DATA UNDERSTANDING REPORT

CHANDANA KOTTA
IMT2014027

DATASETS
The following two datasets were explored:

- Dash currency prices from Feb 4th, 2014 to Oct 3, 2017. 1329 rows, 7 columns
- Ethereum currency prices from Feb 14, 2014 to Oct 3, 2017 790 rows, 7 columns

DATA SOURCE
The data is taken from Kaggle Cryptocurrency Dataset, which is scraped from reliable source coinmarketcap and we are free to use the data.

ATTRIBUTES
Data from both datasets, stored as comma separated files were recorded on a daily basis and have the following attributes:

- Date: Date - datetype
- Open: Opening price - decimal/float
- Close: Closing price - decimal/float
- High: Highest price in day - decimal/float
- Low: Lowest price in day - decimal/float
- Volume: Volume - decimal/float
- Market cap: Market Capitalization - decimal/float

Dimension
Attribute

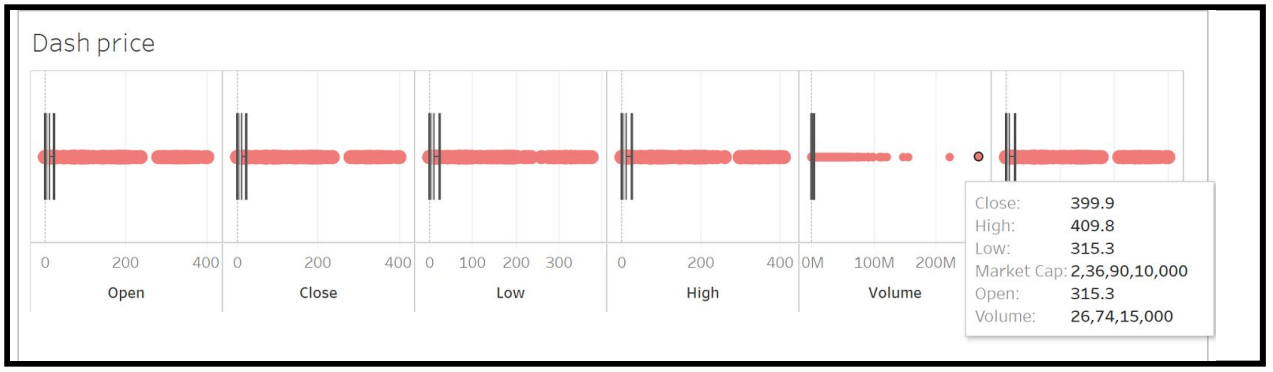
All the attributes are useful for analysis. None of the columns can be dropped.
Since we will be merging various data sources, the common format for date was fixed as YYYY-MM-DD.

DASH CURRENCY DATASET
dash_price.csv , 1329 rows, 7 columns (1 date, 6 numeric)

- There are no missing values/null values in the dataset.
- There are no typographical errors.
- No measurement errors, or bad metadata.
- Coding inconsistencies will arrive while using the date column while aggregating multiple data sources.
- No anomalies were detected in the data distribution.

Supporting charts were generated using tableau.

Box plots for all attributes:



Data preparation: Values in the date column was changed from the format "Month_name DD, YYYY" to a common format of YYYY-MM-DD. The dataset is ready for the next stage of the CRISP-DM framework.

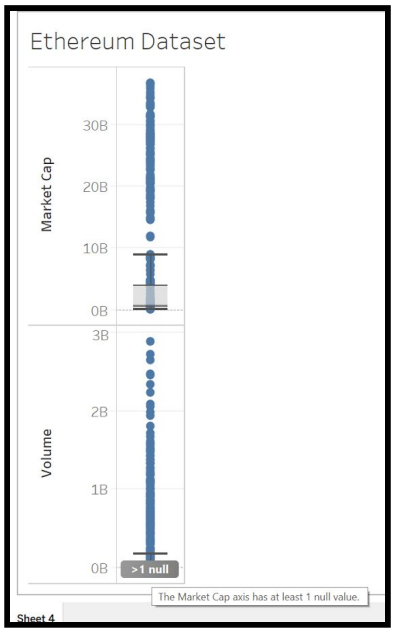
R Script:

```
#loading the dataset
dash_price_data <- read.csv("dash_price.csv")
# Correcting the date format
dash_price_data$Date <- as.Date(dash_price_data$Date, format="%b %d, %Y")
```

ETHEREUM CURRENCY DATASET
ethereum_price.csv , 790 rows, 7 columns (1 date, 6 numeric)

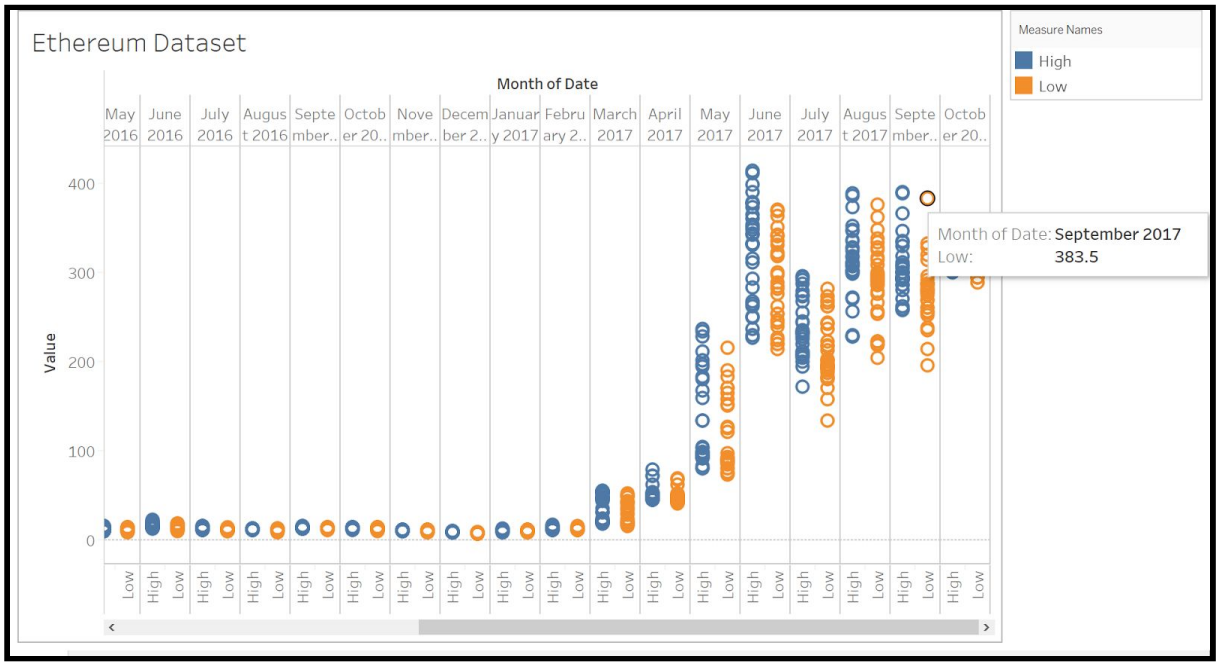
- There was1 missing values/null value in the dataset. The volume column had one missing value.
- There are no typographical errors.
- No measurement errors, or bad metadata.
- Coding inconsistencies will arrive while using the date column while aggregating multiple data sources.
- No anomalies were detected in the data distribution.

Supporting charts are generated using tableau.
Missing value in market cap identified by tableau and corrected in excel using average operator.

