International Institute of
Information Technology, Bangalore

Project Strategy 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 31, 2017

Contents

1 Data Understanding

1.1 Initial Data Collection

The source of the data is **Kaggle**; a featured dataset called Cryptocurrency Historical Prices. It consists of .CSV files of the prices of top cryptocurrencies including Bitcoin, Ethereum, Ripple and Bitcoin cash. For the purposes of exploration, we have initially considered the data on Bitcoin and Ethereum.

1.2 Description of data

The files are basically of two types. One, which contains the seven attributes, captured from `coinmarketcap.com`, namely - Date, Open, Close, High, Low, Volume, Market Cap. They describe respectively date, the opening price of the currency, the closing price of the currency, the lowest and highest prices recorded in a day, the total amount of cryptocurrencies swapped in the period of 24 hours, and total evaluation of the currency on a given day. The attributes capture DateTime and numeric data. This file is the same for both Bitcoin and Ethereum dataset.

$$MarketCap = PriceXCirculatingSupply \tag{1}$$

where Circulating Supply is the best approximation of the number of coins that are circulating in the market and in the general public's hands.
The second bitcoin file consists of the following 24 handcrafted features, which are :

- 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.

The dataset file describing Ethereum consists of the following 19 handcrafted 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

1.3 Data Exploration

Data Exploration involves getting insights from the data using charts and visualizations. As explained in the previous section, there are two types of datasets i.e. one related to daily trading prices and other giving details on the blockchain characteristics. According to the summary statistics below, the lowest open price has been 68.5 whereas the highest is 4901. which implies that currency has surged very rapidly. From exploratory analytics done, we can see that there hasn't been much activity in the years 2013 and 2015 while activity has grown in 2014 and it picked lot of action from traders in the years 2016 and 2017.

