International Institute of
Information Technology, Bangalore

Data Understanding And Preparation
Document
DS 707 Data Analytics

Blockchain understanding and Cryptocurrency Analysis

Akanksha Dwivedi - MT2016006
Hitesha Mukherjee - MS2016007
Nayna Jain - MS2017003
Tarini Chandrashekhar - MT2016144

Instructors :
Prof. Ramanathan Chandrashekhar
Prof. Uttam Kumar

October 29, 2017

Contents

1 Collect Initial Data

1.1 Initial Data Collection Report

The source of data is kaagle.com which has the recently and regularlyupdated cryptocurrency data.The data collected consists of the crptocurrencies like Etherium and Bitcoin. We have the data in the csv format.

2 Describe Data

2.1 Data Description Report

Bitcoin dataset has around 5 years historical data with regular updates. There are many parameters which affect the price of a bitcoin.This dataset has features related to Ethereum. This is very similar to the bitcoin dataset and is available on a daily basis. Data is taken from Etherscan.

Dataset has one csv file for each currency. Price history is available on a daily basis from April 28, 2013.All the data except Date is of numeric and continuous type. The columns in the csv file are

- Date : Date of observation
- Open : Opening price on the given day
- High : Highest price on the given day
- Low : Lowest price on the given day
- Close : Closing price on the given day
- Volume: Volume of transactions on the given day
- Market Cap: Market capitalization in USD

Related to other attributes specific to particular cryptocurrency Eg. `bitcoin_dataset`.
Bitcoin Dataset (`bitcoin_dataset.csv`) :
This dataset has the following features.

- Date : Date of observation
- `btc_market_price` : Average USD market price across major bitcoin exchanges.
- `btc_total_bitcoins` : The total number of bitcoins that have already been mined.
- `btc_market_cap` : The total USD value of bitcoin supply in circulation.
- `btc_trade_volume` : The total USD value of trading volume on major bitcoin exchanges.

- `btc_blocks_size` : The total size of all block headers and transactions.
- btc_avg_block_size : The average block size in MB.
- btc_n_orphaned_blocks : The total number of blocks mined but ultimately not attached to the main Bitcoin blockchain.
- btc_n_transactions_per_block : The average number of transactions per block.
- btc_median_confirmation_time : The median time for a transaction to be accepted into a mined block.
- btc_hash_rate : The estimated number of tera hashes per second the Bitcoin network is performing.
- btc_difficulty : A relative measure of how difficult it is to find a new block.
- btc_miners_revenue : Total value of coinbase block rewards and transaction fees paid to miners.
- btc_transaction_fees : The total value of all transaction fees paid to miners.
- btc_cost_per_transaction_percent : miners revenue as percentage of the transaction volume.
- btc_cost_per_transaction : miners revenue divided by the number of transactions.
- btc_n_unique_addresses : The total number of unique addresses used on the Bitcoin blockchain.
- btc_n_transactions : The number of daily confirmed Bitcoin transactions.
- btc_n_transactions_total : Total number of transactions.
- btc_n_transactions_excluding_popular : The total number of Bitcoin transactions, excluding the 100 most popular addresses.
- btc_n_transactions_excluding_chains_longer_than_100 : The total number of Bitcoin transactions per day excluding long transaction chains.
- btc_output_volume : The total value of all transaction outputs per day.
- btc_estimated_transaction_volume : The total estimated value of transactions on the Bitcoin blockchain.
- btc_estimated_transaction_volume_usd : The estimated transaction value in USD value.

Etherium Dataset has following features.Ethereum Dataset (ethereum_dataset.csv): This dataset has the following features

- Date(UTC) : Date of transaction
- UnixTimeStamp : Unix timestamp
- eth_Etherprice : price of ethereum
- eth_tx : number of transactions per day
- eth_Address : Cumulative address growth
- eth_Supply : Number of ethers in supply
- eth_Marketcap : Market cap in USD
- eth_Hashrate : hash rate in GH/s
- eth_Difficulty : Difficulty level in TH
- eth_Blocks : number of blocks per day
- eth_Uncles : number of uncles per day
- eth_Blocksize : average block size in bytes.
- eth_Blocktime : average block time in seconds.
- eth_GasPrice : Average gas price in Wei
- eth_GasLimit : Gas limit per day
- eth_Gasused : total gas used per day
- eth_Ethersupply : new ether supply per day
- eth_ChainDataSize : chain data size in bytes
- eth_ens_Register : Ethereal Name Service (ENS) registrations per day

3 Explore Data

Data Exploration involves getting insights into the data using charts and visualizations.As explained in Data Description section, there are two types of datasets.One related to daily trading prices and other giving details on its blockchain characteristics.

