2nd Capstone project proposal

Title: **"Cryptocurrency stock market prices prediction"**

1. The Problem:

Cryptocurrencies are forms of digital currency where encryption protocols are used to secure different transactions, and to secure the creation of new units and transfer of funds. One of the most known cryptocurrency assets, and also the first to be created, is Bitcoin. What is remarkable about Bitcoin is the explosive increase in its valued price in the last year. Given this, there might be a growing interest in developing a model able to predict future prices for Bitcoin and other cryptocurrencies.

2. Who is the client?

FBS is an online forex broker company that offers Bitcoin trading to its customers. In order to do so, FBS requires a predictive model able to predict future Bitcoin stock prices, which are used to make appropriate decision regarding trading, thus maximizing the profit for its customers

3. Where can the data be obtained?

The data set that I plan to use, from the "Coin metrics" website (https://coinmetrics.io/data-downloads/). There, one can find daily cryptocurrency data and download it in CSV format.

4. What will the approach be?

Once the data has been acquired, I plan to perform some exploratory data analysis. In particular, I think it would be useful to explore the existence of temporal autocorrelation in the prices fluctuation, as well as periodicity, which could be studied with Fourier transform analysis. It would also be interesting to carry out, for example, Twitter sentiment analysis to correlate people's mood with variability of crypto currencies prices. In order to perform prices prediction, I plan to use methods such as LSTM neural networks or ARCH, which are commonly used in stock market predictions.

5. What are the deliverables?

- A final project report
- Power point slides
- Code

5. Data Wrangling Steps:
(*: the data wrangling mentioned here, are in the first part on the EDA.ipynb, included in this same folder)**

The data can be downloaded from Coinmetrics' website with a series of straightforward steps shown below:

```
#Link to dataset
html_bitcoin = "https://coinmarketcap.com/currencies/bitcoin/historical-data/?start=20130428&end=20180216"
#Load the data into a dataframe
bitcoin = pd.read_html(html_bitcoin + time.strftime("%Y%m%d"))[0]
#Formating the Date column into date-time
bitcoin = bitcoin.assign(Date=pd.to_datetime(bitcoin['Date']))
#Replace volume values of "-", and set the column to int64 type
bitcoin.loc[bitcoin['Volume']=="-",'Volume']=0
bitcoin['Volume'] = bitcoin['Volume'].astype('int64')
#set the index to be the Date column
bitcoin.set_index('Date', inplace = True)
#... and reverse the order of the dataframe
bitcoin = bitcoin.iloc[::-1]
```

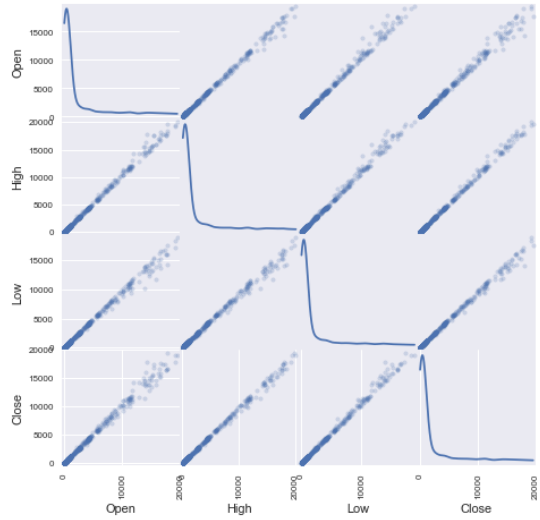
After these steps, the data is ready to use, thus no further data wrangling steps are needed. The data set consists of Open, High, Low and Close prices, as well as trading volume and market cap values over time. Below, is an example how the data set looks like:

	Open	High	Low	Close	Volume	Market Cap
Date						
2013-04-28	135.30	135.98	132.10	134.21	0	1500520000
2013-04-29	134.44	147.49	134.00	144.54	0	1491160000
2013-04-30	144.00	146.93	134.05	139.00	0	1597780000
2013-05-01	139.00	139.89	107.72	116.99	0	1542820000
2013-05-02	116.38	125.60	92.28	105.21	0	1292190000

