Extract

Obtained a file from Kaggle.com on ***World Happiness Report*** and the United States International Census ***Age Specific Fertility***. Data for the World Happiness had a range of years from 2015 to 2017 by country. Data for age specific fertility had a wider range of years by country. We decided to focus our data on the years of 2015-2017 by country. The data from the United States International Census on ***Age Specific Fertility*** was downloadable through google cloud platform's public dataset using the BigQuery API. This data was then downloaded to a csv.

Transform

We began by cleansing the data from the ***World Happiness Report.*** This data was provided in 3 different csv files. Each file represented a different year. Upon inspection of the data we found that some columns had data that were not in the other years' csv's and their titles were not similar but the data content were the same. Using Pandas, we renamed the columns so that all the column names would be aligned. We added a year column to the 3 csv files to prepare the files to be appended into 1 file. We kept similar columns from the 3 files and dropped columns that did not exist in each year of data. We then appended the data for the 3 years of the Happiness report into 1 csv. The final product of the ***World Happiness Report*** included data in 1 csv from the years of 2015-2017.

The ***Age Specific Fertility*** data was formatted by the country name column in order to align it with the country data column in the ***World Happiness Report*** dataset. We then filtered using .loc for years 2015-2017.

At this point we have 2 csv that have been viewed and cleansed in Pandas. These two files were then merged using an outer join on Country and Year and put into a DataFrame. We used the dropna to drop any row of data that did not have all data satisfied in both files. We were left with one file with 447 rows of data. We performed calculations on the data for the Average fertility rate across all ages.

Load

Our created DataFrame was now ready to be loaded into our database. We created an engine to connect to Postgres. Our data was successfully loaded into Postgres.

We chose this data because we came across data about Fertility by country on Kaggle.com. We thought it would be nice to know if there are any factors in the dataset where we can draw a conclusion regarding if you reside in a specific country are more or less fertile. The fertility data also had it broken down by age group which means we would be able to know by age group the countries where you are likely to conceive at a younger or older age. We then searched for an additional dataset that would complement the Fertility dataset. We came across the World Happiness data and began inspecting the columns to determine if there was any way we could join the data to get meaningful information. The data in the Happiness report had information that would tell a great story regarding Fertility and the Happiness score of the country. The Happiness data had additional information regarding Economy (GDP per Capita), Health (Life Expectancy), Trust (Government Corruption), etc by country and year. This data will provide information for trending by country on our topic of interest; Does the Happiness Score of a Country affect fertility?