Project 4 Report: Relationship Between Deforestation and CO2 Emissions

Nasim Attareizy

June 6, 2024

1 Introduction

The central question of this project is: "How do deforestation rates (Net forest conversion) correlate with CO2 emissions in various countries?"

Initially, datasets were selected without a full understanding of the project's requirements, leading to a misalignment with the actual topic. The original dataset did not adequately capture the information about climate change, focusing on other aspects which was not related to the study.

2 Data Sources

2.1 Deforestation Data

The deforestation data is sourced from the Deforestation and Forest Loss dataset on Kaggle. This dataset includes annual change in forest area for various countries.

2.2 CO2 Emissions Data

The CO2 emissions data is sourced from the CO2 Emissions by Country dataset on Kaggle. This dataset provides annual CO2 emissions for different countries.

2.3 Data Structure and Quality

Both datasets are in CSV format and contain structured data. The deforestation dataset includes columns such as Country, Year, and Net forest conversion. The CO2 emissions dataset includes columns such as Country, Year, and CO2 Emissions (kt). The data quality is generally good, but some inconsistencies and missing values were addressed during data cleaning.

2.4 Licenses

The deforestation dataset is publicly available and does not have specific licensing restrictions. The CO2 emissions dataset is provided by the Ulrik Thyge Pedersen and is available under the World Bank's open data terms of use. Both datasets are freely usable for this project.

3 Data Pipeline

3.1 Overview

The data pipeline involves several steps:

- **Loading Data**: The datasets are loaded from CSV files.
- **Data Cleaning**: Renaming columns for consistency, filtering relevant columns, and merging datasets based on country and year.
- **Visualizing Data**: Displaying the cleaned and merged data to analyze the correlation.

3.2 Technology Used

The pipeline is implemented using Python, with libraries such as pandas for data manipulation, sqlite3 for database operations, and matplotlib and seaborn for data visualization.

3.3 Challenges and Solutions

One challenge was ensuring consistency in country names and year formats between the two datasets. This was addressed by careful column renaming and filtering. Another challenge was handling missing or inconsistent data, which was managed by inner merging the datasets to include only the common records.

3.4 Error Handling

The pipeline is designed to handle errors by checking for missing values and inconsistencies. Any records with missing critical data were excluded to maintain the integrity of the analysis.

Dataset	Source	Description
Deforestation Data	Kaggle	Annual change in forest area for various countries.
CO2 Emissions Data	Kaggle	Annual CO2 emissions for different countries.

Table 1: Data Sources Overview

4 Results and Limitations

4.1 Output Data

The output data is a merged dataset containing information on both deforestation and CO2 emissions for each country and year.

4.2 Data Structure and Quality

The merged dataset maintains a structured format with columns for Country, Year, Net Forest Conversion, and CO2 Emissions (kt). The data quality is improved through cleaning steps, though some limitations remain due to the original data sources.

4.3 Data Format

The output data is stored in a database format, chosen for its simplicity and ease of integration with analysis tools.

4.4 Correlation Analysis

To quantify the relationship between deforestation and CO2 emissions, the Pearson correlation coefficient was calculated. The result was a correlation coefficient of -0.676.

4.5 Interpretation

A correlation coefficient of -0.676 suggests a moderate to strong negative correlation between deforestation and CO2 emissions. This indicates that, in general, as deforestation increases, CO2 emissions tend to decrease and vice versa. This negative correlation might seem counterintuitive, as deforestation is often associated with increased CO2 levels due to the loss of trees that absorb CO2. However, this result could reflect the complexity of factors at play, such as different countries' stages of development, energy consumption patterns, and deforestation practices.

4.6 Limitations

- **Data Quality**: The data may have missing or inconsistent entries that could affect the analysis.
- **Temporal Alignment**: The datasets may not cover the same time periods for all countries, leading to potential biases.
- **Other Factors**: The relationship between deforestation and CO2 emissions can be influenced by other factors such as economic activities, policy changes, and natural events.