**DBZ ELT Project**

**Submitted 3/20/21**

**DBZ Team (Jason-2):**

* Sy
* Akadia
* Catherine
* Tristan
* Billye

**Github repository link:** <https://github.com/kaekacha/DBZ>

**Background:**

News outlets report that the birth rate is declining in high income countries, including South Korea. Some explanations for the cause may be linked to economic variables, such as high housing costs, low disposable income, lack of a college education, and debt. The purpose of our project is to create a SQL database that will contain the data points necessary to analyze how fertility rates in the country of South Korea are associated with multiple economic variables.

*Research Question:* How is the declining birth rate in South Korea related to economic factors (namely, household disposable income, nominal housing prices, having a college degree, and having housing debt.

**Extraction:**

The data is obtained from different sections of the OECD website ([https://data.oecd.org/](https://data.oecd.org/economy.htm)). The goal of our ELT project is to clean the datasets for the different factors, merge them, and convert them to a SQL database. Each of the datasets was exported as a CSV file and converted to a dataframe using the Python Pandas library.

**Transformation:**

Each dataset was filtered to the country of Korea. Only the following columns were kept: location, indicator, subject, measure, time, and value (dropping the frequency and flagged codes columns). The indicator, subject, and measure variables are codes for the economic indicators. The variables that will link the multiple datasets to each other will be Time (which is the year of the data). The data was filtered to years 1995 through 2018 (except for debt which was filtered from years 2008 to 2018). Each dataset was converted to a CSV file.

**Loading:**

The CSV files were then converted to dataframes, and the concatenate function in the Python Pandas library was used to append the dataframes together, and the index was reset. Next a table was created in pgAdmin with the same column names and sequence as the appended CSV file, and then the appended data was imported to the table. Below is a picture of our database in pgAdmin.

