**Group 3: Analysis of Bitcoin Blockchain First Draft**

**Group Members**

* Hangyu Kang (hkang98)
* Shengwen Yang (syang382)
* Bodi Yang (byang89)
* Fengxia Dong (fdong6)
* Sam Tauke (sstauke)

**Introduction:**

For the past decade, Bitcoin (BTC) has been the pre-eminent cryptocurrency. Unlike the public banking system we are used to, the recorded transactions in blocks are open to everyone and everyone shares the same transaction records once the transaction had been done between the users. It uses block-chain technology for security. Each block is a record of the transactions that have happened since the last block. And those blocks are connected in a chain, each block points to the blocks before and after it. Thus, we call the cryptographic technique the block-chain. A block is finished once a bitcoin miner has solved the mathematical equation linked to that block. The miner who solves the hash problem associated with a block is then allowed to push the verified list of transactions from the block to the blockchain and is rewarded with BTC as payment. Because the size of the block is dependent on how fast miners can solve the hash problem, the amount of time between new blocks varies. The BTC protocol aims for each block to take about 10 minutes and every 2016 blocks the protocol adjusts the problem difficulty to get closer to this 10-minute average.

Additionally, blocks are limited to at most 1 mb of transaction data. So, it is possible that the amount of transactions since the last block could exceed the amount that can fit into the current working block (in cases of large numbers of transactions or a long time to solve block). To solve this, individuals will pay fees to miners to ensure that their transactions are included in the next block, ensuring that that they don’t have to wait a long time before they can consider their transaction as valid.

**Goal of the research:**

1. What is the average number of transactions/transaction quantity/time elapsed per block?

2. How has the number of BTC transaction / total BTC transacted changed over time? Has it been strictly increasing or have there been dips?

3. Is there a relationship between the number of transactions and price? If daily price is high, is the number of transactions also high?

**Statistical computation:**

**Data cleaning**:

Our group intends to analyze data on the BTC ledger. Our data comes from a user on kaggle who scraped the BTC ledger to pull all transactions. The data is available https://www.kaggle.com/shiheyingzhe/bitcoin-transaction-data-from-2009-to-2018. The data is organized in a series of csv files that each has five columns: block height, input hash, output hash, sum of BTC transacted, and timestamp. The data are stored in 22 files with 19.72 GB in total. We also get the price of bitcoin online.

We cleaned the raw data with Linux Shell. We cleaned the raw data by removing the punctuation characters (quotation, parenthesize, etc.) to make it neat, and sorting the data (based on time).

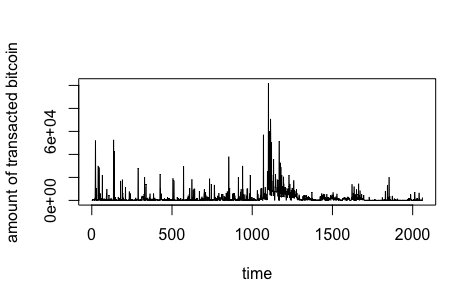
We conducted analysis with R.

First question: after the calculation, we got the average of categories below

Average number of transaction : 7.196507

Average number of transaction quantity : 2109.965

Average time elasped : 23002.9

Second question: we plotted how amount of transacted bitcoin change through time.

Third question: (tentative) we will use the linear regression to find the relationship between price and number of transaction.

Distributed computing:

We use CHTC clusters for distributed computing. .sub script is written to call process.sh script and the 22 data files; process.sh will run the R script to conduct computation for individual dataset.