Data Report:
Impact of Sector-Specific CO2 Emissions on Urban Air Quality in Luxembourg

Main Question

How do sector-specific CO2 emissions from energy combustion influence urban air quality indicators in Luxembourg?

Data Sources

The data for this project comes from the Luxembourg Statistical Office (LUSTAT) and covers the period from 2018 to 2022. LUSTAT provides open data and as such the data is freely available for public use. These datasets can only be used under their license terms that include proper attribution and non-commercial use. In this project, two datasets were used with compliance to those license terms:

1. Quarterly (2018-Q1 to 2022-Q4) **CO2 emissions data** from different sectors: road transport, air transport, and others.
   The dataset originally contains several columns with combined information, including the data flow identifier, emission type, frequency, time period, observed values, and notes on emissions and observations. The reason why this particular dataset was chosen is because it gives a breakdown of emissions by sector in detail which is very important when examining how specific activities are affecting urban air quality.
2. Yearly (2018 to 2022) **air quality data** based on various air pollutants, such as sulphur dioxide, nitrogen dioxide, nitrogen oxides, ozone, carbon monoxide, particulate matter PM2.5 and PM10, lead, arsenic, cadmium, nickel, and benzo(a)pyrene.
   The dataset originally contains several columns with combined information, including the data flow identifier, frequency, EU norms, specifications, time periods, observed values, and multiple notes. The reason why this particular dataset was chosen is because the data here is comprehensive enough to highlight the major contaminants in urban air and can be used for pollution trend analysis.

Data Pipeline

For this project, a data pipeline was created that is capable of downloading, cleaning, transforming and storing datasets useful in conducting an analysis. The implementation of this pipeline was done using Python language and different libraries were employed to ensure quality of information and easy handling. Therefore, the following technologies were used:

- **python**: for automation and scripting;
- **requests**: for accessing and downloading data via APIs;
- **pandas**: for data manipulation and restructuring;
- **os**: for working with file paths and directories..

Pipeline steps:

1. **Data downloading**:
   The data was fetched from the LUSTAT using their REST APIs. Afterwards, the data was downloaded in CSV format and saved locally in the files as "air_quality_2018_2022.csv" or "co2_emissions_2018_2022.csv" accordingly for the further use. Since the data covers the fixed period from 2018 to 2022, it can be treated as static, and, therefore, does not require continuous updates.

2. **Data processing**:

   - CO2 emissions data:
     The downloaded data has a number of columns containing mixed information. The data had to be cleaned and transformed in order to make it useful. At first, the matter with inconsistency arose: some quarters lack values; formatting varies among the columns. It was split into separate columns, pivoted, and annual totals computed for all of these. First of all, the 'DATAFLOW' and 'EMISSIONS: Emissions' columns were broken down so as to get emission types and time periods that are crucial for our analysis. Then the data was pivoted with respect to time periods in order to have an organized structure by periods and sectors for purposes of analysis. Finally, quarterly data was aggregated to obtain annual total per sector since there were no quarterly results provided in the air quality dataset unlike this case where only annual results were available. The final dataset was validated to ensure all rows have consistent formatting and no missing values.

   - Air quality data:
     Same as CO2 emissions dataset above, raw data has multiple columns with mixed information which needed cleaning and transformation for analysis purposes. The following steps were taken: filtering relevant columns, renaming columns, pivoting data. Firstly, only necessary columns were chosen. Some columns from the initial dataset, such as "DATAFLOW" and "FREQ: Frequency", for example, were dropped due to their uselessness for the further analysis. Afterwards, the columns were renamed for clarity and consistency. Therefore, the dataset consisted of 7 columns: type of solution, EU norm, 2018-2022 years. And finally, the data was pivoted in order to be able to organise it by type of pollution and years in chronological order.

3. **Data saving**:
   Once all steps are performed, the processed data was saved for further analysis as "processed_co2_emissions_2018_2022.csv" and "processed_air_quality_2018_2022.csv" accordingly.

Error handling:

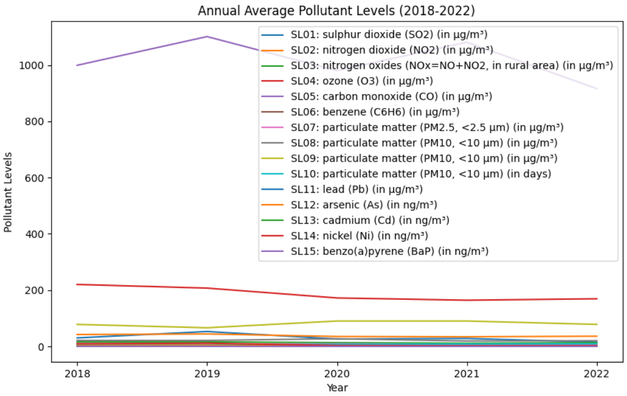
- HTTP errors: For both CO2 emissions and air quality datasets, HTTP errors are managed during API data download using "response.raise_for_status()" to catch and handle issues, with logs kept for later review.

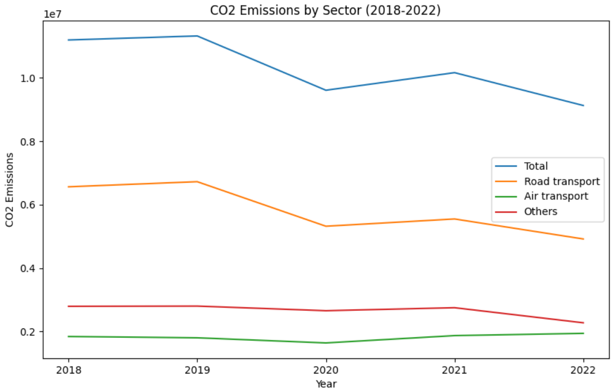
- Data validation checks: Validation checks to ensure data completeness and correctness after downloading and before processing. It is implemented in "def validate_dat"a using "raise ValueError".

- Directory and file paths existence checks: Before saving or processing data, the pipeline checks the existence of the specified directories and file paths and creates the required directories if they are missing. It is implemented in the setup with "os.makedirs".

Results and Limitations

The data pipeline built for this project successfully automated the processes of downloading, cleaning, transforming, and storing the datasets from the LUSTAT. The processed datasets were used to create the graphs below, which illustrate CO2 emissions by sector and annual average pollutant levels from 2018 to 2022. The graphs reveal trends in CO2 emissions and pollutant levels, with sector-specific emissions data showing a decline, especially in road transport, and pollutant levels indicating stability or slight changes over the years. The graphs were created using matplotlib library, and the code can be found in “analysis.py" file.





Further analysis of air quality data could include comparing the pollutant levels in each year to the given EU Norm. In this case, it would be possible not only to see changes in the numbers, but also to understand how extreme the situation is and wether additional measures need to be taken from the policy makers.

The project has several restrictions. Firstly, it uses historical data and may not be able to show the current status of what actually happens or future tendencies. This makes it difficult to use the analysed data for predictions. Secondly, the original datasets were from different time periods; air quality data was collected annually while CO2 emissions data were taken quarterly. This limits how detailed the analysis can be. Furthermore, if both sets had quarterly data, it would help to understand better how emissions affect air quality every year. Similarly, sector specific-approach can provide some useful indications but may overlook cross-cutting impacts between sectors. Finally, basing on publicly available datasets means that the study is limited by the accuracy and comprehensiveness of the provided figures. Mistakes or gaps in the primary information source could impact outcomes of analysis.