Crazy 8 ETL Project

Project Overview

For this project, we wanted to analyze how COVID-19 affected happiness around the world. To accomplish this, we pulled data from the World Happiness Report website and some free international census data from The World Bank to assemble a database that could be used for future analysis on this topic.

Extract

- We extracted our data from the World Happiness Report website: https://worldhappiness.report/ed/2021/#appendices-and-data

- On this website, we were able to download World Happiness Report data and data corresponding to COVID-19 related deaths.

- We downloaded our data as CSV files from this website.

- The international census data was extracted from The World Bank (https://data.worldbank.org/indicator/SP.POP.TOTL). The file format is a CSV file.

Transform

To transform our data into a usable format, we cleaned up the downloaded CSV files to get what we needed, and did away with what we did not need.

WHR Data

The World happiness report data was complete without any unusable data for this report, so we kept it as is.

Covid Data

The Mortality report provided by the WHR website provided an excess of data, so we dropped the columns that were not usable and re-saved the condensed CSV file.

International Census Data

The International Census Data was imported from The World Bank as a CSV file.

Load

Technology and Technique Choices

For our final database, we decided to use PostgresSQL. The flexibility and reliability of this database would allow for greater stability for future analysis over our second choice, the SQLite database.

To upload to the database, a SQL script was used. Following are the steps taken to assemble the final database tables.

Step 1:

- We focused on the World Happiness Report Data first, by importing the 2 separate CSVs from 2008-2019 and 2021. We then merged them into a master table which spanned from 2008-2021. We kept all data points provided in the CSV. We then ordered the master table by descending year.

Step 2:

- Our next step was to load the COVID-19 data. In the script we renamed a few columns to shorthand that could be used for easier Unions later on.

Step 3:

- The population data was loaded in last from the internationalcensus.csv File.

Step 4: Merging Master Tables

- The WHR and Covid data was merged using an outer join on the country name and year.

- The population data was merged with the WHR_COVID data using a left outer join on country name and year. This completes the join process and gives access to master table.

- Temporary tables were dropped and the data is uploaded to the database!