**ETL Project Report**

By Eliot Chern, Sean Galloway, Cora Micsunescu, Peter Vlahos

**Extraction:**

Extracted data from five different CSV files using the pandas read\_csv function.

**Transformation:**

Used the following style points:

* All column names should be lower case and snake case
* "country" is the index for each table/data frame

The wealth transformation involved the following steps:

* The original data had many columns, consolidated down to "country," "year," "income\_per\_person" to save space in the database.
* The original data had records for many years' worth of wealth data; decided to use the latest year, which was 2016.
* The column names were succinct and followed the snake case style agreed upon, so nothing needed changing.
* For a gross data cleanup, we used drop\_duplicates and dropna.
* Since this is a country-specific database, set "country" to the index.

The COVID mortality rate transformation involved the following steps:

* The original data had many columns, consolidated down to "Country," "Deaths," "Mortality Ratio" to save space in the database.
* The column names had spaces, capital letters and are extended. Decided to do the following name changes:
  + "Country": "country",
  + "Deaths": "deaths",
  + "Mortality Ratio": "mortality\_rate"
* For a gross data cleanup, we used drop\_duplicates and dropna.
* Since this is a country-specific database, set "country" to the index.

The adolescent and infant datasets from the World Health Organization involved the following steps:

* I chose to use Kaggle as this has a variety of datasets that have already been formatted and prepared for data visualization.
* Initially I wanted to use the maternal mortality ratio dataset but since this dataset represents ratios and the infant mortality rate dataset represents a rate, I decided to use the adolescent birth rate so as to compare like with like. In researching rates and ratios, I found that rates and ratios are different measurements with a rate being a comparison of numbers with different units and a ratio being a comparison of two numbers with the same units.
* The datasets did not require much cleaning or formatting as they had already been cleaned up by the creator of the dataset.
* The column titled “Indicator” had to be removed from both datasets as it was the definition of what was being measured repeated for every row. In the case of the Infant Mortality dataset, I also decided to take out the column titled “Dim1”, which contained the sex of the infant. I decided to do this to make the table simpler and easier to analyze. I also had to parse the numbers in the “First Tooltip” so as to only print out the whole number and not a range of numbers.
* The column titles were confusing and did not reflect the data being represented so they needed to be renamed:
  + “Location”: “country”
  + “Period”: “year”
  + “First Tooltip”: “birth\_rate” and “mortality\_rate”
* I chose to select the range of 2016-2019 in order to coincide with the date range of the other datasets from the team.
* Even though I set the index to country in the pandas dataframe, I needed to add an additional primary key column in postgres sql in order to allow the database to upload duplicate country values. This did not affect the subsequent joins and views as the tables could still be joined by country.

**Loading:**

Created the base "etl-project\_db" database in PostgreSQL. The database is empty at this point. Ran the schema.sql file in PostgreSQL to create the tables. In the Jupyter Notebook, use sqlalchemy to create an engine and connect to the database. Used the pandas to\_sql to load the cleaned data into the PostgreSQL database. Now that the data is in the database used a sequence of increasingly complex SELECT and JOIN commands to test the data.

**Sources:**

**Wealth Distribution**

<https://www.kaggle.com/psterk/income-inequality?select=combined_final_last_10_years.csv>

This analysis focuses on income inequality as measured by the Gini Index\* and its association with economic metrics such as GDP per capita, investments as a % of GDP, and tax revenue as a % of GDP. One political metric, EIU democracy index, is also included. The data is for years 2006 - 2016

**COVID Mortality by Country**

<https://www.kaggle.com/paultimothymooney/coronavirus-covid19-mortality-rate-by-country?select=global_covid19_mortality_rates.csv>

The 2019–20 coronavirus pandemic is an ongoing pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Coronavirus COVID-19 confirmed cases, deaths, case mortality ratios, country, latitude, and longitude.

**Adolescent Birth Rate by Country**

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

This is a complete dataset from the World Health Organization from the years 1950-2019. This dataset contains adolescent birth rate per 1000 women aged 15-19 years by country and year.

**Infant Mortality Rate by Country**

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

This is a complete dataset from the World Health Organization from the years 1950-2019. This dataset contains the infant mortality rate, which is probability of dying between birth and age 1 per 1000 live births. The dataset breaks down by male infants, female infants, and both sexes.