Project Report

Extract:

The two datasets used were obtained from Kaggle at the following web addresses:
https://www.kaggle.com/datasets/meeratif/indian-cities-by-population-dataset
https://www.kaggle.com/datasets/kkhandekar/carbon-monoxide-concentration-indian-cities

The first dataset, “Indian Cities by Population”, is a comparison of the population of Indian cities in 2001 and 2011. There are five columns: “Rank” contains an integer for each city, indicating how it compares to other Indian cities, with “1” being the highest population. The “City” column contains the name of the city in the form of a string. The “Population (2011)” column contains the population of the city in the year 2011. Similarly, the “Population (2001)” column contains the population of the city in 2001, both of these contain commas every three digits and so they are stored as strings. The last column, “State or Union Territory”, indicates the name of either the state or union territory that the city is located in, stored as a string.

The second dataset is “Carbon Monoxide Concentration - Indian Cities”, which contains data on the air carbon monoxide concentration in Indian cities. The columns here are: “State” - the name of the state (or union territory) that the city is in, as a string. “City”, the name of the city, as a string. The last three columns “Avg(ppb)”, “Max(ppb)”, and “Min(ppb)” provide the average, maximum, and minimum air carbon monoxide concentrations in parts per billion, respectively.

These were downloaded as .csv files, then read into Pandas dataframes in a Jupyter notebook file.

Transform:

Once loaded into dataframes, data cleaning began. First, column names were changed to be simpler and more SQL-friendly, and the less useful columns were dropped. The dropped columns were “Population (2001)” from the population data, and the “Max(ppb)” and “Min(ppb)” from the pollution data. Next, rows with missing data were dropped. Then, some of the city names in the population dataset had what appears to be citations at the end of them (e.g. “Bettiah[30]”), these would get in the way of any joins between these datasets on their “city” column. So, a str.replace() statement was used to remove these. Finally, duplicate rows were removed from the Carbon Monoxide dataframe, no duplicates were present in the Population dataframe This was the extent of the cleaning that was needed, as the datasets were already of a decent quality.

Refer to the “data_etl.ipynb” file for the exact code used for this section.

Load:

A database named indian_cities_db was created in PostgreSQL, then queries were used to create two empty tables with column names and data types matching those used in the cleaned dataframes, one named “indian_population” and the other named “indian_pollution”. Refer to the “indian_cities.sql” file for the exact queries used here, as well as the final join query used to join the two tables on the “city” column.

SQLAlchemy was used to connect the “data_etl.ipynb” file containing the dataframes to the indian_cities_db database. From there, Pandas was used to load the dataframes into their respective tables in the database. Pandas’ “read_sql_query()” function was then used to query the database to confirm that this had been successful, and then finally to join these two tables via their “city” column; this is featured in the last cell of the notebook file. As this was an inner join, cities not matching between the two were dropped.

So, the final schemata was two tables in a PostgreSQL database. The first, “indian_population”, contained the columns: “city”, “population”, and “state”, all data types were text. The second, “indian_pollution”, contained the columns: “state”, “city”, and “average_ppb”, with the first two being text and “average_ppb” using the float data type.

These datasets were chosen because, though we weren’t actually required to perform analysis for this project, we wanted to use the ETL process to prepare datasets that could actually be used to answer interesting questions. So, we decided to investigate if there is a correlation between the population of a city and its pollution levels. The motivation behind choosing Indian cities was to look into another country, one which we aren’t very familiar with. Also, there is far more data for this sort of thing for Indian cities than Australian, as well as there of course being far more cities in India too, making for a larger dataset.