**Project # 2**

**Group 6**

**Monica Wang**

**Shahab Araghinejad**

1. **Data**

Two sets of data were used to practice an ETL (Extract, Transform, Load) process.

* World Life Expectancy Indicator in country-wide scale

(<https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete?select=WHOregionLifeExpectancyAtBirth.csv>)

* World Birth Rate in country-wide scale

(<https://www.kaggle.com/utkarshxy/who-worldhealth-statistics-2020-complete?select=adolescentBirthRate.csv> )

Both datasets cover the most recent and updated health statistics of the world (countries recognized by WHO) in csv formats.

Major data fields for the above datasets are country, year, gender, and indicators but in different formats that needed to be cleaned up before using.

1. **Summary of the ETL process**
   1. **Extract Data**

Data were selected from Kaggle database. All data were in csv format so the selected data were simply downloaded from the database. No specific challenges existed to extract the data.

* 1. **Transform**

To transform data, a process a cleaning up were carried out. There were countries with multiple/different time periods. Only data of the years when both life expectancy and birth rate existed were kept and used to create the new database. VBA were used to clean up the csv files.

Files had similar titles of “tooltip”. It was renamed and organized in a way to make the recognition of data origins better.

* 1. **Load**

The cleaned-up/transformed data were loaded to the PostgreSQL using Python. The data frames were exported as csv files. They were merged (through Python) to create a unified database.

1. **Conclusion**

One of the findings by looking at the new merged database was that countries with the highest life expectancies tend to have a lowest birthrate and vise versa. The countries with the highest birthrate (200 per 1000 woman) are associated with life expectancy of about 50 years.

Republic of Korea has the has the lowest birth rate of all countries.