BME AUT Automatizálási és Alkalmazott Informatikai Tanszék

Homework Documentation

Business Intelligence

2022 Autumn

Pollution Impact in People and the Environment

Pedro Jorge Fonseca Seixas – PQDI43

pjfseixas.17@gmail.com

MŰEGYETEM 1782

Introduction

The focus of this work is to understand the impact of air pollution on the environment and the people themselves. The data has historical information about controlled substances (such as CFCs) usage, CO2 emissions and air pollution levels in different countries, its deaths, the earth's temperature and ozone hole size. The goal is to assess if CO2 emissions directly impact those other things through different types of data visualization.

Technologies

For this implementation, multiple technologies were used, each for its own purpose:

- Pentaho was used for ETL tasks
- Docker / Docker-compose were used to handle the virtualisation to simplify the installation and management process for the database
- PostgreSQL was used as a data warehouse
- pgAdmin was used to manage the PostgreSQL database
- Microsoft PowerBI was used as the reporting tool

Usage

To properly instantiate and access the data and report, in the necessity of making any changes, you should clone the repository of the following link and follow the README.md file instructions: https://github.com/pedrojfs17/BME-BI. The repository already contains all files needed for the project, including the datasets. The report can be seen at this link: PowerBI Report.

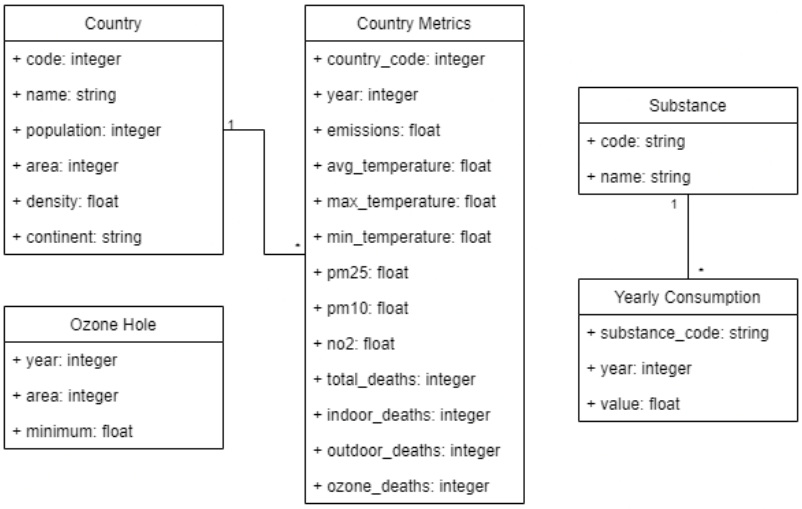
Data

For this project, six datasets were used:

- Ozone Hole Data : contains information about the ozone hole size and the minimum ozone recorded throughout the years.
- Controlled Substances : contains information about the consumption of controlled substances through the years worldwide.
- CO2 Emissions : contains information about the countries (population, area, density) and the CO2 emissions over the years.
- Surface Temperatures : contains the monthly average temperatures for each country.
- Air Pollution : contains information about the air pollution levels around the globe through the years.
- Air Pollution Deaths : contains the number of deaths due to air pollution in each country.
- Continents : contains the mapping between the countries and the continent they belong to.

Besides the Ozone Hole data, which was already in the desired format, every dataset needed to be changed to meet the project needs. These changes compose the ETL process and are described in the next section in detail.

After the ETL process, this data is stored in a relational PostgreSQL database, that will serve as a data warehouse for the reporting tool. As the datasets were not all in the same format, there was a clear need of building a relational model that could store all data to be used later:



ETL

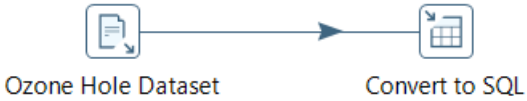
The ETL process is composed of five main jobs, each one with its own purpose:

- *"Setup Database Job"*: Resets the database. Removes everything from it, and creates the necessary tables, without any data in them
- *"Ozone Hole Job"*: Creates and populates the database with the information about the ozone hole
- *"Controlled Substances Job"*: Creates and populates the database with the information about the controlled substances
- *"Country Metrics Job"*: Creates and populates the database with the information about each country and their different metrics
- *"Full Database Job"*: Merge of the above jobs for the initial setup.

These different jobs were created so that whenever a change is necessary in the ETL tasks, only the job corresponding to it should be run to update the database, thus being more efficient.

