Project explanation

Data Sources

In this project we took two datasets found on Kaggle.com. Both files contained data from measurements taken for the year 2019.

One was a CSV file containing a listing of countries and their score based on a measured of economic freedom. This score is determined by the US-based Heritage Foundation. The URL is as following: https://www.kaggle.com/lewisduncan93/the-economic-freedom-index. The whole file had the following columns with data: CountryID, Country Name, WEBNAME, Region, World Rank, Region Rank, 2019 Score, Property Rights, Judical Effectiveness, Government Integrity, Tax Burden, Gov't Spending, Fiscal Health, Business Freedom, Labor Freedom, Monetary Freedom, Trade Freedom, Investment Freedom, Financial Freedom, Tariff Rate (%), Income Tax Rate (%), Corporate Tax Rate (%), Tax Burden % of GDP, Gov't Expenditure % of GDP , Country, Population (Millions), "GDP (Billions, PPP)", GDP Growth Rate (%), 5 Year GDP Growth Rate (%), GDP per Capita (PPP), Unemployment (%), Inflation (%), FDI Inflow (Millions), Public Debt (% of GDP).

The second CSV we used was a measurement of happiness surveyed by The World Happiness Report. The URL is: https://www.kaggle.com/unsdsn/world-happiness. This file contained the columns: Country Name, the Happiness Score, GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust in the Government, Corruption.

Purpose of data

Ultimately, we would use this data to study the correlation between factors that define economic freedom, and factors that define happiness. We would therefore not only include the actual scores of economic freedom and happiness, but also some of the factors that were measured and used to calculate them by the respective institutions that do these studies.

Steps (Extract, Transform, Load)

1. First we downloaded the CSV and opened them. Here we already decided to take out several columns which we felt where not as relevant to goal. In the economic freedom dataset, we took out CountryID, WEBNAME, Region, Region Score, Country. These were not categories which were factors used to calculate economic freedom. In addition, we kept the Country Name category, which would be used to correlate the data with the other dataset.

In the other CSV, we took out the rank, as such an indicator was not contained by the economic freedom csv. For the rest, we kept the Country, for obvious reasons, and all the other columns as they represented factors used to calculate happiness.

In both cases, we renamed some of the columns (variables) to make them easier to transcribe to Postgres and easier to understand ourselves.

2. We then opened PostgresSQL and designed the schemes for both csv files. In both cases, defined all the numeric variables as Floats (having multiple decimals), and the strings as VARCHAR.

3. We opened Jupyter Notebook, and loaded both CSV files into Dataframes.

4. Then in the same Jupyter Notebook page, we made a connection to Postgres using a string.

5. We loaded each of the Dataframes with the function to sql.

6. In Postgres, we did an inner join with the Happiness table of the Economic table on Country Name. This helped assure that we were left with the countries which did have all the data. We also turned all the numeric variables to numeric, and converted the decimals to two for each numeric variable. Postgres also helps us verify that the data has loaded successfully from Python.

Result

Now we have a relational database which contains the country names, and numerous variables which we could deploy in data analysis and visualization.