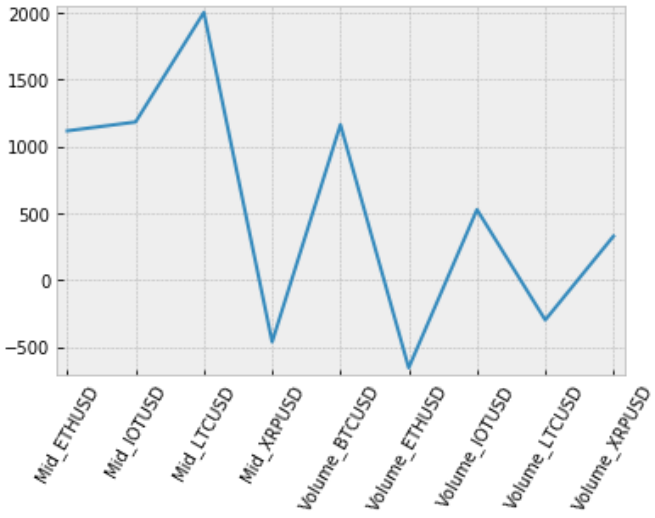
Mid_IOTUSD	0.921704	0.840506	1.000000	0.906808	0.672821
Mid_LTCUSD	0.899781	0.891562	0.906808	1.000000	0.737545
Mid_XRPUSD	0.625674	0.879821	0.672821	0.737545	1.000000

This was a similar case with volumes:

	Volume_BTCUSD	Volume_ETHUSD	Volume_IOTUSD	Volume_LTCUSD	Volume_XRPUSD
Volume_BTCUSD	1.000000	0.618239	0.556679	0.629883	0.262252
Volume_ETHUSD	0.618239	1.000000	0.392737	0.558947	0.344295
Volume_IOTUSD	0.556679	0.392737	1.000000	0.392757	0.290177
Volume_LTCUSD	0.629883	0.558947	0.392757	1.000000	0.296894
Volume_XRPUSD	0.262252	0.344295	0.290177	0.296894	1.000000

These correlations were substantial enough to build a model in efforts to predict Bitcoin prices.

After normalizing the data and further splitting into a training and test set, I utilized a Lasso and Linear Regression model side-by-side. In both models, Litecoin prices and Bitcoin volumes were adjusted to be overweight while underweighting Ripple prices and Ethereum Volumes as evidenced:



With the Lasso model, I achieved and accuracy score of 89.39%. With the Linear Regression model, I achieved a similar accuracy of 89.37%.

4. Limitations and further research

4a) Technical Limitations

Although minor in the broader analysis, I faced a technical limitation with Pandas and its Datetime functionality. This proved an obstacle with making the comparison with Bitcoin 2017 prices and the Tulip price data where the tulip data falls out of range.

Timestamp Limitations

Since pandas represents timestamps in nanosecond resolution, the time span that can be represented using a 64-bit integer is limited to approximately 584 years:

```
In [66]: pd.Timestamp.min
Out[66]: Timestamp('1677-09-21 00:12:43.145225')

In [67]: pd.Timestamp.max
Out[67]: Timestamp('2262-04-11 23:47:16.854775807')
```

4b) Data Limitations

The tulip data itself was also a concern as I pulled 14 data points of inconsistent frequency from an academic paper(a few sparse monthly data points and then a grouping of daily prices). I could not locate the actual index itself. This may be worth further exploration to conduct a proper comparison.

The Bitcoin data is also only from the My Wallet Users and may not be an accurate representation of the broader population of market participants.

The Bitcoin and cryptocurrency data was also at such a scale that made visual analysis a challenge. For example, attempting to visualize Bitcoin in 2010 (at $0.07) and in 2017 (at $14,165.57) on the same plot.

4c) Further research

The Quandl Bitcoin data is only a selection of all the measure available. There may be other features worth exploring to improve the Bitcoin Price Regression models.

I also suspect other features may also improve this model including the broader stock market (we saw a firm correlation with the S&P) and other digital coins.

5) Recommendations and Actions (more work here)

1. <u>Bitcoin is a source of diversification for a risk-tolerant traditional portfolio.</u>

(write more here)

2. <u>Bitcoin is presently in treacherous territory, likely still in “Greater Fool” territory.</u>

My analysis supports the idea that recent Bitcoin activity is entirely distinctive from the rest of Bitcoin's life. The model further suggests that we remain in this particular life-cycle stage despite the massive correction. Proceed with caution.

3. Diversifying with other coins is unlikely.

We see many correlations among Bitcoin and other coins that achieving diversification may be difficult. Rather, I propose treating Bitcoin and other coins as a single digital asset class with a focus on diversifying across asset classes. In other words, buying a variety of different coins isn't diversification but adding Bitcoin and other coins to a portfolio mixed in with traditional classes is recommended.

Bitcoin is likely a storage-of-value than an medium of exchange.

As such, Bitcoin should be treated as an investment vehicle and not currency for the purposes of classifying Bitcoin into an asset class. Would not recommend using Bitcoin for purposes of Cross-Currency exchange for portfolio management.