Capstone Project

Data Science Nano Degree

Table of Contents

Introduction

Project Overview

Bitcoin (BTC) is proposed to provided a modern era store of value and be a potential successor to Gold (GLD) in this respect (Ammous, 2018).

Ammous (2018) illustrates that one of the key determinate factors of both BTC and GLD in conferring this property is that they both demonstrate a high stock-to-flow ratio.

Stock-to-flow refers to the property by which the current global stock or holding of an asset compares to the potential incoming flow of that asset. For example, the global stock of GLD compared to the potential new production of GLD is very high - i.e. its a rare metal on Earth and the effort of extraction of new GLD only accounts for some 2% of the stock. This is compared to the potential stock-to-flow of fiat currency whereby government issued paper (no longer backed by anything like a gold standard) can be printed / minted readily (with little effort) thus large quantities of new flow can devalue currently held stock, leading to inflationary periods and the consequent socio-economic results.

The economic model of BTC is purported to resemble that of GLD as BTC miners are effectively expending a larger amount of resource and effort in minting new Bitcoins. In addition the total global supply of BTC is capped to 22 million; this should result in a deflationary currency.

This study will make heavy use of the methods of Hilpisch (2020), who writes as a specialist in the domain of data-driven finance. So as such it is both a learning exercise of personal interest in the crypto currency markets though also a vehicle by which the author can learn and test the methods of Hilpisch.

Problem Statements

1.Investigate the potential correlation between BTC and GLD returns in timeseries, assuming a positive correlation indicates some representation whereby BTC is also considered a store of value. The assumption here is that when investors opt for gold as a store of value they could also opt for BTC.

2.Machine learning will be applied to determine if the trades into BTC or GLD can be recommended over just holding a long position (it is suspected that the recent activities driving the price of BTC higher could result in an extremely simple, hold only strategy)

Metrics

The metrics for this project are relatively simple, given the strong connection to the financial domain and are either the price of the assets (GLD, BTC), the returns or relevant statistical metrics such as correlation.

Data

Data for crypto currency prices where obtained for free from https://www.cryptodatadownload.com and data is from the crypto currency exchange Binance. Data for gold prices was obtained from Yahoo Finance https://uk.finance.yahoo.com.

Environment

The analysis was carried out using Jupyter and Python 3.x. The analysis notebook and list of required libraries in use are noted in the accompanying GitHub repository at https://github.com/rsiwicki/rock_dex/.

Notebook: https://github.com/rsiwicki/rock_dex/blob/main/blockchain_labs_rockdex.ipynb.

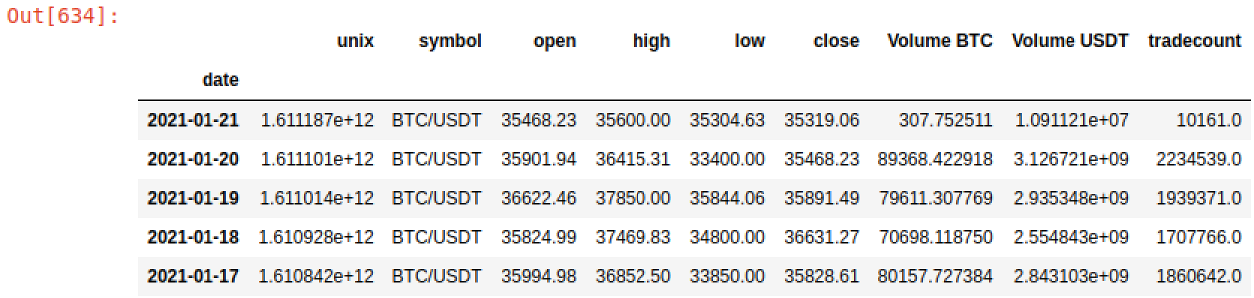
Data Exploration

The project does not require a full Extract Transform Load (ETL) pipeline, however there as a need to extract and clean the data after a period of discovery. The extraction and initial analysis of daily BTC (from the exchange Binance (2021) follows).

The raw data was in the form of a time-series of Open, Low, High, Close and additional information such as trade volumes and counts.

It appeared that there were null values for trade count. Upon inspection it seems that the missing values for trade count occur near the start of the data, assuming that Binance where not publishing these values during this time period. This was ameliorated later.

That data was visualised and analysed using Pandas upon import as a time-series. The first column (index_col=0), whilst representing the time-series is a Unix epoch time, we therefore reimported the data using Pandas indicating that the date column (index_col=1) was the true indicator of the time series. Pandas will automatically account for this and correctly build the time-series index (see Fig 1).



Out[634]:

date	unix	symbol	open	high	low	close	Volume BTC	Volume USDT	tradecount
2021-01-21	1.611187e+12	BTC/USDT	35468.23	35600.00	35304.63	35319.06	307.752511	1.091121e+07	10161.0
2021-01-20	1.611101e+12	BTC/USDT	35901.94	36415.31	33400.00	35468.23	89368.422918	3.126721e+09	2234539.0
2021-01-19	1.611014e+12	BTC/USDT	36622.46	37850.00	35844.06	35891.49	79611.307769	2.935348e+09	1939371.0
2021-01-18	1.610928e+12	BTC/USDT	35824.99	37469.83	34800.00	36631.27	70698.118750	2.554843e+09	1707766.0
2021-01-17	1.610842e+12	BTC/USDT	35994.98	36852.50	33850.00	35828.61	80157.727384	2.843103e+09	1860642.0

Figure 1: Example of BTC Raw Data

Visualisation of data often benefits from domain specific representations. In this case illustrating OHLC trade data using Candlestick charting (this originally was a Japanese method, used to help illustrate potential patterns in data that could result in new insight and trading strategies).

