**Documentation**

**Activity 1**

**Data Screening, Manipulation, and Normalization Documentation**

**Objective:** This documentation aims to transparently describe the steps undertaken for screening, manipulating, and normalizing the datasets used in the analysis of the relationship between CO₂ levels and global temperature anomalies. This ensures reproducibility and maintains the integrity of the original datasets while allowing accurate scientific analysis.

**1. Original Datasets**

The following publicly available datasets were used:

1. **NOAA Mauna Loa CO₂ Data:**
   * Source: NOAA Global Monitoring Laboratory
   * Original File: co2\_mm\_mlo.csv
   * Description: Monthly CO₂ concentrations measured at Mauna Loa Observatory.
   * Temporal Range: 1958–2024
   * Units: CO₂ in parts per million (ppm).
2. **NASA GISTEMP Temperature Anomalies:**
   * Source: NASA Goddard Institute for Space Studies (GISS)
   * Original File: GLB.Ts+dSST.csv
   * Description: Global temperature anomalies relative to the 1951–1980 baseline.
   * Temporal Range: 1880–2024
   * Units: Temperature anomalies in °C.

**2. Data Screening**

Before analysis, both datasets were inspected for structure, completeness, and potential issues:

1. **CO₂ Data Screening:**
   * **Issues Identified:**
     + Metadata rows were present at the top of the file.
     + Placeholder values (-9.99) were used to indicate missing or invalid data.
   * **Actions Taken:**
     + Removed metadata rows using the skiprows parameter.
     + Replaced placeholder values with NaN to indicate missing data.
     + Aggregated monthly data into annual averages to align with the temperature dataset.
2. **Temperature Data Screening:**
   * **Issues Identified:**
     + Multi-row headers caused misalignment of column names.
     + Data included multiple metrics (e.g., monthly anomalies, annual means, etc.).
     + Some values were stored as strings instead of numeric.
   * **Actions Taken:**
     + Skipped header rows using the skiprows parameter.
     + Extracted the "Year" and "J-D" (annual mean temperature anomalies) columns.
     + Converted all numeric columns to the appropriate data type.
     + Removed non-numeric rows and missing values.

**3. Data Manipulation**

The following steps were performed to prepare the datasets for analysis:

1. **CO₂ Dataset:**
   * Renamed the "Average CO2" column to "CO2" for clarity.
   * Aggregated monthly values into annual averages by grouping by year.
   * Dropped rows with missing CO₂ data to ensure consistency.
2. **Temperature Dataset:**
   * Renamed the "J-D" column to "Temperature\_Anomaly" for clarity.
   * Dropped rows with missing temperature anomalies to ensure complete data.
3. **Merging Datasets:**
   * Both datasets were merged on the "Year" column using an inner join, retaining only years present in both datasets.

**4. Normalization**

Normalization was applied to facilitate visual comparison of trends in CO₂ levels and temperature anomalies, given their vastly different scales:

A math formula with black text

AI-generated content may be incorrect.

* **Variables Normalized:**
  + **CO₂ Levels (ppm)**
  + **Temperature Anomalies (°C)**
* **Reasoning:**
  + Normalization scales both variables to the same range (0–1), making trends visually comparable without altering their intrinsic patterns or relationships.

**5. Transparency and Data Integrity**

* **Original Values Preserved:**
  + The raw CO₂ and temperature values were neither altered nor interpolated.
  + All transformations (e.g., normalization) were for visualization purposes only.
* **Documentation in the Manuscript:**
  + All manipulations (e.g., column renaming, aggregation, normalization) are described in the methodology section of the manuscript.
* **Access to Cleaned Datasets:**
  + Cleaned datasets (cleaned\_co2\_data.csv and cleaned\_temperature\_anomalies.csv) are stored permanently in a public repository or Google Drive, ensuring accessibility for independent verification and future research.

**6. Summary**

The data manipulation and normalization steps were necessary to:

* Ensure consistency and compatibility between datasets.
* Facilitate accurate scientific analysis and visualization.
* Maintain the integrity of the original data by clearly documenting all transformations.

**Activity 2**

To ensure the integrity and transparency of the data preprocessing workflow, the following steps document how the original dataset ("GLB.Ts+dSST.csv") from NASA GISTEMP was transformed into the cleaned dataset ("cleaned\_monthly\_temperature\_anomalies.csv"). These steps align with the objective of **Activity 2: Creating a Heatmap to Analyze Temperature Anomalies Across Years and Months**.

**1. Objective**

The preprocessing steps were undertaken to create a dataset suitable for visualizing **monthly temperature anomalies** across years. The aim is to provide students with a clear and meaningful representation of temperature patterns over time using a heatmap.

