- **Introduction**: Introduces and motivates the question you answer in the report
- **Used Data**: Describe the used data you used for the analysis (the output of your data pipeline). Briefly discuss the structure and meaning of the data (such as domain-specific value types), and implement the obligations to comply with the data licenses of your data sources if necessary.
- **Analysis**: Present the executed analysis: method, result, and interpretation. This section doesn't need to show code, but the reader should understand what you did and why it is appropriate what you've done to answer the question. Focus on the results (positive and/or negative) but leave out any failed attempts.

Conclusions: Explicitly answer the question you posed to yourself. Critically reflect whether the question could be answered completely and if there are any remaining uncertainties or limitations.

The link between health status and net greenhouse gas emissions in the EU: a cross-country analysis

1. Research questions

1. How do net greenhouse gas emissions influence the health status of EU countries?
2. Which EU countries exhibit the highest levels of greenhouse gas emissions?
3. How do net greenhouse gas emission values of the EU countries change over time?

2. Data Sources

For this report, Eurostat was selected as the data source due to its comprehensive and wellstructured datasets on various topics. Eurostat data is of high quality because it is consistent in format, regularly updated, and relevant to the user (in this case me) and the stated research questions. The quality assurance is performed in accordance with the European Statistics Code of Practice. The chosen proxies and respective Eurostat data sources (CSV files) for each of the two domains are:

2.1.1 Health domain

Heathy life years: is an aggregate expectancy indicator that measures how many years a person (at birth) is expected to live without severe or moderate health problems. Health expectancies are calculated using the Sullivan method, which combines mortality and health status data. Mortality data includes age-specific death rates, probabilities of dying and surviving and, among other, life expectancy. Health status data is measured using a special variable from the survey, which asks about limitations in daily activities due to health problems for at least the last six months.

The overall accuracy of this dataset is considered high, as this indicator receives significant attention and is closely monitored by policymakers, specifically the Directorate-General for Health and Consumer Protection (SANCO) and the Directorate-General for Employment and Social Affairs (EMPL). There have been no major corrections requested, with the primary issues relating only to comparability over time and between countries.

2.1.2 Environmental domain

Net greenhouse gas emissions: is used as a proxy for the environmental status of a country. The net greenhouse gas emissions indicator represents a basket of the emissions of national greenhouse gases, covering CO2, CH4, N2O, and F-gases. There are 2 options of the indicator to be chosen: (1) an index to 1990 or (2) absolute amount in tonnes of CO2 equivalent per capita. The indicator is presented in two forms: (1) net emissions including land use, land use change and forestry (LULUCF) and (2) excluding LULUCF.

This data is not sourced directly by Eurostat; the original source is the European Environmental Agency. The accuracy of the indicator is considered high, as the data is reported under the United Nations Framework Convention on Climate Change.

2.2 License

Both datasets are published by Eurostat under a standard open-data license with permission for both non-commercial and commercial re-use. To fulfill the obligations of the license, the correct source (Eurostat) is indicated. Details on the license can be found here.

3. Data Pipeline

The data pipeline is structured using the ***Extract-Transform-Load (ETL)*** methodology in Python, as detailed in the script.py file. Initially, we import the necessary libraries, such as `pandas` and `numpy` for data manipulation and `sqlalchemy` for connecting to the SQL database to store the final SQLite file.

3.1 Extract

In the Extract stage, the data for each dataset is directly downloaded from a URL containing a CSV file. The Pandas function `read_csv` is used to retrieve the data.

3.2 Transform

The transformation of the data starts with initial inspection (glance at it). It was observed that some unnecessary, irrelevant for analysis information is present in both columns and rows.

Clearing rows: In the health dataset, rows indicating gender are not needed (only rows with the value 'total' in the 'sex' column are relevant), so they were deleted using the `dataframe.isin` function. In the net greenhouse gas emissions dataset, rows with alternative indicators were removed, specifically the index indicator and the second form of the net emissions indicator excluding LULUCF. This leaves only the relevant indicator forms within the dataframes: net greenhouse gas emissions measured in tonnes per capita, including LULUCF.

Clearing columns: Unnecessary columns include DATAFLOW, LAST UPDATE, freq, unit, indic_he, OBS_FLAG, sex, and src_crf. For analysis, the DATAFLOW and LAST UPDATE columns are not needed since they provide meta-information about the dataset. Other columns contain additional information not addressed in the research. The deletion of columns is done via the `dataframe.drop` function.

Pivoting years: After removing unnecessary rows and columns, it was found that the data was not clearly organized by time span, i.e. year data was stored in rows. To make the data more representative, the dataframe was pivoted to turn year data into columns using the `dataframe.pivot` function.

Null values: The pipeline also addresses incompletness issue, i.e. missing values, specifically in the health dataset. Assuming a normal distribution, missing values were filled using the linear interpolation method of neighboring years' values (`dataframe.interpolate` function). If the first observation year's data was missing, the following year's data was used to fill the gap (`dataframe.bfill` function). Given the nature of the dataset and the type of the variable (numeric float) interpolation method minimizes bias when filling the missing values.

3.3 Load

After the transformation step, the data is prepared to be loaded into a specified directory. Since the target directory is in another folder, the loading process is carried out using the SQLAlchemy engine. First, a connection to an SQLite database is established using the `create_engine` function. Then, the datasets are loaded into the SQLite database files located in the ../data directory (`dataframe.to_sql`).

4. Result and Limitations

The output of the data pipeline consists of two cleaned datasets, fully prepared for research analysis. Both datasets are structured as panel data, combining time series (the evolution of variables over time) and cross-sectional data (multiple groups representing different countries). This structure is well-suited to the research question, allowing for analysis of both health and environmental domains, their evolution over time, and differences between countries. The SQLite file format was chosen for its convenience in further processing and analysis. The data quality was enhanced by: (1) deleting unnecessary information, (2) structuring the data so that columns represent years, (3) filling in missing values using a linear interpolation.

4.1 Limitations

It is important to acknowledge some limitations in the data. The primary limitation is the original datasets themselves: both indicators are complex computational indexes subject to

biases and computational errors. Additionally, the base values used to compute these indicators are self-reported by countries, which may allow for potential manipulation. Also, the data is not fully up-to-date, with the most recent release representing data from 2022, resulting in a lag of two years.
Furthermore, some (irrelevant to the research) data was dropped, however, some information, such as, for example, last update date, other forms of the indicators and distribution by gender is now lost. But more importantly, new data was added: missing values were corrected using linear interpolation, which relies on certain assumptions, in this case normal distribution. This, as any other addition of the data introduces some validity concerns. However, the dataset is still considered reliable because the assumption of a normal distribution suits the nature of the data well, reassuringly minimizing distortion in the results.