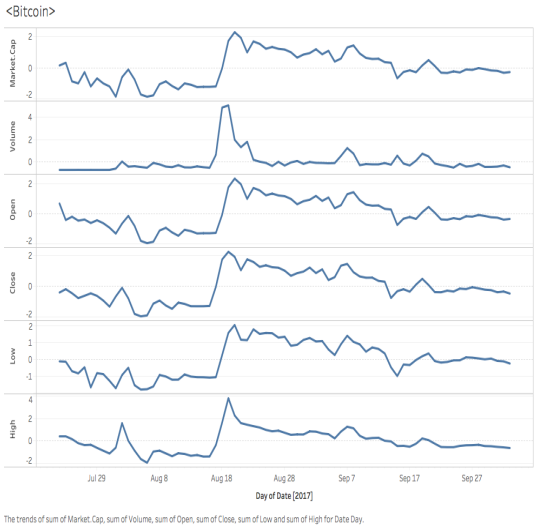

Figure 1: Variation of Bitcoin attributes on a daily basis

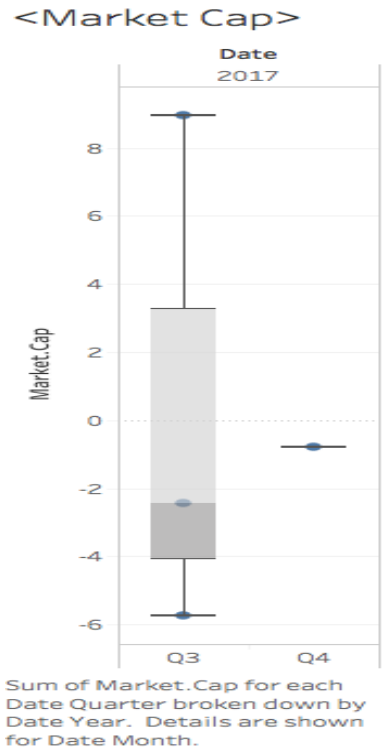

Figure 2: Box plot showing Market cap of Bitcoin

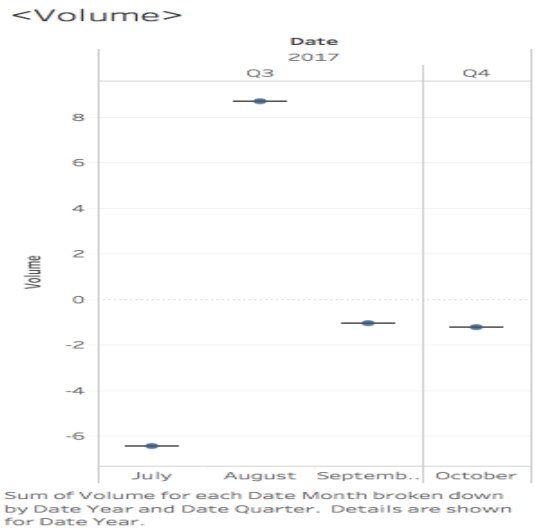

Figure 3: Box plot showing Volume trends of Bitcoin

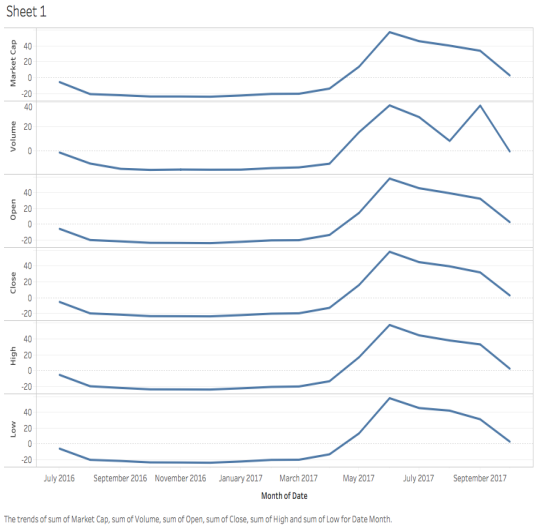

Figure 4: Variation of Ethereum attributes on a monthly basis

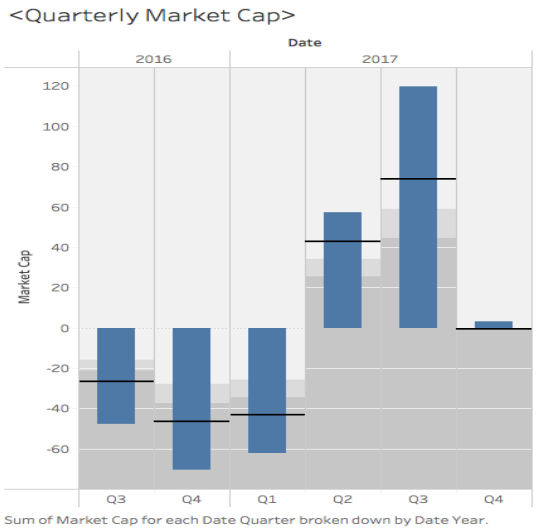

Figure 5: Box plot showing quarterly market cap on Ethereum

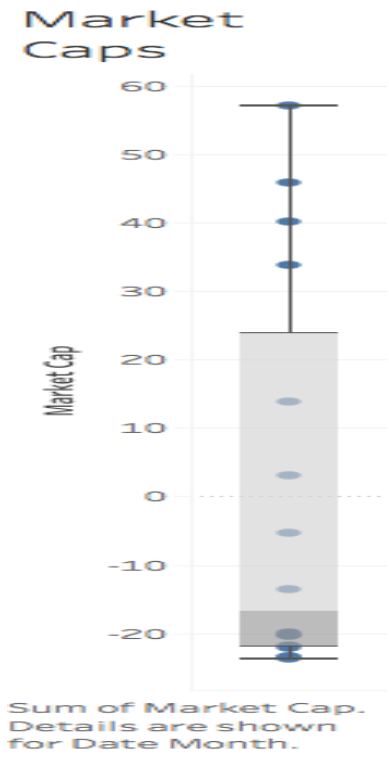

Figure 6: Box plot showing Volume trends of Ethereum

Figure 1 - shows that all the 6 attributes vary almost similarly with time in case of bitcoin.
Figure 2 - This box plot shows the variation of market cap values, while there lie 2 outliers for the first quartile, and there isn't enough data to derive a range for the second quartile.
Figure 3 - This box plot shows that that volume values don't exist as a range for Bitcoin, but exist as almost absolute values for the months of the year.
Figure 4 - Much like in case of Bitcoin, the above graph confirms that Ethereum attributes vary almost similarly with each month.
Figure 5 - This box plot shows the variation in the market cap of each quarter of Ethereum and there seem to be no outliers.
Figure 6 - This box plot shows that that volume values don't exist as a range for Ethereum, but exist as almost absolute values for the quarters.