The summary statistics and the chart both shows the similar trend that open/low/high/close has been trending in similar way. Further, where the lowest open price has been 68.5, the highest is 4901. which implies that currency has surged very rapidly.

From charts we can see that there hasn't been much activity in 2013, activity has grown in 2014, then was slow in 2015, but recently 2016 and 2017, it has picked lot of action from traders.

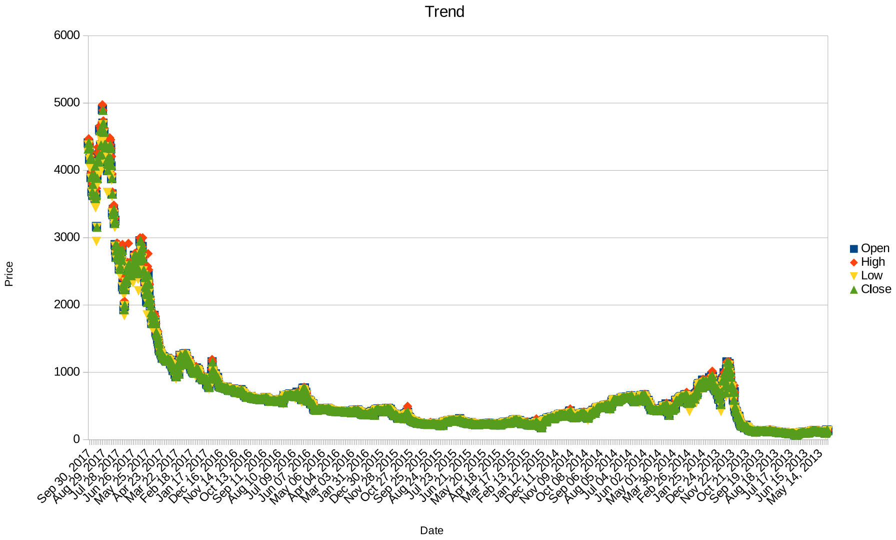

Figure 1: Daily Price Trend

Figure 1. show the chart plotted on bitcoin daily trading prices.

This data can be used to:

- Predict cryptocurrency price in the future.
- This also helps to analyse the surge in the market.
- The comparision of this chart between different cryptocurrencies will help us to compare them and find the popular cryptocurrency and the one which has got the highest price.

  We have also plotted the box plot and see there are lot of outliers, that is because there has been surge recently in crypto currency especially bitcoin trading activities.These outliers also help to analyse any anamolies. Box Plot on Market Price for BitCoin Dataset.csv.

  For a given continuous variable, outliers are those observations that lie outside 1.5 * IQR, where IQR, the 'Inter Quartile Range' is the difference between 75th and 25th quartiles. Look at the points outside the whiskers in below box plot.

  Below figure shows the Univariate approach(BoxPlot On single Column or Attribute)

4 Verify Data Quality

Data has been verified to identify:

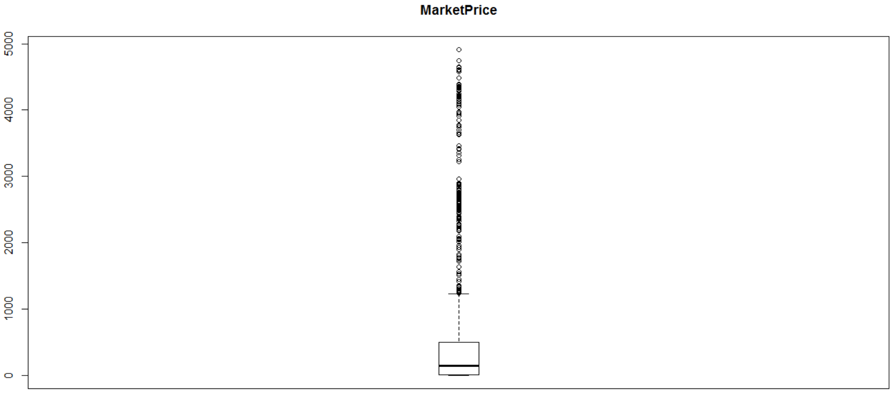

Figure 2: Daily Price Trend

- Missing data: Using R, We identified that bitcoin price dataset has missing values for Volume for 7 months of Year 2013. It amounts to around 15% of the data. We also found that bitcoin dataset contains around 27 missing values, which was only around 0.92% of the total dataset.
- Data Errors: The dataset is mostly numeric and so doesn't have any errors as such. Further there is no text or factor data, so there are no typographical errors.
- Measurement Errors: There is single source of data and is based on single measurement scheme, so there are no measurement errors recorded.
- Coding inconsistencies: Since the data is from single source, there are no coding inconsistencies. Further we have single format of files i.e. csv, all follow the same delimiter scheme.
- Bad Metadata: Metadata is from standard terminology, so there were no bad metadata issues.

