ETL Project – Meghana Bhimasani, Gina Kim, Andres Gomez-Potdevin

**Extract**: We focused on two datasets from Kaggle – US Baby Names (used StateNames.csv source) and NYC Dog Names (used dogNames2.csv).

**Transform**: Using pandas and jupyter notebook, we read the csv files into pandas data frames. We cleaned the data by filtering by year (the state names csv file had far greater range of years compared to the dog names csv file) and dropped duplicate values. We decided to transform the state names data by selecting data from years 2011 and onward as well as data exclusive to state of New York (as the dog names dataset was specific to New York). Following this, we dropped the “Year” column and kept just “Name”,”Gender”, “State”, & “Count” columns. Finally, we merged the two datasets by name and count columns. This allowed us to count the overlapping names among the two datasets.

**Load**: We made the sql connection and loaded the final database onto MySQL workbench as new\_df. This can then be used in the future to update counts of shared names for dogs and humans.