The box plots on volume and market cap for both Bitcoin and Ethereum give us a valuable insight. The Market Cap value exists as a range for different months of the year, whereas Volume values are absolutes. This shows that while Market cap differs and is volatile because of the volatility in the pricing of the cryptocurrencies, the Volume remains constant for a given time frame, because it is the average amount of crytocurrency swapped in a day. This value doesn't differ rapidly for a given time frame.

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.

  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.

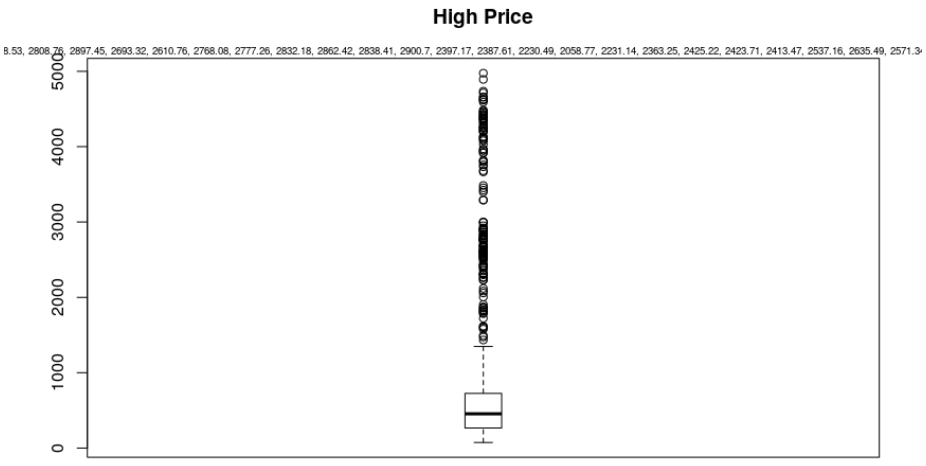

Figure 7: Univariate Approach : High _Price

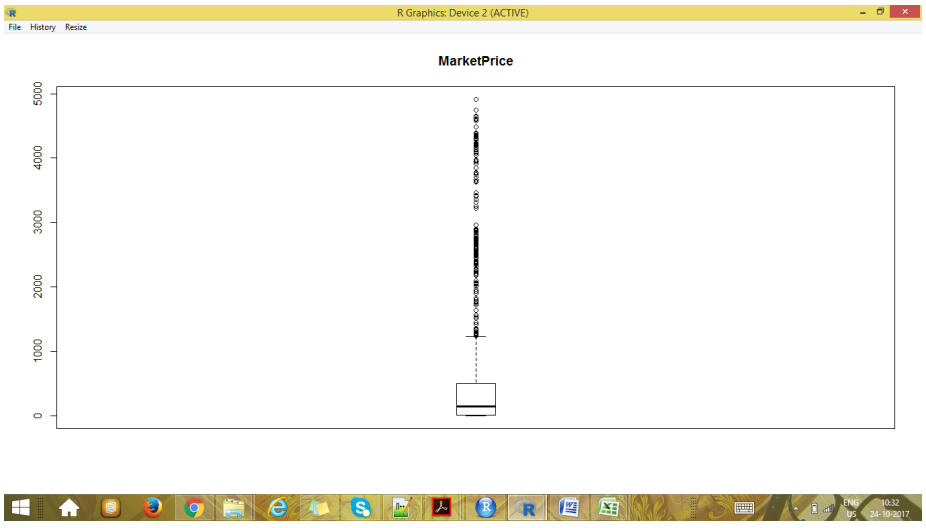

Figure 8: Bitcoin_Market_Price_Outliers

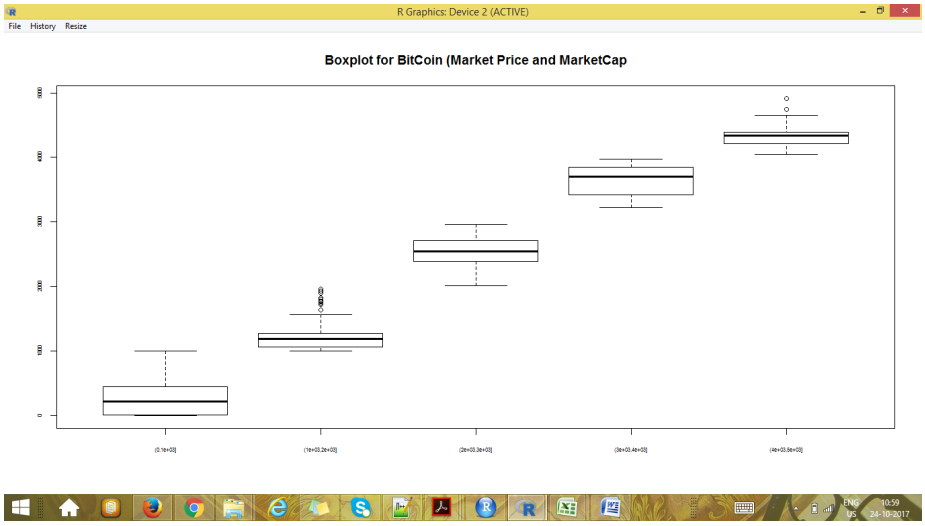

Figure 9: Bivariate Approach : Bitcoin Market Price and Market Cap Outlier

