Group 4 - Ashwini Jha, Divya Palasamudram, Zoe Brentzel

Homework 5B - Project Part 1

Github: https://github.com/zpenguin19/Intro-to-Data-Science/tree/main

About the Data

Dataset 1 - Global Emissions

- The data is from *Our World In Data*, and it has been aggregated and wrangled by the CORGIS Dataset Project
- © 2023 CORGIS Datasets Project. Project by Austin Cory Bart, Dennis Kafura, Clifford A. Shaffer, Javier Tibau, Luke Gusukuma, Eli Tilevich.
- The data has been collected, aggregated, and documented by Hannah Ritchie, Max Roser, Edouard Mathieu, Bobbie Macdonald and Pablo Rosado
- https://corgis-edu.github.io/corgis/csv/global_emissions/
- According to the project's GitHub page, "Our complete CO_2 and Greenhouse Gas Emissions dataset… includes data on CO_2 emissions (annual, per capita, cumulative and consumption-based), other greenhouse gases, energy mix, and other relevant metrics."
- There are many datasets provided by *Our World In Data* that are all linked to the GitHub page: https://github.com/owid/co2-data/blob/master/README.md#our-source-data-and-code

Dataset 2 - Global Development

- The data is from the World Bank and includes information from 1980 - 2013
- © 2023 CORGIS Datasets Project. Project by Austin Cory Bart, Dennis Kafura, Clifford A. Shaffer, Javier Tibau, Luke Gusukuma, Eli Tilevich
- https://corgis-edu.github.io/corgis/csv/global_development/
- The link under "Overview" on the page above is incorrect, it should be https://fdc.nal.usda.gov
- According to the "The following data contains records collected on different countries and geographic locations from 1980 - 2013 from the World Bank. Included is different data about urban development, agriculture and rural development, health, and infrastructure."
- The data seems to have been taken from the National Agricultural Library. It also sources from the World Bank.

Cleaning and Wrangling

- Data-Wrangling is reshaping and transforming the data to make analysis easier
- Our data was already very organized to begin with. Our data also looked to be very clean, with very few missing values, so we used a command provided by Dr. AM to 'dirty' the data before we could clean it.
- We cleaned the data by first checking which columns had null values (3 columns for Development, and 2 for Emissions), and removing all rows that contain these null values.
- Once we merged the data, we removed the "Data.Health.Total Population" column, as it contained the same information as the "Country.Population". We also removed both 'Year' columns and both 'Country' columns, as we had added a new column to contain both the country and the year for a given row.

Data Merging

- One challenge we had was that we didn't have a particular ID for each row in our dataset. Each row had a country and year listed. These were two separate columns, but every entry had a different combination of country and year, so we combined the two columns under the standardized format `Country_YYYY` to make merging easier.
- After we combined the columns, we merged the datasets horizontally by the `Country_YYYY` column. The code is in the Python notebook in the Homework 5 folder in the GitHub repository.

Contributions

Ashwini: Found the datasets, wrote code to 'dirty' and clean the dataset, final edits, submission
Divya: Took meeting minutes, wrote code to clean and merge the dataset
Zoe: Formatted this document, created, shared, and organized the GitHub repository, recorded steps taken to clean, wrangle, and merge, and uploaded datasets and added data dictionaries to GitHub Wiki