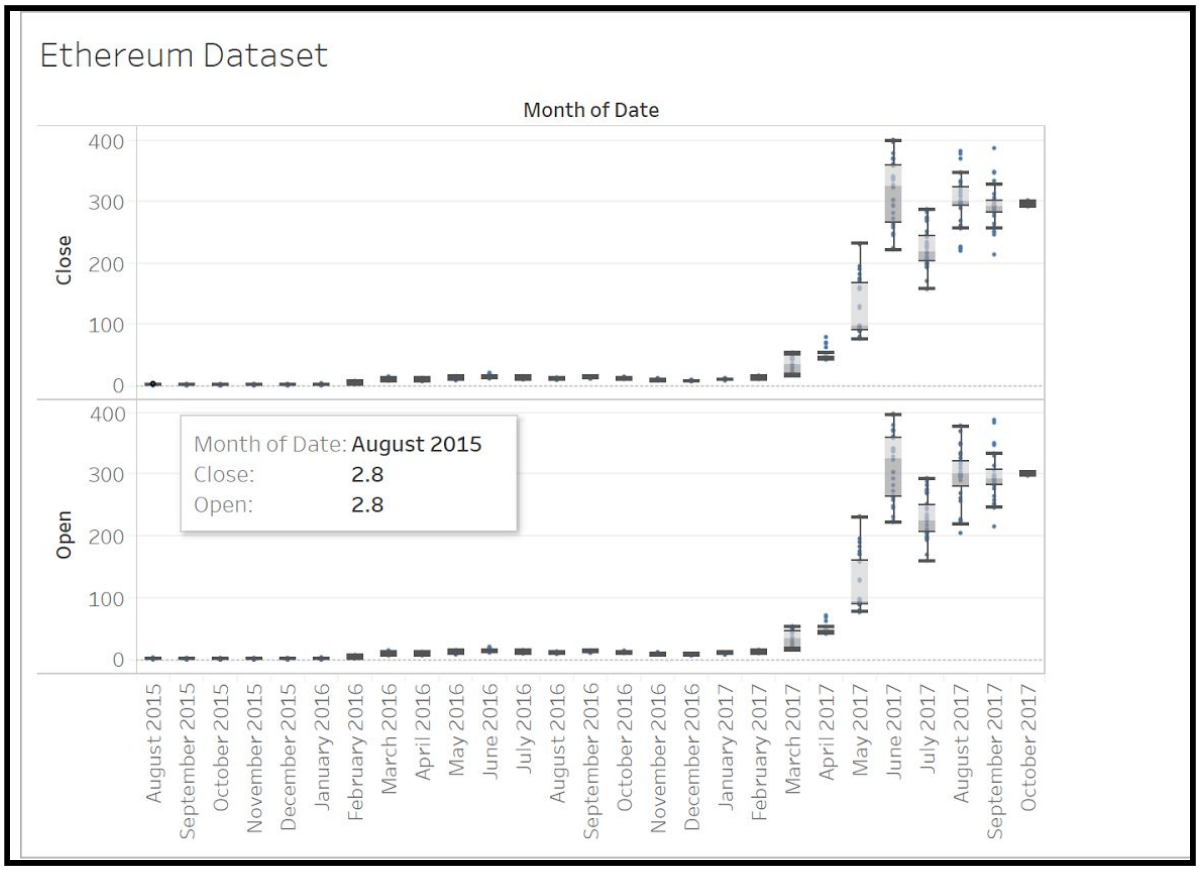


786	Aug 11, 2015	0.708087	1.13	0.663235	1.07	14,63,100	4,27,96,500
787	Aug 10, 2015	0.713989	0.729854	0.636546	0.708448	4,05,283	4,31,30,000
788	Aug 09, 2015	0.706136	0.87981	0.629191	0.701897	5,32,170	4,26,37,600
789	Aug 08, 2015	2.79	2.8	0.714725	0.753325	6,74,188	16,79,11,000
790	Aug 07, 2015	2.83	3.54	2.52	2.77	1,64,329	
791							

High and Low columns against month shows expected increasing trend, and few outliers.



Box plots for easily identifying outliers: Open and Close column against month shows expected increasing trend, and few outliers.



Data preparation:

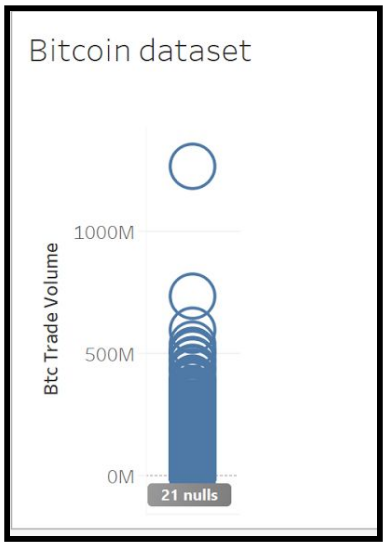
- The single missing value for market capitalization was edge value and we decide to delete that value.
- Values in the date column was changed from the format “Month_name DD, YYYY” to a common format of YYYY-MM-DD. The dataset is ready for the next stage of the CRISP-DM framework.

R Script:

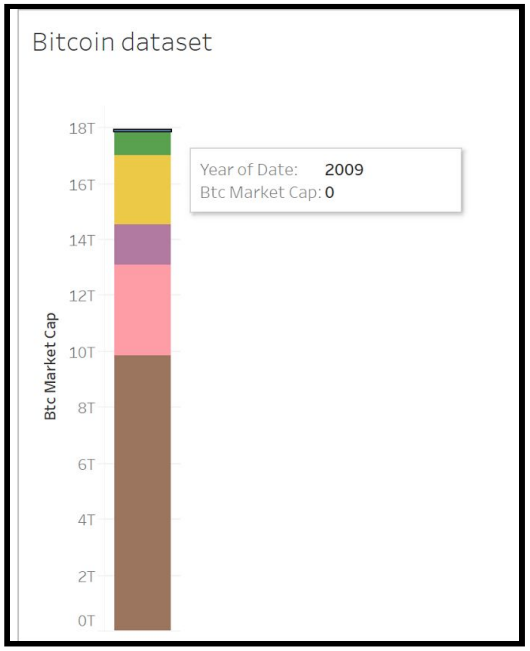
```
#loading the dataset
ethereum_price_data <- read.csv("ethereum_price.csv")
# Correcting the date format
ethereum_price_data$Date <- as.Date(ethereum_price_data$Date, format="%b %d, %Y")
# The - value is replaced by NA in market cap and then deleted the attribute
ethereum_price_data$Market.Cap[ ethereum_price_data$Market.Cap == '-' ] <- NA
ethereum_price_data<- na.omit(ethereum_price_data)
```

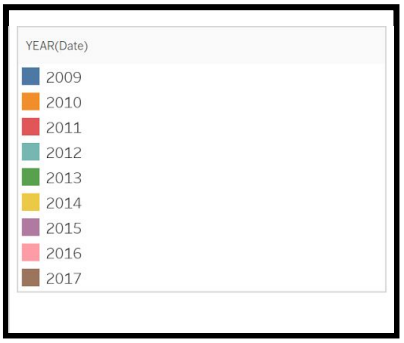
BITCOIN DATASET

In addition to the above two currencies, I have explored the bitcoin dataset. There were 21 null values in the trade value column. The nulls were scattered across different dates that were non-consecutive.



Bitcoin market cap was 0 in the first year of its existence. Data is consistent with true values.





