**Overview**

**US Data and merging US-MX-CA datas and clean up**

* USvideos\_df is csv file US\_category\_df is json file, retrieved from kaggle.com,
* A regular store of the JSON data into a DataFrame resulted in missing data as the JSON was not fully traversed. As a result an alternate approach was used to parse out those non-traversed data elements into list and applied to a dataframe,
* In order to merge the csv and JSON dataframes on category id an integer conversion was made on the id\_list in the JSON dataframe. This prevented issues/errors with nt64 and object columns when attempting the merge. The dataframes were "left" joined on the csv dataframe,
* USvideos\_df has category ID, US\_category\_df file has ID columns,
* "for" loop used to retrieve the ID from jsons dictionary, and created etag, title, id list,
* ID column from US\_category\_df renamed as category id to be used for merge files,
* "combined\_USdata" is created as new dataframe,
* Added in a country code of 'US' to the combined dataframe to distinguish a country of origin
* title\_x and title\_y columns renamed as video title and category\_title, data frame renamed as renamed\_combined\_USdata
* title\_x and title\_y columns renamed as video title and category\_title, data frame renamed as renamed\_combined\_MXdata
* title\_x and title\_y columns renamed as video title and category\_title, data frame renamed as renamed\_combined\_CAdata
* new renamed\_combined\_USdata data, renamed\_combined\_MXdata, renamed\_combined\_CAdata frames saved as csv file in the resource folder.
* Created new notebook to merge all csv files, used concat function to merge 3 csv files.
* New merged\_all data has 122281 rows × 18 columns named ‘video\_id', 'trending\_date', 'video\_title', 'channel\_title', 'category\_id', 'tags', 'views', 'likes', 'dislikes', 'comment\_count', 'comments\_disabled', 'ratings\_disabled', ’video\_error\_or\_removed', 'kind', 'etag', 'items', 'category\_title', 'country'.
* Saved this merged\_all dataframe as combined.csv file for team to use for SQL and ERD processes.