In this visualisation method the Open and Close prices are represented by the body of the *candle* (forming a box for each daily price quad). If the closing price is higher than the opening price the body of the candle is white, if the closing price is lower than the opening price the body colour is darker (in this case blue). The *wicks* of the candle represent the daily extremes of pricing. So the bottom line extending from the body is the lowest price experienced during the time period and the top line extending from the body is the highest price experienced for the period.

We will respresent the OHLC data for BTC daily here using the Cufflinks Python library (recommended by Hilpisch, 2020) (see Fig 2).

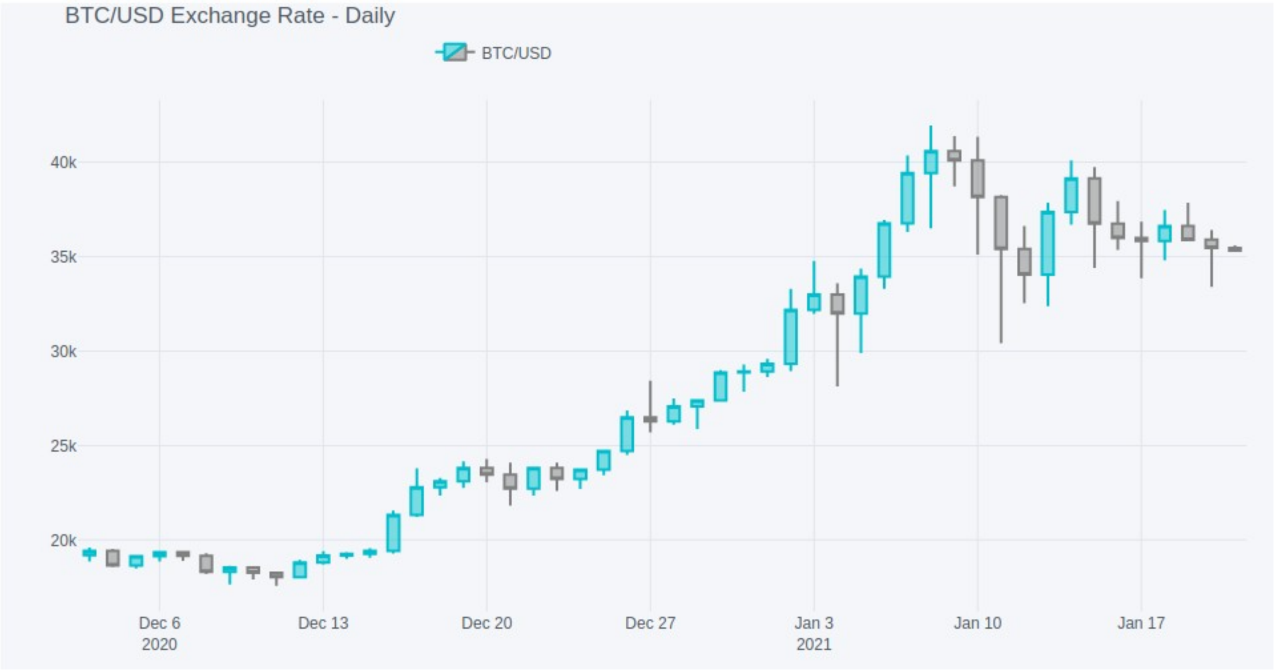


Figure 2: Candlestick Charting Representation of BTC Daily Data

In this next section we borrow some further techniques from Hilpisch (2020) to further illustrate relationships in the data to help convey information. These also explore our ability to use rolling window calculations in Python Pandas (see Fig 3).

1. plotting mean max and average of closing prices.
2. sub-plots of closing price, volumes and tradecounts.
3. applying a very simple fast-slow EMA strategy - to see if anything we produce using Machine Learning (ML) can out perform one of the most simple non-fundamentals analysis strategies.

```
Out[643]: <AxesSubplot:xlabel='date'>
```

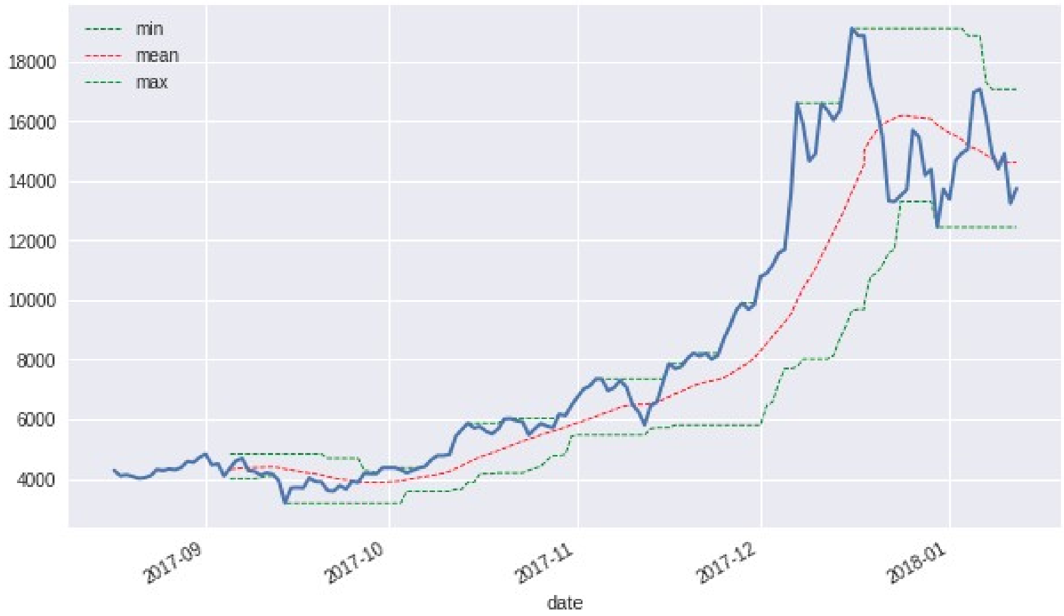


Figure 3: Rolling Window Metrics for BTC Daily

Subplots of the closing price, trade count and average trade size were also plotted. It was of interest to see if the average trades ize was increasing with price.

