Exploring Agricultural Impacts on Greenhouse Gas Emissions

Agriculture is essential for feeding the world's population and supporting economies. However, it also contributes significantly to greenhouse gas emissions, which drive climate change. The USA, China, and India are the three largest crop producers, and their agricultural practices have a substantial impact on the environment. This project aims to explore the relationship between crop production and greenhouse gas emissions in these countries over a period of 30 years. The key question is: **How does crop production in these countries influence greenhouse gas emissions?**

0.1 Data Sources

0.1.1 Crop Production Dataset from OECD[1]

This dataset provides detailed information on crop production metrics around the world for the years between 1990 and 2030 , focusing on major crops: Wheat, Maize, Rice, and Soybean. It includes data measured in tonnes per hectare, and total production quantities Measured in thousand tonnes. this dataset is in CSV format and can be downloaded from kaggle.

0.1.2 Historical Emissions Gases Dataset

Source: Climate Watch, 2024. Washington, DC: World Resources Institute. This dataset is a compilation of greenhouse gas (GHG) emissions data for all countries for the years between 1990 and 2020 , covering multiple sectors and including emissions of various gases also it is in the CSV formation and can be downloaded from Kaggle. It is provided under the Creative Commons CC BY 4.0 license, permitting unrestricted reuse with proper attribution. The crop production dataset is provided by OECD , which generally allows usage for academic purposes [1].

These datasets were chosen because they provide comprehensive and relevant information to study the impact of crop production on greenhouse gas emissions over time. The crop production data offers detailed metrics on key crops, while the emissions data covers the necessary environmental impact metrics.

Academic Use Disclaimer: The purpose of this project is strictly for academic reasons and not for commercial use. The datasets used in this project are intended to provide educational insights and facilitate learning on the impact of crop production on greenhouse gas emissions.

0.2 Data Pipeline

In this project, we developed a data pipeline utilizing **Python** and the data manipulation library, **Pandas** also it written in **OOP**. The primary objective was to process the Historical Emissions Gases Dataset, transform it for further analysis, then apply a similar procedure to the Crop Production Dataset. After processing both datasets, we merged the resulting data frames and stored the final merged dataframe into an SQLite database. Our data originated from CSV files, which we meticulously cleaned and reshaped to fit our analysis needs.

Our journey began with reading the emissions data from a CSV file. We had to ensure that the data was correctly interpreted, especially concerning missing values. Some emission values were recorded as ’0.0’, ”, ’0,000’, and ’0’, which needed to be treated as missing data. To handle this, we specified these values as missing while reading the CSV. This step was crucial to prevent any erroneous data from skewing our analysis.

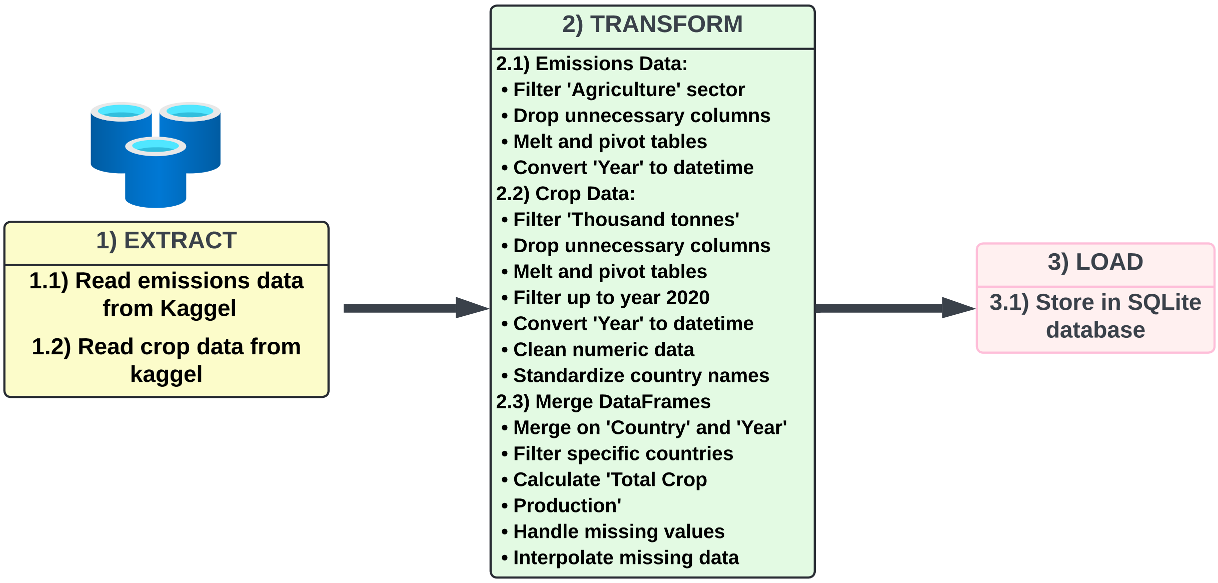
1) EXTRACT
1.1) Read emissions data from Kaggel
1.2) Read crop data from kaggel
2) TRANSFORM
2.1) Emissions Data:
• Filter 'Agriculture' sector
• Drop unnecessary columns
• Melt and pivot tables
• Convert 'Year' to datetime
2.2) Crop Data:
• Filter 'Thousand tonnes'
• Drop unnecessary columns
• Melt and pivot tables
• Filter up to year 2020
• Convert 'Year' to datetime
• Clean numeric data
• Standardize country names
2.3) Merge DataFrames
• Merge on 'Country' and 'Year'
• Filter specific countries
• Calculate 'Total Crop
• Production'
• Handle missing values
• Interpolate missing data
3) LOAD
3.1) Store in SQLite database

