Scraping Coinbase for Crypto Data

Adam Gao

July 8, 2019

Summary:

Coinbase.com is a website displaying cryptocurrency prices of different cyrptoassets. Using the BeautifulSoup Python library, I scraped pricings of the 50 top cryptocurrencies (Bitcoin, Ethereum, etc) from Coinbase.

The data is converted from its HTML/JSON origins into a Python-friendly Pandas dataframe. The dataframes, which contain pricing information is saved as Excel files which can be re-opened by Python or Excel for easy visualization and analysis.

The python files are organized as follows:

- "Scrape-Coinbase.py" serves to scrape and save pricing data.
- "Read-Data.py" plots cryptocurrency pricing data.
- "Analyze-Data.py" analyzes cryptocurrency pricing data using statistical methods imported from Python libraries.

Data Format:

"Scrape-Coinbase.py" creates a folder containing 52 Excel files and "datetime.txt" containing the date and time at the end of the extraction.

Each Excel file's name corresponds to a cryptocurrency. In each file, there are 5 spreadsheets corresponding to different time durations of pricing. These durations are "hour", "day", "month", "year", and "all". I opted to use Excel instead of csv file formats to utilize multiple spreadsheets per file.

Each row contains the following data:

- Price, in the currency of the country from which the website was accessed
- Time, at which the price was taken. The units are always 1/1000 seconds and the times are equally spaced. The number itself gives no indication what the date/time of the price is.

The date and time (datetime in python) during which the data is taken is recorded to a txt file as well.

Analysis:
Each cryptocurrency's pricing data is analyzed for a value called an ADF statistic, (taken from the augmented Dickey-Fuller test).

The ADF statistic value describes whether the cryptocurrency's price over time is a random-walk (non-stationary) or not.
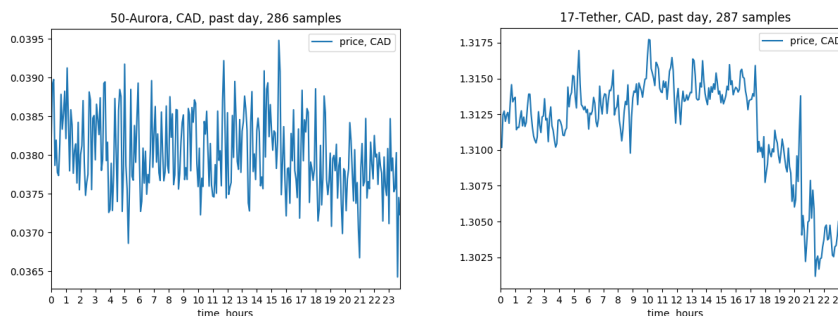

ADFstatistic, past day
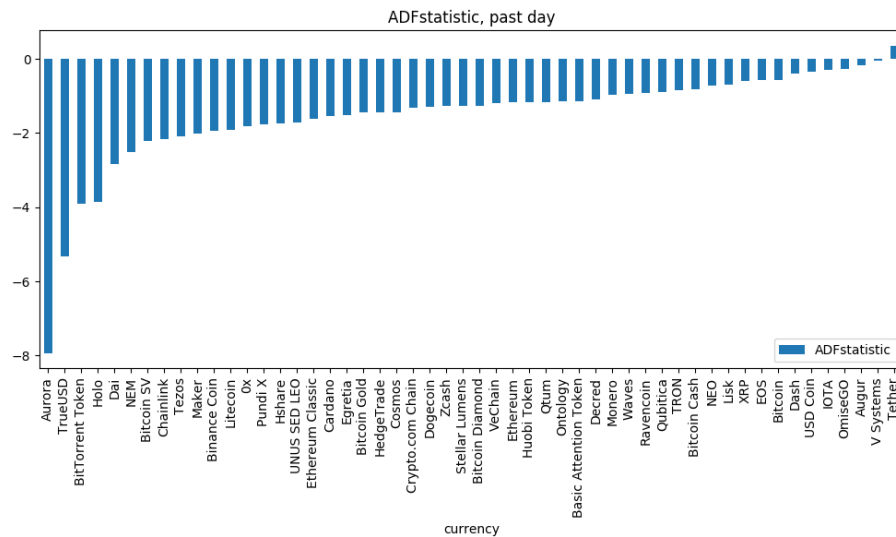
Above is a bar graph of the ADFstatistic for each cryptocurrency taken from the past day (July 4, 22:47:35), from most negative to most positive. A negative value of the ADFstatistic indicates that the model is non-stationary, i.e. the null hypothesis is rejected.

p-value, past day
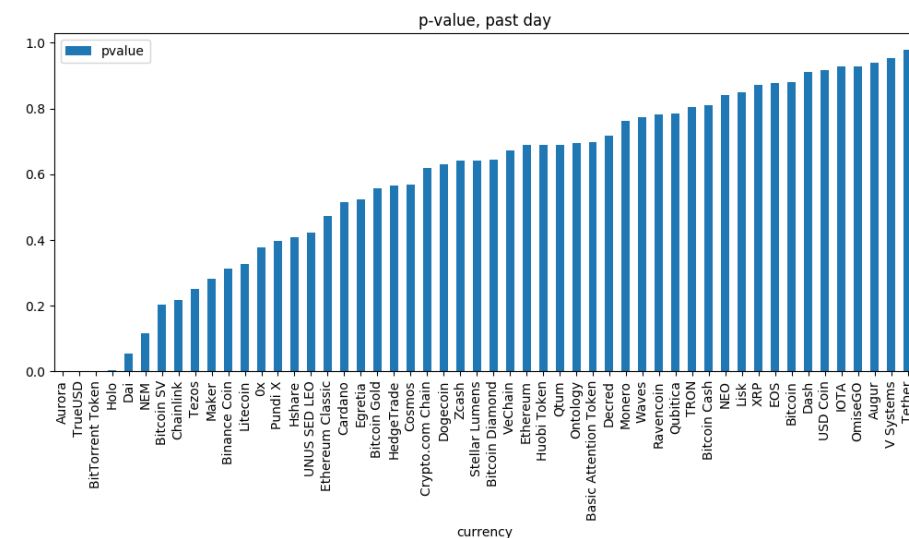
A small p-value (<0.05) indicates strong evidence against the null-hypothesis (the price is non-stationary).

Analysis pt. 2

To illustrate what the ATF statistic really means, we can examine the pricing data of the cryptocurrencies with the highest and lowest ATF statistic, Aurora and Tether:

On visual inspection, more positive ATF statistics are associated with more 'drifting' in the stock price, or some sort of trend unassociated with noise.

Note that this particular set of data was analyzed over the past day. The "stationarity" of the data can potentially change on a day to day basis. The stationarity of the above data is only accurate for the day preceding July 4 at 10:47 pm.

Furthermore, not all the listed cryptocurrencies have sufficient samplings for the chosen duration, (e.g. a cryptocurrency is 2 months old under an analysis over the past year.)

It is critical that the analyst takes note of the timespan and the datetime during which the data was taken.