**Date:** 01/15/2019

**Members:** Marlon Schieber, Jasmine and Dawit Hailu

**ETL Project:** Fertility Rate and Life Expectancy at Birth

**Data Source:** Kaggle/World Bank Data

**Extract:**

* Reviewed **Kaggle** to find a dataset for our ETL Project and decided to work on the World Bank Data on *“Countries Population, Fertility Rate and Life Expectancy”*
* **Two .csv files** were downloaded from: <https://www.kaggle.com/gemartin/world-bank-data-1960-to-2016>
* **Files:** 
  + fertility\_rate.csv (265 x 61)
  + life.expectancy.csv (265 x 61)
* **Goal:**
  + The database will be used to find out the ***relationship between fertility rate and life expectancy at birth*** worldwide in the last six decades.
  + **Fertility Rate** is defined as the total number of children that would be born to each woman
  + **Life expectancy** at birth is defined as how long, on average, a newborn can expect to live, if current death rates do not change.
* Data, the two .csv files, **extracted** into Pandas data frame for transformation.

**Transformation**:

* The entire two .csv data for 265 countries (1960-2018) extracted to dataframe for cleaning and transforming;
* Three irrelevant columns for our ETL Project ('Country Code', 'Indicator Name', 'Indicator Code') dropped;
* There is missing data for more than 30 countries, which has been a challenge to aggregate the data as necessary…
* Transposing the dataframe for further aggregation were a challenge as well as resetting the index by country;
* The data would be organized and analyzed in MySQL by decades to show the relationship between fertility rate and life expectancy;

**Load:**

|  |  |
| --- | --- |
|  | * The entire dataframe loaded to SQL; * connection string set “root:......@localhost/world\_etl\_db”; * engine created using (f'mysql+pymysql://{connection\_string}') |