Figure 1: ETL Diagram of the project

0.2.1 Transformation and Cleaning Steps

For the emissions data, only rows where the `Sector` is `Agriculture` are kept to focus on relevant data. Unnecessary columns like `Unit`, `Data source`, `ISO`, and `Sector` are removed to simplify the dataset. For the crop data, only rows with `Unit` as `Thousand tonnes` are retained to ensure consistency in measurement. Columns like `Unnamed: 44` to `Unnamed: 46` are removed to eliminate irrelevant data.

The emissions dataframe is transformed from a wide to a long format using `melt` to standardize the data structure, making it easier to analyze. It is then pivoted to have each gas as a separate column, facilitating focused analysis on specific emissions. Similarly, the crop data is melted and pivoted to organize crop production values into separate columns for each crop, which standardizes the dataset for analysis.

The `Year` column in both dataframes is converted to `datetime` format to handle dates correctly and enable time-based operations. Crop data is filtered to include only rows up to the year 2020 to ensure the dataset is up-to-date.

Non-numeric characters in crop production values are removed, and these columns are converted to `float` to ensure numerical operations can be performed accurately. Zero values in the `Wheat` column are replaced with NaNs to handle missing data correctly, preventing zero values from skewing the analysis. Country names are standardized, like replacing China (People's Republic of) with China, to ensure consistency in naming conventions.

0.2.2 Problems Encountered and Solutions

Issues such as missing data and inconsistent data formats were addressed by replacing zero values with NaNs, using interpolation, cleaning non-numeric characters, and standardizing names. Handling different units in the crop data was managed by filtering to include only rows with Thousand tonnes.

0.2.3 Error Handling and Adaptability

The pipeline checks for necessary columns, raises errors if columns are missing, and manages missing data by converting values to NaNs and using interpolation.

0.2.4 Final Data Preparation

After cleaning and transforming the data, the emissions and crop production data are merged on Country and Year. Remaining missing values are interpolated using time-based methods to ensure a continuous and robust dataset for further analysis and modeling.

0.3 Result and Limitations

The final dataset from the pipeline includes agricultural emissions and crop production data, merged and stored in an SQLite table named merged_crop_emission.

```
#   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Year                   124 non-null    datetime64[ns]
 1   Country                124 non-null    object
 2   All GHG                124 non-null    float64
 3   CH4                    124 non-null    float64
 4   N2O                    124 non-null    float64
 5   Maize                  124 non-null    float64
 6   Rice                   124 non-null    float64
 7   Soybean                124 non-null    float64
 8   Wheat                  124 non-null    float64
 9   Total Crop Production  124 non-null    float64
dtypes: datetime64[ns](1), float64(8), object(1)
```

The data is structured in a table with nine columns, each representing a specific metric of emissions or crop production. The dataset covers multiple countries over 30 years (1990-2020), providing a comprehensive view of agricultural impacts on greenhouse gas emissions and crop yields. The data quality is high, with extensive cleaning to remove inconsistencies and standardize formats. Missing values are handled using interpolation to ensure continuity and accuracy.

Overall, the pipeline effectively processes and integrates emissions and crop production data, providing a robust dataset for analysis. Despite potential issues, the data structure and quality are suitable for examining agricultural impacts on emissions, but findings should consider the identified limitations.

Bibliography

[1] "OECD (2024), Crop production (indicator). doi: 10.1787/49a4e677-en (Accessed on 31 May 2024)." [Online]. Available: https://data.oecd.org/agroutput/crop-production.htm#indicator-chart