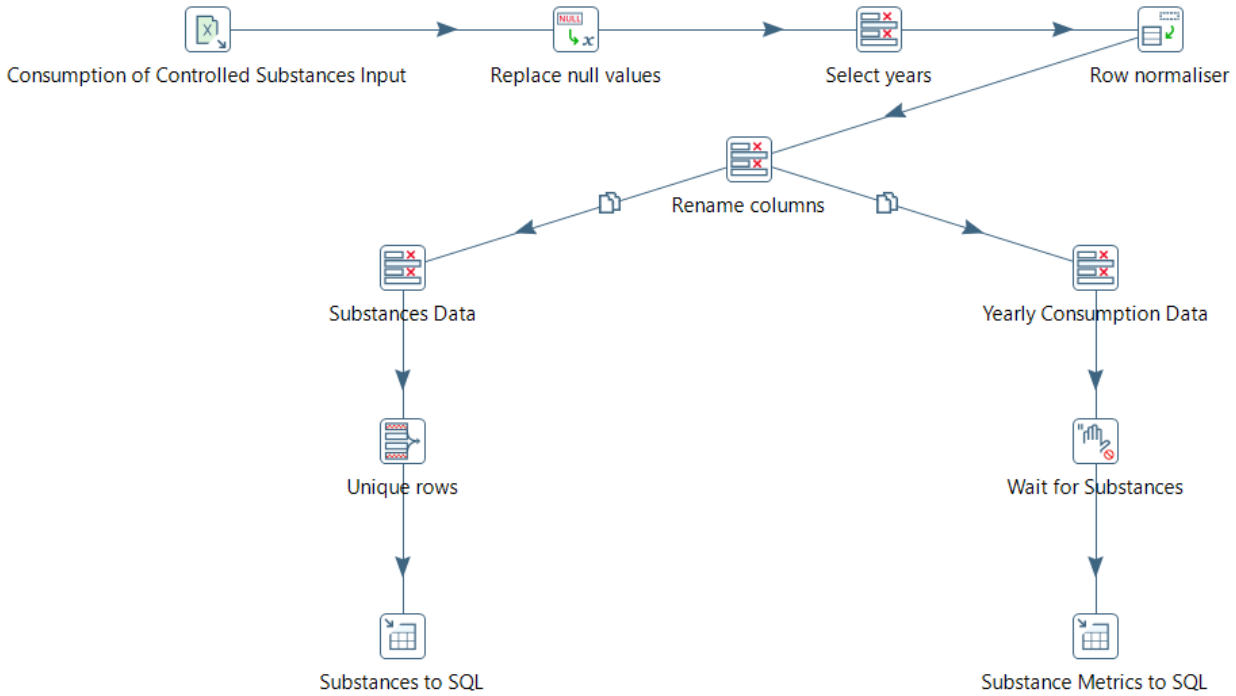
There are three main transformations that are necessary for the previous jobs. The following sub-sections will briefly describe each one of them:

Ozone Hole Dataset



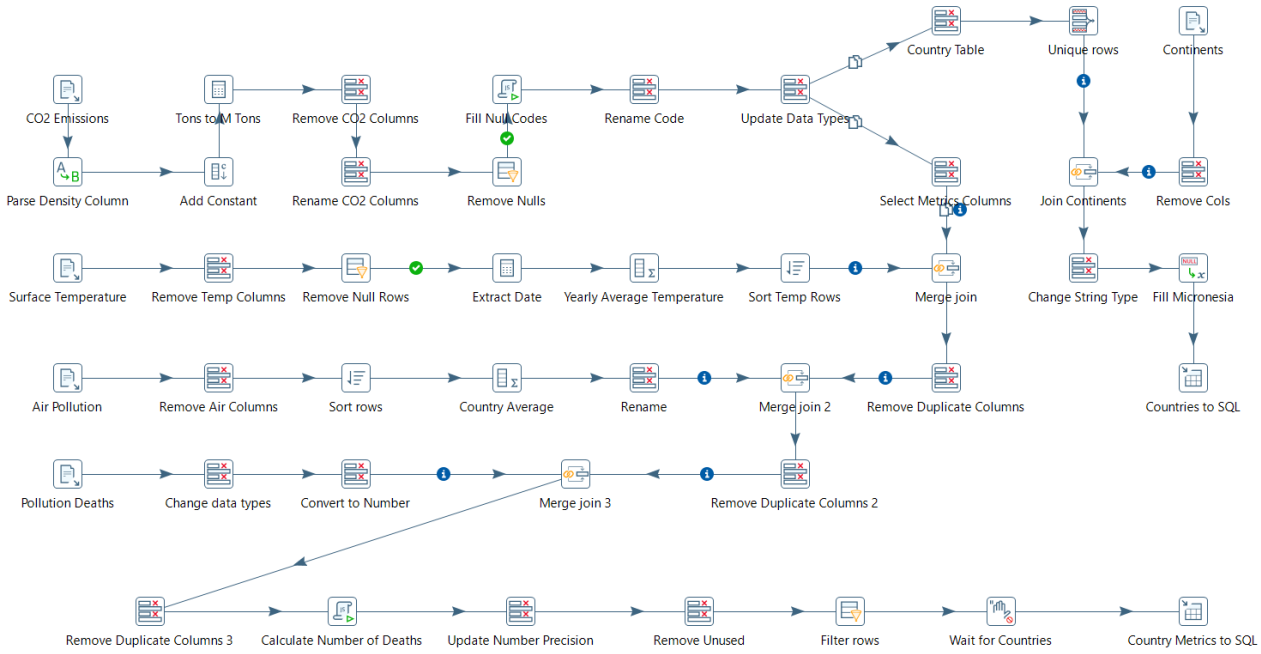
This first dataset was already very clean and only contained the necessary information in the proper format. The only step was to map the dataset fields to the database ones and populate the **Ozone Hole** table with them.

