Data Report

Project Title: Correlation between Amazon deforestation, Forest fires and CO_2 emissions in Brazil.

Main Question

How does deforestation in the Amazon correlate with forest fires and CO_2 emissions in Brazil?

Introduction

The Amazon rainforest, the largest on Earth and often referred to as the planet's lungs, plays a crucial role in regulating global carbon dioxide levels. However, over time, extensive deforestation has severely disrupted its ecosystem, making it increasingly susceptible to forest fires. These fires not only devastate the forest but also contribute significantly to CO_2 emissions. This project aims to explore the correlation between deforestation, forest fires, and CO_2 emissions in Brazil (specifically in the states within the Amazon Rainforest Basin), using statistical analysis and data visualization techniques.

Data Sources

1. Datasource 1: Amazon_Deforestration
Metadata URL: Deforestation in Amazon (Brazil) Metadata
Data URL: Deforestation in Amazon (Brazil) Data
Data Type: CSV
Description: This dataset provides annual deforestation data from 2004 to 2019 for Brazil (Amazon), sourced from PRODES using satellite imagery.
Data Structure and Quality: The dataset is structured as a CSV directory, with separate files containing data for different variables in tabular format (CSV files). It reflects data from reliable Brazilian government sources, ensuring its accuracy in representing real-world conditions. The dataset includes all the necessary information, such as Year, State, and Deforestation Area. It is of high quality, consistently formatted, and suitable for detailed analysis.
License and Obligation: The dataset is available under the CC0 Public Domain, allowing unrestricted use for both non-commercial and commercial purposes without copyright. For full details, click on License.

2. Datasource 2: Burned_Area (Forest Fires)
Metadata URL: Burned Area in Brazil Metadata
Data URL: Burned Area in Brazil Data
Data Type: Zipped CSV
Description: This dataset provides monthly data on burned areas from 2002 to 2023, categorized by landcover classes in all the countries including Brazil (Amazon). It allows for

an analysis of forest fires (Burned Area) in the Amazon, offering insights into the correlation between deforestation and forest fires.
Data Structure and Quality: The data is provided as zipped CSV files, structured with columns for Year, Month, Country, State, and Burned Area/type of land. It is comprehensive, high-quality, and formatted consistently.
License and Obligation: The dataset is available under the Creative Commons Attribution 4.0 International (CC BY 4.0) license (see the second page of the PDF for details). This license allows reuse with proper credit and disclosure of any modifications. This reuse policy is implemented by the European Commission. To comply with the obligations, appropriate credit will be given to GWIS along with a link to the license.

3. Datasource 3: Carbon_Emissions
Metadata URL: CO₂ emissions in Brazil (Amazon Forest) Metadata
Data URL: CO₂ emissions in Brazil (Amazon Forest) Data
Data Type: Zipped CSV
Description: The dataset provides monthly emissions (Tons) data from 2002 to 2023, categorized by pollutant and covering all countries including Brazil.
Data Structure and Quality: The data is provided as zipped CSV files, structured with columns for Year, Month, Country, and emissions/pollutants. It is comprehensive, high-quality, and formatted consistently.
License and Obligation: Same as for Datasourse 3.

Data Pipeline

The automated data pipeline is build on ETL format which includes the following stages:

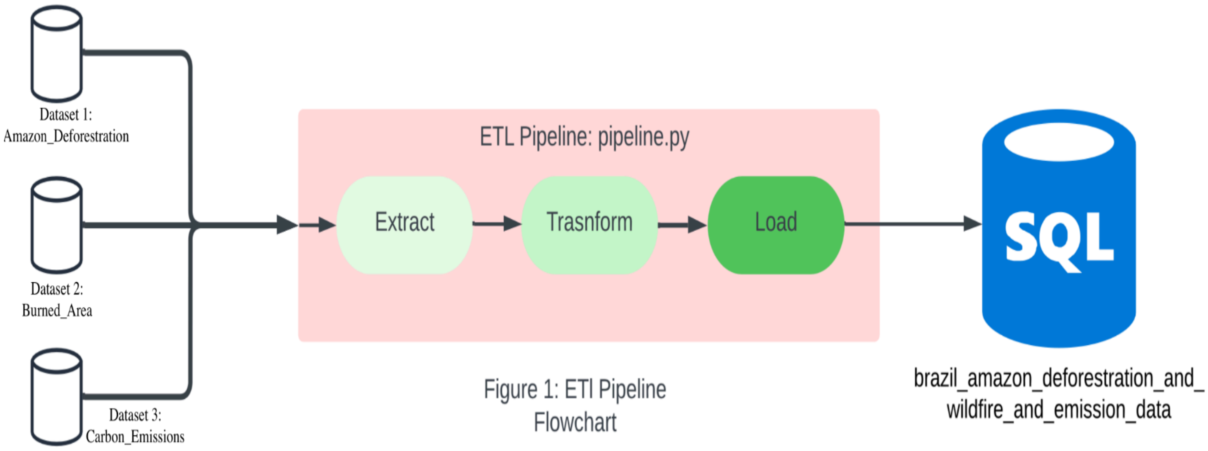

Figure 1: ETl Pipeline Flowchart

Extraction: The data extraction process involves downloading datasets on Brazilian Amazon Rainforest degradation (from a Kaggle dataset by Mariana Boger Netto) and wildfire data (from the Global Wildfire Information System). These datasets are in CSV/Zipped CSV format, which are then loaded into pandas dataframes for analysis.

Transformation: The transformation stage prepares the datasets for analysis by ensuring they are clean, consistent, and ready for integration. Data is loaded from CSV files into pandas DataFrames, and missing values are removed. ***Irrelevant Columns*** are dropped, such as region-specific ones for deforestation and emissions data. ***Renaming Columns*** are done for clarity, e.g., 'Ano/Estados' to 'Year' and 'AMZ LEGAL' to 'Deforested Area.' ***Handling Zero Values*** in critical columns, like 'Forest_BA' and 'CO2,' are excluded. The datasets are filtered to include only Brazilian data and records from 2004–2019 to align temporal coverage. ***Aggregating Monthly Data*** for wildfire and emissions datasets is done to yearly totals, and ***Data Type Consistency*** is ensured for compatibility during integration. These transformations ensure seamless integration and accurate analysis.

Loading: The final stage involves loading the transformed data into a structured SQL database named “brazil_amazon_deforestration_and_wildfire_and_emission_data”, ensuring efficient storage for easy retrieval and analysis in subsequent stages of the project.

Problems Encountered, Solutions, and Error Handling

During the transformation process, the datasets exhibited inconsistencies such as redundant columns, zero values, and non-standard naming conventions. To address these issues, the following steps were taken: filtering the data for relevant years (2004–2019) and Brazil-specific records, dropping unnecessary columns, renaming columns for clarity, and aggregating monthly data into yearly totals. Subsequently, validation checks using Pandas were performed to ensure correct data types and confirm the absence of missing or invalid values.

Result and Limitations

Data Output: The final pipeline output is stored in the “amazon_merged_data” table, integrating “Burned Area (Forest_BA)”, “CO_2 emissions (CO2)”, and “deforestation” data (Deforested Area), all aligned by year and country. The data is tabular, with consistent data types, and the 'Year' column aligns across datasets, with numeric measurements for easy comparison.

Data Quality: The datasets are highly accurate, with consistent records from 2002 to 2023, and are complete with no gaps. Data consistency is maintained through standardized formats for each country, while unique records ensure reliable analysis without duplication. The datasets are directly relevant to the project’s objectives. The data is cleaned, aggregated, and free of missing or zero-value rows, with temporal consistency maintained between 2004 and 2019. Values are aligned across datasets for reliable correlation analysis, and column names are standardized for clarity and ease of use.

Limitations: The datasets are of high quality with no missing values, but the Burned Area in Brazil dataset contains zero values for burned area attributes. The final dataset is aggregated to annual data for 2004–2019, which limits the number of rows available for correlation analysis and reduces the validity of the findings. Aggregating the data to yearly totals also removes the ability to analyze finer temporal patterns, such as seasonal or monthly variations in wildfire activity and CO_2 emissions. Additionally, the dataset may underreport economic and human impacts due to incomplete records or reporting discrepancies, limiting the comprehensiveness of the analysis.