It seems that the number of trades increased with price as did the average trade size. Though earlier trade sizes were also larger (driven by an unknown factor).

It seems that increased number of trades could represent liquidity and this liquidity could also increase the value of the asset.

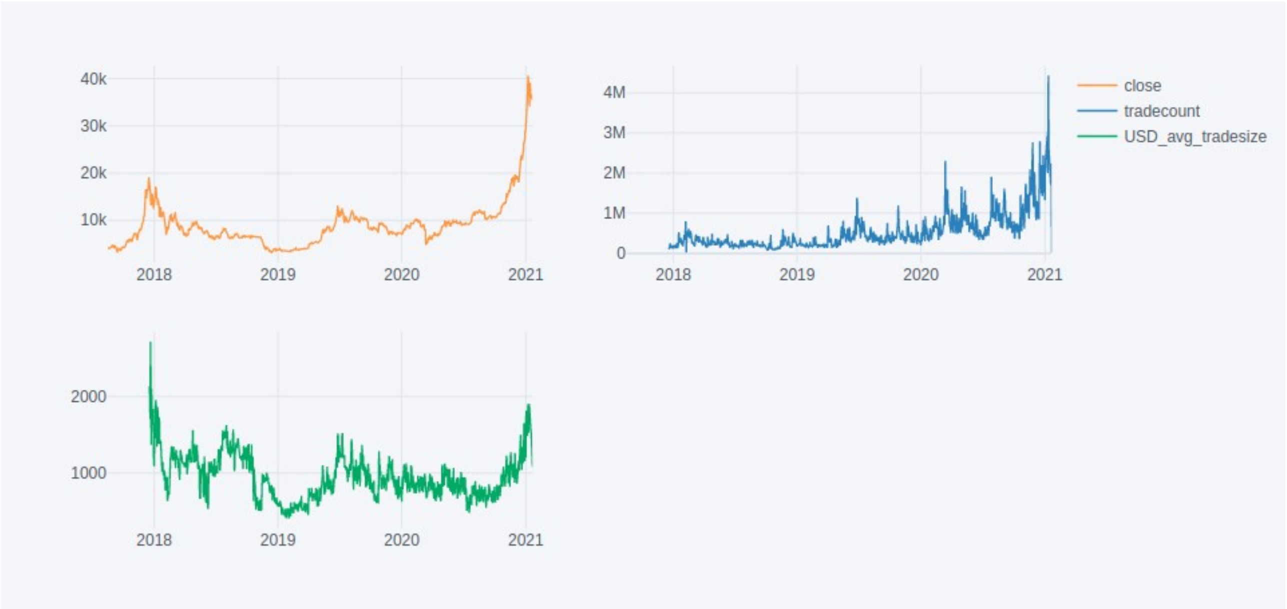


Figure 4: Price, Trade Volume and Average Trade Size Subplots

In the section that follows we utilised a simple method from Hilpisch (2020) that utilises a fast and slow Simple Moving Average (SMA) to indicate Buy Sell signals. In this case when the Fast SMA trend-line moves over the Slow SMA trend-line a long position in BTC is indicated. The general trend of BTC has been to move upwards within this time frame, though as a basic signal this indicator looks like it would have been positive if what simplistic.

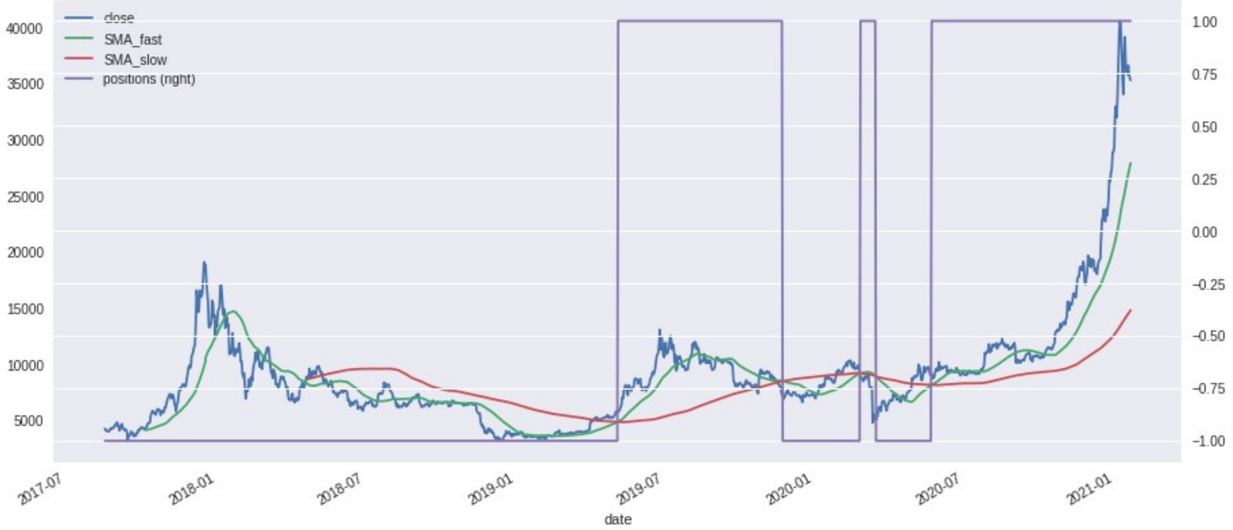


Figure 5: Fast and Slow SMA Trade Strategy for BTC

We then followed to load the gold prices from Yahoo Finance (2021) and explore. Some further pre-processing of this data was required, particularly null values for certain days of the week.

