ETL Project

By: Aliona Reselian Mizelle and Gina Nelson

Date: January 30, 2020

**EXTRACT:**

We collected population (total) and gdp (current US$) data from Worldbank.org. We were looking for changes in population initially and chose to concentrate on Italy and Japan as there was a noticeable decline in the two countries population in the recent years. We wanted to see how does a declining population affect a country’s gdp and the significance of that affect. We decided to limit the years of analysis to 2008-2018 decade.

**Population, total**

Total population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship. The values shown are midyear estimates.

**GDP (current US$)**

GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars. Dollar figures for GDP are converted from domestic currencies using single year official exchange rates. For a few countries where the official exchange rate does not reflect the rate effectively applied to actual foreign exchange transactions, an alternative conversion factor is used.

TRANSFORM:

We downloaded two separate datasets from the World Bank website comprising annual population per country and annual GDP per country between the periods of 1960 to 2019. We are fortunate to have found databases that were relatively clean and required minimum of cleaning. However, we decided to not include our analysis on the year 2019 due to missing population information of many countries on the list. To drop 2019 column, we used Excel but we used Python-Pandas to help us drop the data that we aren't using in our analysis starting from 1960 till the year 2007.

After importing both data sets to our jupyter notebooks, we changed our data frame to our chosen years. One of the challenges we encountered while transforming the data was extracting the two countries we chose to use from the data frame. One of us chose to filter the data to include only Italy and Japan and overcame the challenge by using the .isin function which checks to see if the values are in the series.

On the other hand, another chose to keep the country data intact in the dataframe and filter the countries on the database level. Both methods yielded the same output in the database. Another challenge was transforming the numbers in the data sets to a format that sql would accept. Initially, we used thousand separators in the .csv file to overcome the numbers outputting as scientific numbers. However, when we tried to load it into the database, we ran into a double precision error thus we had to change the numbers back to a number and simply rounded to whole numbers to eliminate the decimals. This transformation led to the dataframe successfully loading into the database.

We chose to use a relational database, postgres, to store our information. Postgres provides an efficient way to store the tables and the data in it as well as conveniently connects Pandas to SQL. This allows to manipulate data either in SQL or Pandas. Our data was organized neatly in columns and rows and using a relational database, which allows it to maintain that structure. We also were not building a complicated data model, therefore the use of a relational database was appropriate.