|  |  |
| --- | --- |
| **ETL TECHNICAL REPORT** |  |

# Team Members: Christin Whitlock, Isha Iqbal February 13th, 2020

**EXTRACT**

The data sources utilized in our ETL project were found on Kaggle.com. Originally both datasets were from the census bureau and provide international health and population metrics of countries from 2014 to now, and projections through 2050. For our purposes, we chose the following two CSV files:

**DATA SOURCES**

1. Age Specific Fertility Rates
   1. Format: CSV File
   2. Link: <https://www.kaggle.com/census/international-data#age_specific_fertility_rates.csv>
2. Mortality Life Expectancy
   1. Format: CSV File
   2. Link: <https://www.kaggle.com/census/international-data#mortality_life_expectancy.csv>

**TRANSFORM**

Similar processes were used for the transformation of both CSV files. The steps were as follows:

1. CSV file was loaded onto jupyter notebook
2. Dataframe was created from the CSV file with pandas
3. Filtered dataframe was created through removal of columns
   1. We also filtered out rows eliminating all other countries besides the United States
4. Column headers in dataframe were renamed
5. Connected to PostgreSQL

**LOAD**

A relational database (PostgreSQL), was chosen to store the data we transformed in Python into the database US\_Census. SQL Alchemy was utilized in order to create the engine to connect jupyter notebook to PostgreSQL for the final creation of our tables: fertility and life expectancy. The final database can be used to look at projections of fertility rates and life expectancy of US from 2014 to 2050.

Through joining the two tables within the database on the primary key, year, we can easily compare the two sets of data to see if, for example, a low fertility rate would correlate with a lower life expectancy within that specific year.