Controlled Substances Dataset



The consumption of controlled substances dataset needed to be adapted to the tool requirements. The null values were replaced by 0, and the first year was removed since there was no information on the years between the first and the second (five years later). The dataset had a column for each year, but to be stored in the database, there was a need to make each column a different row. With the data in the correct format, the dataset was divided into two streams: build the **Substance** table and build the **Yearly Consumption** table.

Country Metrics



This transformation is clearly a lot more complex than the other ones. To populate the **Country** and **Country Metrics** table, the four datasets needed to be merged into one. This transformation includes both the transformation of each dataset and the process of joining them.

For the CO2 Emissions dataset, as the number of emissions was huge, the value was cast to millions of tons. After that, some countries didn't have the country code (which is used as a primary key), so there was a need to fill those missing values. Since all codes had two letters, these values were filled with the first three letters of the country's name so that there were no intersections. Although one of the datasets had more countries than this one, the only countries stored are the ones from this dataset since it contains information about the population, area, and density, contrary to the others. These countries were merged with the dataset that had the mapping between countries and continents. After the merge, only one *null* value, Micronesia, needed to be filled. The Country table was easily populated from this resulting dataset.

In the Surface Temperature dataset, the data was stored by month, so there was a need to aggregate those values into years. After removing the null rows and extracting the year from the date, the values were aggregated by year, country-wise. From this dataset, the minimum, maximum and average yearly temperatures were saved for each country for each year.

As for the Air Pollution dataset, something like the last dataset was made, but instead of aggregating months to years, there was information about multiple cities in each country that was averaged. This gave the average pollution levels of each country, year-wise.

After merging these three datasets, the only one left was the Pollution Deaths dataset. The dataset contained the deaths per 100 thousand people in each country due to pollution. For the desired reporting, this number had to be translated into the total number of deaths in each country. This calculation was made with the information about each country's population (from the CO2 Emissions dataset).

Between all the described transformations, it was also necessary, multiple times, to either remove some columns, change some data types, or sort the rows. To finalize this ETL transformation, the rows were filtered to have only data since 1800, since most metrics before that date were not available and loaded to the database to populate the **Country Metrics** table.

Presentation

As stated in the Usage section, the built report can be seen in this link. It is composed of four main pages and two extra helpful pages, with different types of information and different types of presentations that go from simple bar charts or line graphs to more complex maps. The conclusions taken from the data are available in the report pages where they are taken. Each of the pages is described next:

Ozone Hole and Temperature

This page relates the ozone hole area with the maximum, average and minimum temperatures of the earth surface. There is the ability to filter the continents, to show the data specific to the desired set of continents. In the temperature graphs, a prediction can also be seen for the temperatures of the next 10 years.

Deaths due to Air Pollution

In this page there is some visuals to show the number of deaths related to the different types of air pollution and the countries that have been mostly affected. There is a range filter to select the desired years. There is also the possibility to sort the countries either alphabetically or by the number of deaths due to pollution. The option to drill-through is available in the countries, by right-clicking and selecting "Country Details", to find out more information about them.

Causes of the Ozone Hole

This page was created to assess the causes of the ozone hole. There are some graphs present that relate the ozone hole area to the minimum recorded ozone concentration in a year as well as graphs about the emissions of CO2 and other more powerful greenhouse gases worldwide or divided by the different continents.

Pollution Levels and Deaths

In this page some graphs are presented to evaluate the relation between the air pollution metrics and the deaths that occur due to them. The distinction between developed and developing countries is very well seen in this page.

Country Details

This page shows the data for a specific country. It has a map, shows which continent is from, population, area, CO2 emissions and data related to air pollution levels, deaths, and the recorded temperatures.

Help

This page's purpose is to clarify some doubts that the viewer may have. It explains the meaning of some air pollution metrics, as well as the existence of negative values in a consumption graph.