**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 COVID global mobility change rate transformation involved the following steps:

* The original data had many columns, consolidated down to "country\_region," "date", "retail\_and\_recreation\_percent\_change\_from\_baseline," "grocery\_and\_pharmacy\_percent\_change\_from\_baseline ", "parks\_percent\_change\_from\_baseline", "transit\_stations\_percent\_change\_from\_baseline", "workplaces\_percent\_change\_from\_baseline", and "residential\_percent\_change\_from\_baseline" to save space in the database.
* Filter the rows to reflect the latest data "2021-01-05" only.
* For a gross data cleanup, we used drop\_duplicates and dropna.
* Since we are only using the data from "2021-01-05", I dropped "date" column.
* The column names are lengthy and may not necessarily reflect the data we want to present. Decided to do the following name changes:
  + "country\_region\_code": "country",
  + " retail\_and\_recreation\_percent\_change\_from\_baseline ": "retail\_and\_recreation",
  + " grocery\_and\_pharmacy\_percent\_change\_from\_baseline": "grocery\_and\_pharmacy",
  + " parks\_percent\_change\_from\_baseline ": "parks",
  + " transit\_stations\_percent\_change\_from\_baseline ": "transit\_stations",
  + " workplaces\_percent\_change\_from\_baseline ": "workplaces",
  + " residential\_percent\_change\_from\_baseline ": "residential"
* Since this is a country-specific database, set "country" to the index.

The adolescent and infant dataset transformation involved the following steps:

* I chose to use Kaggle as it had a variety of datasets that were already prepared for data visualization.
* Initially, I selected the maternal mortality ratio dataset along with the infant mortality rate dataset but as the rates and ratios are different measurements, I decided to select the adolescent birth rate dataset instead so as to compare like with like.
* I did not have to do a lot of data cleaning as the dataset had no missing values or out of place values.
* The column titles were confusing and had to be renamed:
  + “Location”: “country”
  + “Period”: “year”
  + “First Tooltip”: “birth\_rate” for adolescent dataset and “mortality\_rate” for infant dataset
* Both datasets had a column titled “Indicator” that had the definition of the dataset repeated for each row of data. This column was eliminated. In addition, the infant mortality dataset had a column titled “Dim1” which recorded the sex of each infant mortality rate. I decided to eliminate this column to make it simpler and easier to upload into sql.
* My datasets covered the years 1950-2019. I only included data from 2016-2019 so as to match the date ranges of the other datasets.

The Internet individual use as a percentage of population transformation involved the following steps:

* The original data had many columns, consolidated down to "Country," "Country Code," and the years 2015 to 2019. 2020 was not available within this dataset.
* Drop first 4 rows that contained no or header information.
* Following predetermined naming rules:
  + The column names had spaces, capital letters and are extended.
  + This data set kept and changed column names:
    - "Data Source": "country",
    - “World Development Indicators": “country\_code”
    - "Unnamed: 2": "%\_pop\_use\_internet"
    - “Unnamed: 59”: “2015”
    - “Unnamed: 60”: “2016”
    - “Unnamed: 60”: “2017”
    - “Unnamed: 61”: “2018”
    - “Unnamed: 62”: “2019”
  + Keeped all “null” or “zer0’ information for future analysis.
* Creating years 2015-2019 as column names in the pgAdmin tables, required double quotes.
* To Select in pgAdmin query, the year selected had double quotes ie i.”2019”

**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 and display 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.

**Infant Mortality Rate by Country**

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

Comprehensive dataset from the World Health Organization that contains the infant mortality rate, which is the probability of dying between birth and age 1 per 1000 live births. Dataset covers 1950-2019 and is broken out by country and year.

**Adolescent Birth Rate by Country**

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

Comprehensive dataset from the World Health Organization that contains the adolescent birth rate, which is the birth rate per 1000 women aged 15-19 years. Dataset covers 1950-2018 and is broken out by country and year.

**Covid-19 Community Mobility Reports (Global)**

[https://www.gstatic.com/covid19/mobility/Global\_Mobility\_Report.csv](https://www.kaggle.com/paultimothymooney/coronavirus-covid19-mortality-rate-by-country?select=global_covid19_mortality_rates.csv)

The community mobility reports show movement trends by region, across different categories of places such as, recreation, parks, workplaces, and residential. The reports log the change in percentage daily against a baseline. The baseline is made of days in the recent period, before widespread disruption as communities responded to Covid-19, which is the median value from the 5-week period Jan 3 – Feb 6, 2020.