**Project 2 – Extract, Transform and Load**

**Life Expectancy and Countries GDP**

**Objective**

The goal for the project was to find data for country’s populations along with GDP and Life expectancy to determine whether a countries GDP impacted life expectancy. Once obtaining the data sources, the data will be cleaned as required and loaded onto the necessary database. The database chosen was SQL, which then allowed conclusions to be made from the combined data sets.

**Extract**

For this project 2 sets of data were extracted from two different data platforms; Kaggle and Data.World on a countries GDP and life expectancy.

The data included Country, Population, GDP, Deathrate and Life Expectancy and was obtained from the following sources all as CSV files:

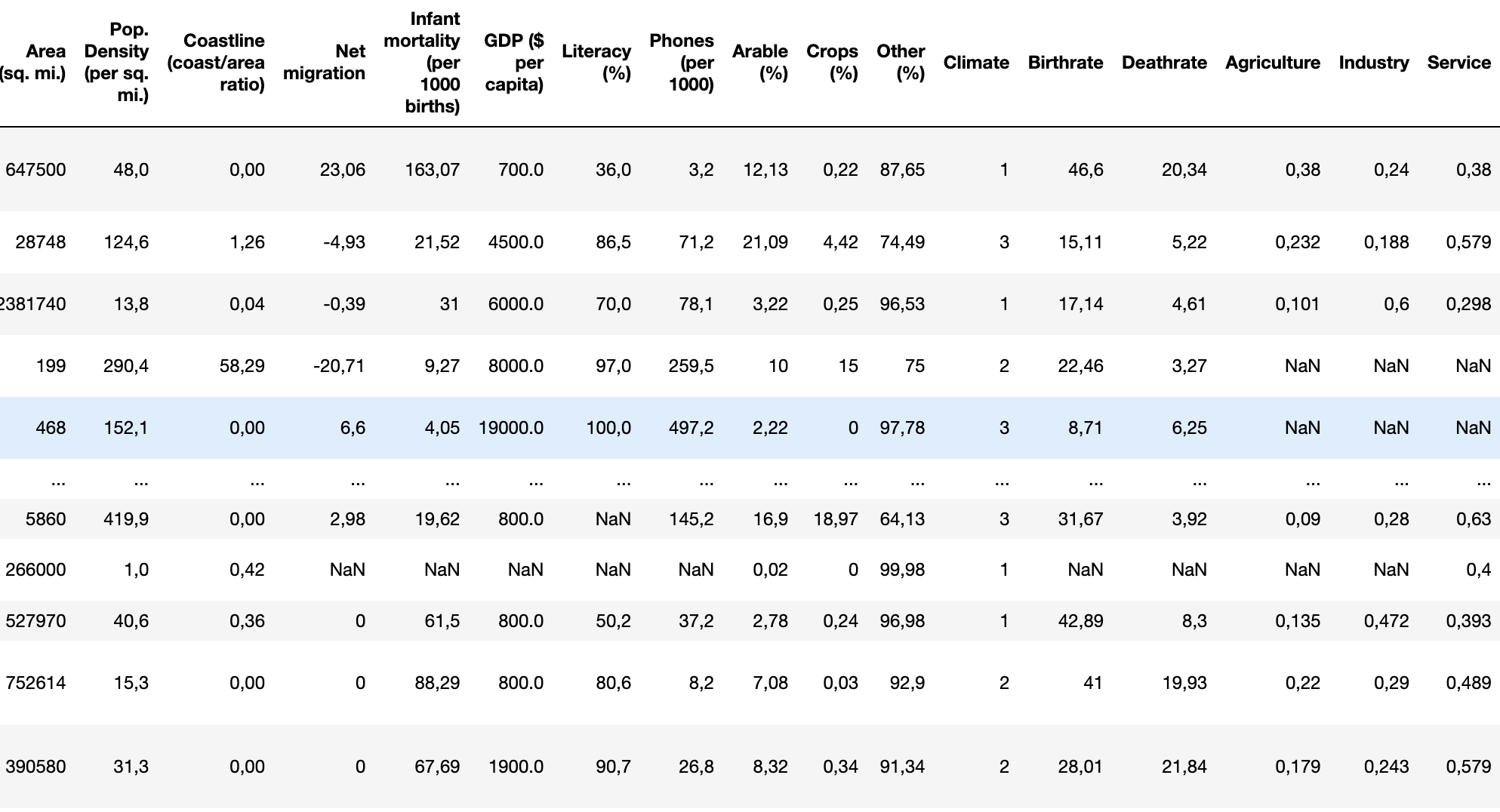
<https://www.kaggle.com/datasets/fernandol/countries-of-the-world>