Bitcoin_dataset Data Understanding and Quality

IMT2014048

Shashank Motepalli

About Bitcoin_dataset

No. of records= 2920

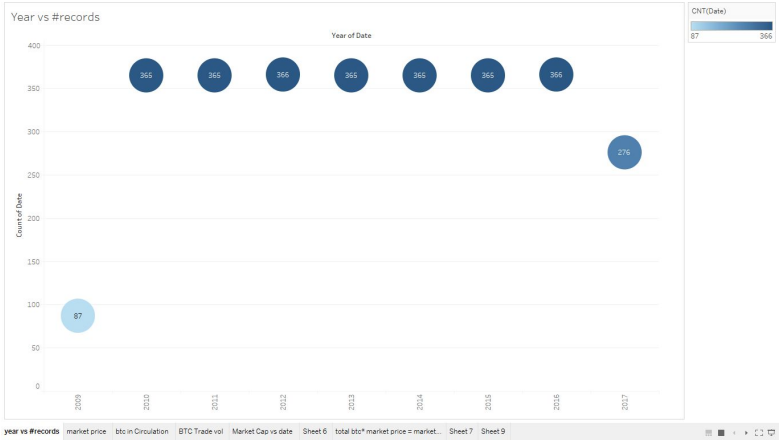
Dataset scraped from: Blockchain Info

Dataset obtained from: Kaggle

Column Attributes

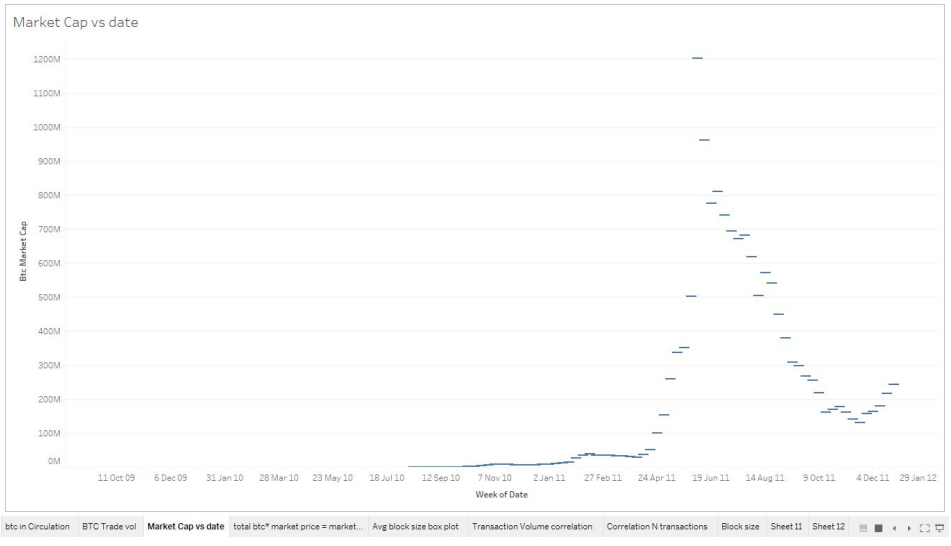
- **Date:** Date time. dimensional, primary key in continuous time series bitcoin dataset.It is continuous with no missing values. All the records are captured at 00:00 of each day starting 6/10/2009 to 3/10/2017.

Issue 1: The date attribute is of form 06-08-2011 00:00:00 but in all our other datasets it is of form 2011-08-06. So to maintain consistency in system, we need to change format.



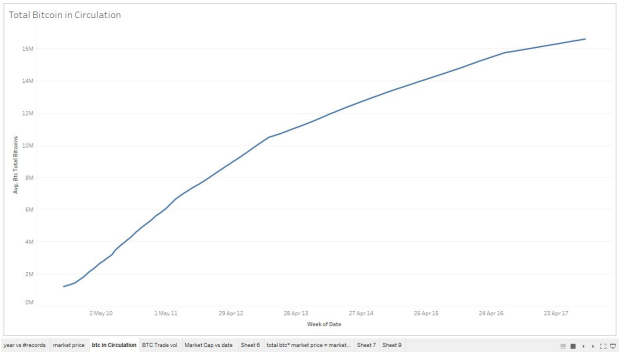
- **Market price:** numeric. Average USD market price across major bitcoin exchanges.

Found that market price is $0 during first year of existence.



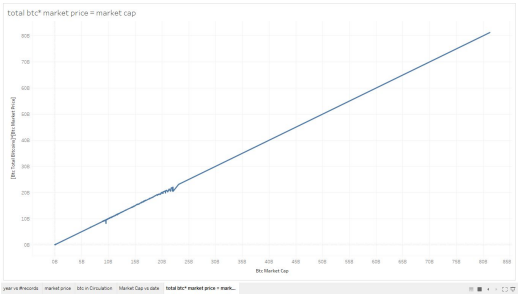
- **Total BTC:** Numeric. The total number of bitcoins that have already been mined.

Avg Total Btc vs Week kept on increasing, so verified to be correct.

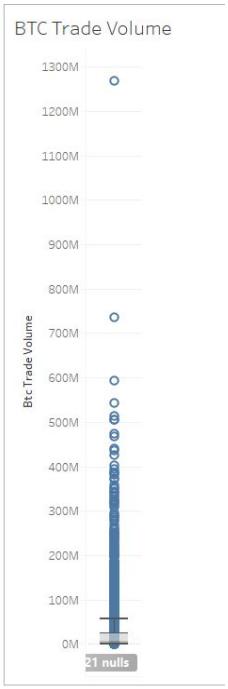


- **Market Cap**: numeric. This attribute has value of 0 during first year of existence.

Issue 2: The market cap attribute is redundant, it is just aggregated attribute. total btc* market price = market cap

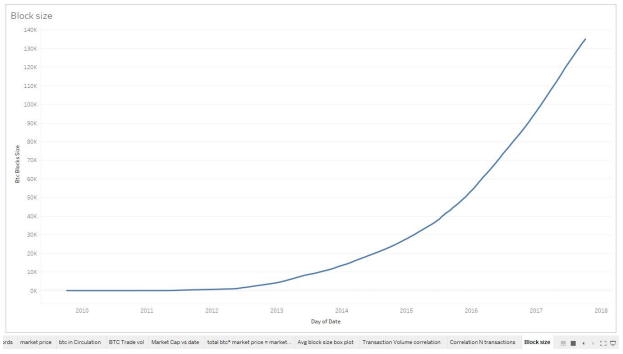


- **BTC Trade Volume:** Numeric. The total USD value of trading volume on major bitcoin exchanges.

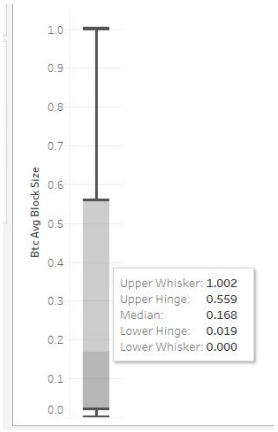


*Issue 3:*21 null values are present randomly in data.

- **Block Size**: The total size of all block headers and transactions. Numeric. The value seem correct as it is always increasing with time. This attribute is related to memory considerations, so might not be useful for our analysis.



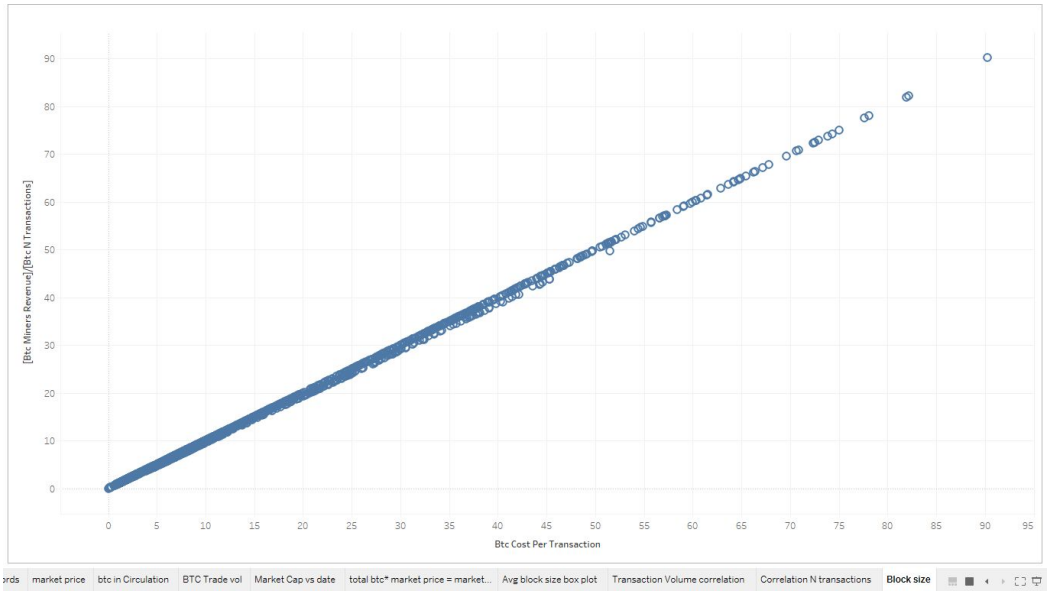
- **Avg block size:** The average block size in MB. Numeric.



- **N Orphaned Blocks:** The total number of blocks mined but ultimately not attached to the main Bitcoin blockchain. Numeric in range(0,7).
- **N Transactions per Block:** Numeric. The average number of transactions per block.
- **Median confirmation time:** The median time for a transaction to be accepted into a mined block. Numeric (in min).
- **Btc Hash Rate:** The estimated number of terahashes per second the Bitcoin network is performing. Numeric.
- **Btc Difficulty:** A relative measure of how difficult it is to find a new block. Numeric.
- **Btc Miners Revenue:** Total value of coinbase block rewards and transaction fees paid to miners. Numeric
- **Btc Transaction fees:** The total value of all transaction fees paid to miners. Numeric
- **Btc cost per transaction percent :** miners revenue as percentage of the transaction volume. Numeric

- **Btc cost per transaction :** Numeric.

