ETL Project: Popularity of Baby Names 2014

Data Sources:

* <https://www.ssa.gov/oact/babynames/limits.html>
  + Txt files from the National Level
  + Txt files from the State Level (NY)
* <https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf>
  + CSV file for the city of New York

Data Cleanup:

* We converted the txt files from the Social Security data to csv files.
* We only pulled the year 2014 from the New York State File.
* The NYC csv needed duplicate rows to be dropped, an entire column dropped (Ethnicity), and similar rows combined by count. This was done using SQL.
  + Since each name and year was broken down by Ethnicity, we needed to create a total count of each name within the year without the separator, Ethnicity.

A screenshot of a cell phone

Description automatically generated

* After reading in the three csv files to Python, we changed the column headers to make sure they were the same across all files.
* The Rank column was added into each of the data frames in Python because they were not originally a part of the CSV files

Combining Data:

* Within Python we combined the three csv files to create a single Data Frame by doing a merge with the column headers; Name and Gender.
* Each name in the data has a Count and Rank so when combining the data, we needed to ensure that the Count and Rank were presented for each level of data; National, State and City.

Final Data Frame Includes:

* Using the csv created from Python the final Data Frame was uploaded to a data base in PGAdmin.

A screenshot of a cell phone

Description automatically generated

Analysis:

* Based on count and rank;
  + Which names have the most popularity between:
    - New York City
    - New York state
    - National Level