Foundation of Data Science

Cryptocurrency Analysis

Project Report

Authors: Kunal Jayesh Modi (kjm597)
FNU Karan (kx361)

Background and Overview:

We now live in the age of digital currencies, with cryptocurrencies birthed within the last decade. Already, there are over a thousand cryptocurrencies in the market. As we take on this new, proliferous market, it is important that we try to understand what's going on. There are too many risks at both the micro-level (e.g., personal investments) and the macro-level (e.g., prevention of market crashes and major loss of capital). This is where proper analysis of various cryptocurrency data comes handy.

The below blog was a driving point in developing interest in this topic. It seamlessly explains some trends in the cryptocurrencies and also gives some datasets to work on.

https://blog.timescale.com/analyzing-ethereum-bitcoin-and-1200-cryptocurrencies-using-postgresql-3958b3662e51

The first thing that took our attention was the following statement "***Turns out that if you had invested $100 in Bitcoin in July 2010, it would be worth over $5,000,000 today***." This amount is quite staggering and raised some quick questions in our mind.

- Why is there a sudden surge in the interest in recent days? Is it due to the increase in the price in the last few days? etc.
- How many such cryptocurrencies are there and what are their prices and valuations?

Some good historical data could help a lot to answer these questions. We have collected the data from the following sources.

- https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory

We also looked at these links to understand more about the currencies and their analyses:

- https://www.linkedin.com/pulse/blockchain-absolute-beginners-mohit-mamoria/
- https://medium.com/@meganrisdal/cryptocurrency-datasets-on-kaggle-3852259265c1

Target and predictor variables:

- Target Variable: The target variable for our model is the '**future closing price'** of the cryptocurrencies. We have implemented the model and predicted the prices for Bitcoin, Ethereum and Litecoin.
- Predictor Variables: The predictor variable is '**close price**' over the time series.

Problem Statement:

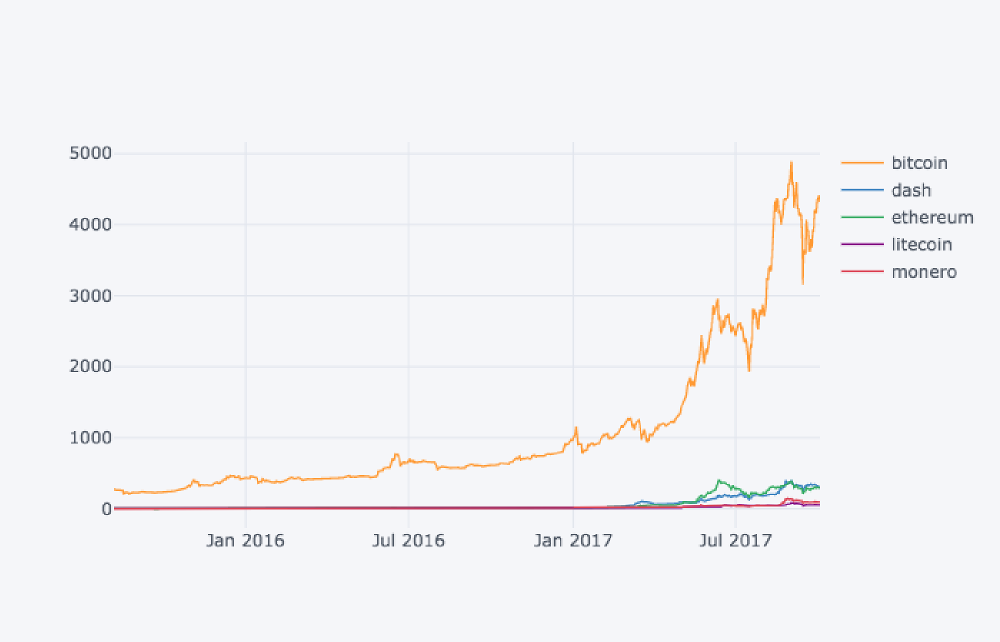
The goal of this project is to first analyze the trend of the cryptocurrencies over time and see whether the price fluctuations of currencies correlate with each other or not. After the exploratory analysis over the data, we would want to predict the future prices of the cryptocurrencies which would help us decide our investment decisions and also compare the predicted close prices with the actual close prices.