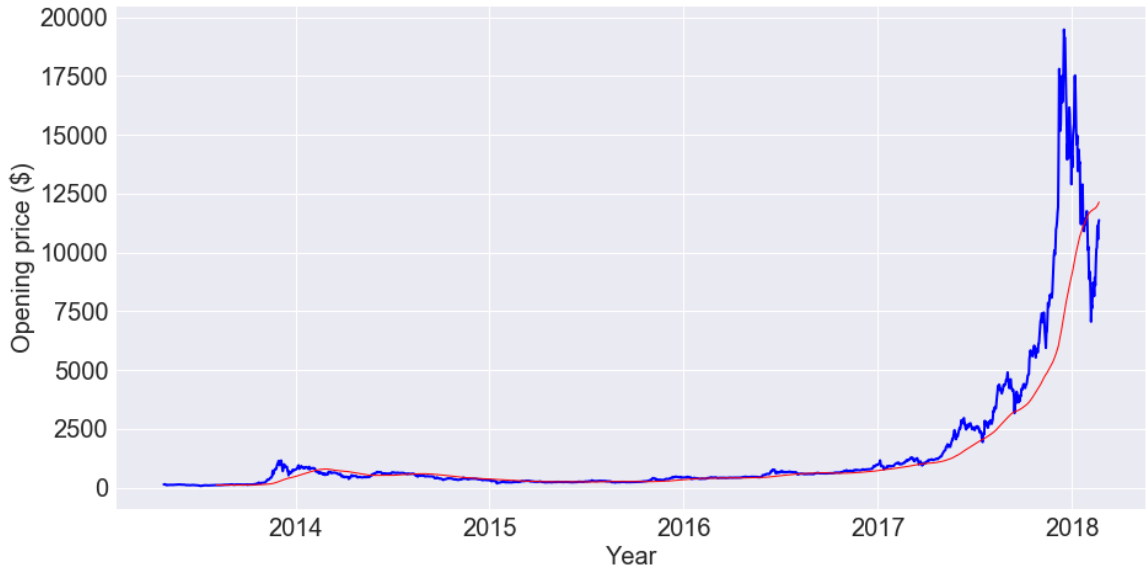
6. Exploratory data analysis:

The purpose of this exploratory data analysis is to identify possible trends in the historical bitcoin market data that could be of use in predicting future market closing prices. As a starting point, a good way to plot this type of data is by using candlestickplots. In the example below (last 5 months of bitcoin's data) each rectangle represents one day. The Bottom border represents opening price. The Upper border represents closing price The upper and bottom sticks represent the highest and lowest price during that particular day. Green color means that the closing price was higher than the open price, while the black color means the closing price was lower than the open price:

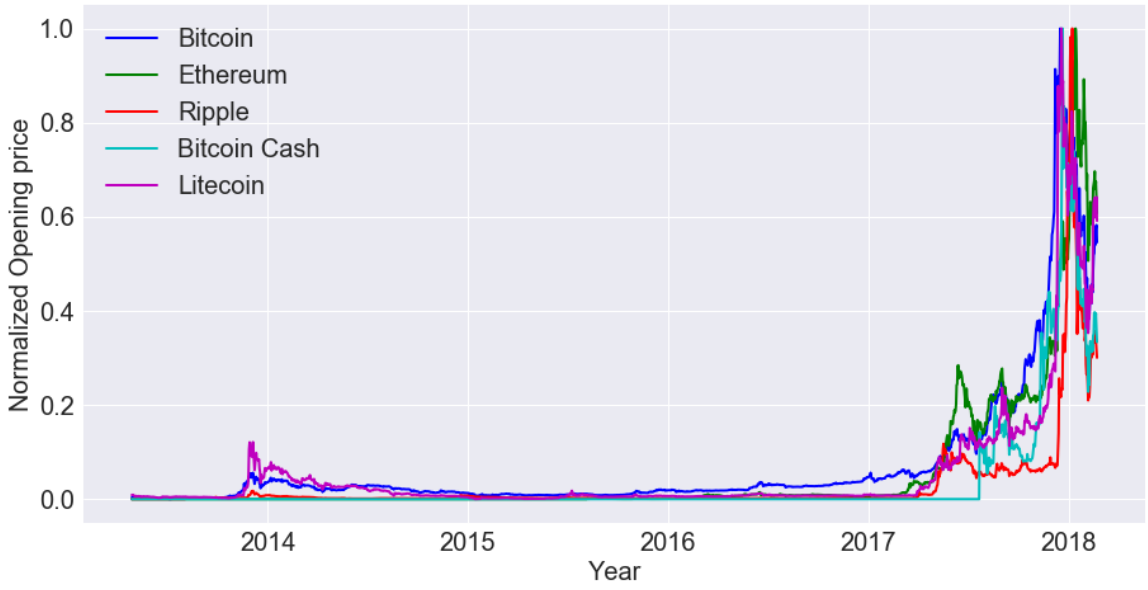
One could also look at the correlation between the Open, hi, low and closing prices, using a scatter plot matrix:



As expected, the Open, High, Low and Close prices are highly correlated to one another. Thus, for predicting future trends in bitcoin's stock market we could simply use the Open price (plotted below). The redline represents a smoothed version of the data, calculated as a moving average with a window of 100 data points.



