Michael Hissey

ETL Project

11/6/19

**Data Source:** Education Statistics from World Bank Open Data

* Found via Kaggle – link: <https://www.kaggle.com/theworldbank/education-statistics/download>

**Process:** Using Education Data from Kaggle, I was able to collect data surrounding educational attainment levels among differing demographics across multiple countries. However, the data surrounding those education statistics was separated among three different CSVs – EdStatsCountry.csv, EdStatsCountry-Series.csv, and EdStatsSeries.csv. This presented a challenge in mapping educational data back to the original country, as data about the countries was located in EdStatsCountry.csv (including country code), while data about attainment levels was located in EdStatsSeries.csv. This was solved by using EdStatsCountry-Series.csv, which mapped each country code (which was associated with country data) to a series code (which was associated with attainment and demographic data), and thus served as a key that could connect the two other CSVs.

Using the ETL process, I extracted data from each CSV, eliminated extraneous columns that weren’t useful using Pandas in Jupyter Notebook. I then transformed them, again using Pandas in Jupyter Notebook, to have a standard naming format for each column that would properly match up to the schema and databases I had established in PostgreSQL. Then, after establishing a connection between my PostgreSQL server and my Jupyter Notebook, I loaded the data from the CSVs into PostgreSQL, and performed an inner join operation to connect the three CSVs, using series code and country code as the columns by which the join could be performed. This resulted in a final table which mapped education data, such as a country’s income group, to demographic data, such as annual population growth rate.