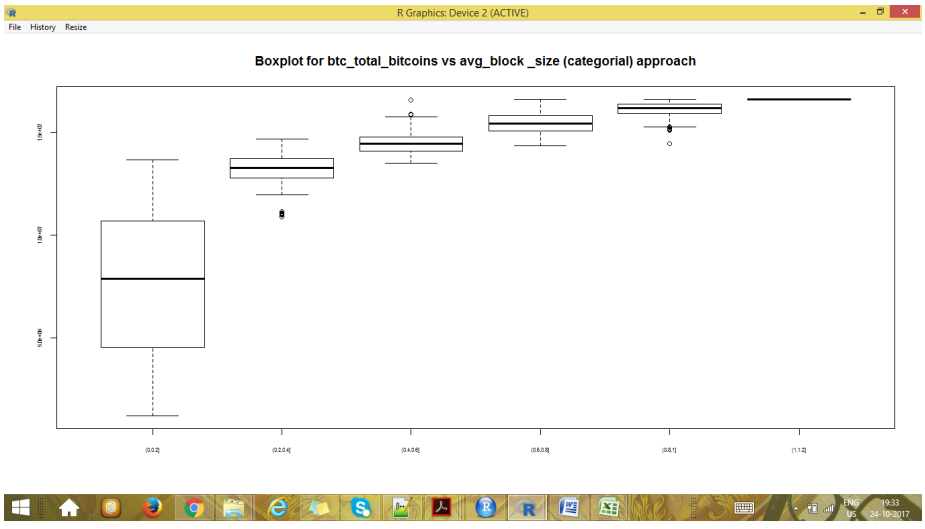

Figure 10: Bivariate approach : Boxplot for Total Bitcoins and AvgBlockSize

Figure 7 - The figure captures outliers on High Price for BitcoinPrice.csv. We are plotting only one column or attribute "High" present in the BitcoinPrice.csv.
Figure 8 - The figure captures outliers on Market price for BitcoinDataset.csv. We are plotting only one column or attribute "MarketPrice" present in the BitcoinDataset.csv
Figure 9 - This figure captures outliers on Market Price and Market Cap for bitcoin datset.csv. We observe outliers which show up as dots.
Figure 10 - This figure captures outliers on Total Bitcoins and Average block size for bitcoin datset.csv. We observe outliers which show up as dots.BoxPlot for btc_total_bitcoins versus avg _block_size for bitcoin dataset.csv. We observe there are two outliers which show up as dots outside the whiskers of box plot.

