PROJECT REPORT - GLOBAL QUALITY OF LIFE INDICATORS

TUESDAY, 05 MAY 2020

REPORT BY:

Gresa Murati, Sue Maltz, Mia Curtopelle

Executive Summary:

Our objective is to obtain datasets referencing global country information. Using these datasets, we are able to design queries to compare and measure the quality of life for countries around the world.

ETL Process:

EXTRACTION

We extracted our first dataset from World Bank at this url https://databank.worldbank.org/source/world-development-indicators, where we were also able to format the data as desired. Our second dataset was extracted from the Climate API of the World Bank by installing “wbpy” API library and following the documentation at the following link: https://pypi.org/project/wbpy/. Our third and fourth datasets were extracted from Kaggle at the following urls: https://www.kaggle.com/unsdsn/world-happiness,https://www.kaggle.com/fernandol/countries-of-the-world.

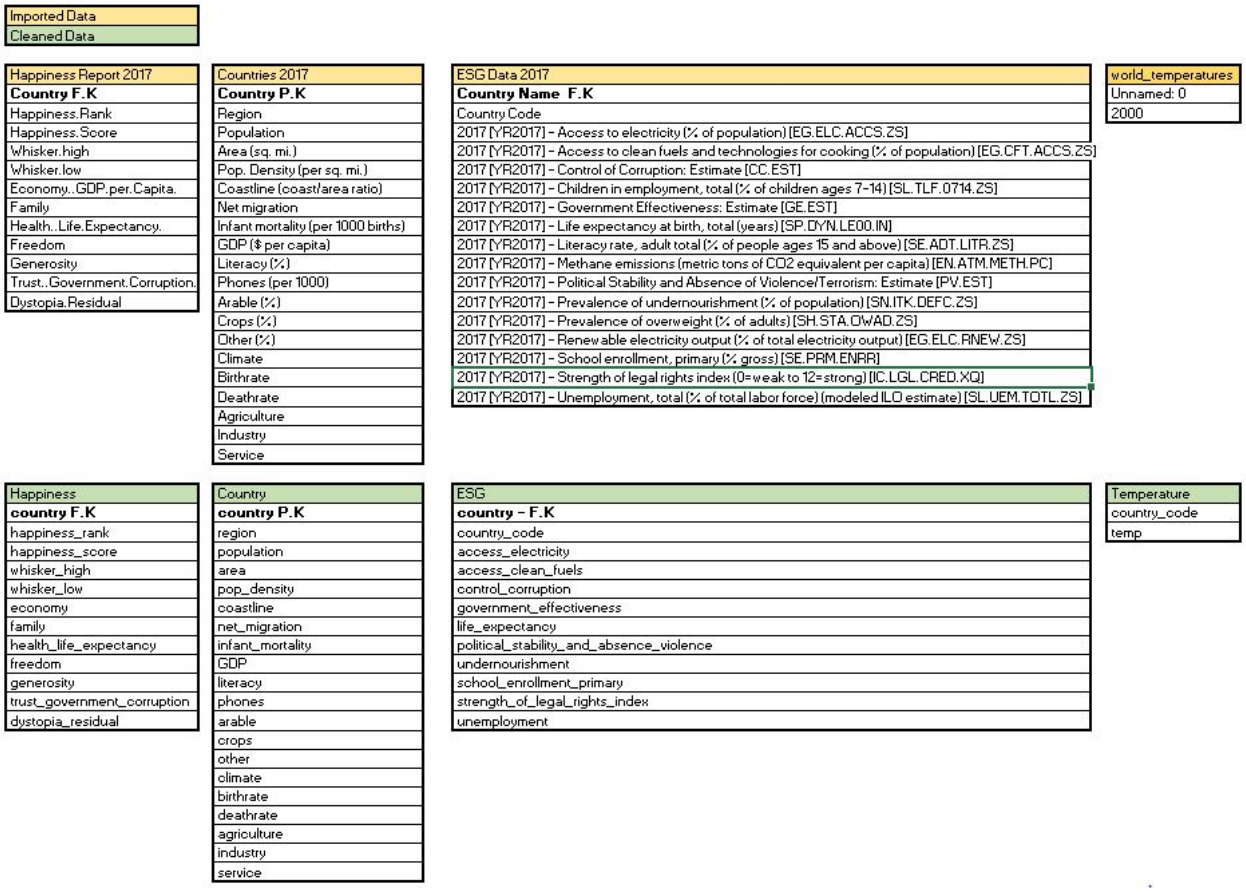
TRANSFORMATION

