Project 2 – Pounds

ETL Project - Final Report

**Summary:** For this project I pulled two data sets from https://data.world/. These two sets were formatted as CSVs and I saved them in a resource folder within my directory. A sample of the CSV in excel is shown below. The first data set was the smoking prevalence as a percent per country. The second was the life expectance in years per country. I was hoping to find a correlation between the two sets by pulling into one table both pieces of data from two different years and seeing if smoking trending down correlated to life expectancy trending up.

A picture containing photo, wooden, room, hanging

Description automatically generated

**Extract:** To extract this data, I imported the two CSVs as data frames in my jupyter notebook using the code below.

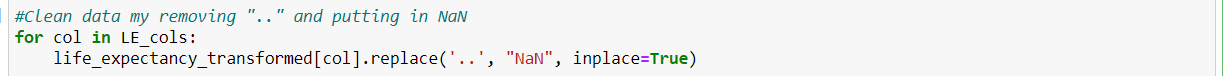
A screenshot of a cell phone

Description automatically generated

**Transform:** To clean and transform this data I did three distinct steps. The first step was to put the data into Third Normal Form. Both data sets had both a country code and country name column. I create a master country code and country name data frame and when I created the transformed dataframes, I only pulled in the country code. The second step was creating the transformed tables and pulling in only the columns I wanted. For both tables, this included the country code and the years. Lastly, the data set utilized two periods (“..”) for cells with no data and the columns needed to be renames for clarity and simplicity. I used the rename fun used one loop each and replaced all instances of two periods with the string “NaN”. I encountered issues with this step early on when I tried to load my data in the SQL server Postgres. I had set the column data type as REAL and when I initially ran the code with a frequently used “NULL” string in place of the two periods, I got an error. Looking into this issue showed me that the string “NaN” is the acceptable null string for the Postgres REAL variable. Below are some snippets of the code I wrote to transform the data.

A screenshot of a cell phone

Description automatically generated



**Load:** Once the data was extracted and loaded, I made a connection to the Postgres server, and loaded the data. To do this I created a schema.sql document to create my tables. See below for code snippets.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Running the schema above initialized my table and when I ran the code below, the data was added to the SQL server.

A screenshot of a social media post

Description automatically generated

The first table selects the smoking prevalence and the life expectancy for the USA vs the world in the year 2000.

A screenshot of a social media post

Description automatically generated

The first table selects the smoking prevalence and the life expectancy for the USA vs the world in the year 2016.

A screenshot of a cell phone

Description automatically generated

From the data above, from 2000 to 2016, the US smoking prevalence went down ~10% and the world went down ~7%. Additionally, the life expectancy went up ~2 years in the US and ~4.5 years worldwide.