1.4 Data Quality Verification

Data has been verified to identify:

- Missing data: 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 does not have any numeric errors. 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, thereby 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 following the similar delimiter scheme.
- Bad Metadata: Metadata is from standard terminology, hence no bad metadata issues.

We observed that the Price data is fit to see the volatility trend of crypto currency. Volume and Market capitalization data will help in creating models for predicting future prices. Since the data in different attributes are of different units, it needs normalization. The data is clean and consistent, however one of the attributes in every file has missing values. So, we employed **mice** package to predict the missing values. We did not encounter any data errors, spelling inconsistencies or bad metadata in the the dataset.

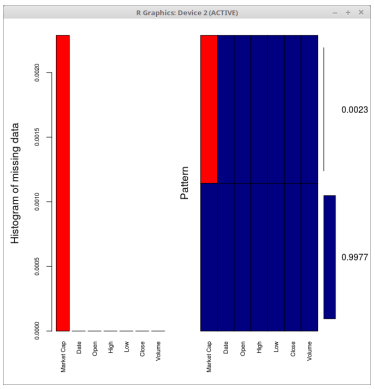

Figure 11: Percentage of missing values in Ethereum Classic Price Dataset

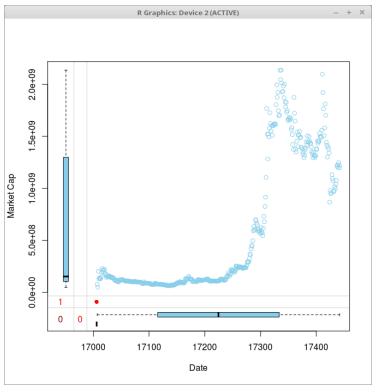

Figure 12: Box plot of missing attribute with Date column

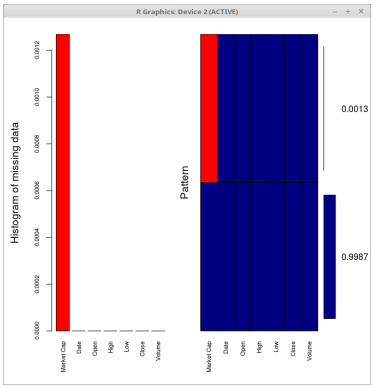

Figure 13: Percentage of missing values in Ethereum Price Dataset

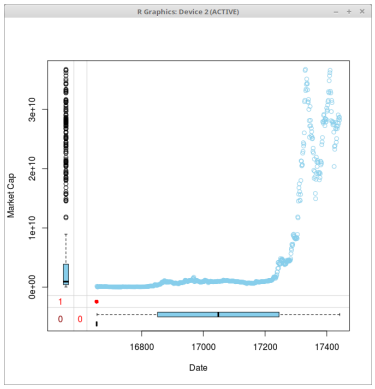

Figure 14: Box plot of missing attribute with Date column