<https://data.world/ndsouza/alcohol-v-life-expectancy>

**Countries GDP Dataset:**

Table

Description automatically generated



**Life Expectancy Dataset:**

Table

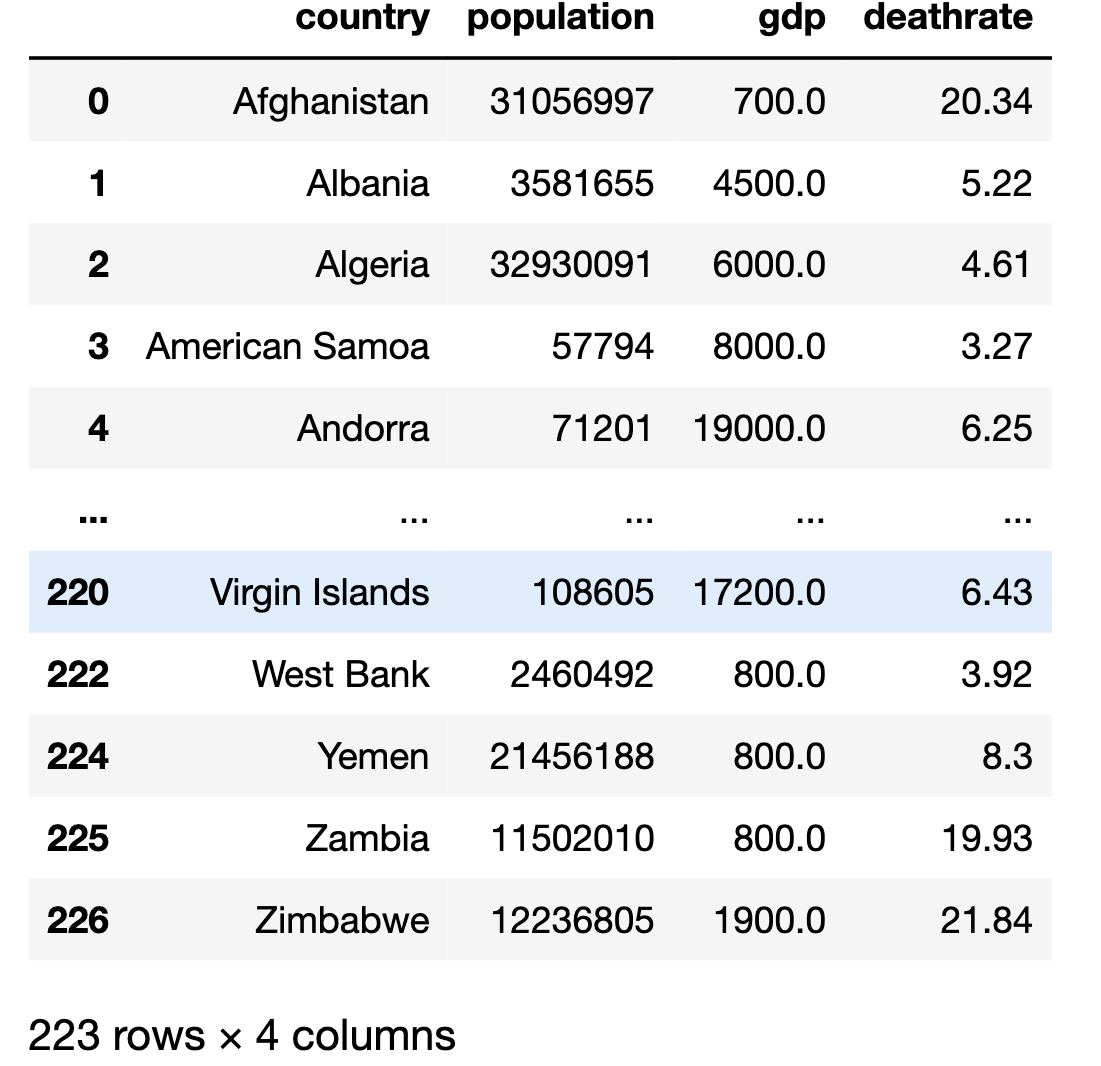
Description automatically generated

**Transform:**

* Pandas in Jupyter Notebook was used to load both the CSV files as a Pandas data frame using Python.
* The data was then created as a visualisations into separate data frames to study the data.
* The columns that were not required were then dropped using Python.
* The following columns were dropped from the Countries GDP dataset; Region , Area (sq. mi.), Pop. Density (per sq. mi.), Coastline (coast/area ratio), Net migration, Infant mortality (per 1000 births), Literacy (%), Phones (per 1000), Arable (%), Crops (%), Other (%), Climate, Birthrate, Agriculture, Industry and Service.
* The column “alc” was dropped from the Life Expectancy dataset as this was not required for the analysis.
* As there were spaces in the “country” column for the countries GDP dataset, these had to be removed in order to successfully load and merge the datasets.
* Following on, two empty tables were created on Postgres SQL with the same column names as the data sets being used.
* The columns were then renamed to match the column names for the SQL tables.
* Any rows that had missing information were dropped.
* The country id was used as the primary key in the SQL query tables that were created.