We could also ask if bitcoin's opening prices for are correlated with the opening prices of other cryptocurrencies. Fort that, we can plot the normalized Opening prices for different cryptocurrencies.



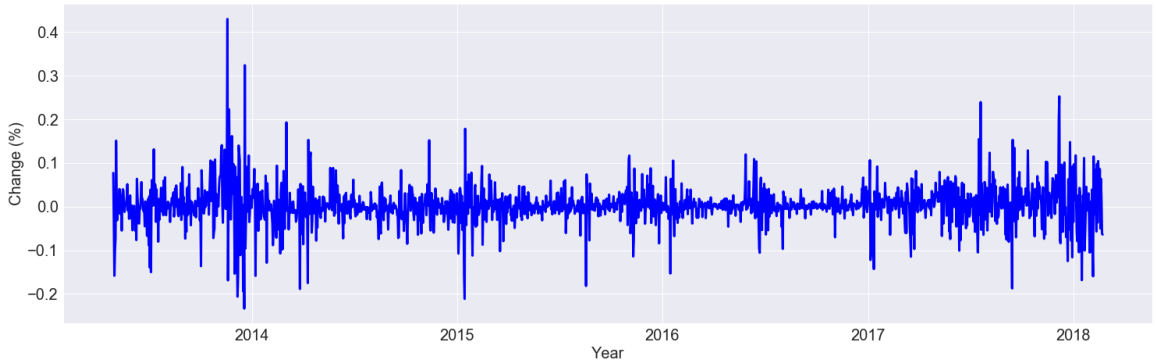
We can see that all of the cryptocurrencies listed show a similar pattern of growth, which is most evident from 2017 onward.

To get a better idea about the correlation of these patterns, we can plot a scatter matrix of the normalized open prices. This matrix shows the correlation coefficient, calculated with the entire open price time series.

	bitcoin	ethereum	ripple	bitcoin_cash	litecoin
bitcoin	1.000000	0.917312	0.817799	0.945870	0.955455
ethereum	0.917312	1.000000	0.885665	0.893533	0.923178
ripple	0.817799	0.885665	1.000000	0.854794	0.865852
bitcoin_cash	0.945870	0.893533	0.854794	1.000000	0.941806
litecoin	0.955455	0.923178	0.865852	0.941806	1.000000

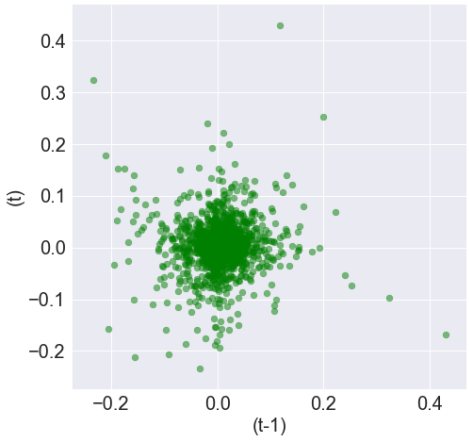
Looking at the correlation matrix from above, we can confirm the high degree of correlation among the opening prices of the cryptocurrenices listed here.

Another way to look at the Open price data is to express it as the percentage of change over time. The plot below shows the percentage of change in open prices of Bitcoin.

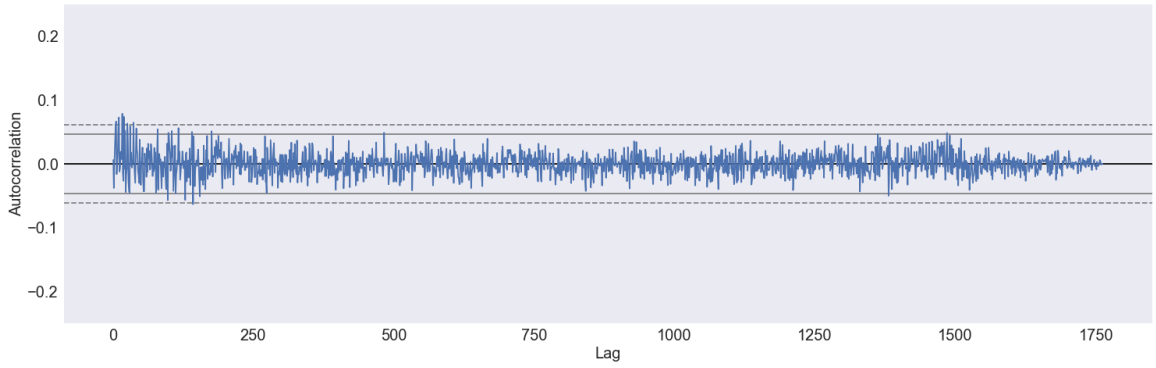


Although the percentage of change varies around 0, there are multiple "peaks", which may suggest that this time series is autocorrlated.