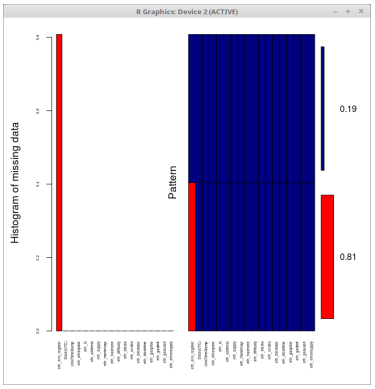

Figure 15: Percentage of missing values in Ethereum Dataset

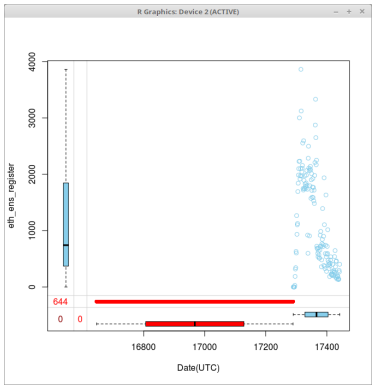

Figure 16: Box plot of missing attribute with Date column

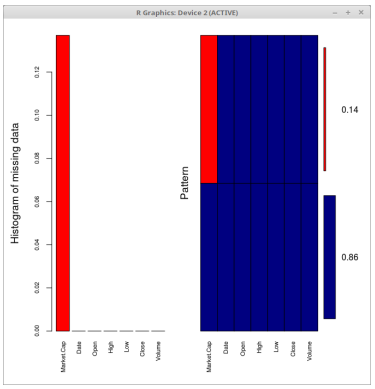

Figure 17: Percentage of missing values in Bitcoin Cash Price Dataset

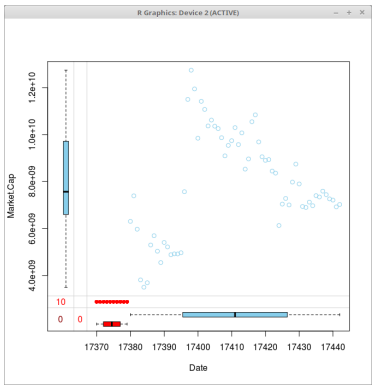

Figure 18: Box plot of missing attribute with Date column

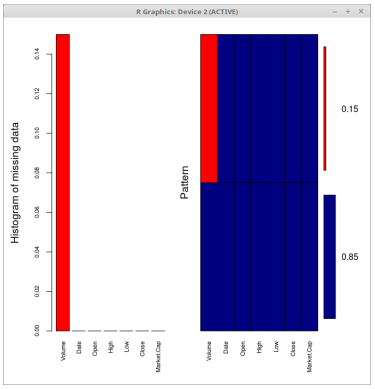

Figure 19: Percentage of missing values Bitcoin Price Dataset

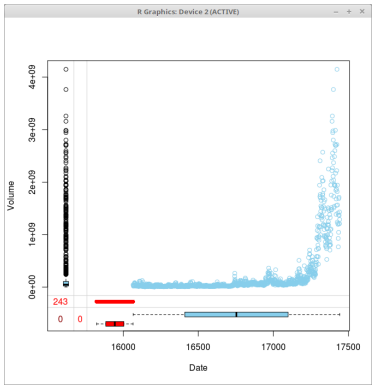

Figure 20: Box plot of missing attribute with Date column

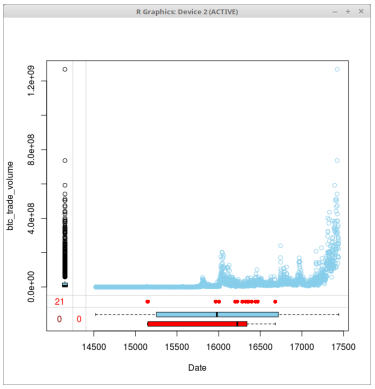

Figure 21: Percentage of missing values in Bitcoin Dataset

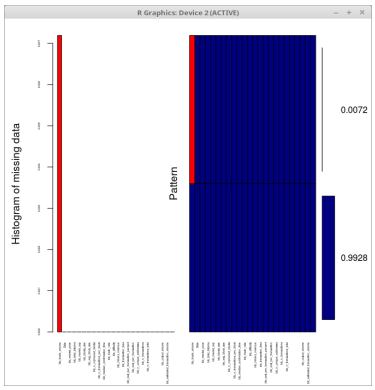

Figure 22: Box plot of missing attribute with Date column

Figure 11-24 describes the overall percentage of missing values for all the six datasets used for this exploration endeavour. The Histogram shows the different attributes contributing in the missing values. The box plot takes two attributes in account with their corelation with each other, one of which has been fixed as Date feature. The other attribute is chosen, contributing most to the missing values. As the two box plots (red and blue in color) are not similar in size with each other, thereby suggesting that the NA values in the dataset are not missed at random.