5 Data Understanding

5.1 Data Cleaning

5.1.1 Missing Values

There are following possible ways to handle missing values:

- Ignore the tuple: This method can be used when the percentage of missing data is very less. We have used this technique for handling missing data in bitcoin dataset as we had only 0.92% of missing data.

- Fill in the missing value manually: Since this is time series data, this option is not possible.
- Use a global constant to fill the missing value: Again, since this is a time series data, a single global constant may not give the right value and affect the overall analysis.
- Use a measure of central tendency for the attribute(mean or median): Since this is time series data, further has very high volatility since cryptocurrency trading is comparatively very new, mean/median would result in biases.
- Use the attribute mean or median for all samples belonging to the same class: This mechanism can be used in context of time. We can fill the missing values with the average of its nearest set of time period, eg. weekly or monthly or yearly average. It can be decided to take monthly/yearly by analysing the trend in the visual charts.
- Use the most probably value to fill in the missing value: This technique can also be used to predict the missing value based on its most correlated attribute. The correlated attribute can be identified based on domain knowledge or by plotting the charts. Since timeseries data is continuous data, we can use Linear Regression technique to predict the missing related value.

Handling Missing Values in Volume in bitcoin price data

Figure 2. shows that volume has been consistent across the whole year of 2014 with slight peaks.

We have used two ways to identify the missing values. After that we plotted again to verify our prediction.

1. From our domain information, we consider that volume of the price might get impacted based on its average daily price.

We calculate the average daily price by taking average of daily High and Low. Using linear regression model then we build a model between Volume and Average Daily Price. And then we predict the missing values for Volume based on respective average daily prices.

Figure 3 shows the chart with filled missing values using this method.

2. Looking at the Figure 2, it can be seen that the trend in volume of trading has been consistent across year 2014. So, we applied the method where we can take the mean of simliar class and use that mean value to fill the missing values. Thus, we calculate the mean for Dec 2013 to Dec 2014 and then use that mean to fill the missing value. This will avoid any biases as we take the mean from the similar class and not overall.

Figure 4 shows the chart after filled missing values using this method.

If we compare Figure 3 and Figure 4, we can see that in this particular case, regression mechanism didn't predict the values so correctly, but using the mean from nearest neighbour was more consistent