Type of Model:

A time series is a collection of data points that are collected at constant time intervals. The data we obtained is also a time series data. What differentiates a time series from regular regression problem data is that the observations are time dependent and, along with an increasing or decreasing trend, many time series exhibit seasonality or trends. We chose the ARIMA model to forecast the time series since the data shows Auto Regression as well as Moving Average trends.
Also, it captures a suite of different standard temporal structures in the time series data.

Approach:

1. **Data Cleaning:** The data did not contain any missing values, or even outliers.
2. **Exploratory Analysis:** We plot the data to observe the data visually. Also, we look for any kind of trends or seasonality that the data might show.



3. **Checking the Stationarity of the Time Series:** A common assumption that is made to perform time series modelling is the stationarity of the data. A *stationary* time series is one whose statistical properties such as mean, variance, autocorrelation, etc. are all constant over time. To whether the data is stationary or not we perform the Unit Root Test also known as the **Dickey-Fuller Test**.

   We first perform the logarithm on the close prices, since there were some values which were

   In our case the data was not stationary, so we made it stationary using differencing, which is the most common technique to make time series stationary.

4. **Build the ARIMA Model:** We can obtain the parameter ‘p’ (number of autoregressive terms), ‘d’ (number of non-seasonal differences needed for stationarity), ‘q’ (number of moving average terms) by observing the ACF and PACF plots as shown in Figure below.

   We divided the data into training and test data. We can then build the ARIMA (Auto Regressive Integrated Moving Average) model using the chosen parameters on the training data.

In the figure below, the 'p' and 'q' values can be determined as follows:

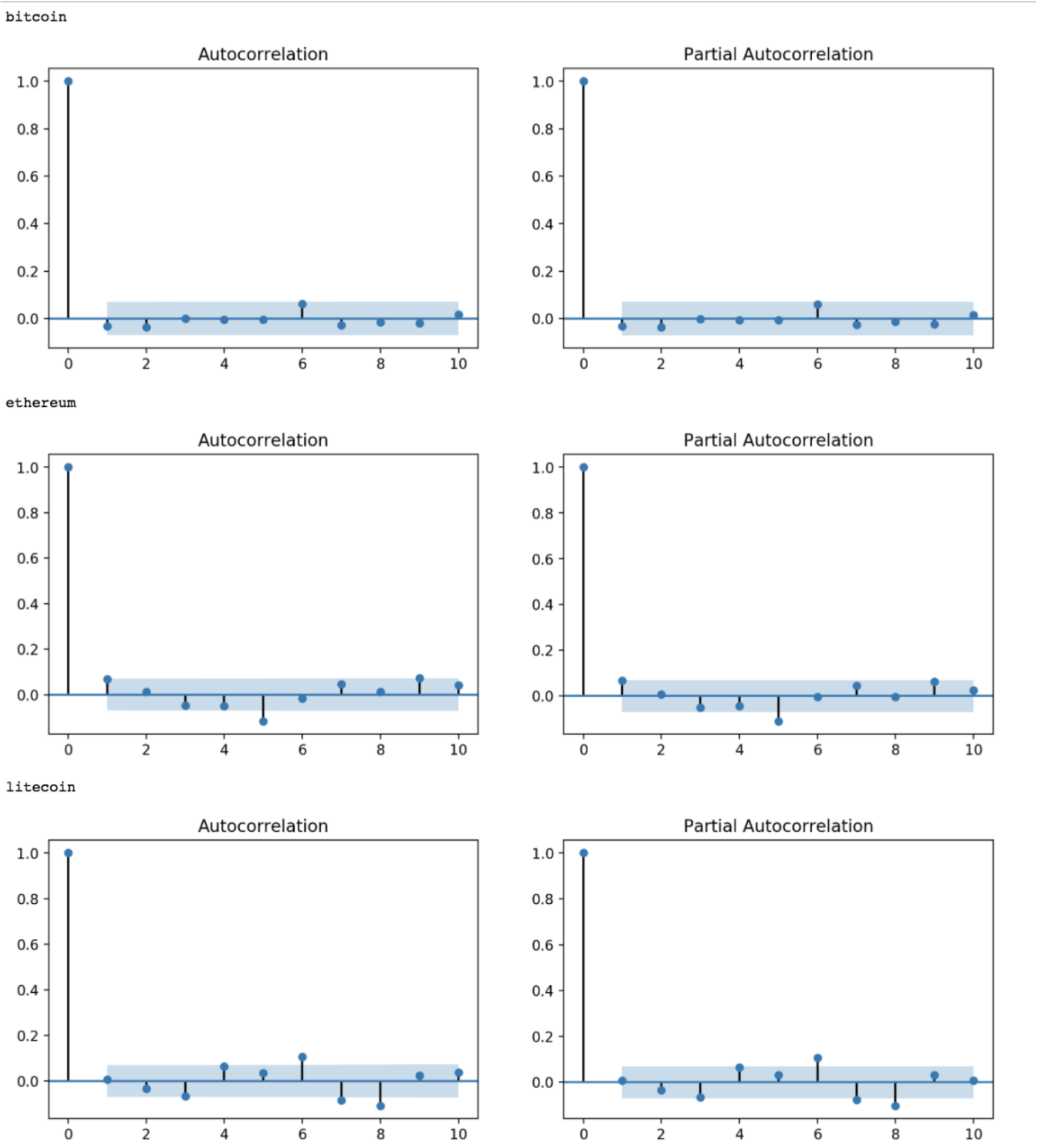
p: The lag value where the PACF cuts off (drops to 0) for the first time. If you look closely, p=1.