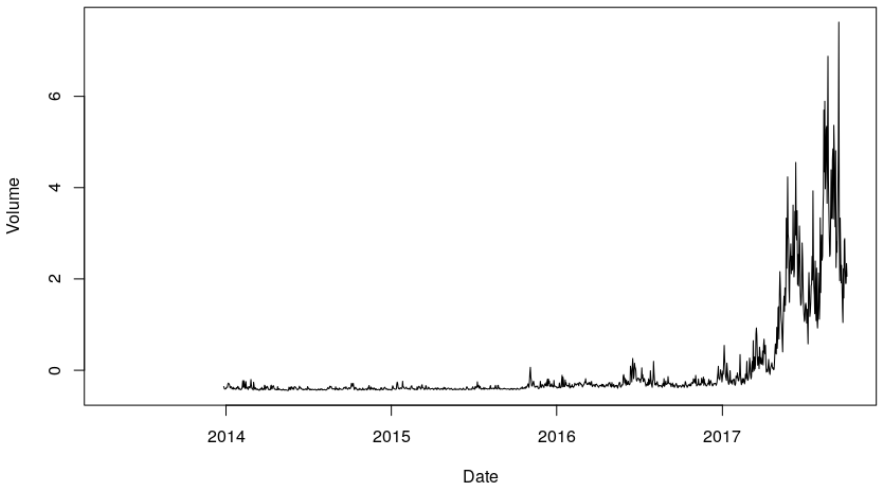

Figure 23: Bitcoin Volume Against Date

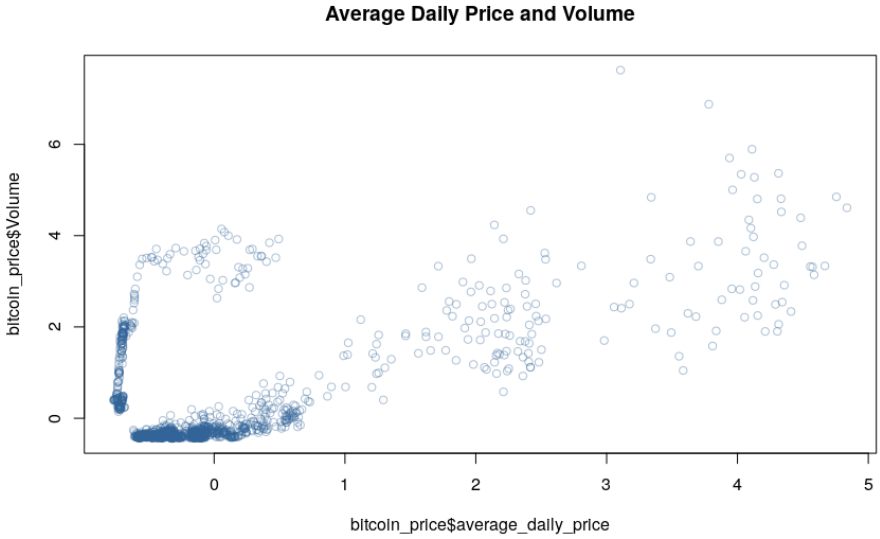

Figure 24: Using Linear Regression

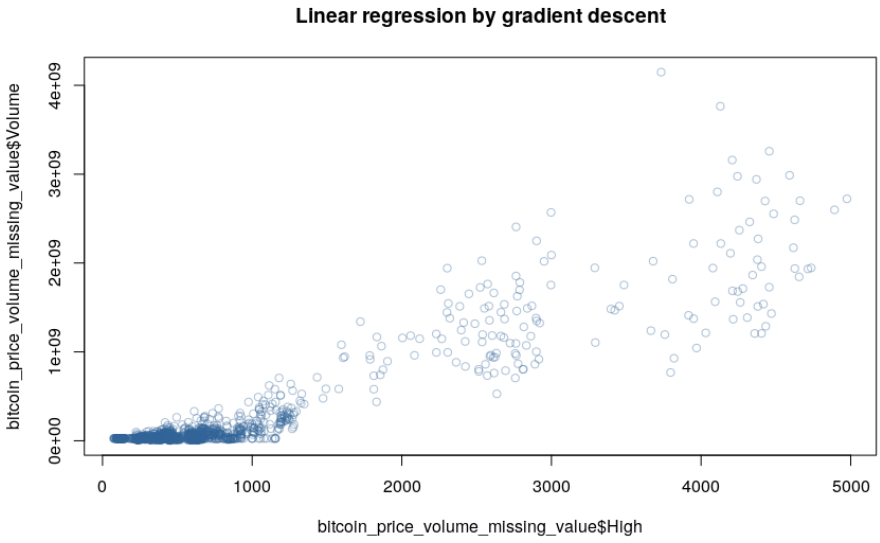

Figure 25: Using Nearest Neighbour Mean

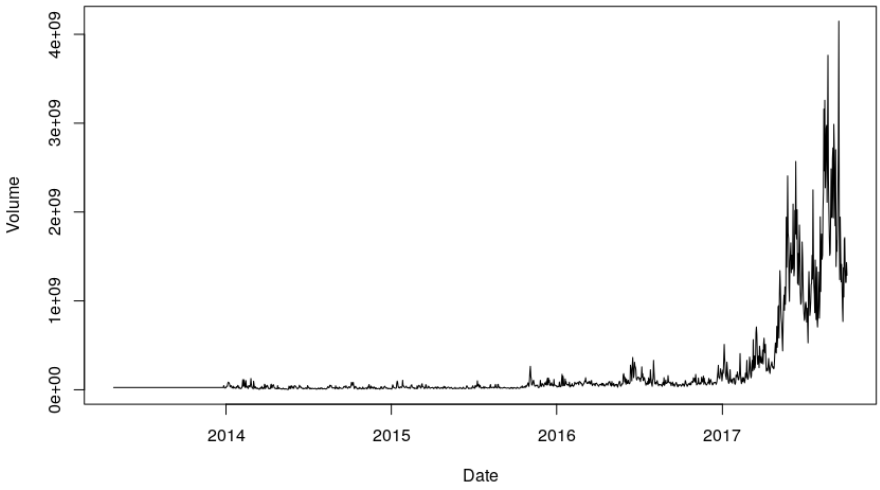

Figure 26: Missing Values Against Date after Filling

The above figure shows that missing values are in trend after filling.
Figure 25 - This figure 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.
Figure 26 - 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.
Figure 27 - 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 similar class and use that mean value to try and 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 in Excel.
Figure 28 - This shows the chart after filled missing values using this method. If we compare above figures, we can see that in this particular case, simple regression mechanism didn't predict the values so correctly, but using the mean from nearest neighbour was more consistent. Hence, we used hybrid regression and decision trees mechanism to predict the missing values, excluding the Date column.

2 Data Preparation

2.1 Selection of Data

There are two types of missing data:

- MCAR: Missing Completely At Random. This is the desirable scenario in case of missing data.
- MNAR: Missing Not At Random. Missing not at random data is a more serious issue and in this case it might be wise to check the data gathering process further and try to understand why the information is missing. For instance, if most of the people in a survey did not answer a certain question, why did they do that? Was the question unclear?