To test this, we can first look at the correlation (see the plot below) between the data at times (t-1) and (t).

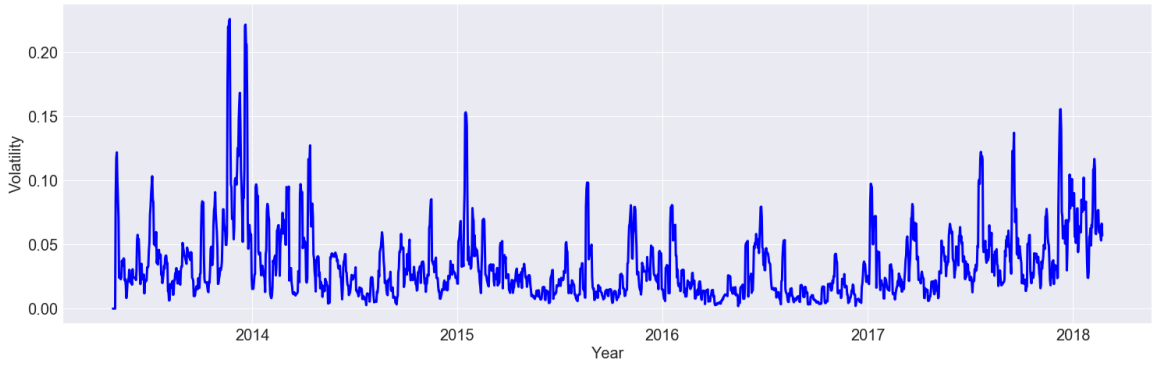


The previous plot shows the lack of correlation ($r \sim 0.006$) for a lag of 1. To get a more general view of autocorrelation of this time series, we can use pandas autocorrelation_plot function to calculate the autocorrelation for other lag values.

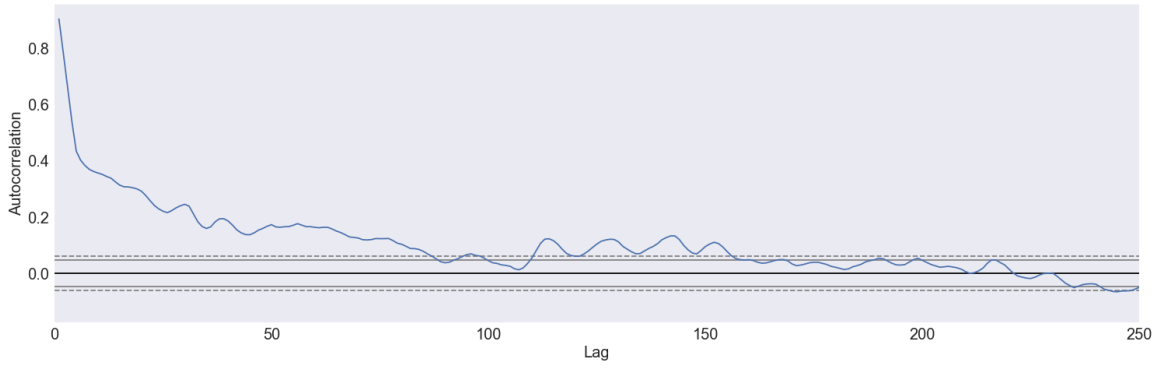


In the previous graph, the dashed line corresponds to the 99% confidence interval. Any values above that line, would point to those lag values for which the signal (percentage change over time) is autocorrelated. This plot shows that there's no autocorrelation in the time series of percent changes of bitcoin's open prices.

A third way to look at this type of data would be to calculate the Volatility, which is the standar deviation of the percentage of change in open prices over time. We can calculate it with pandas using a rolling window of size 5.



At first glance, there seems to be a certain periodicity in the volatility, which suggests that this signal is autocorrelated. Below, is the autocorrelation plot for this signal:



As before, the dashed line corresponds to the 99% confidence interval. Any values above that line, would point to those lag values for which the volatility is significantly autocorrelated. The plot shows that for lag values lower than 80, there is a significant correlation. Thus, this shows that there are periodic changes in the volatility of bitcoin's open prices.

7. Conclusion:

After performing this EDA, it was determined that there are no patterns/periodicity in the percentage of change of bitcoin open prices, that could be used to predict future variations of bitcoin's stock market prices. On the other hand, the Volatility was found to show autocorrelation, which shows the presence of periodic changes. Predicting volatility might be useful to determine times when a sudden big change in bitcoin's prices is expected. In future analysis one could also look at other types of data, such as people's mood measured by sentiment analysis, to see how they could affect Bitcoin's stock prices.