**Group 4 – ETL Project:**

***YouTube Video Trends – United States, Canada, and Great Britain***

**Step 1: Extract –** *your original data sources and how the data was formatted (CSV, JSON, pgAdmin 4, etc.)*

Our main and only data set was a subset of what seemed like clean google queries, which summarized key video performance stats for YouTube video entries. These statistics label the sum of videos as “trending” and allow an insight on which videos are performing better in contrast to others.

The data set was found via Kaggle, [here](https://www.kaggle.com/datasnaek/youtube-new?select=CAvideos.csv). We considered this data set moderately reliable due to its site popularity and frequency of use and additional notebook content. The data set is separated between country CSV files which include multiple columns of performance metrics for each unique video. Along with the CSV files congregated by country division, there are also corresponding category codes via JSON files.

We decided to utilize the main 3 English-speaking countries for a more expedited clean up process. Individually we downloaded each of the three CSV’s and three corresponding JSON files for further joins, as the premise of our data was to review video performance in conjunction with category descriptions.

**Step 2: Transform -** *what data cleaning or transformation was required.*

* 1. Read in initial CSV sheets for each applicable country
     + My initial data cleaning started on the single US country sheet. It was during these first transformations we ran into the importance of object type as there were a few columns that although they appeared as datetime or another string, they were not and would require object conversions
  2. Read in corresponding category JSON files for each country
  3. Do a reflective drop of columns in CSV to ensure a cleaner sheet
  4. Loop through JSON layers to identify the data desired – our category JSON had a subset of a dictionary within a dictionary. To overcome this we needed to loop through the dictionaries of the JSON and then append to a new list
  5. Using the newly created list we populated a new data frame in that we could merge onto our existing country table
  6. We ran into issues with merging and attempted to use the concat function – this however created errors and populated NaN values. We eventually realized that the merge was being obstructed by an object type conflict and we adjusted accordingly
  7. Once a category merge was done, we needed to assign a unique “country” code to each sheet prior to combining them into a master sheet
  8. Once a new column value was created we were able to merge to a master sheet, and from here create “new” sheets
  9. One sheet that I used the master to populate was the category sheet – within this sheet I populated new fields of interest such as “total\_feedback” which is just a way to summarize the multiple columns in prior sheets
  10. Some additional transformation I was working on a lot was datetime and found this to be the most problematic. I ended up a regular expressions conversion for the “trending\_date” field, but was still unable to find the final solution for having pandas recognize the out of order format as a date time format

All cleaning was done within multiple jupyter notebooks. I found it simpler to utilize multiple books throughout the creation of multiple tables – aside from the master sheet. This also really struck the importance of notating and the order of syntax within each of my notebooks.

**Step 3: Load -** *the final database, tables/collections, and why this was chosen.*

Along with familiarity, we found it best to utilize PgAdmin and a Postgres database because our data overall seemed “relational” more than anything. Between measuring relations between the individual country data, our understanding of the project was to represent the data in other usable ways – such as the category table for summary data or the video performance data. Some questions posed by the data that was collected and cleaned could be – how does one channel, or user’s videos perform against others? Which category of videos has the most views? When are certain categories of videos posted during the day?