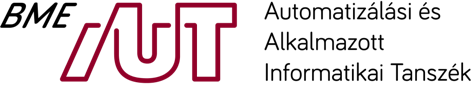
**Homework Documentation**

Business Intelligence

2022 Autumn

**Pollution Impact in People and the Environment**

**Pedro Jorge Fonseca Seixas – PQDI43**

**pjfseixas.17@gmail.com** 

# **Introduction**

The focus of this work is to understand the impact of the air pollution both in the environment and in the people themselves. The data used has historical information about the controlled substances (such as CFCs) usage, CO2 emissions and air pollution level in different countries, and information about the deaths related to it, the earth temperature and ozone hole size. The goal is to assess if CO2 emissions have direct impact on 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.

# **Data**

For this project, 6 datasets were used:

* [Ozone Hole Data](https://www.kaggle.com/datasets/varunsaikanuri/ozone-hole-data) : contains information about the ozone hole size and the minimum ozone recorded throughout the years.
* [Controlled Substances](https://ozone.unep.org/countries/data-table?report_type=0&output_type=odp-CO2e-tonnes&party_grouping=total&period_start=1986&period_end=2021&ignore_zero=1&group_by=group&op=GENERATE+REPORT&form_id=ozone_country_data_form__report_table_form) : contains information about the consumption of controlled substances through the years, worldwide.
* [CO2 Emissions](https://www.kaggle.com/datasets/moazzimalibhatti/co2-emission-by-countries-year-wise-17502022) : contains information about the countries (population, area, density) and the CO2 emissions over the years.
* [Surface Temperatures](https://www.kaggle.com/datasets/berkeleyearth/climate-change-earth-surface-temperature-data?select=GlobalLandTemperaturesByCountry.csv) : contains the monthly average temperatures for each country.
* [Air Pollution](https://www.kaggle.com/datasets/totoro29/air-pollution-level) : contains information about the air pollution levels around the globe through the years.
* [Air Pollution Deaths](https://www.kaggle.com/datasets/akshat0giri/death-due-to-air-pollution-19902017?resource=download) : contains the number of deaths due to air pollution in each country.

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

Diagram

Description automatically generatedAfter 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 5 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 3 main transformations that are necessary for the previous jobs. The following sub-sections will briefly describe each one of them:

A picture containing arrow

Description automatically generatedOzone Hole Dataset

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

Diagram

Description automatically generatedControlled 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 of the years in between the first and the second (which was 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, one to build the **Substance** table, and another to 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 4 datasets needed to be merged into one, so this transformation includes both the transformation of each dataset and the process of merging them.

For the CO2 Emissions dataset, as the number of emissions were very large, the value was casted 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 2 letters, these values were filled with the first 3 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. The Country table was easily populated from this 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, for each country, for each year, the minimum, maximum and average yearly temperature were saved.

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 were averaged. This gave the average pollution levels of each country, year wise.

After merging these 3 datasets, the only one left was the Pollution Deaths dataset. The dataset contained the number of deaths, per 100 thousand people of 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).

**Diagram

Description automatically generated with low confidence**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**

TODO – Talk about PowerBI report

# **Conclusion**

TODO – Conclusions taken from this report