It seems likely that the missing dates, predominantly Sundays are holidays and the prices were not available to collect. Regardless of the reason to plot a chart, for example, would display breaks in continuity for these days. It seems reasonable to fill the missing data with that of the previous day to retain time series continuity. To achieve this we used the Pandas forward fill method of the fillna() function.

It was now possible to plot another simple SMA trading strategy chart for the GLD data to see how it differed from that of BTC.

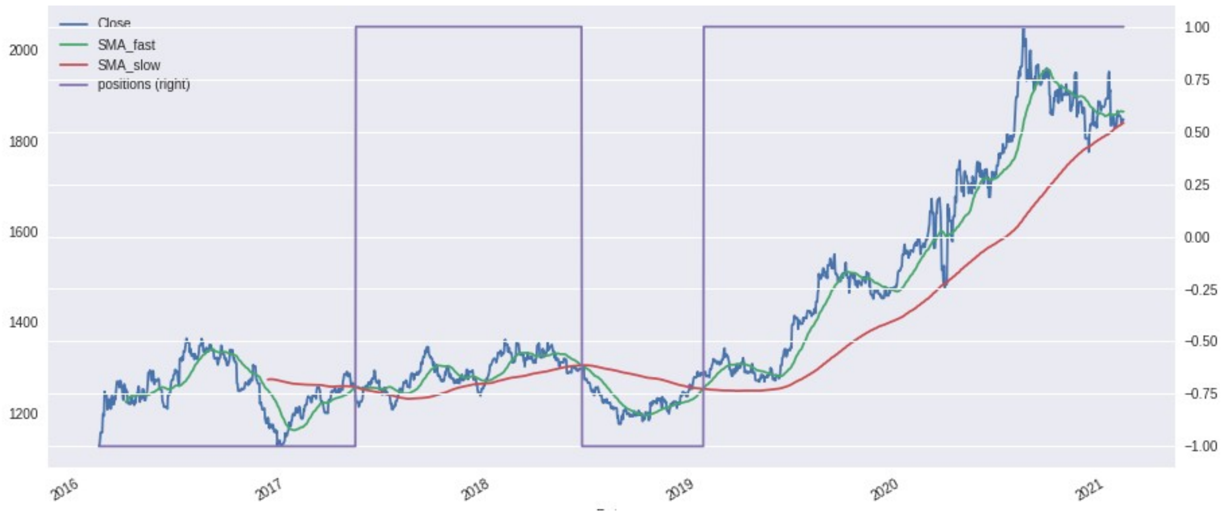


Figure 6: Fast and Slow SMA Trade Strategy for GLD

So already we can see a similarity between BTC and GLD. We then followed to plot the log of the return values for each asset (see Fig 7).

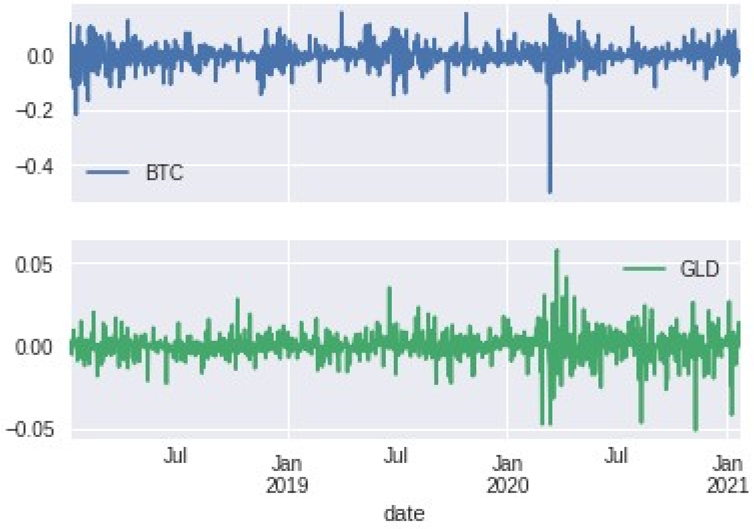


Figure 7: Log Returns by Time for BTC and GLD

We can also follow the lead of Hilpisch (2020) and utilise ordinary least-squares (OLS) regression to determine the extent of correlation.

