Introduction

Air emissions refer to the gases and particles released into the atmosphere from various sources. Among these, the increase in greenhouse gases such as carbon dioxide and methane poses a significant risk to the environment and is a major cause of climate change. In addition to greenhouse gases, airborne particles also threaten human health. Analyzing targeted data can yield valuable insights that help mitigate the effects of climate change.

The Federal Statistical Office of Germany provides valuable data detailing the emissions from various economic sectors and their evolution over the years 1996-2021.

Question

Which industries emit the most air pollutants and greenhouse gases in Germany?

The focus of this project is to understand the amount of air emissions produced by each economic sector in Germany.

Data Sources

Provider and Quality

All three data sources are provided by the Federal Statistical Office of Germany (Statistisches Bundesamt). According to information in here, after 2020, the source of publication has changed to GENESIS-Online database, which can be accessed in here. There are some differences in terms of economic sectors and gas types between the data from 1995-2019 and 2020-2021. Despite these differences, the data still offers valuable insights into the trends in air emissions.

The Federal Statistical Office of Germany provides quality data: with clear methodologies and documentation. More information about the quality can be accessed from their website.

Motivation Behind the Data

The air emissions data provides valuable information on the extent to which domestic economic actors cause emissions of greenhouse gases and pollutants into the air. Identifying these trends can help steer mitigation efforts in the right direction.

Structure of the Data

Data is presented in tabular form. It includes year information, types of gases, and different economic sectors.

Data License

According to Genesis-Online, data is licensed under the "Data Licence Germany - Namensnennung - Version 2.0", licence text available at www.govdata.de/dl-de/by-2-0. Data can be used, altered, processed, and merged as long as the user ensures the name of the provider, the link to the dataset, and refers to the license text, which are fulfilled in the project-plan.md file and scripts.

An Overview of the Transformed Data

In [ ]:

```
import pandas as pd
df = pd.read_csv("../data/Luftemissionen_2000.csv", delimiter=";").head()
df
```

Out[ ]:

	year	economic_sector	Kohlendioxid (CO2)	Methan (CH4)	Distickstoffmonoxid (N2O)	Stickoxide (NOx)	Schwefeldioxid (SO2)	Kol
0	2000	Erz.d. Landwirtschaft u. Jagd sowie damit verb...	9751762	1345969	100198	169012	3782	
1	2000	Forstwirtschaftl. Erzeugnisse und Dienstleistu...	411683	1284	9	2875	47	
2	2000	Fische und Fischereierz., Aquakulturerz., DL	53113	1	2	669	148	
3	2000	Kohle	1036480	662952	5	1116	1551	
4	2000	Erdöl und Erdgas	2511279	7258	15	1495	336	

Data Pipeline

The data for the years 1995-2019 and 2021 can be downloaded directly. However, the data for 2020 must be accessed through the Genesis-Online website or via an API because it does not have a direct download link. The link is obtained via Genesis-Online website's UI, Genesis-Online has the ability to share a link to the source (again a web page), users can visit this link and download the data. To automate this, the Selenium package is used to visit the Genesis-Online link and click the download button for the 2020 data.

The pipeline begins with a script that uses Selenium to download the 2020 data to the "data" directory without any transformation. Next, the "extract_transform.py" script downloads the remaining data, stores it in memory, applies the necessary transformations to all the data (including 2020), and then saves it in the "data" directory. The Pandas library is used for these transformations.

After the pipeline processes the data, it is stored in the 'data' directory, organized by year. Storing the data this way reduces the amount of data manipulation needed during processing and makes it easier for humans to interpret and navigate.

Transformation Steps

Data from 1995 to 2019 is structured with years as columns, rows are economic-sectors, gas types and their values. However, for the years 2020 and 2021, the structure is reversed with gas types as columns and economic sectors as rows. Since the latter is easier to interpret, the data till 2019 is transformed to have the same form as 2020 and 2021. In addition, there exists a small difference between the untransformed 2020 and 2021 data. Therefore, the 2020 data first transformed to be in the same format as 2021. Transformation for 1995-2019 is done by Pandas's "pivot_table" function which has the similar logic as "group by" functionality. In this case, economic sectors are used as indexes and gas types are used as columns. For 2020 and 2021, simple column manipulation was sufficient.

Since there are no missing or untrustable values, the data does not require further processing.

Problems Encountered and Error Handling

Since the data from all of the years don't have the same format, some tedious reformatting was required. While using the Selenium, several cautions were taken to avoid unexpected errors. To ensure that the web application is in a state to execute the Selenium commands, "WebDriverWait" is used. By using this strategy, elements are waited to be clickable. In addition, some buffer time is used to ensure that the file is downloaded before proceeding to the next steps. Currently, the Genesis-Online Database displays a warning pop-up indicating that the website is in beta. To ensure the code remains functional if this warning is removed in the future, the Selenium code for handling this pop-up is kept separate from other parts.

For the other data processing steps, "request" library is used to check whether or not the given URL returns the status 200. The code attempts multiple requests before exiting if the status is not 200.

If the source's link changes, the pipeline should be updated with the new links also.

Results and Limitations

As the result of this pipeline, air emissions data for each economic sector, organized by year, is extracted to the 'data' directory. Each year's data is stored in a separate CSV file because this format is easy to work with when tabular data is used. Storing each year separately makes it quicker for humans to access the information. Since the data is current only up to 2021, it is not the most recent source for interpreting air emissions. Additionally, because the data covers all regions of Germany, a more detailed analysis by individual cities cannot be performed. Despite these limitations, it is still worth working on.