**Cleaned Data**

**Countries GDP Dataset:**



**Life Expectancy Dataset:**

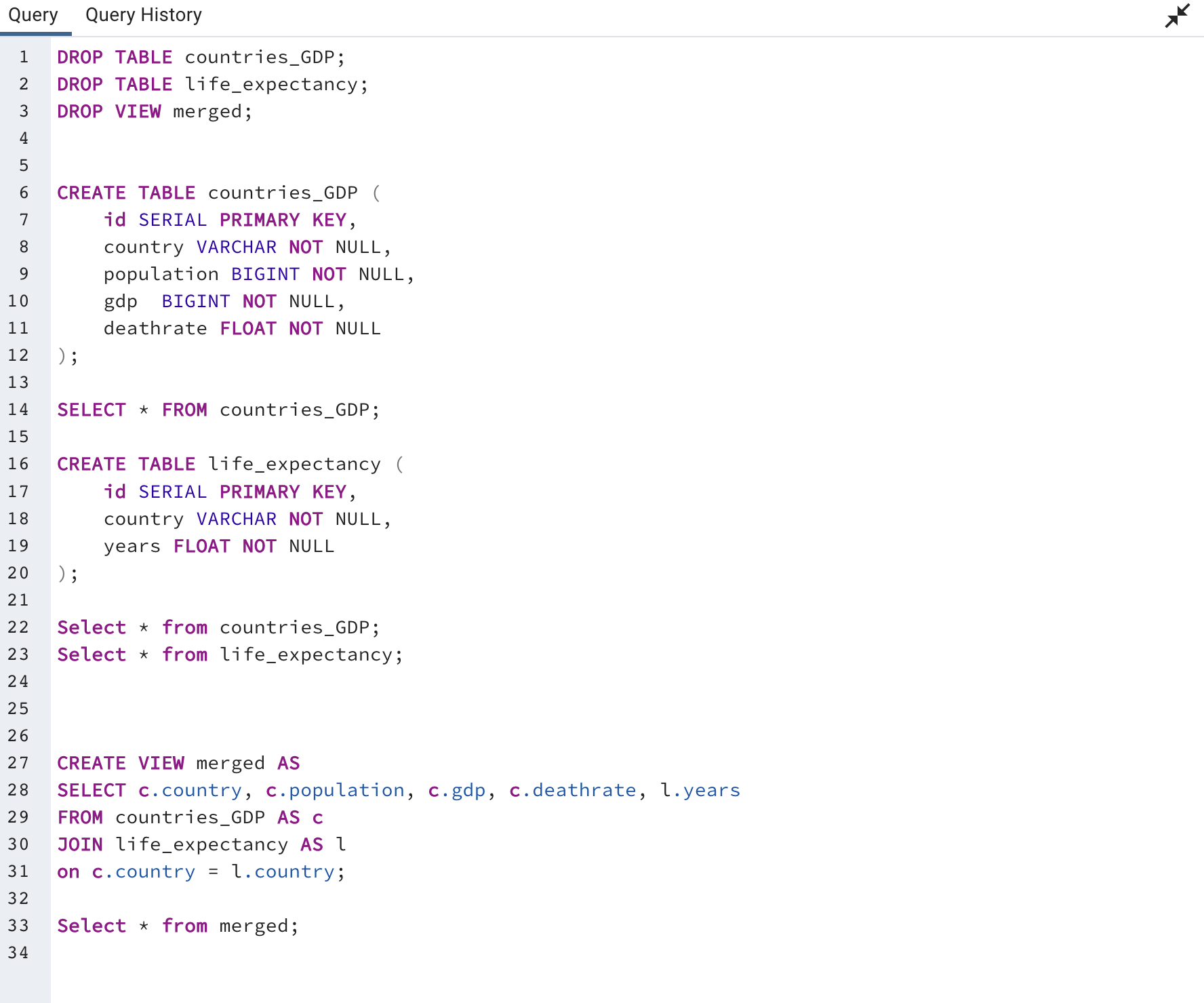
Table

Description automatically generated

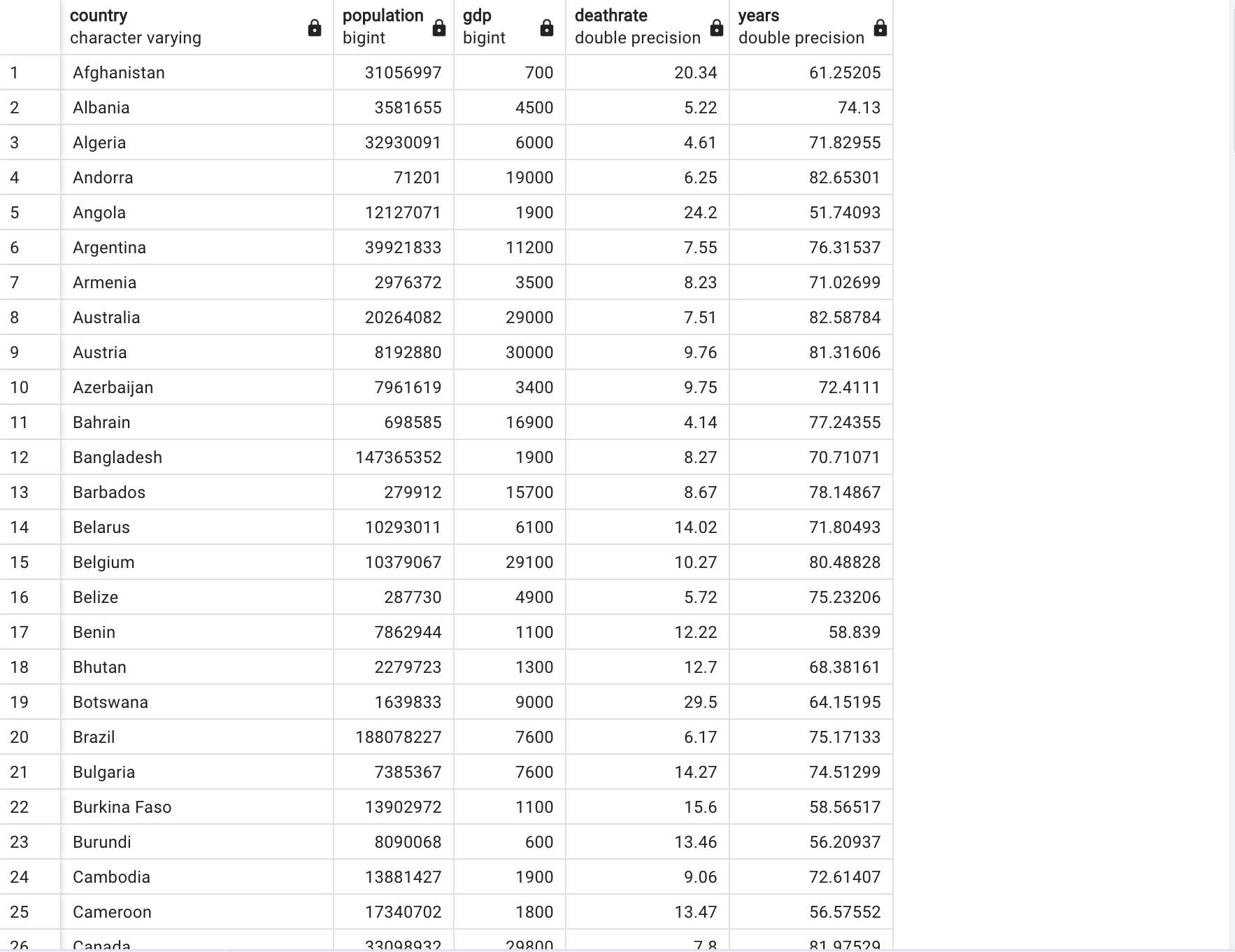
**Load:**

Once I had pulled the CSV files using Python in Jupyter Notebook and loaded each data frame for both datasets, a database connection was created from Jupyter Notebook to Postgres SQL. PG admin was used in order to store the data sets in the created tables. Once the data had been loaded and merged correctly with the required information, the data was loaded and read onto Python where the tables were altered and updated to allow a final analysis of the data.