q: The lag value where the ACF chart crosses the upper confidence interval for the first time. If you look closely, q=1

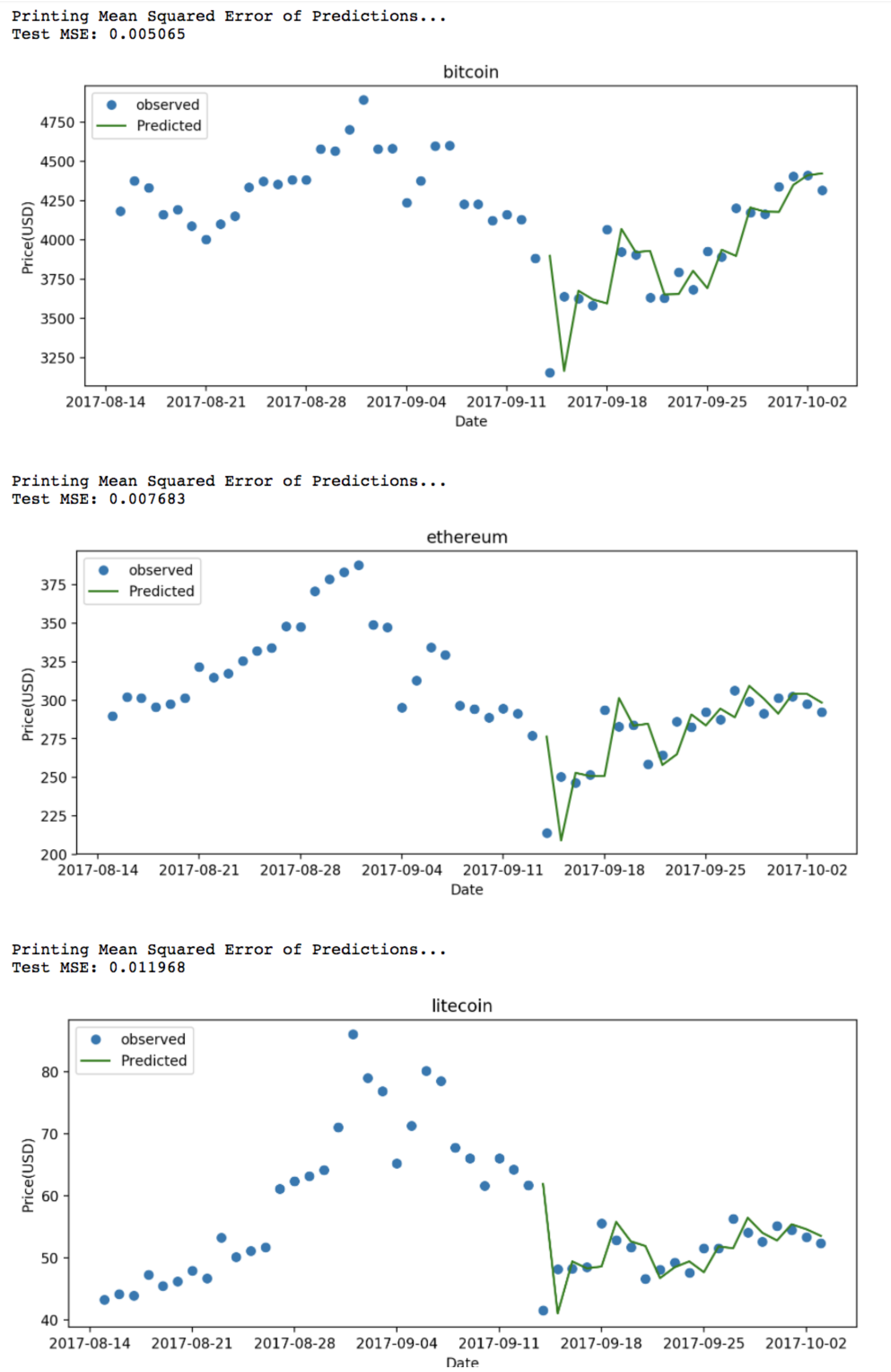
d: Since we used first order differencing to make our data stationary, d = 1

