**Project 2-ETL Process**

Group 9 Team Members:

Chad Dubiel, David Martinez, Katy Fuentes

Scope of Research:

Cryptocurrency pricing and Covid case counts.

Github Repo:

<https://github.com/cdubiel08/ETL-Project-Group-9>

Data Sources:

* Source
  + Covid-19 Dataset- <https://www.kaggle.com/imdevskp/corona-virus-report>
    - Full\_grouped.csv
  + Bitcoin Historical Data- <https://www.kaggle.com/mczielinski/bitcoin-historical-data>
    - bitstampUSD\_1-min\_data\_2012-01-01\_to\_2020-09-14.csv

Write Up Process Summary:

* ● What data sources you chose, and why?

Kaggle included a comprehensive world count of Covid cases and a bitcoin historical chart from 2012 through 2020.

* ● Detailing the process of the extraction, transformation, and loading steps.

The bitcoin historical data was read in Pandas and condensed to 2020 historical information only. This new file was used as a new data file and the data was updated with a revised column name of 'Unnamed: 0' to 'Block'. The data was further condensed to include only 'Block','Timestamp','Volume\_(BTC)', 'Volume\_(Currency)', and 'Weighted\_Price' columns. The 'Volume\_(Currency)', 'Weighted\_Price' columns were formatted to currency styles. In addition, the ‘Timestamp’ was transformed to extract only transactions at occurred at 12:00:00 (noon) to obtain a unique value for each day at that specific time since the Bitcoin data is by the second. Lastly, the ‘Timestamp’ was modified by dropping the time and only leaving YYYY-MM-DD format to match the Covid date information.

The Covid-19 dataset included various data sets and the “full\_grouped” and “day\_wise” was reviewed to determine which would be used. The “day\_wise” csv was selected because it was more comprehensive and provided a better overview. This file was read in Pandas and the columns were condensed to 'Date', 'Confirmed','Active', 'New cases', and 'No. of countries'.

● Explain why you have performed the types of transformation you did.

Data from the Covid and bitcoin historical data were merged on date for consistent formatting and integration.

* ● Why you chose the type of final database

Mongo was used since it efficiently skips null values related to missing date information in the two datasets. In addition, it would easily capture Covid/Bitcoin date overlaps information.

* ● Schema of the tables/collections in the final database

The “covidcrypto\_db” collection includes 257 records with 2020 daily bitcoin price and worldwide Covid information. Each document includes a id, date, bitcoin price, Covid counts for : world new case, world new deaths, world new recovered, world confirmed, world deaths, world recovered, world active, and countries which is an array of country information that includes name, confirmed, deaths, recovered, active, new cases, new deaths, and new recovered.

The Mongo “\_id” was “TypeError: Object of type ObjectID is not JSON serializable” so a loop through the data set removed the object id from Mongo with the dictionary method .pop(), which rendered the data in JSON format. Specific date queries also included a .pop() to remove the “\_id” to get specific date data sets.

* ● Hypothetical use case(s) for your database

Use the output and compare to markets, commodities, or US dollar to determine trends between various financial markets and Covid cases.