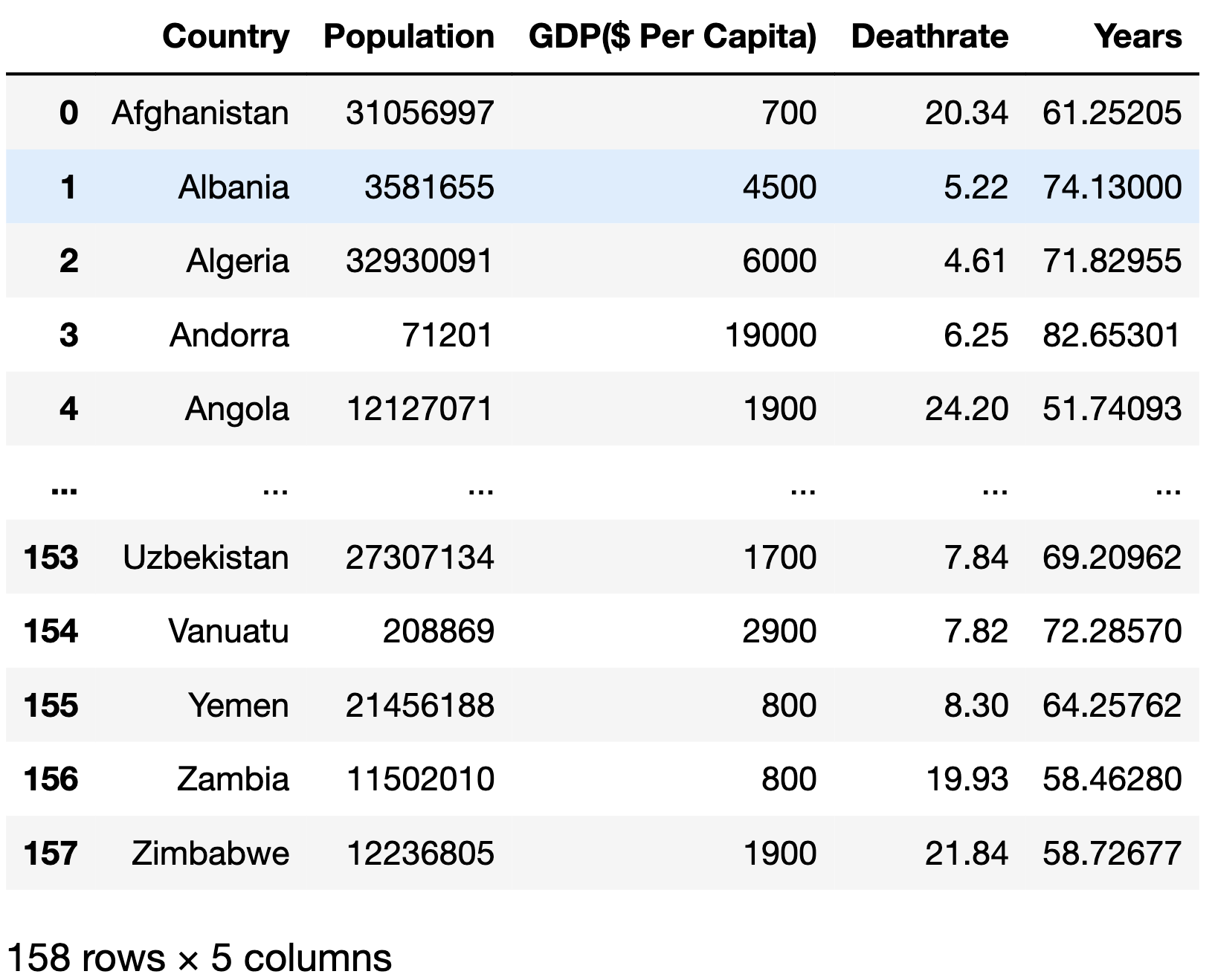
**Postgres Loaded Data:**



**Loaded merged Dataset in SQL**



**Merged Datasets**



**Questions and Analysis**

1. Which countries have the highest GDP?
2. Which countries have the lowest GDP?
3. Which countries have the greatest Life Expectancy?
4. Does a countries GDP have an impact on a countries life expectancy?

**Hypothesis**

* The country with the highest GDP will be China.
* The country with the lowest GDP will be Afghanistan.
* The country with the greatest life expectancy will be Switzerland.
* GDP does not directly affect a countries life expectancy.

**Conclusion**

The final and merged dataset exhibits the GDP and life expectancies for 158 countries which provide an insight into the Gross Domestic Product for countries and whether life expectancy is impacted by a countries GDP. The final dataset allows analysis on life expectancy and provides information as to whether there may be other factors impacting a specific countries average human life expectancy.

Table

Description automatically generatedTable

Description automatically generated

Table

Description automatically generated Table

Description automatically generated

As shown in the above tables, the data did not support my hypothesis as the country with the highest GDP was Luxembourg and not China.

The country with the lowest GDP was Sierra Leone, although Afghanistan was in the bottom 10 countries regarding the GDP, it does not support the hypothesis.

The country that had the highest life expectancy was Japan with Switzerland being 2nd and so this does not support my hypothesis.

The data shows that GDP does have a slight impact on a countries life expectancy as shown in the bottom 10 countries for GDP and life expectancy where Sierra Leone are bottom for both tables.

In conclusion, this signifies that although GDP may not be the sole factor in impacting a country’s life expectancy, it does have a small overall impact and therefore does not support my hypothesis.