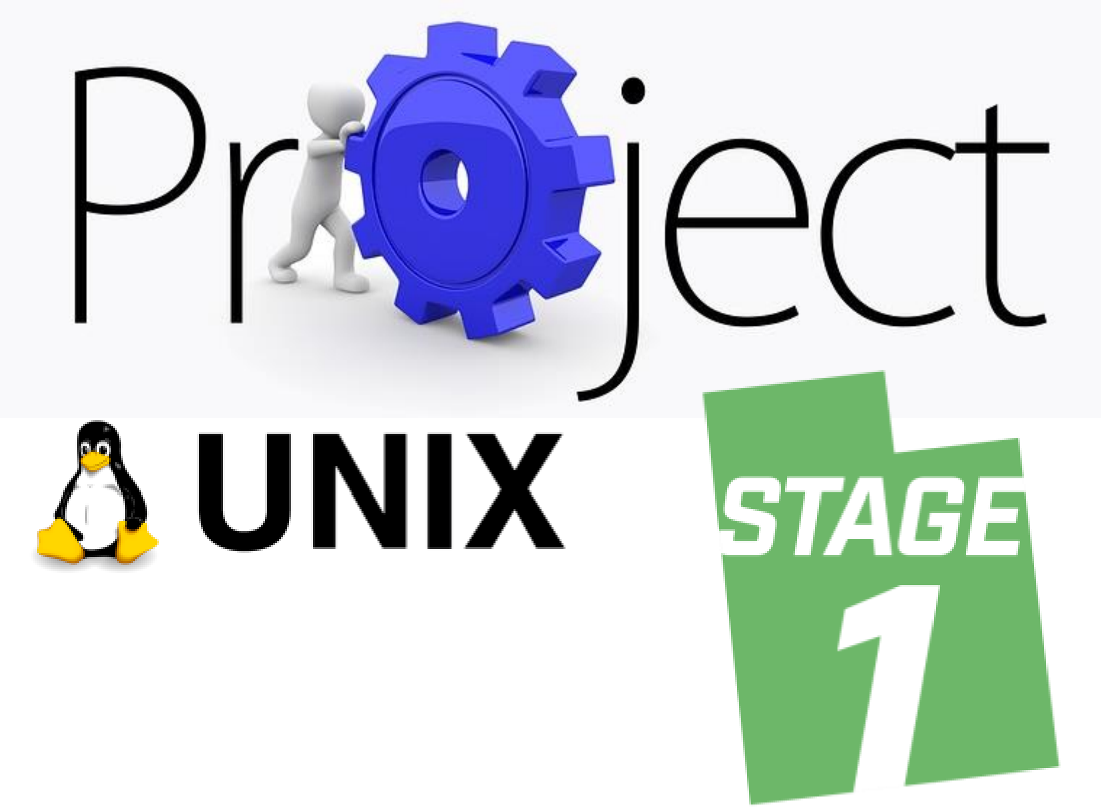
KARAN GODA



Project Stage 1

Data Finding, Cleaning and Loading

Prepared by: **Karan Goda**, Undergraduate Bachelors in Computer science and technology

SID: 460496371

Section 1

Introduction

My interest in cryptocurrency

Cryptocurrency has always fascinated me since I discovered it in 2014 when I noticed its rapid rise from a mere $100 to $1200 a couple weeks later. This had encouraged me to trade and speculate on cryptocurrencies as well as doing a little bit of forex trading. I was only 16 when I discovered Bitcoin. Since it was an unregulated market I could move in and out small funds that were liquid enough to not get squeezed by the market. I have done qualitative analysis such as technical analysis of the structure of the charts, however Unix gave me an opportunity to go even deeper in order to analyse Bitcoin thoroughly as I was able to create complex pipelines which allowed to me discover the dates a range Bitcoin had displayed, the highs and lows of bitcoin, the type of linear correlation between Bitcoin and other cryptocurrencies (however for this project, we will only consider Litecoin).