Issue 4: Redundant, miners revenue divided by the number of transactions.

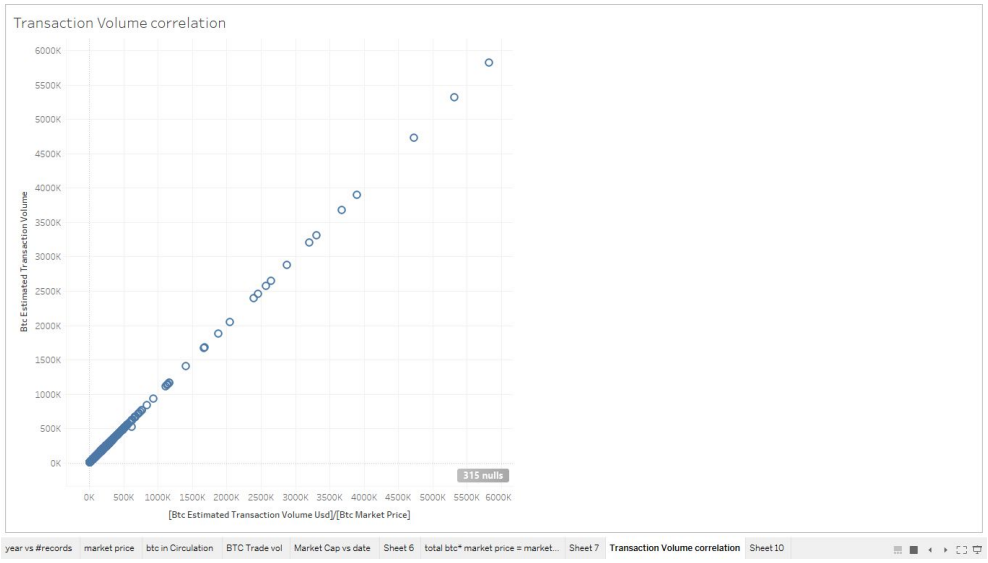


- **Btc N unique addresses :** The total number of unique addresses used on the Bitcoin blockchain. Numeric.
- **Btc N Transactions** : The number of daily confirmed Bitcoin transactions.
- **Btc N Transactions Total** : Total number of transactions.
- **Btc N Transactions Excluding Popular** : The total number of Bitcoin transactions, excluding the 100 most popular addresses. There is some correlation with total transactions and daily confirmed transactions too.

- **Btc N Transactions Excluding chains longer than 100 :** The total number of Bitcoin transactions per day excluding long transaction chains. Numeric.
- **Btc output volume :** The total value of all transaction outputs per day.

- **Btc estimated transaction volume :** Numeric. The total estimated value of transactions on the Bitcoin blockchain.
- **Estimated transaction volume usd** : The estimated transaction value in USD value.

Issue 5: This attribute is redundant. Its product of transaction volume and market price USD. Null values are due to division by 0 price during 2009.



Bitcoin dataset Data Preparation

R Scripts:

```
#loading the data.
bitcoin_dataset <- read.csv("bitcoin_dataset.csv")
#Issue1: Date is of wrong format. To fix it, we need to change to date data type.
bitcoin_dataset$Date<- as.Date(bitcoin_dataset$Date, format="%Y-%m-%d")
#Issue 2: deleting the market cap column.
bitcoin_dataset$btc_market_cap <- NULL
#to check which columns have null values
colnames(bitcoin_dataset)[colSums(is.na(bitcoin_dataset))>0]
#Issue 3: The na values in trade volume are filled with linear interpolation.
library("zoo")
na.approx(bitcoin_dataset$btc_trade_volume)
#Issue 4: Btc cost per transaction attribute is redundant.
bitcoin_dataset$btc_cost_per_transaction <- NULL
#Issue 5: Btc estimated transaction volume USD attribute is redundant.
bitcoin_dataset$btc_estimated_transaction_volume_usd <- NULL
```

Data Analysis Project

Cryptocurrency historical prices

Pranav.S

IMT2014039

Litecoin cryptocurrency analysis

1) Data Understanding Report

Data understanding is the knowledge that one has about the data, the needs the data will satisfy, its content and location.

a) Data collection

The Litecoin dataset was collected from a Kaggle dataset containing historical cryptocurrency prices, which in turn was scraped from (https://coinmarketcap.com).

b) Data Description

The Litecoin dataset consists of many rows, with a row entry for each day. Most of the value types are numeric, with only the date attribute being in the date format.

Row description

The database comprises of litecoin data from 28th April 2013, to 3rd October 2017, with a row entry for every day. Thus, there are 1620 rows in the dataset.

Column fields

The dataset consists of 7 attributes

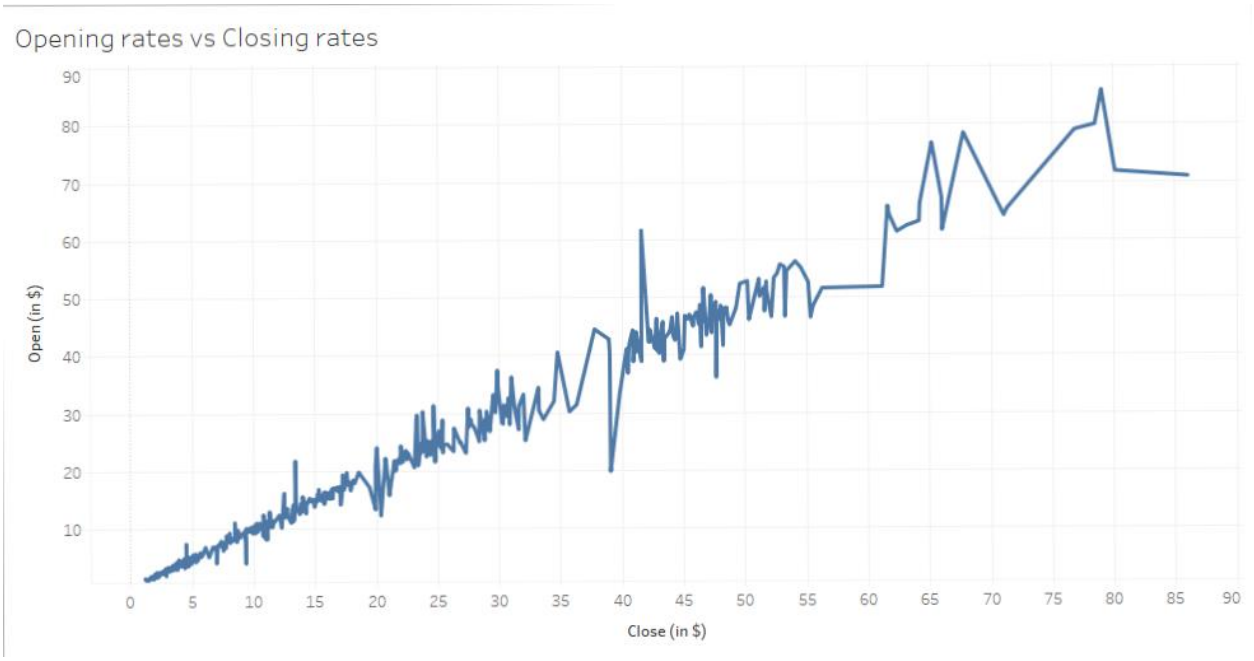
Column Attribute	Description	Metric
Date	Date of observation is of the form MM/DD/YYYY	DateTime
Open	The opening price of litecoin on the given day	Numeric
Close	The closing price of litecoin on the given day	Numeric
High	Highest price of litecoin on the given day	Numeric
Low	Lowest price of litecoin on the given day	Numeric
Volume	Volume of transactions on the given day	Numeric
Market Cap	Market capitalization in USD ($)	Numeric

c) Visual reports

As part of data understanding, visual tableau reports were constructed.

As we can see, from the timeseries graph, litecoin's market cap has been increasing over the last few months reaching upto $4.1B in September 2017, before dropping down and currently(3rd October 2017) having a market cap of $2.8B

Line-plots were plotted for Market-cap, Opening value, Closing value, Lowest value and Highest value. We observe that there is a similar pattern among the attributes. A plot between opening and closing rates were plotted to see the correlation between them.