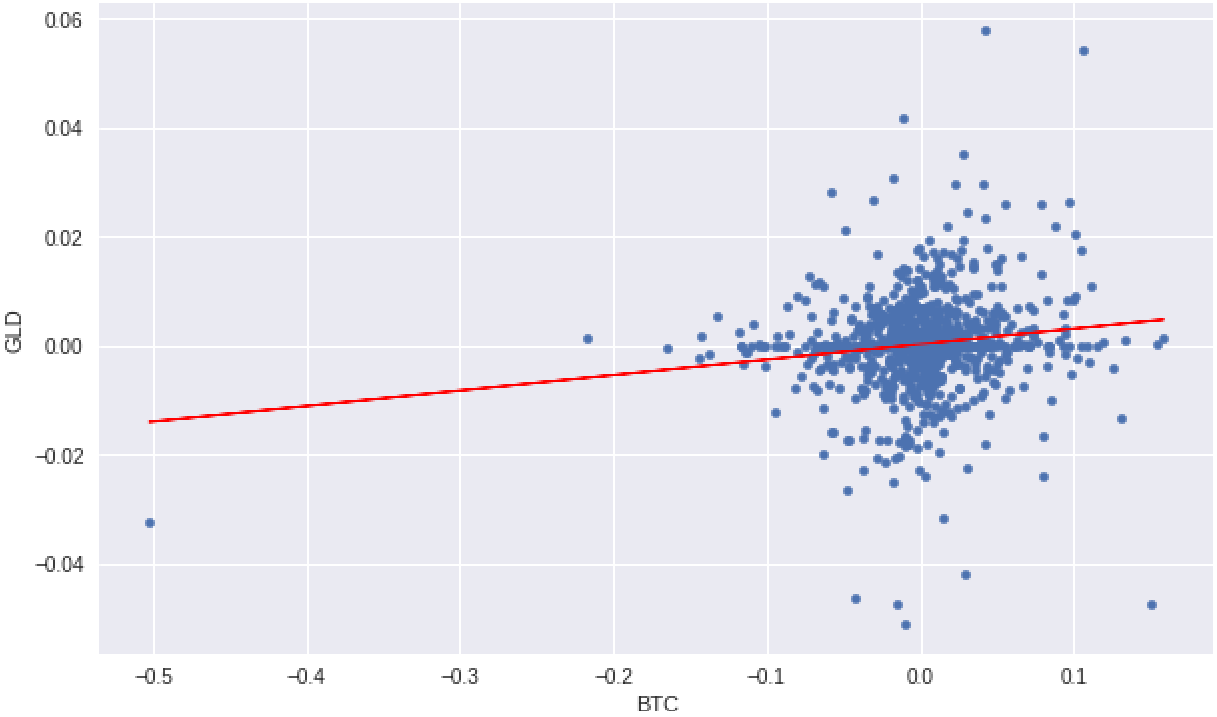


Figure 8: OLS Regression of GLD and BTC returns

The correlation matrix can be demonstrated, thus:

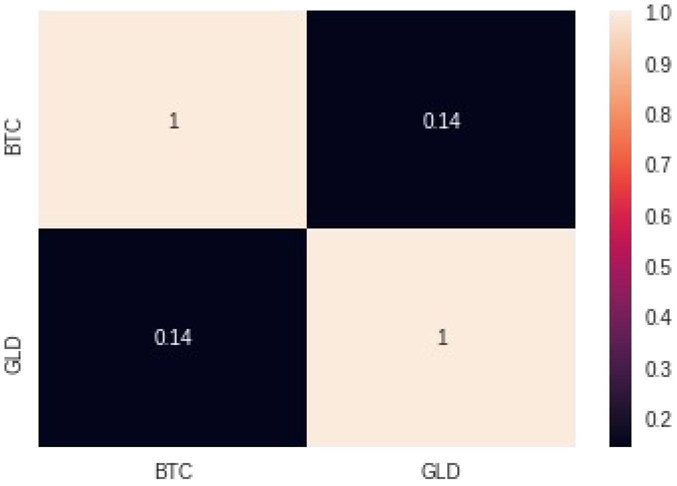


Figure 9: Correlation Matrix of BTC and GLD returns

We can see that the correlation seems negligible.

Because the data is time series the correlation can be plotted overtime to determine if there are periods where this correlation is stronger (See Fig 10).