The data sources I have used

I have **used two datasets** to do a more complex and meaningful analysis as it allowed me to display objectivity and allowed me to do more complex pipelines that gave me some nice data by using some columns of each dataset in the calculation. The data sources are:

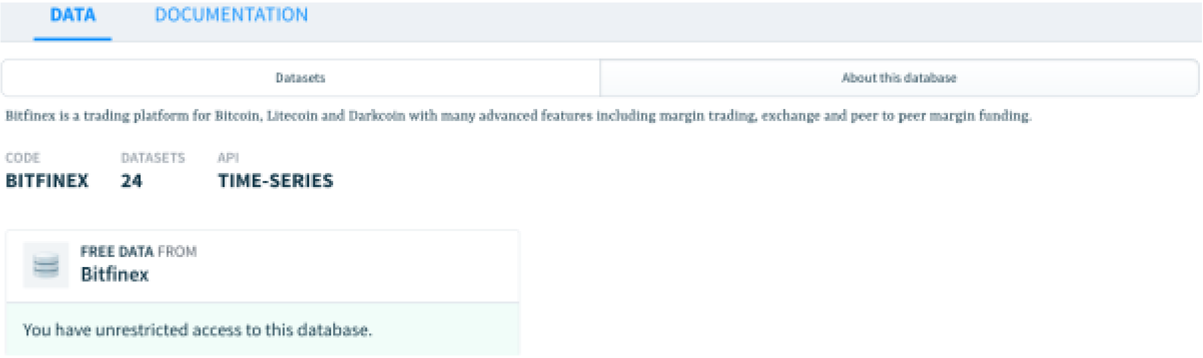
- http://bitcoincharts.com/charts/btcmarketsAUD
- https://www.quandl.com/collections/markets/bitcoin-data

Bitcoin charts has been a very amazing tool to use to obtain the BTCAUD rates because the data source had given me the power on what dates I can choose as well as the currency involved. Therefore, I chose the Australian dollar vs Bitcoin pair because it would make more since Australia conducts its dealings in the Australian dollar, and not the US Dollar. However the lack of cryptocurrency pairs was this website’s limitation as I wanted my dataset to be a theme on pairs involving BTC.

The other data source **Quandl** I had used was more limited in terms of Bitcoin pairs because it did not have the specific pairing (BTCAUD) I had chosen. However, Quandl was a collection of knowledge when it came to providing altcoin pairs in the cryptocurrency environment, especially the more important ones which are ETHUSD, ETHBTC, ETCBTC, LTCBTC, XMRBTC, XMRUSD, ZECBTC, DSHUSD, and DSHBTC. Out of these, I had chosen only LTC because the other currencies are newer in general and I was basing my research on the years 2014-2016 where altcoins like ZECBTC did not exist for example, or there may not be enough data to do reliable analysis on it. LTCBTC provided the only viable option.

Permissions to use this data

Both of these websites **have given me appropriate rights to use the data** as mentioned in their about pages.

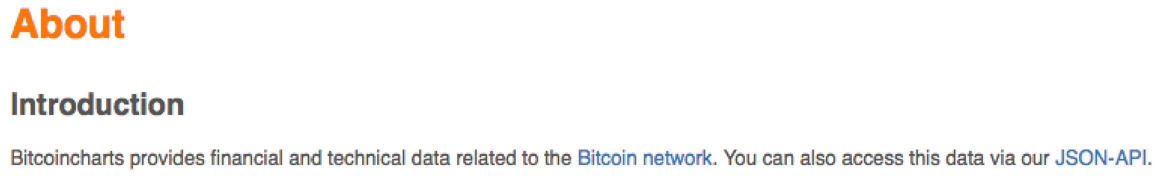


Bitfinex is a trading platform for Bitcoin, Litecoin and Darkcoin with many advanced features including margin trading, exchange and peer to peer margin funding.

CODE	DATASETS	API
BITFINEX	24	TIME-SERIES

FREE DATA FROM
Bitfinex

You have unrestricted access to this database.



About

Introduction

Bitcoincharts provides financial and technical data related to the Bitcoin network. You can also access this data via our JSON-API.

The contents of the data

One of my pipelines from the ‘cleaning_data.txt’ file (more information on other pipelines, cleaning and filtering will later be mentioned in the report) outputs a ‘metadata.txt’ file which contains the columns of the merged dataset file.

The columns are:

1. **Date:** The date is the common element between the cryptocurrency pairs LTCBTC, and BTCAUD. It is to track the prices of either currency at any given point in time between 2014-2016.
2. **LTCBTC:** LTCBTC shows the pair’s prices on the 1D (daily) timeframe between 2014-2016 in order to provide more accuracy in finding a day with a specific price point for example.
3. **Open:** The opening price for BTCAUD on the market
4. **High:** The high price in the range for BTCAUD on the market
5. **Low:** The low price in the range for BTCAUD on the market
6. **Close:** The closing price of the daily candle for BTCAUD on the market
7. **vol_btc:** The volume of BTCAUD in which the quantity of dollar flow is measure in terms of BTC
8. **vol_aud:** The volume of BTCAUD in which the quantity of dollar flow is measure in terms of AUD
9. **weighted_price:** The weighted price of BTCAUD by averaging the high, low and close with most weigh upon whichever had the highest volume (which would give it the highest importance).

