ETL Project

Group 3

Samantha Gilmore

Veronica Hamilton

Helena O’Farrow

INTRODUCTION:

Our group chose to research pollution data as this is a common interest of ours. Our goal was to find publicly available data sources pertaining to pollution that we could extract, transform, and load.

EXRACT:

The deaths data we used came from a source called "Deaths Due to Air Pollution" found on kaggle.com. Due to the size of the file, we exported the data from the source and opened it in Excel where we removed all data not pertaining to the US prior to saving the file as a CSV.

|  |
| --- |
|  |

The population data we used came from and a source called "US Air Pollution Data" found on data.world This file had a few rows of text that we removed by opening the data in excel and deleting those rows. There also appeared to be some errors in the data at the far end of the table so we removed those columns as well. Finally, we save the file as CSV.

TRANSFORM:

Using Jupyter Notebook, we manipulated the "us-death-rates-from-air-pollution" in which we dropped all columns except "Year" and "Air Pollution (total) (deaths per 100,000" we then eliminated data from any year not within the 2000-2009 range we needed and renamed the "Air Pollution (total) (deaths per 100,000" to "Air\_Pollution\_Deaths".

Still using Jupyter Notebook, we began eliminating superfluous data from the population file, by using a location function to request the US summary line. We then converted the data contained within the population columns to float and remove the commas.

LOAD:

In PGAdmin we created a new data base called pollution\_db which contains two tables, one for each data source, called population and deaths. We connected both of our transformed resources to our SQL database by creating a new super user within PGAdmin and plugging that information into a connection string within each respective Jupyter Notebook file. From there we directed the connection to the corresponding table.