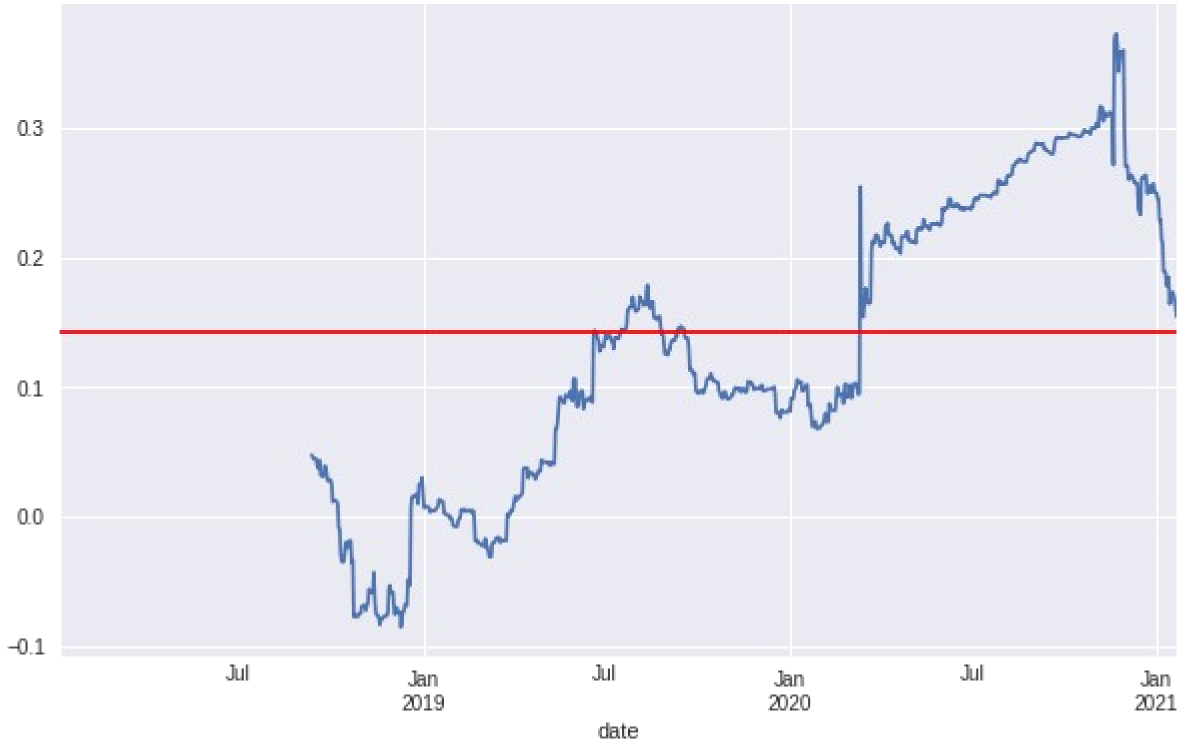


Figure 10: Correlation of BTC and GLD by Time

Correlation analysis over time series points to negative correlation prior Jan 2020, thereafter positive. This however could be attributed to some other hidden factors or other global economic situations (speculation though possibly the coronavirus pandemic). The hypothesis at the outset were true, that GLD and BTC should show some correlation if they are both considered stores of value. This does not appear to be the case prior to Jan 2020 though correlation seems to have increased after Jan 2020 though a continual trend upwards has been shown in the data set from the origin time. Regardless, the correlation is still weak at under 0.3, it would be interesting to follow the trend over time.

Does this increasing correlation really mean that BTC is becoming more like gold? Time could demonstrate that this correlation could potentially demonstrate the start of monetary competition with other stores-of-value.

Part 2. Trading Strategies and Machine Learning

In this section we return to the SMA example from earlier. We attempt to calculate the:

1. Return of the simple SMA strategy over just holding a long position in the BTC asset.
2. The possible enhancement of strategy that can be achieved using a simple machine learning model.

(We assume a trading period of 2018-01-03 to 2021-01-21 and cut the data this way to ensure a relevant cross over and complete inner join other than missing quoted days within the GLD price).

We calculated the log return of a hold only trading strategy for BTC and named this simple_returns. We then created a new measure, sma_returns to calculate the return determined by taking the short or long position indicated by utilising the SMA_slow and SMA_fast crossover (this is plotted as a dashed red line where -1.0 indicates short and 1.0 indicates long. The returns were also simulated as placing a trade at the close of day t0 and earning the returns of day t+1. The strategies are assuming no effect of trading costs, commissions or spreads (see Fig 11).

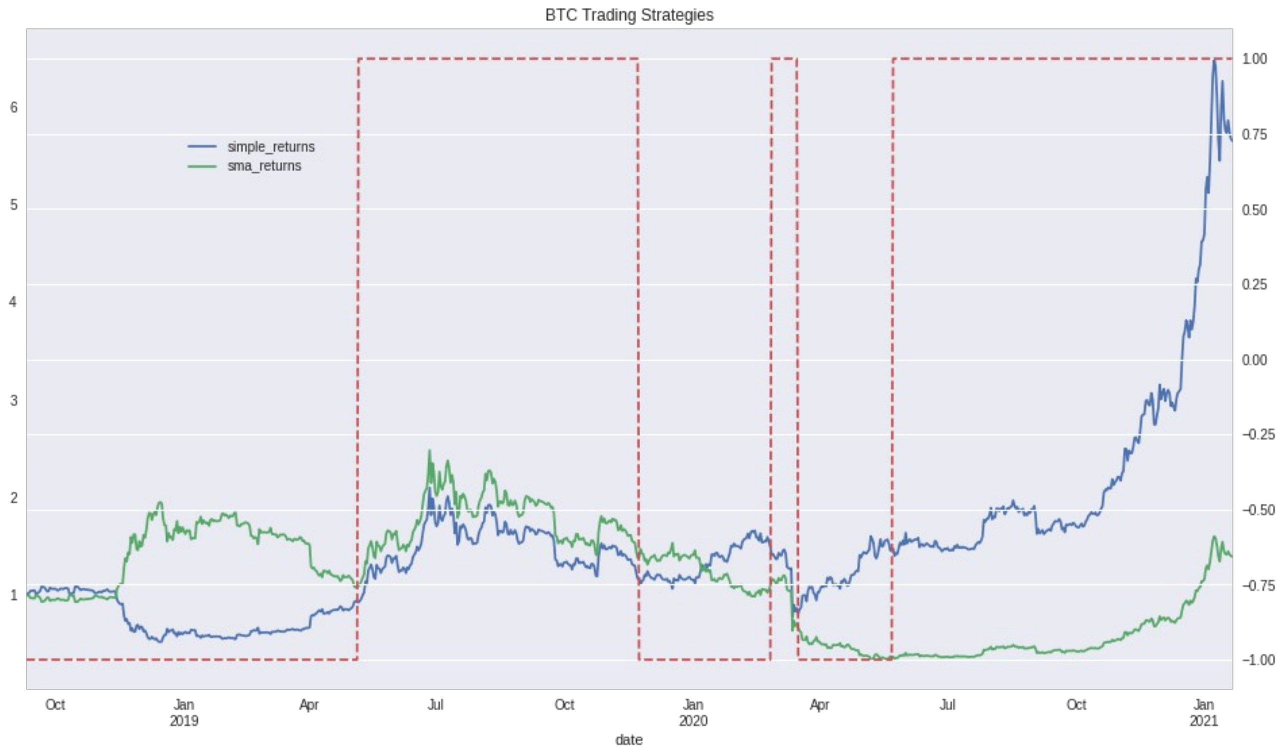
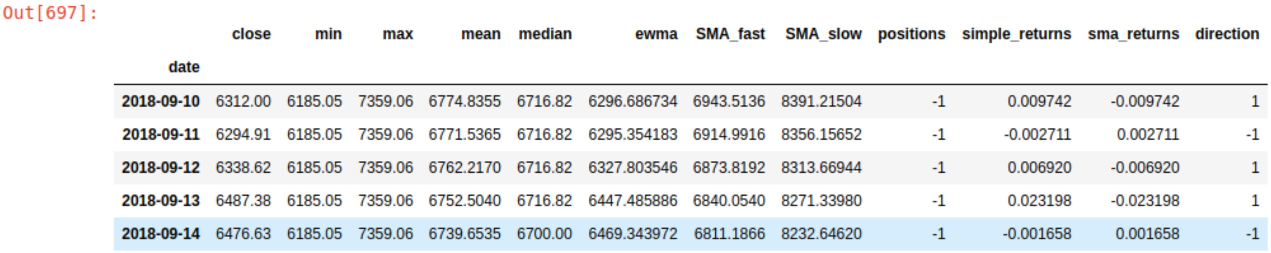


