ETL Project

September 3, 2020

Molly Oesterling, Mary Whistler, & Chris Wycott

Extract

We chose to examine a 2019 World Happiness Report dataset and a World Environmental Variables dataset. Both of these datasets were CSVs extracted from Kaggle.

We chose these datasets because we were interested in creating a dataset that allows users to examine the effects that various environmental variables may have on countries' happiness.

The world happiness dataset included included the following fields:

- Overall rank
- Country
- Happiness score
- GDP per capita
- Social support
- Healthy life expectancy
- Freedom to make life choices
- Generosity
- Perceptions of corruption

Meanwhile, the environmental variables dataset included the following fields:

- Country
- Accessibility to cities
- Elevation
- Aspect
- Slope
- Cropland cover
- Tree canopy cover
- Isothermality
- Various rainfall statistics
- Various temperature statistics
- Wind
- Cloudiness

Transform

I. Step 1: From each dataset, choose what variables we would like to examine.
II. Step 2: Load the datasets into Jupyter notebook
III. Step 3: Delete the undesired data columns
IV. Step 4: Rename some of the columns so that they would be compatible with SQL tables

V. Step 5: Delete all rows containing NA values from the datasets
VI. Step 6: Create new DataFrames with the cleaned datasets
VII. Step 7: Set the index for both DataFrames.
VIII. Step 8: Created table schemata to hold the data in PgAdmin.

Load

I. Step 1: Create database connection within Jupyter notebooks to PGAdmin
II. Step 2: Confirm table connections in Jupyter
III. Step 3: Load DataFrames into database
IV. Step 4: Re-run the databases in PgAdmin and confirm.
V. Step 5: Confirmed data successfully loaded into two tables named environment, and happiness.
VI. SQL made the most sense for what we were trying to accomplish, i.e., import csv data into tables which we could then query in order to see if there is a relationship between happiness and environmental factors.