By observation, one can see that the relation between open and close rates, is not very strong, and hence we cannot do away with either of them.
The same was observed for other possible pairs too.

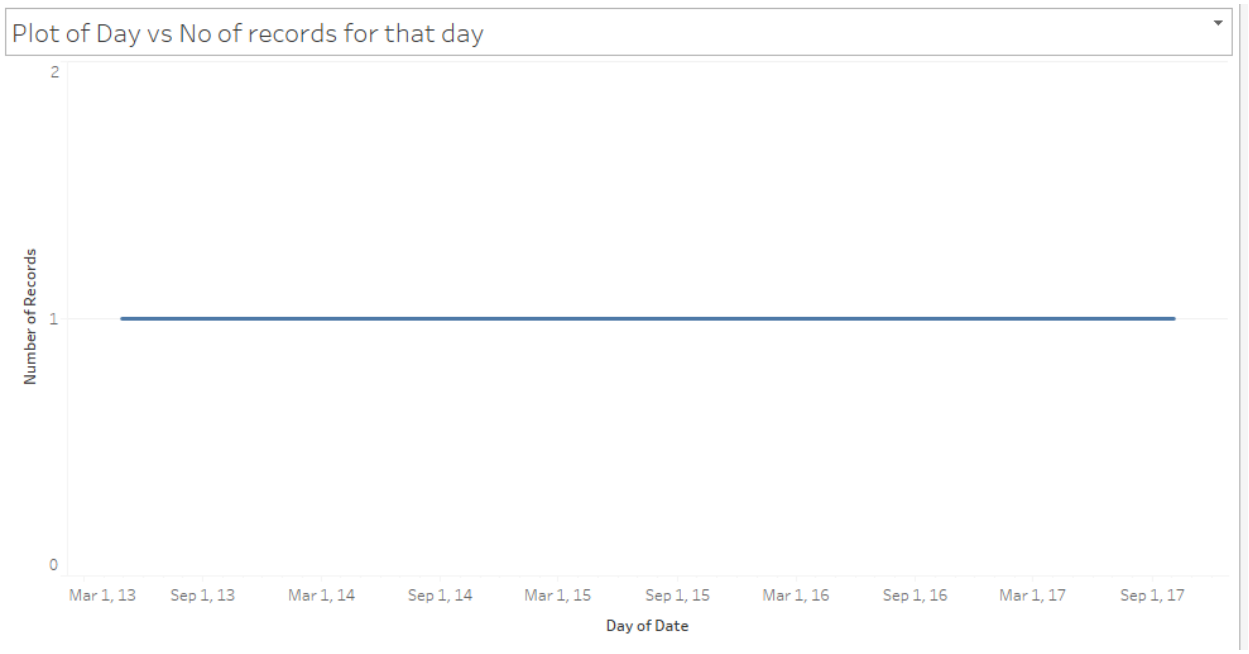
2) Data Quality Report

The data is got from a reliable source (i.e) https://coinmarketcap.com, and is hence assumed to be reliable.

The litecoin dataset has around 243 null values for the volume attribute from, 28/4/2013 to 26/12/2013. We will either replace the values with 0's or delete those rows. As part of preparation we do both separately. Firstly, the null values for volume during this time period are replaced with 0's. And secondly, in another copy, the rows containing the null values are deleted.

The date format in the dataset is of the form, *MM/DD/YYYY*, while the default format for datetime in R is *YYYY-MM-DD.* Also, having a standard date format for all the datasets that we use for the project, is better. So, the date format in the dataset should be changed to this format.

To check if there are any missing values, a line plot was drawn between day and the number of records for that day.



This shows that there's exactly only one entry for each day. Thus, there is no missing data , and also no data has been duplicated. We can also check for null values in a column using R.

3) Data Preparation

In this part, R scripts are written to correct the data quality issues of the dataset.

Issue 1: Null values

R-script

```
# Shows the number of missing/NA values for each column
colSums(is.na(litecoin_price))
#We find that only the volume columns has NA values

# Changes the NA values of Volume to 0's
litecoin_price$Volume[is.na(litecoin_price$Volume)] <- 0
```

We do the above replacement, or we delete the rows containing null values

```
litecoin_price <- na.omit(litecoin_price)
```

Issue 2: Changing the date format

R-script

```
# Changes the date format of the dataset to a standard date format
litecoin_price$Date <- as.Date(as.character(litecoin_price$Date), format = "%b %d, %y")
```

Bitcoin cryptocurrency analysis

A similar analysis that was done for the litecoin dataset was performed on the bitcoin dataset too, as the datasets were similar in almost all aspects.

1) Data understanding report

The bitcoin dataset was collected from a Kaggle dataset containing historical cryptocurrency prices, which in turn was scraped from (https://coinmarketcap.com).

The dataset consists of 1620 rows, with each row entry corresponding to a day. The valuetypes of the columns are mostly numeric, with only the date attribute being in the date format. The column fields are the same as litecoin's, totaling to 7 columns.

The data visualization plots were similar to the litecoin plots.

2) Data quality report

The data is got from a reliable source (i.e) https://coinmarketcap.com, and is hence assumed to be reliable.

The bitcoin dataset has around 243 null values for the volume attribute from, 28/4/2013 to 26/12/2013. The null values for volume during this time period are replaced with 0's.

The date format in the dataset is of the form, *MM/DD/YYYY*, while the default format for datetime in R is *YYYY-MM-DD.* Also, having a standard date format for all the datasets that we use for the project, is better. So, the date format in the dataset should be changed to this format.

3) Data Preparation

In this part, R scripts are written to correct the data quality issues of the dataset.

Issue 1: Null values

R-script

```
# Shows the number of missing/NA values for each column
colSums(is.na(bitcoin_price))
#We find that only the volume columns has NA values

# Changes the NA values of Volume to 0's
bitcoin_price$Volume[is.na(bitcoin_price$Volume)] <- 0
```

We do the above replacement, or we delete the rows containing null values

```
litecoin_price <- na.omit(litecoin_price)
```

Issue 2: Changing the date format

R-script

```
# Changes the date format of the dataset to a standard date format
bitcoin_price$Date <- as.Date(as.character(bitcoin_price$Date), format = "%b %d, %y")
```

Data Understanding Report for cryptocurrencies

Ripple and Monero

Armitha(Imt2014060)

Data Collection:

We have taken the cryptocurrencies dataset from kaggle which in turn had been scraped the data from the website https://coinmarketcap.com/

Dataset:

Ripple cryptocurrency dataset : There are 1522 rows (Here, each row constitute a day) and 7 columns in the dataset.

Monero cryptocurrency dataset : There are 1232 rows (Here, each row constitute a day) and 7 columns in the dataset.

