Chenhui Zheng

Lingyu Li

**Bitcoin Comparisons Project**

**Project Background**

The purpose of this project is to compare bitcoin prices from 2014 to 2019 with other commodities such as Gold and Foreign exchange rates to identify if there is any correlation. Besides the prices of bitcoin from two exchanges, Bitstamp and Coinbase, our team has added a third dataset which includes Gold prices and exchange rate between U.S dollars and Euro. In order to set a consistent standard, all prices are denoted in U.S dollars.

**Project Steps:**

**Process 1: Extract Data**

To achieve our goal, we have gathered three different datasets (Bitstamp dataset, Coinbase dataset, and Gold dataset) from Kaggle. The sources of these datasets are listed below:

Dataset Sources:

* + Bitstamp & Coinbase
    - <https://www.kaggle.com/mczielinski/bitcoin-historical-data>
  + Gold & USD/EUR
    - <https://www.kaggle.com/altruistdelhite04/gold-price-data>

**Process 2: Transform Data**

**Step 1: Remove NaN**

Though both the Bistamp and Coinbase dataset are over 25 MB, much of the information inside are NaN, so the first step we did is to remove all the NaN cells with the code #pd.dropna() command.

**Step 2: Scoping**

To further clean up the bitcoin data, we have removed some information that are not crucial to our comparison projects, which include: open, high, low, volume and weighted price. Instead we narrow the dataset down to date and closing price of that day.

On the other hand, we have also narrowed down the Gold.csv data to only include Gold price and Euro currency exchange rate.

**Step 3: Date Modification and Drop Duplicated Columns**

Step 3 is the hardest part of this project because not only different dataset has different date range but also different data types and data format. Therefore, we have created #translate\_function and #translate\_function\_2 function to clean up the date time to "%Y-%m-%d” format. Moreover, we’ve only kept the first closing price of each day to make sure we don’t have duplicated values.

**Step 4: Merging the tables**

Since we are using the dates as primary key, once we have standardized the date across different tables, we use pandas to merge them into one dataset. Inner join was used to include only data points that exists in all three tables.

**Step 5: Rename Columns**

Due to the ambiguity of the naming we have named all columns in a consistent way.

**Process 3: Load Data**

In the final process, we have created a SQL database bitcoin\_comparison\_db as well as the table bitcoin\_comparisons. The final table demonstrated the fluctuation of bitcoin, gold, euro, in relation to U.S Dollars from 2014 to 2018. The reason why we chose PostgreSQL instead of MongoDB was because the final dataset only contained numbers, therefore, does not require non-relational database such as MongoDB.

**Conclusion**

This assignment was relatively easy in comparison to the previous one since we only need to Extract, Transform, and Load the dataset. The most difficult parts of this assignment were to change the unix timestamps to regular dates and to load the final dataset into the SQL database.