By Sam Black, Bashier Mazid, Andrew Park, and Kwasi Yeboah-Afihene

**Extract:**

Our team used two csv files from Kaggle to build a data set that will allow the user to analyze the changes in the value of bitcoin and the S&P 500 Index on a daily basis from January 2012 to November 2018. We read the files into jupyter notebook using pandas

\* \*\*T\*\*ransform: what data cleaning or transformation was required.

**Transform:**

1. Dropped all NaN rows from the bitcoin data set using .dropna() on the data frame
2. Converted the unix timestamp to a readable date/time so that we could compare time frames with the other data set. We then split out the date (excluding the time) to a separate column.
3. Removed all duplicates in the “Date Only” column, keeping only the last (closing) price for the day.
4. Removed unnecessary columns from the data set.
5. Renamed columns so the data is more legible.
6. Rounded S&P 500 prices to two decimal places
7. Reformatted date so that the format matches that of the bitcoin data (for example: 2012-01-03)
8. Dropped unnecessary columns and renamed some so that they would make more sense once the tables were merged (for example: Volume was changed to Volume(S&P\_500))
9. Merged the two data sets on the ‘Date’ column
10. Created two columns tracking the percent change from row to row (day to day) using the .pct\_change() function, changed the data types of those columns to float, and reformatted so that it displays as a percentage.
11. Renamed columns again to make the column names compatible with SQL (removed special characters including parentheses).

**Load:**

We connected to the mySQL database and used the .to\_sql function to push the data. (User will need to create database with their own password and create database to connect to). We then used the pd.read\_sql query function to verify the connection worked. We chose to put all of the data into one mySQL table so that the data set is ready to use and query as desired.