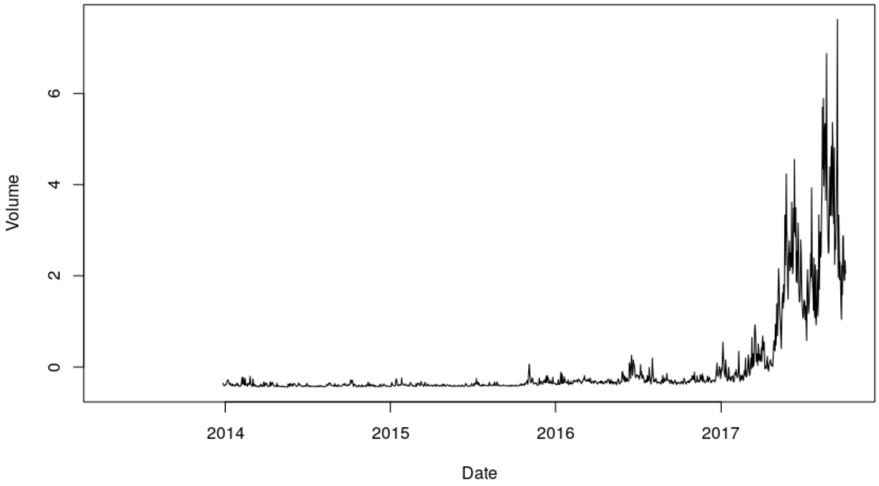

Figure 3: Bitcoin Volume Against Date

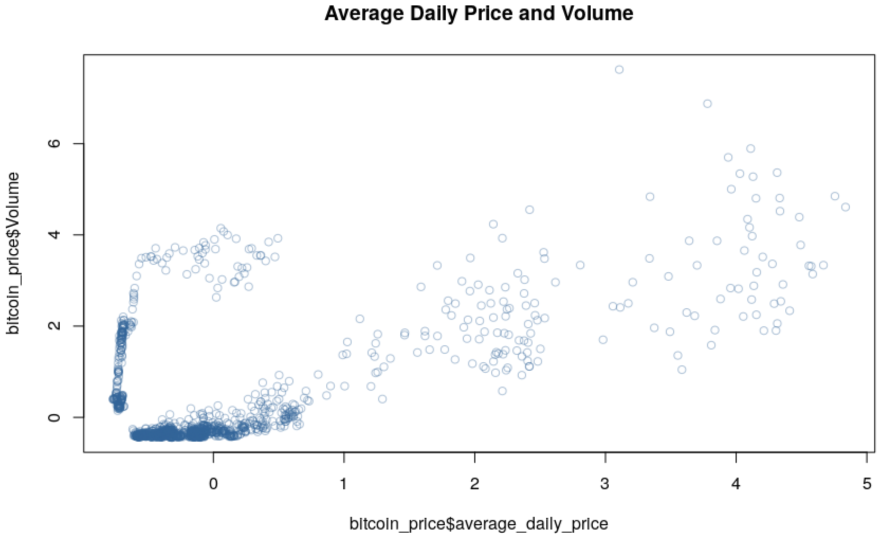

Figure 4: Using Linear Regression

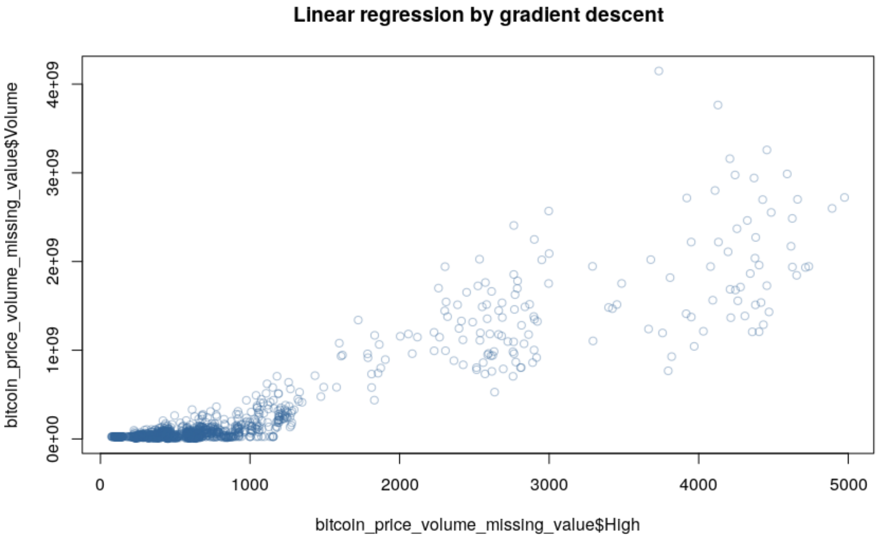

Figure 5: Using Nearest Neighbour Mean

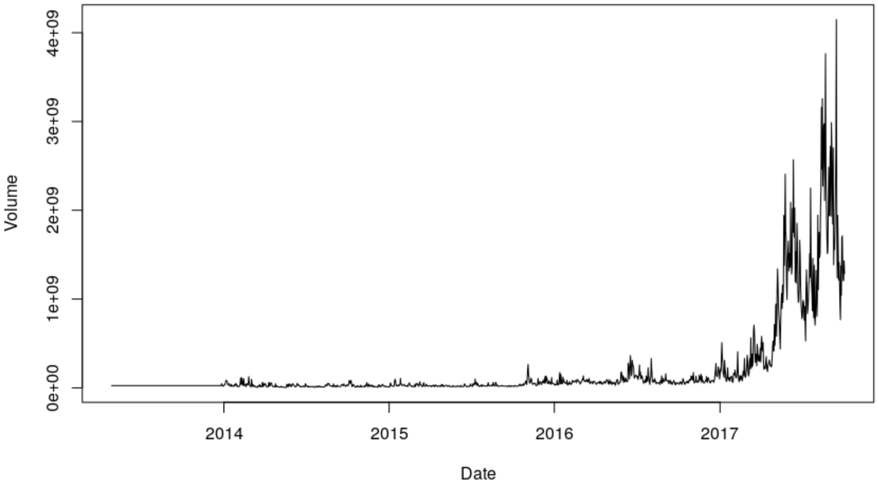

Figure 6: Missing Values Against Date after Filling

So, we go ahead with technique 2 and use that for filling our missing values.

Figure 5 shows that missing values are in trend after filling.

Handling missing values in the bitcoin_dataset.

Bitcoin Dataset has only 27 missing values, which amounted to around 0.92% of the total dataset. Since this is very small amount and ignoring them may not impact the overall analysis, we used the technique to ignore the missing values.

We thus clean the data by removing the rows with missing values.