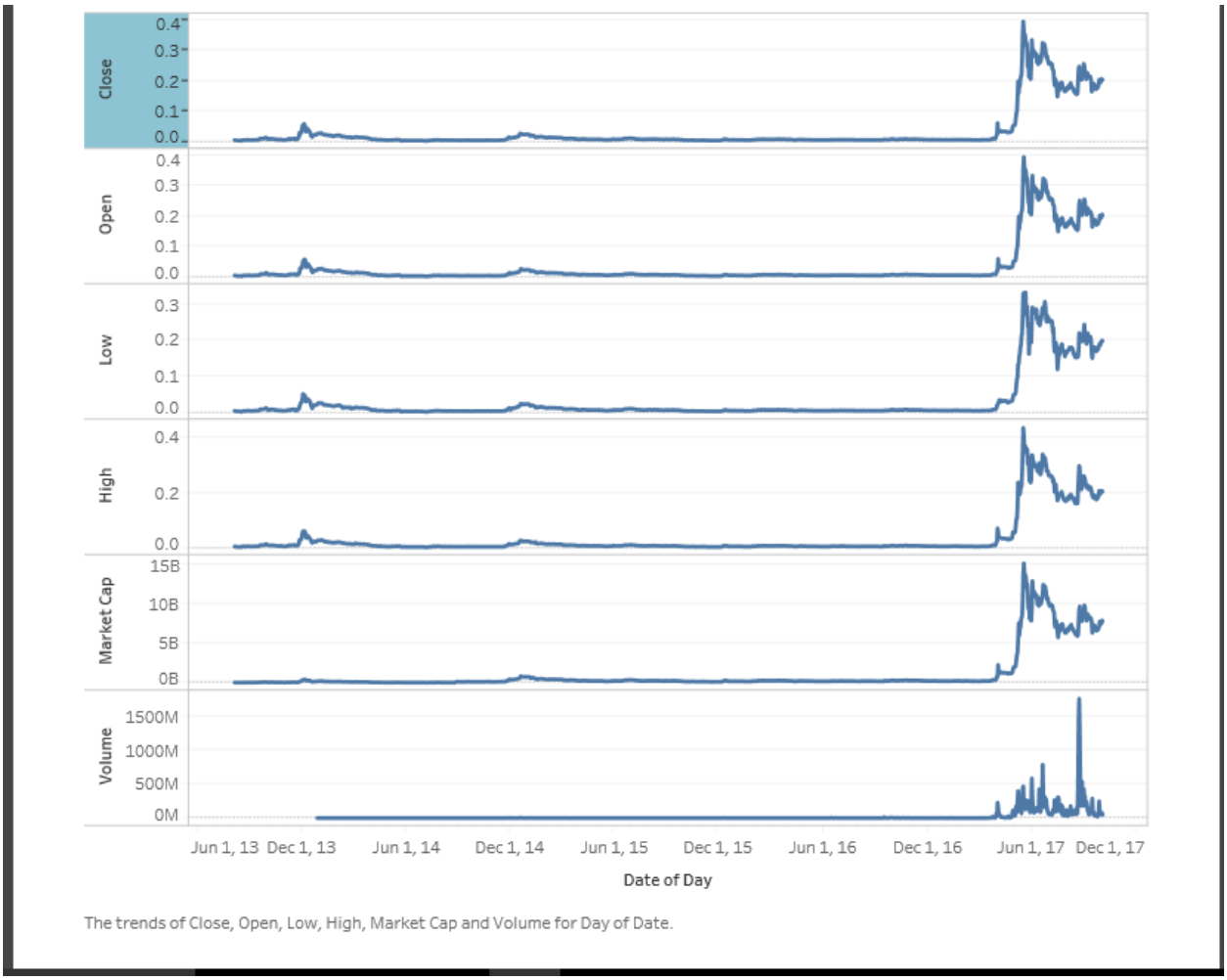
These 7 columns are same in both the datasets. They are:

1. Date : Type - Date , Description : Date of the observation.
2. Open : Type - Numeric, Description : Price of Cryptocurrency at 12:01 AM UTC of a given day
3. Close : Type - Numeric, Description : Price of Cryptocurrency at 11:59 PM UTC of a given day
4. High : Type - Numeric, Description : Highest price of Cryptocurrency on a given day
5. Low : Type -Numeric, Description : Lowest price of Cryptocurrency on a given day
6. Volume : Type-Numeric, Description : Total sum of the transactions, including buying and selling.
7. Market Cap : Type -Numeric , Description : Market capitalization in USD.

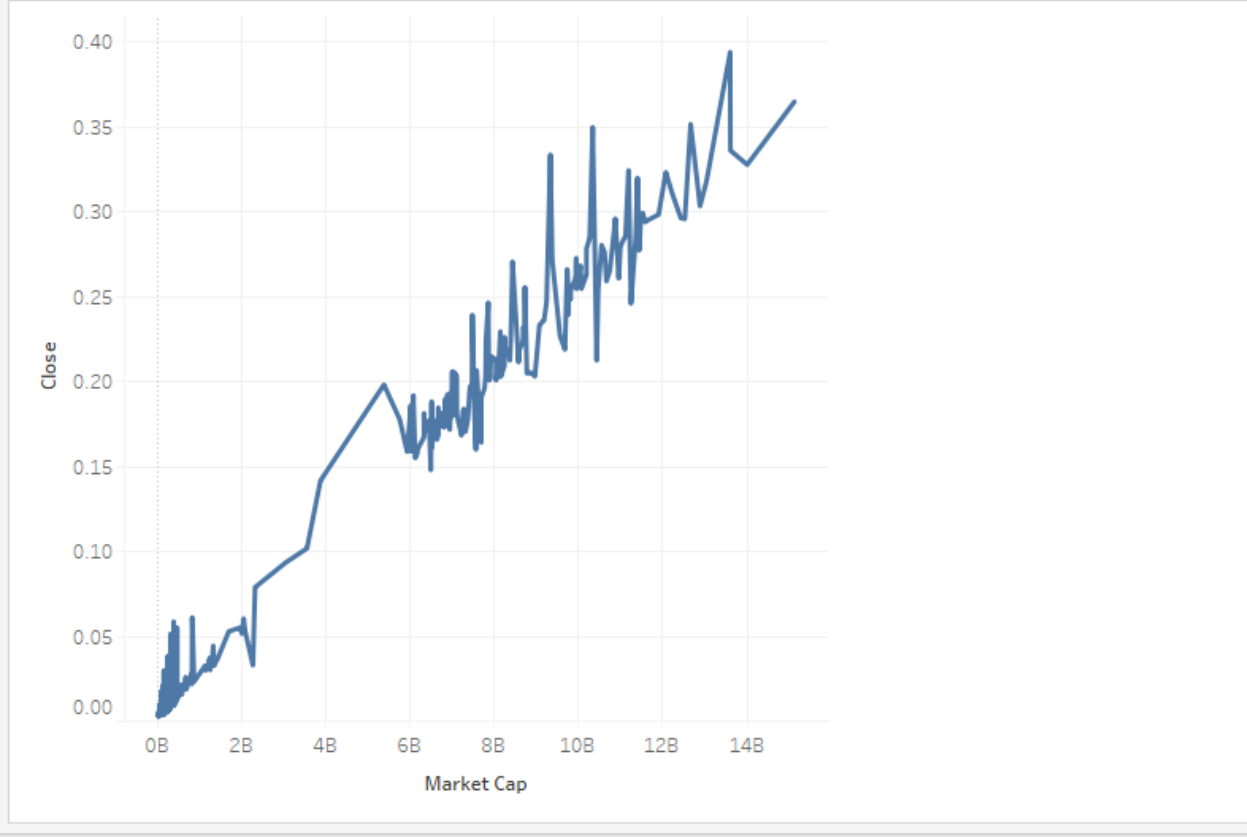
All the attributes above except Date are measured in USD.

Data Understanding For Ripple:

Plot showing the day trends for the attributes Open,Close,High,Low,Market Cap and Volume.