Figure 11: Fast Slow SMA Trading Strategy Returns for BTC

We can see that a simple SMA trading strategy would have been inferior to an even simpler holding strategy during the time period with the opportunity cost of 5.650317 - 1.389344 times the return.

In this section we prepared the model fit and evaluation teachnique as per Hilpisch (2020).

Data preparation to use the above strategies required that the returns in the data are classified as the log returns as per the previous calculations for the SMA trading strategy. The direction of the return also needs to be classified as to whether it is positive or negative. To do this we began to adjust the existing data from the SMA strategy above (see Fig 12).



Out[697]:

date	close	min	max	mean	median	ewma	SMA_fast	SMA_slow	positions	simple_returns	sma_returns	direction
2018-09-10	6312.00	6185.05	7359.06	6774.8355	6716.82	6296.686734	6943.5136	8391.21504	-1	0.009742	-0.009742	1
2018-09-11	6294.91	6185.05	7359.06	6771.5365	6716.82	6295.354183	6914.9916	8356.15652	-1	-0.002711	0.002711	-1
2018-09-12	6338.62	6185.05	7359.06	6762.2170	6716.82	6327.803546	6873.8192	8313.66944	-1	0.006920	-0.006920	1
2018-09-13	6487.38	6185.05	7359.06	6752.5040	6716.82	6447.485886	6840.0540	8271.33980	-1	0.023198	-0.023198	1
2018-09-14	6476.63	6185.05	7359.06	6739.6535	6700.00	6469.343972	6811.1866	8232.64620	-1	-0.001658	0.001658	-1

Figure 12: Example Data Including Returns Direction

Now the directions were classified according to the sign of the return.

We then engineered features for our algorithms input by using a lag of the returns. We choose six lags as the number of features; this is equivalent to a trader using six consecutive historical data points to predict the next movement direction of the currency. The output is illustrated below (see Fig 13)

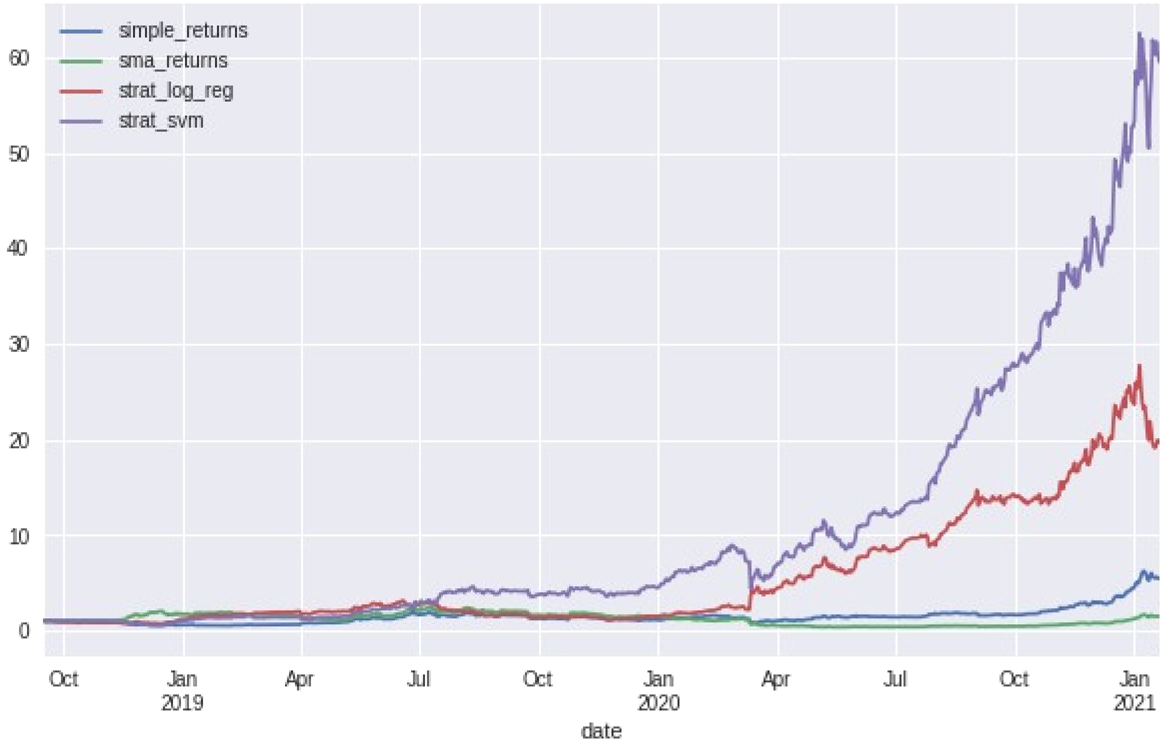


Figure 13: Returns of Strategies for Entirety of Data for BTC Daily

The SVM is clearly outperforming other methods, quite substantially. Though this method is somewhat artificial as the entire data set it utilised. To test more realistic scenarios a test and train split of the data can be used. This will simulate training the data on historical returns and then using that to predict possible future returns on unseen data.