**2. Original Dataset: GLB.Ts+dSST.csv**

* **Source**: NASA GISTEMP
* **Structure**:
  + **Rows**: Each row corresponds to a year from 1880 to the most recent year.
  + **Columns**:
    - **Year**: The year for each data row.
    - **Months (Jan–Dec)**: Monthly temperature anomalies (relative to the 1951-1980 baseline).
    - **Annual and Seasonal Means**:
      * **J-D**: January to December annual mean.
      * **D-N**: December to November mean.
      * **DJF, MAM, JJA, SON**: Seasonal means (e.g., DJF = December, January, February).

**3. Preprocessing Steps**

**a. Loading the Original Dataset**

The dataset was loaded using Pandas, and all columns were preserved during the initial inspection to maintain data integrity.

**b. Dropping Unnecessary Columns**

* Columns unrelated to the monthly anomalies, such as:
  + **Annual Means** (J-D, D-N)
  + **Seasonal Means** (DJF, MAM, JJA, SON)
* These columns were removed to focus exclusively on monthly anomalies (Jan–Dec) for the heatmap.

**c. Handling Missing Values**

* Missing or placeholder values (e.g., "NaN") were checked and verified to be minimal.
* Rows with missing monthly anomaly data were removed to ensure a complete dataset for visualization.

**d. Ensuring Proper Data Types**

* **Year**: Ensured to remain as int64.
* **Monthly Temperature Anomalies** (Jan–Dec): Converted to float64 for numerical calculations and visualizations.

**e. Consistency Check**

* Verified that the data accurately represents temperature anomalies relative to the 1951–1980 baseline.
* Cross-referenced values in the cleaned dataset with the original NASA GISTEMP dataset to ensure no data distortions.

**f. Saving the Cleaned Dataset**

* The cleaned dataset was saved as "cleaned\_monthly\_temperature\_anomalies.csv" for further analysis and heatmap generation.

**4. Rationale for Preprocessing**

1. **Educational Objective**:
   * The cleaned dataset focuses on monthly anomalies, aligning with the activity's objective to create a heatmap showing monthly temperature variations across years.
2. **Data Integrity**:
   * Preprocessing steps, such as removing unrelated columns and handling missing values, ensured the data remained accurate and undistorted.
3. **Simplified Analysis**:
   * Removing extraneous columns allowed for a more focused and accessible analysis, particularly for students new to data visualization.

**5. Resulting Dataset: cleaned\_monthly\_temperature\_anomalies.csv**

* **Rows**: 145 (corresponding to years from 1880 to 2024).
* **Columns**: 13
  + **Year** (1880–2024).
  + **Monthly Temperature Anomalies** (Jan–Dec) as floating-point values.
* **No Missing Values**:
  + Verified that all monthly anomalies are available for each year.

**6. Future Accessibility**

To ensure academic transparency and reproducibility, the cleaned dataset, "cleaned\_monthly\_temperature\_anomalies.csv," will be:

* Hosted in a **permanent location** (e.g., Google Drive or repository).
* Accompanied by this preprocessing documentation for context and proper citation.

By documenting this workflow, we uphold the integrity of the original dataset while meeting the educational needs of **Activity 2**. Let me know if additional details or refinements are needed!

**Activity 3**

**Original Dataset Information:**

* **Source URL**: <https://raw.githubusercontent.com/owid/co2-data/master/owid-co2-data.csv>
* **Description**: The dataset, provided by Our World in Data, contains global and country-level CO₂ emissions data across multiple years. It includes emissions from various sources (e.g., coal, oil, gas), emissions per capita, cumulative emissions, and other socio-economic indicators (e.g., population, GDP).
* **Total Columns**: 79
* **Total Rows**: 50,191

**Data Screening and Preprocessing:**

To ensure the dataset meets the objective of analyzing carbon emissions by region and identifying the top-emitting countries, the following steps were performed:

1. **Data Loading**:
   * The dataset was downloaded from the provided URL.
   * The raw file (owid\_co2\_data.csv) was loaded into Python using Pandas for inspection.
2. **Initial Inspection**:
   * Checked the dataset structure, including column names, data types, missing values, and summary statistics.
   * Verified the key columns required for the analysis:
     + country: Name of the country or region.
     + year: Year of the emission record.
     + co2: Total CO₂ emissions (in million metric tons).
3. **Filtering Relevant Data**:
   * Filtered the dataset to include only the most recent year available (year = 2023) to focus on the latest emissions data.
   * Excluded non-country entries (e.g., "World," "Europe (GCP)").
   * Dropped unused columns to reduce the dataset size and focus on relevant variables:
     + Retained only country, year, and co2 columns.
4. **Handling Missing Values**:
   * Verified that the filtered dataset for the year 2023 did not have missing values in the co2 column for the selected countries.
   * No imputation was required, as the subset for 2023 had complete records for top emitters.
5. **Sorting and Selecting Top Emitters**:
   * Sorted countries based on total CO₂ emissions in descending order.
   * Selected the top 10 emitting countries for visualization purposes.
6. **Normalization (Not Applied)**:
   * No normalization was performed as the objective was to visualize absolute emissions by country for the year 2023.