This means that the optimal values for the ARIMA (p, d, q) model are (1,1,1).

It is possible for an AR and an MA term to cancel each other's effects in a mixed ARIMA model. So, let's try a model with one fewer MA term, particularly because of convergence warning as seen above. Thus, we used the parameters (1, 1, 0).



5. **Making Predictions:** Using the model formed in the above step, we can make predictions on the close prices of the test dataset and see how the model performs. Finally, we created the model and following was the result on the test data.



```
Printing Mean Squared Error of Predictions...
Test MSE: 0.005065
```

```
Printing Mean Squared Error of Predictions...
Test MSE: 0.007683
```

```
Printing Mean Squared Error of Predictions...
Test MSE: 0.011968
```

Evaluation:

The evaluation metric we used is the **Mean Square Error**. As we can see from the above figure the MSEs for the predicted values of close prices are pretty close to zero, indicating a good fit for the model.

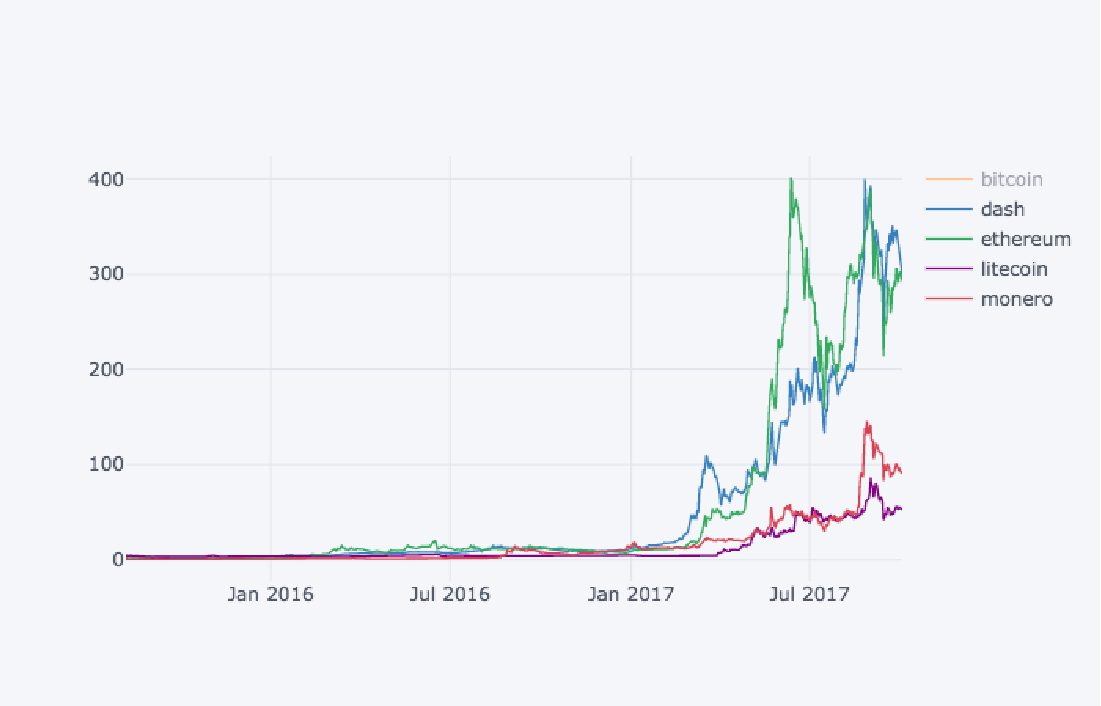
Cryptocurrency	MSE
Bitcoin	0.005065
Ethereum	0.007683
Litecoin	0.011968

