ETL Project Report

Members: Chris, Victor, Shannon

The Ask:

• The sources of data that you will extract from.

* We are extracting data from:
  + <https://www.kaggle.com/jessevent/all-crypto-currencies>
  + <https://www.kaggle.com/philmohun/cryptocurrency-financial-data>

• The type of transformation needed for this data (cleaning, joining, filtering, aggregating, etc).

* We found a csv file from Kaggle, which was formatted to correct US numbers.
  + Date column was inverted to date/time type
  + Data was filtered and cleaned
* We converted a csv file into a .json using Python in Jupyter notebook.
  + We parsed the data from the .json file and renamed the columns we need to match up with the columns of our csv file.
  + We needed to join the tables in MySQL.

• The type of final production database to load the data into (relational or non-relational).

* We used MySQL (relational database) to load the csv and .json files.

• The final tables or collections that will be used in the production database.

* Final collection to include "Currency", “System”, "Date\_open", "Open\_price", "High", "Low", "Close\_price", "Volume", "Market\_Cap", “Close\_Ratio”, “Ranknow”, and “Spread” as combined from our two datasets.

Our ETL Process:

Our first step was to find two datasets that showed a history of cryptocurrency. We decided to use two files from Kaggle which showed consolidated financial information.

\*\* Datasets Used:

1. A total history of Cryptocurrency prices, including opening and closing numbers, highs/lows, and market values.

<https://www.kaggle.com/jessevent/all-crypto-currencies>

2. This data was pulled from coinmarketcap.com to collect information on the trends of cryptocurrency with the intention of predicting future trends.

<https://www.kaggle.com/philmohun/cryptocurrency-financial-data>

ETL Project Report

Logic: We decided to use historical price information pulled from CoinMarketCap.com using an API request. By joining this dataset with the set showing the rank and spread, we expect to do exploratory analysis to answer the following questions for investors:

- What are the trends for the top 10 currencies?

- Which day of the week is the best time to buy/sell?

- Which currencies are the least volatile?

- How does the shape of the trendline open prices correlate with that of Bitcoin?

- Which models can we use for forecasting highs/lows?

The ETL Process

(E)xtract

Kaggle’s “Consolidated Cryptocurrency Financial Data” was used, which had 200 different new currencies with a CSV file including the open, close, high, low, volume, and market cap for years from 2013 to 2019. The information was scraped from Coin Market Cap ( <https://coinmarketcap.com/>). The other dataset was also a CSV file. It included the following variables: coin slug (the commonly used name), the system, , close ratio is the daily close rate, min-maxed with the high and low values for the day (Close Ratio = (Close-Low)/(High-Low)), and the spread, which is the $USD difference between the high and low values for the day.

(T)ransform

Table columns were dropped, renamed, and reformatted as necessary for ease of use. Both datasets were converted into two data frames using Pandas. Each dataset was also cleaned, uploaded to MySQL, and filtered according to currency, which was used as the primary key for all datasets.

(L)oad

We first attempted to use the import wizard in a relational database (MySQL), however it was ruled out as it took a long time to load and work with the two tables. Instead, we used Pandas to clean and load our data.

Production Database Final Tables

We made the final data table through a join on the primary key “Currency”. SQL Alchemy was used to import the data into MySQL, and we had to use this method for working with the combined dataset of more than 10,000 entries. For data visualization, we might select the top 10 ranked cryptocurrencies to analyze and look for forecasting opportunities.