Assuming data is MCAR, too much missing data can be a problem too. Usually a safe maximum threshold is 5% of the total for large datasets. If missing data for a certain feature is more than 5% then we probably should leave that feature out. We therefore check for features (columns) and samples (rows) where more than 5% of the data is missing. Here, the missing value percentage is less than 5% for all the attributes except one, so we keep them. However, for the attribute eth_ens_register approximately 81% values are missing. So, we ignored this dataset until we gather more measurements.

2.2 Cleaning & Formatting of Data

The cleaning and formatting of data follows a certain order:

- We first change the date into proper format, according to the input file.
- We calculate the number of NA(missing) values and get to know their pattern.
- Then we employ imputation. Imputation of missing values refers to replacing missing data with substituted values. In R, we use the mice package to do the same. The mice() function takes three parameters - number of imputed datasets (default value is 5), method of imputation, maximum number of iterations, and seeds(for random number generation).

2.3 Construction & Integration Of Data

- The attributes Open, Close, High and Low attributes are inherently useful in price prediction. However, we have derived an attribute from the former called average price, which can be used in the prediction task. Average price suggests investor to decide whether to trade or not, which might impact thevolume. However, if the volatility is very high, it may not give right prediction. We have calculated average daily price using (Low + High)/2 for bitcoin dataset.

- We merged the above imputed dataset with the original dataset.
- Furthermore, for proper formatting, we add the proper date column and reorder the columns once again.
- Lastly, we normalised the dataset.