Section 2

Transformation and cleaning

My cleaning process

The data was a bit unorganized and messy when I had obtained it from the previously mentioned sources, however with some effort in building pipelines in Unix to clean the data, it was foreseen that the data would display immense potential in obtaining the results that I needed for my research. This mean that I had to remove extra whitespaces, commas before and after the rows in the data, check for null or nonsensical values in the datasets, verifying whether the csv files were formatted in ASCII encoding or not, and many other processes in the Unix code which also has comments for easy understanding.

My cleaning was an automated process which was done purely using Unix pipelines. I did not feel the need to use other tools such as Python or SQL due to the efficiency, short length of code and speed of Unix.

My transformation of the file

I had also transformed the data in the sense that it was now much cleaner to read and two different datasets had been merged into a dataset with the common factor being the date. This was my most fulfilling moment during the assignment because I was joining two files for a more important purpose, and I could do much more analysis that involved both currencies after the joining of the files.

The code I had used to do the automated cleaning as well as transformation in Unix was:

```
#This deletes the 2 extra comments after the rows
cut -d',' -f'1,2,3,4,5,6,7,8' 'BTCAUDOrig.csv' |\
#Checking that the data did not have any empty lines
grep -v '^ $' |\
#Change the word 'Infinity' to 'N/A' to indicate unknown values
sed 's/Infinity/N\/A/g' |\
#Delete the line numbers where date does not match the LTCBTC pair
sed '2,228 d' |\
#Equalize the lines of LTCBTC and BTCAUD in order to prepare a merge
head -n+592 > btcaud.csv
#Merging the files
paste ltcbtc.csv btcaud.csv | cut -d',' -f'1,2,4,5,6,7,8,9,10,11' > mergeLBAUD.csv
#Checking that the file encoding is ASCII
file 'mergeLBAUD.csv'
#Checking the number of rows of the file
wc -l 'mergeLBAUD.csv'
#Checking the number of columns of the file
head -1 mergeLBAUD.csv | sed 's/[^,]//g' | wc -c
#Checks the header of the file to know different columns of the data
head -1 mergeLBAUD.csv
#Gets the header into a new file to know the columns of the data
head -1 mergeLBAUD.csv | tr ',' '\n' > metadata.txt
cat metadata.txt
#This process has now cleaned and verified that the data is valid
```

Section 3

Analysis conducted

My analysis and output

The analysis was originally planned to do in Python, however after again observing that Unix pipelines was much easier to the job in I just finished the analysis I wanted to do in Unix. My data was very enormous, however after the cleaning and transformation; there were only 592 lines left in my merged data file.

Since my data was quantitative in nature, and solely involved the exchange rates, volume and the date that the transactions took place; majority of my analysis was conducted with the sole purpose to find number values relating to movements in price or correlations between other pairs for example:

- The date of the maximum price of BTC
- The date where Litecoin and BTC were at their highest
- The value of BTC where Litecoin was an ‘x’ amount
- The value of Bitcoin at a specific date
- The median price of BTC
- The range of BTC where it was an ‘x’ amount up or down, and the number of days it was or the date where it was the highest for example. This was very interesting to do as most of the exchanges do not allow you to do such analysis after which you can zoom in the dates the price points are given.

```
#The number of lines in the file including the header
wc -l mergeLBAUD.csv
#The number of rows do not have a N/A value for BTC
tail -n+2 mergeLBAUD.csv | egrep -v 'N/A' | wc -l
#The highest weighted price of BTCAUD and the date it was in
cut -d',' -f'1,9' mergeLBAUD.csv | sort -t',' -k'2n' | tail -1
#The total number of rows where the opening price of BTCAUD is greater than $700
cut -d',' -f'3' mergeLBaud.csv | awk '$0 > 700 {print $0}' | wc -l
#Gets the date where LTC was worth 0.008BTC and the BTC high (4rth column)
#at the exchanges was over $315
grep '0.00800' mergeLBAUD.csv | awk -F',' '$4 > 315 {print $1}'
#Gets the largest 'close' value recorded for Bitcoin where
#the 'high' is greater than $350 and the 'low' is less than $300
cut -d',' -f'4,5,6' mergeLBAUD.csv | awk -F',' '$1 > 350 && $2 < 300 { print $3 }' |\
sort -n | tail -1
#The date where both BTCAUD's weighted_price and LTCBTC's were at their lowest together overall
cut -d',' -f'1,2,9' mergeLBAUD.csv | tail -n+2 |\
sort -t',' -k'3,3n' -k'2,2n' | head -1 | cut -d',' -f'1'
```