We began our data transformation process by loading our datasets into Python and using Pandas to clean and merge our datasets.

We used four different datasets, which were imported into pandas as csv files. The first step was to inspect the data by looking at the columns and checking

for missing values. Then we dropped unnecessary columns or columns containing empty values. Our next step was to rename the columns for better legibility.

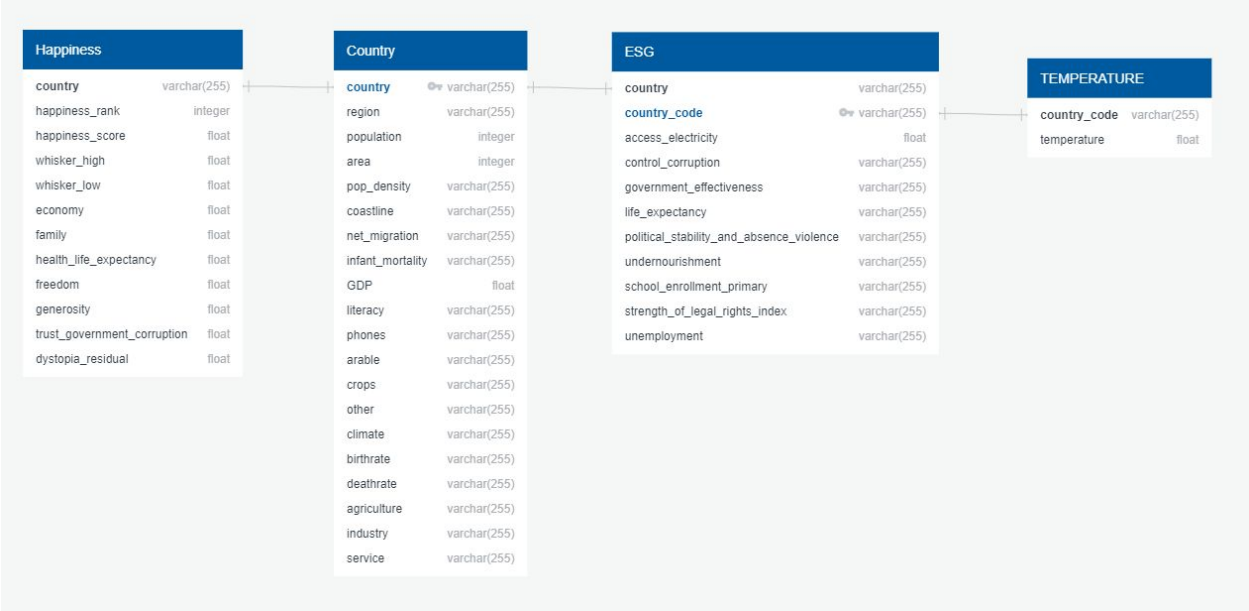
The table schema below shows the imported data on top and the cleaned data below.



To show the connection between the datasets and their primary and foreign keys, we created an ER diagram.

ER DIAGRAM

The four datasets are connected by the primary keys ‘country’ and ‘country_code’. We used quickdatabasediagrams.com to create it.



TRANSFORMATION

After having transformed the data, we used SQLAlchemy to create and load our datasets into Postgres, as we are working with relational datasets.