**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.

**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 inequailty 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 polical 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.

**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.