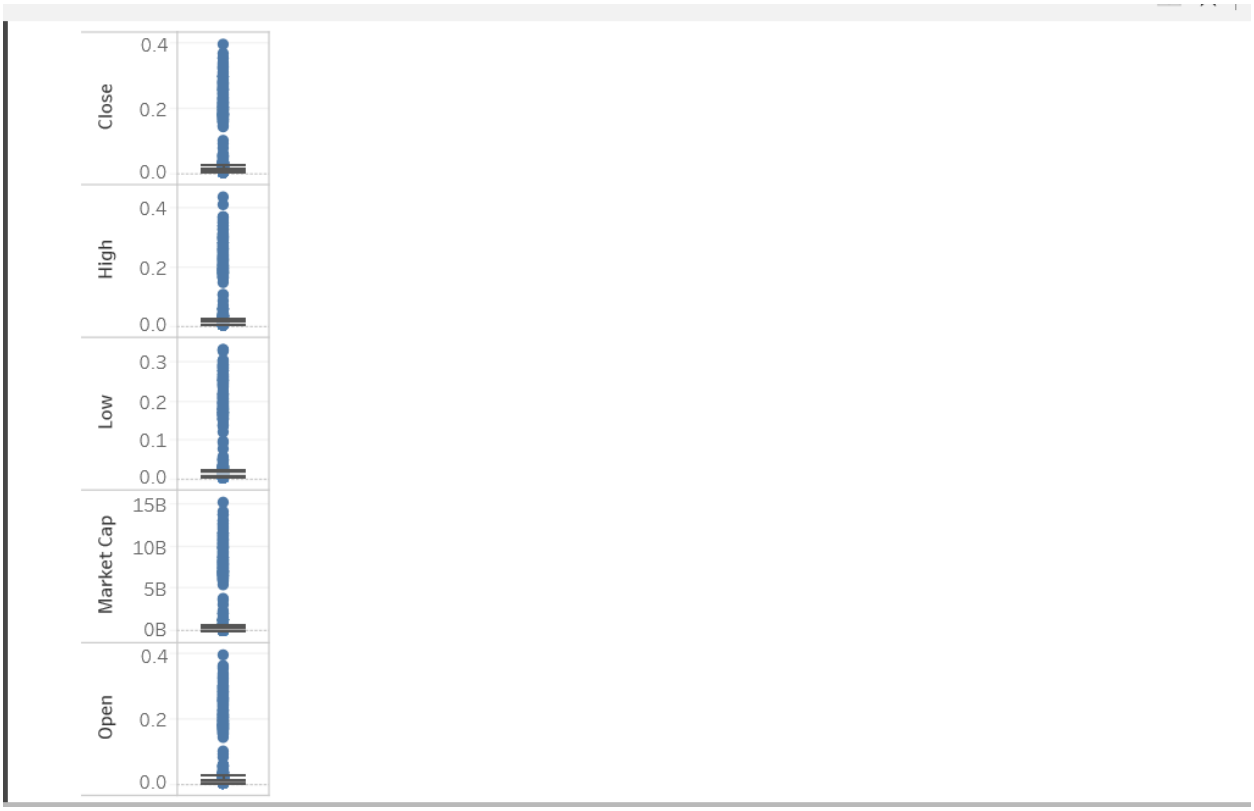
Close VS Market Cap

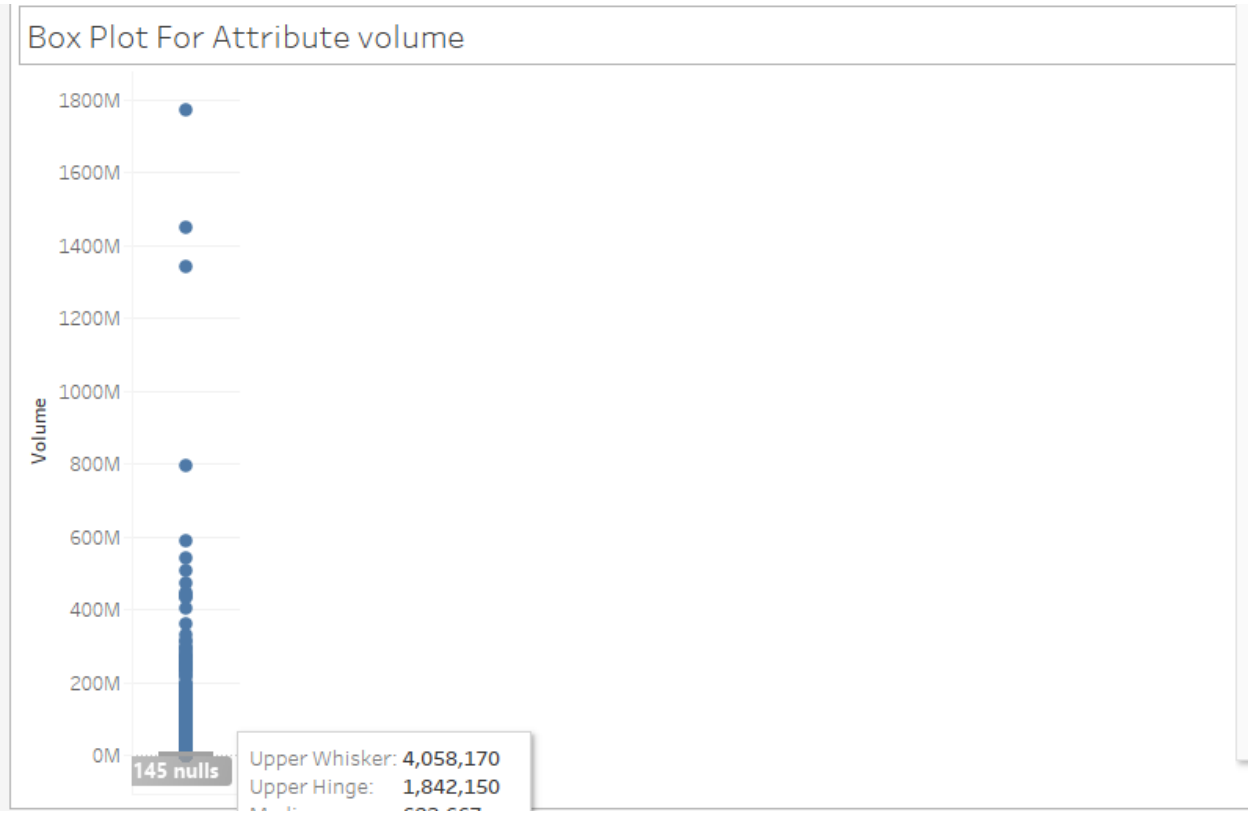


Observation : The attributes Close, High , Low, Open, Market cap follows a similar day trends but are not correlated. So, these attributes are not redundant and we need all these attributes for our analysis. All the attributes have a hike in 2017 .

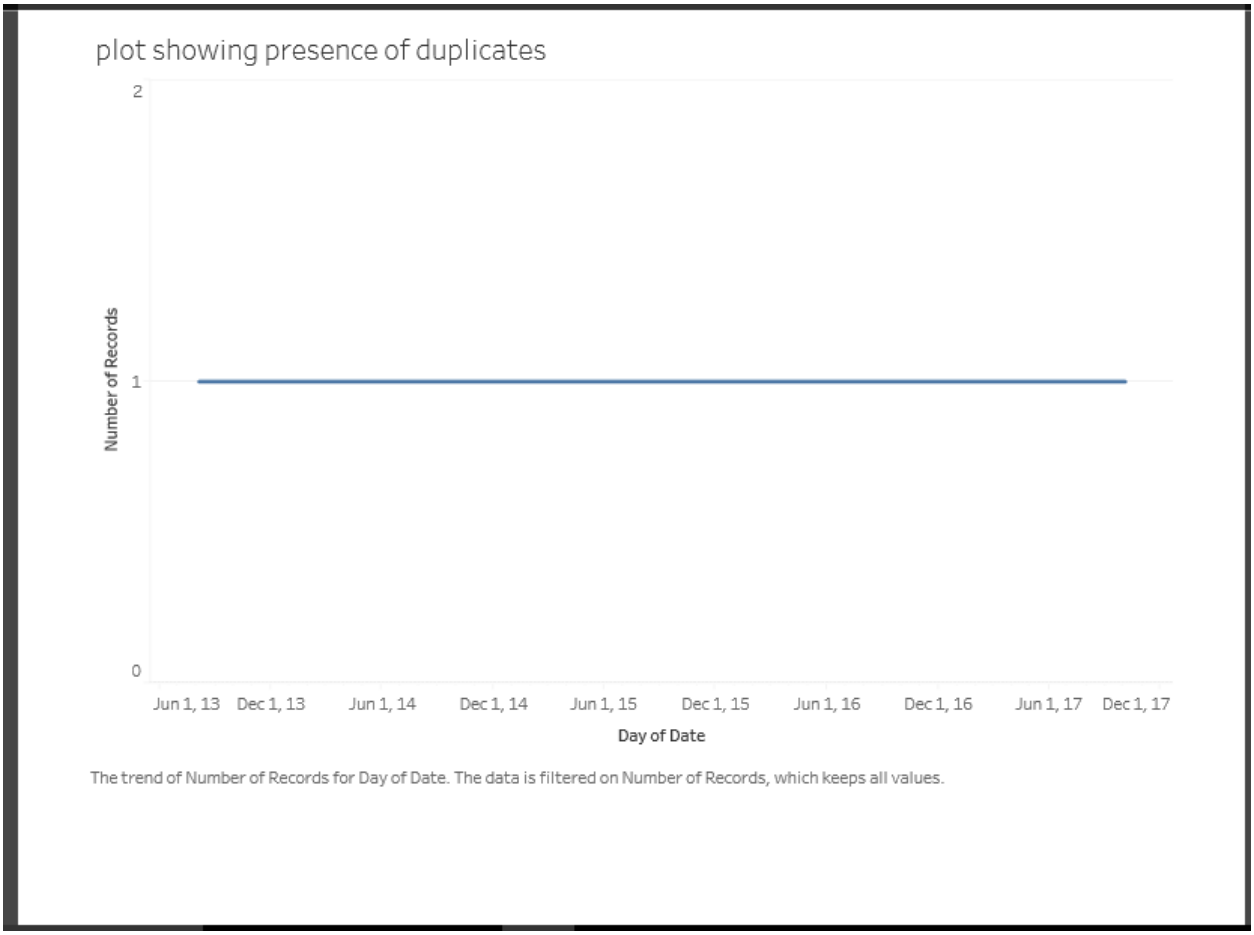
Data Quality For Ripple:

Box plots for Attributes (Open, High, Close, Low and Market Cap)


Close
0.4
0.2
0.0
High
0.4
0.2
0.0
Low
0.3
0.2
0.1
0.0
Market Cap
15B
10B
5B
0B
Open
0.4
0.2
0.0


Box Plot For Attribute volume
1800M
1600M
1400M
1200M
1000M
800M
600M
400M
200M
0M
Volume
145 nulls
Upper Whisker: 4,058,170
Upper Hinge: 1,842,150

There are 145 nulls in the attribute volume. There are no null values in other attributes.



There are no duplicate rows in the data.

The attribute ‘Date’ in the dataset is not in standard date format.

There is no inconsistency in the data because the data is collected from the single source.

There are no dummy values in the dataset.

There are no anamoly and structural issues.

There are no outliers.

<u>Data Preparation For Ripple:</u>

1)Converting the attribute ‘Date’ to standard date format “YYYY-MM-DD”.

R script:

ripple_dataset <- read.csv("ripple_price.csv") #loading the dataset

ripple_dataset$Date <- as.Date(as.character(ripple_dataset$Date), format = "%b %d, %y") #changing the attribute date to standard date format.

2) There are 145 null values in the attribute 'Volume' ranging from days 4-aug-2013 to 26-dec-2013. These nulls have occurred because cryptocurrency ripple had no transactions during these period. So, we have decided to replace these null values with zero.

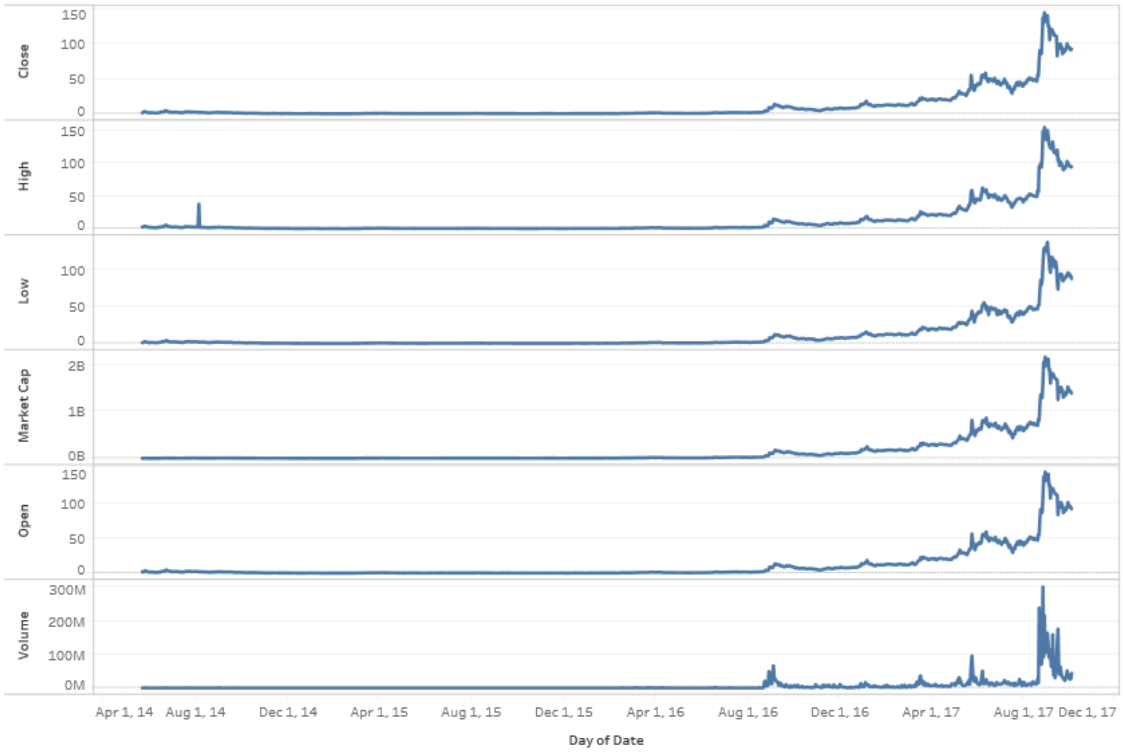
ripple_dataset <- read.csv("ripple_price.csv") #loading the dataset

levels(ripple_dataset$Volume) <- c(levels(ripple_dataset$Volume), 0) # increasing the levels

ripple_dataset$Volume[ripple_dataset$Volume == '-'] <- 0 #changing the '-' or null values to 0.

Data Understanding For Monero:

Plot showing the day trends for the attributes Open,Close,High,Low,Market Cap and Volume.

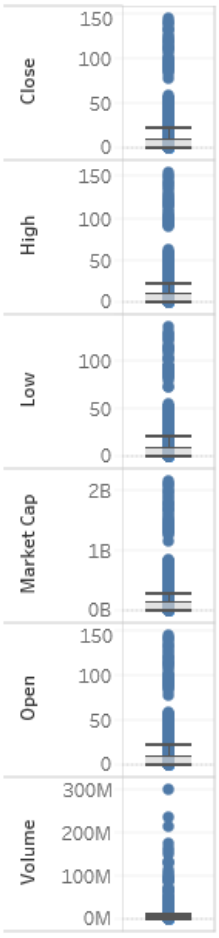


The trends of sum of Close, sum of High, sum of Low, sum of Market Cap, sum of Open and sum of Volume for Date Day.

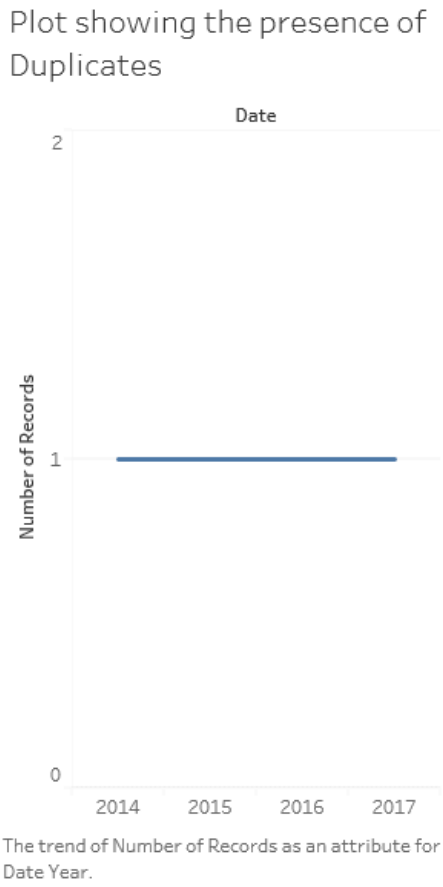
Observation : The attributes Close, High , Low, Open, Market cap follows a similar day trends but are not correlated. So, these attributes are not redundant and we need all these attributes for our analysis. All the attributes have a hike in 2017 .

Data Quality For Monero:

Box plots for Attributes (Open, High, Close, Low , Market Cap and Volume)



There are no null values in the attributes.



There are no duplicate rows in the data.

The attribute ‘Date’ in the dataset is not in standard date format.

There is no inconsistency in the data because the data is collected from the single source.

There are no dummy values in the dataset.

There are no anamoly and structural issues.

There are no outliers.

Data Preparation For Monero:

1)Converting the attribute ‘Date’ to standard date format “YYYY-MM-DD”.

R script:

monero_dataset <- read.csv("monero_price.csv") #loading the dataset

monero_dataset$Date <- as.Date(as.character(monero_dataset$Date), format = "%b %d, %y") #changing the attribute date to standard date format.