Assumptions and Limitations:

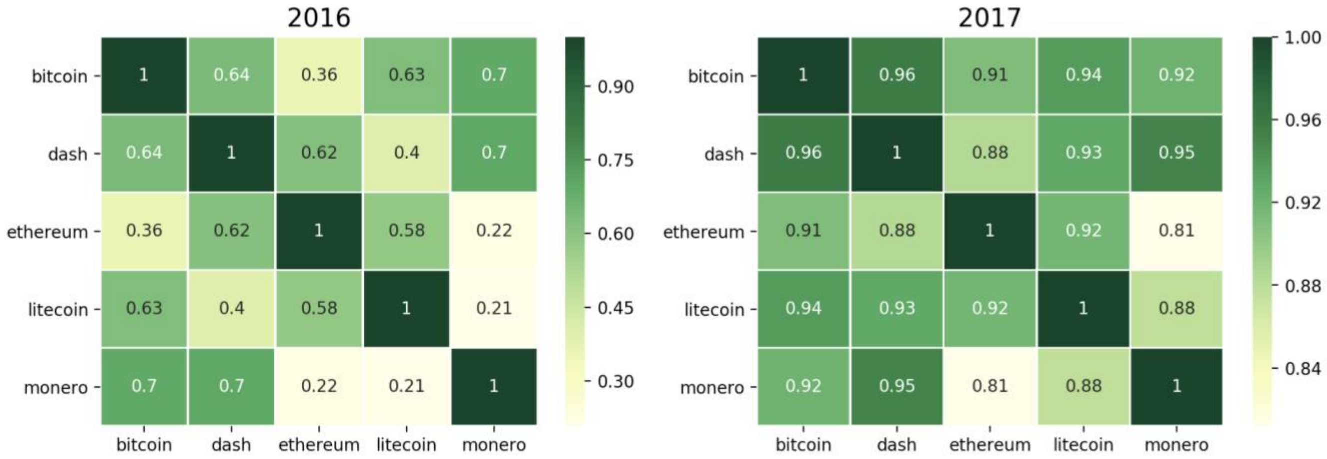
We have made an assumption that the future prices are dependent on historical price trends of the currencies and not on other factors. One limitation of our implementation is the prices of the currencies is dependent on many other factors other than previous prices. Also, there is a lot of volatility in the prices and some sudden changes in prices like that of the last one week may affect the model adversely.

Inferences:

- Bitcoin is the most prominent bitcoin, and that too it became in 2017, very recently. Due to the growth of Bitcoin, we also observed the growth in other cryptocurrencies as we can see in the figure below.



- One major observation was there was a drastic difference in correlation among the cryptocurrencies for 2016 and 2017.



The most immediate explanation to the above figure that comes to mind is that hedge funds have recently begun publicly trading in crypto-currency markets. These funds have vastly more capital to play with than the average trader, so if a fund is hedging their bets across multiple cryptocurrencies, and using similar trading strategies for each based on independent variables (say, the stock market), it could make sense that this trend of increasing correlations would emerge.

What did you change from your original proposal and why?

- One change we did from the original proposal is that in the proposal we had given an assumption that the data is stationary, which was not correct. We observed this during exploratory part of the project.