Splitting the data for this case is extremely simple compared and does not require the test_train_split function as it is merely time linear data. We chose a simple strategy outlined below to partition test and train data whereby train data was prior in the time-series to the split point and test data was future data by reference to the split point.

```
split_point = int(len(df_BTC_close_daily) * 0.3)
```

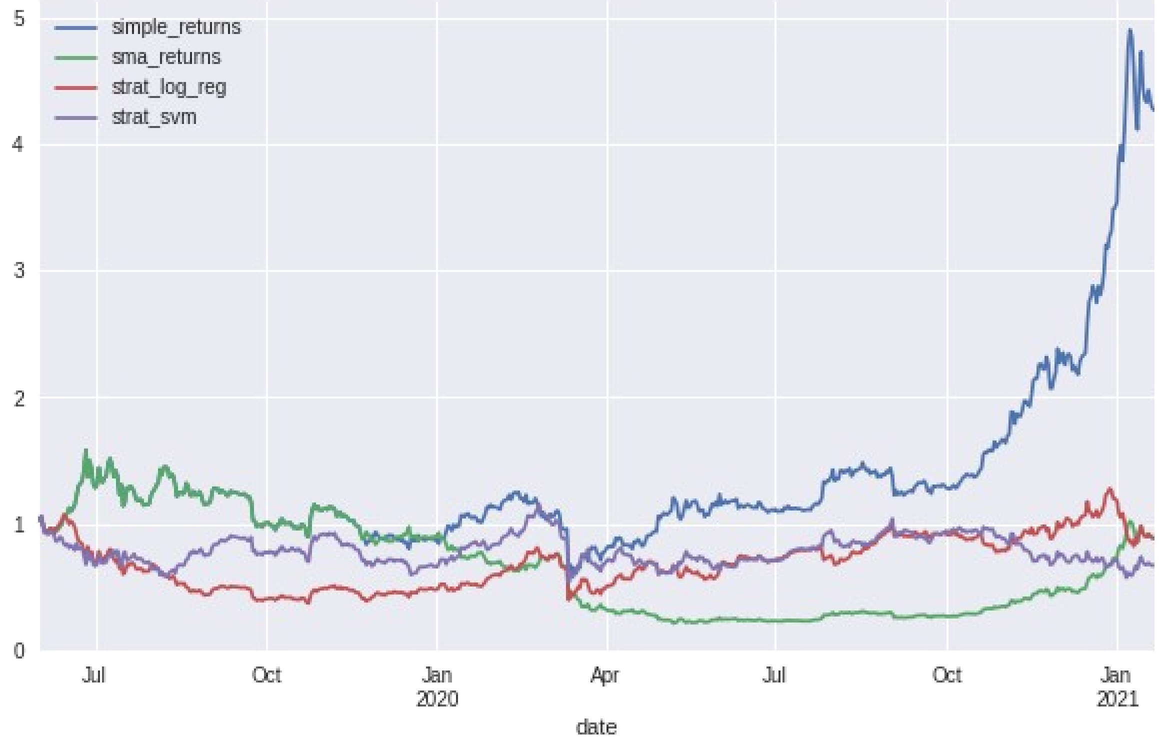


Figure 14: Returns Performance of Machine Learning Strategies for BTC Daily

In this case were the algorithms are confronted with previously unseen data, no algorithmic machine learning strategy works better than a simple hold strategy.

Conclusion

Part 1 of this study, investigating how BTC could potentially relate to GLD as a store of value demonstrates that there is a possibility that there is an increasing correlation over time between BTC price and GLD price movements. To further improve this study, time is required to further assess the relationship in our existing simplified model; though, the addition and study of additional financial metrics that impact the GLD price, though applied to BTC could be considered - e.g. inflation metrics.

Part 2 of this study deeply assessed the methods Hilpisch (2020) demonstrates in order to start identifying a potential machine learning based strategy for increasing hypothetical future returns of trading BTC, through only technical analysis and not fundamental analysis. The simple hold strategy was pitted against a second reference strategy of Fast and Slow Simple Moving Averages (SMA). Further to this Support Vector Machine and Logistic Regression strategies were added with simplified engineered features of return lags and whether a position should be short or long. The machine learning strategies all appeared to be weak once tested with a train / test split of the data and compared to a simple hold strategy.

Though our findings that simply holding BTC generates effective returns and in reality with costs accrued per trade is certainly more efficient is perhaps itself a very fitting answer as to how best to trade BTC in the current environment, essentially buy and hold.

It is likely that the unusual nature of BTCs recent ascent could be confounding the training and testing of our algorithms: i.e. that the real driver for BTC prices is something more fundamental than technical analysis; such as the store-of-value effect studied in part 1.
Future improvement could include attempting the same strategies on hourly data instead of daily, perhaps here there are technical analysis relationships that are more subtle and intra day trading could benefit from a machine learning approach.

References

Ammous, S (2018). *The Bitcoin Standard: The Decentralized Alternative to Central Banking*. Wiley.

Binance (2021). *Crypto Currency Data: Bitcoin USD*. From https://www.cryptodatadownload.com.

Hilpisch, Y (2020). *Python for Finance: Mastering Data Driven Finance*. O'Reilly.

Yahoo (2021). *Gold Prices Daily*. From https://uk.finance.yahoo.com.