ETL Project: Technical Report

**Project Proposal:**

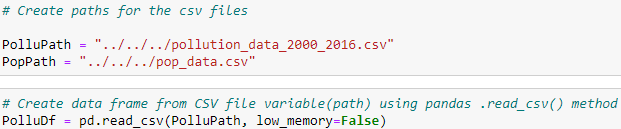
For this project we proposed to analyze pollution and population data from US cities. The pollution data was acquired from Kaggle.com while the population data was from the US Census Bureau.

This proposal was approved, and our group was instructed to limit the scope of the project to two major cities. The cities of choice were among the largest possible, being Los Angeles and New York City.

**Extract:**

The extraction process we chose to use was quite simple given the nature of the data. Both the pollution data extracted from Kaggle.com and the population data extracted from the US Census Bureau were both in the from of csv files.

The csv files were read directly into the Jupyter Notebook by storing the local path of the csv files in a python variable and then reading them into Jupyter Notebook using Pandas .read\_csv method. A clip of this process can be seen below.



**Transform:**

As expected, the transformation process was the most labor-intensive and time-consuming part of the project, by far. Upon receiving the raw data it was immediately recognized that much manipulation was needed in order to complete any sort of analysis of the datasets.

Tasks including but limited to data type transformation, column manipulation, value checks, removal of incomplete records, pivoting, grouping, and table joins were performed in order to prepare that data into a form in which it could be analyzed.

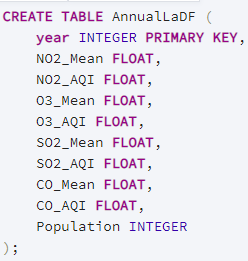
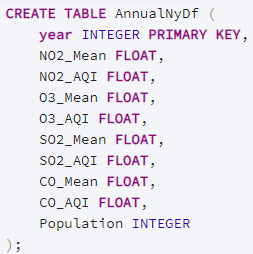
The transformations mentioned above can be seen step-by-step in the Jupyter Notebook provided in this repository and diligent notes were provided within each cell so that any reader can follow along the work with ease.

It should be noted that because the census data was only provided on an annual basis, the population data had to be manipulated in such a manner that it could also be analysis in the same way. This challenge was overcome by averaging the variables over the course of each year in comparison and then grouping by the year. These tables were then joined on the year columns.

**Load:**

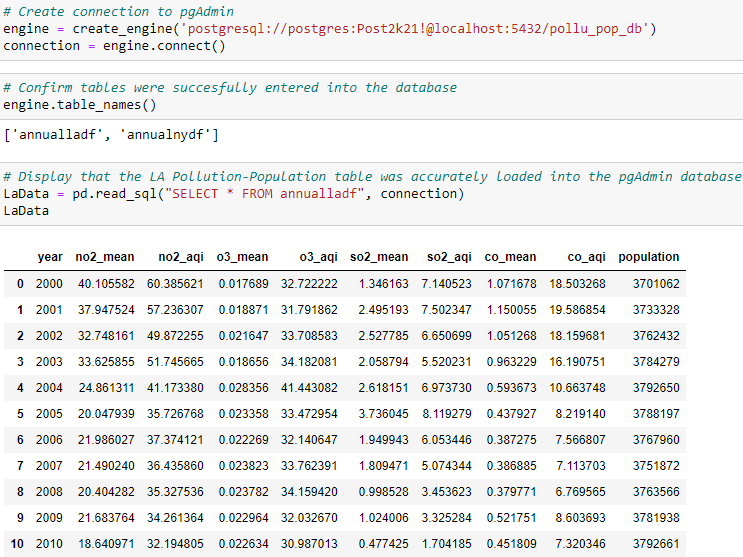
Once the final tables were produced, they then needed to be loaded into a database. Although, either mongodb or pgAdmin were both appropriate methods for uploading this specific data set, our group chose to upload the data into a postgres database using pgAdmin.

The first part of this process used the query editor to create the tables in pgAdmin as seen below.

The next step was to export the tables produced in the Jupyter Notebook to csv files and then manually import them into the tables created by the query editor.

Confirmation of the loading process was shown at the end of the provided Jupyter Notebook. This was done by making a connection to the SQL database using the sqlalchemy module and then importing them as seen in the clip below.



**Analysis:**

The ETL process required for this project was successfully complete as per outline in the project instructions, as described in the outline above. Data was extracted from reliable sources, transformed into a form necessary for analysis, and then uploaded into an external database for further use and analysis.

The data displayed in the database could give insights on how the change in population for the city in interest could be correlated to the change in concentrations of pollutions for that same city. This analysis could be done for any city that the two datasets have in common. Although only the annual mean was analyzed for this specific database, several different aggregates could be performed on the original data frame to analyze other parameters of the pollution data.