**Ensuring Data Integrity:**

1. **Preservation of Original Data**:
   * The raw dataset was not modified directly. All transformations were performed programmatically, and the processed dataset is stored separately as owid\_co2\_data.csv.
   * Original column names and data values were preserved for the selected columns.
2. **Transparency**:
   * The processed dataset reflects only the filtered and sorted results based on the original data for the year 2023.
   * There is no distortion or alteration of the original numerical values.

**Processed Dataset Information:**

* **File Name**: owid\_co2\_data.csv
* **Columns Included**:
  + country: Name of the country.
  + year: Year of the record (filtered to 2023).
  + co2: Total CO₂ emissions (million metric tons).
* **Total Rows**: 10 (Top 10 emitting countries for 2023).

**Objective Alignment:**

The processing steps align with the objective of the activity by ensuring:

1. The analysis focuses on the most recent data to reflect current emission trends.
2. The visualization of top emitters highlights global disparities in emissions, reinforcing the need for international collaboration in addressing climate change.

**Activity 4**

**Purpose**

The purpose of this documentation is to ensure transparency, reproducibility, and integrity in the data preprocessing workflow for Activity 4, which involves building a predictive model of future CO₂ concentration levels based on historical NOAA data. The process was designed to clean and structure the original dataset (co2\_mm\_mlo.csv) into a usable format for analysis and modeling (cleaned\_co2\_modeling\_activity\_4.csv) while maintaining consistency with the original data source.

**Initial Dataset: co2\_mm\_mlo.csv**

* **Source**: NOAA Global Monitoring Laboratory's Mauna Loa CO₂ Monthly Mean Data.
* **Description**:
  + The dataset contains monthly measurements of CO₂ concentrations in ppm.
  + Includes raw, seasonally adjusted, and trend data along with uncertainties.
  + Metadata and headers were present in the first 57 rows, requiring parsing adjustments.

**Data Screening**

1. **Parsing Issues**:
   * The original dataset included metadata, comments, and headers at the beginning, causing parsing errors during loading.
   * Rows and columns were misaligned due to mixed formatting (e.g., missing column headers in the raw data).
2. **Structure Analysis**:
   * Columns were identified as:
     + Year, Month, Decimal Date, Average CO₂, Interpolated CO₂, Trend CO₂, and Uncertainty CO₂.
3. **Missing Data**:
   * Missing values were flagged as -9.99 and -0.99 in several columns (standard placeholders in NOAA data).
   * These values were replaced with NaN to handle them effectively during analysis.

**Data Manipulation**

1. **Column Naming**:
   * Explicit column names were assigned to ensure clarity and consistency:
     + Year, Month, Decimal Date, Average CO₂, Interpolated CO₂, Trend CO₂, and Uncertainty CO₂.
2. **Handling Missing Values**:
   * Missing or placeholder values (-9.99, -0.99) in the Uncertainty CO₂ column were replaced with NaN.
   * Rows with critical missing values in Interpolated CO₂ were retained, as they were not found to contain NaN.
3. **Data Type Conversion**:
   * Numeric columns were converted to appropriate data types (e.g., Year as int, Interpolated CO₂ as float) for accurate modeling.
4. **Aggregation**:
   * Monthly data was aggregated to compute the annual mean of Interpolated CO₂ using a group-by operation on the Year column.

**Final Dataset: cleaned\_co2\_modeling\_activity\_4.csv**

* **Structure**:
  + The final dataset contains two columns:
    - Year: Integer values representing the year.
    - Interpolated CO₂: Annual mean CO₂ concentration in ppm.
  + Example:
  + Year Interpolated CO₂
  + 0 1959 315.87
  + 1 1960 316.91
  + 2 1961 317.64
  + ...
* **Improvements**:
  + Removed unnecessary columns and metadata.
  + Ensured consistency and usability for predictive modeling.

**Quality Assurance**

1. **Validation**:
   * Ensured that the processed dataset matches the original values for corresponding rows and columns.
   * Confirmed the accuracy of aggregation (monthly to annual).
2. **Documentation**:
   * Clearly documented all transformations and cleaning steps for reproducibility.

**Key Considerations**

* **Data Integrity**:
  + All transformations preserved the original dataset's accuracy and meaning.
  + No distortions were introduced during preprocessing.
* **Reproducibility**:
  + The cleaning script and resulting dataset are ready for validation and reuse by other researchers or students.

This documentation ensures that the dataset prepared for Activity 4 aligns with the objective of modeling future climate scenarios and maintains the integrity of the original NOAA data.