Project Plan

Analysis of CO Emissions in the Americas Over the Past Two Decades

Florian Merlau

January 7, 2025

1 Introduction

1.1 Research Question

How have CO emissions changed over the last two decades in North and South America, and which countries are contributing the most to these changes?

2 Data Sources

2.1 Description of Data Sources

The data originates from the *World Development Indicators (WDI)* database provided by the *World Bank*. This source was selected for its comprehensive and reliable information on economic, social, and environmental indicators.

The data includes, among others:

- **Economy**: GDP, trade balances.
- **Society**: Education and health data.
- **Environment**: CO_2 emissions, energy consumption.
- **Metadata**: Definitions, units of measurement, sources.

This dataset was chosen because it includes all relevant global information, particularly data on CO_2 emissions worldwide. The data source consists of multiple datasets.

Additional Details on the Datasets:

- **Data URL**: `https://datacatalog.worldbank.org/search/dataset/0037712`
- **Data Format**: CSV, Excel

2.2 Licensing Information

Licensing:

The data is licensed under the *Creative Commons Attribution 4.0 International License (CC-BY 4.0)*, allowing public access and usage under open data standards.

The licensing terms set by the World Bank include:

- **Attribution**: Properly credit the data source.
- **Usage**: Data is used solely for statistical and research purposes.
- **No Redistribution**: Redistribution or sale of the data requires prior written agreement from the World Bank.

Further details are available at: `https://datacatalog.worldbank.org/public-licenses`.

2.3 Data Structure and Quality

The data is divided into multiple datasets:

- **WDICSV**: Core dataset with country indicators (1960–2023).
- **WDICountry**: Metadata about countries (e.g., region, income group).
- **WDISeries**: Metadata about indicators (definitions, aggregations).

The data structure follows a dimensional model: WDICSV as the fact table, with countries and indicators as dimensions. The quality reflects typical challenges of real-world datasets, particularly in terms of gaps and sparsity.

3 Data Pipeline

3.1 Technology Used

The data pipeline was implemented using Python, leveraging libraries such as `pandas` for data processing and `matplotlib` for visualization.

3.2 Steps for Data Transformation and Cleaning

1. Removal of irrelevant columns.
2. Normalization of date formats.
3. Handling of missing values through imputation or removal.
4. Aggregation of data by year and country.

3.3 Challenges and Solutions

The data processing faced challenges such as missing data for certain years or countries and inconsistent country codes. Missing data was handled using imputation techniques or excluded when estimation was not feasible. Inconsistent country codes were resolved by matching the data to a standardized list and correcting discrepancies.

4 Results and Limitations

4.1 Output Data

Data Structure: The dataset consists of 326,128 rows and 68 columns. Metadata includes *Country Name*, *Country Code*, *Indicator Name*, and *Indicator Code*. Time-series data spans 1960–2023 and is stored as *float64*.

Data Quality:

- *Completeness:* Many missing values, especially in early years.
- *Consistency:* Uniform structure and correct data types.
- *Accuracy:* Extreme outliers and wide value ranges.
- *Timeliness:* Data up to 2023, but declining coverage.
- *Usability:* Well-structured but requires preprocessing.

4.2 Output Format

The output format is SQLite, chosen for its efficiency in handling structured data, support for complex queries, and compatibility with data analysis tools, making it ideal for storing annual CO emissions per country.

4.3 Critical Reflection

Despite high data quality, limitations include gaps due to missing original data and comparability issues caused by differing collection methods. Future work could integrate additional sources to validate the analysis and improve robustness.