**Data Cleaning and Transformation Documentation**

**1. Project Overview:**

Data cleaning and transformation is an essential step in any data analysis project. Raw datasets often contain inconsistencies, missing values, or duplicate records that can compromise analysis results. This project focuses on cleaning and preparing **two datasets** for analysis:

1. **Dataset Deforestation** – covering data such as country-level measurements, thresholds, area, and tree cover loss.
2. **CO₂ Emissions Dataset** – covering annual CO₂ emissions data for various countries.

The main objective is to produce **clean, consistent, and analysis-ready datasets** that accurately reflect the information without errors or inconsistencies.

**2.Dataset Descriptions:**

**Dataset Deforestation:**

This dataset contains records of environmental measurements for multiple countries. Key columns include:

* country – the name of the country or region
* threshold, area\_ha – area-related measurements
* extent\_2000\_ha, extent\_2010\_ha – coverage metrics over time
* tc\_loss\_ha\_2001–2012 – tree cover loss over several years

**Observed Issues:**

* Duplicate entries for the same country/region
* Missing or incomplete entries for smaller countries
* Inconsistent formatting of country names and numeric values

**CO₂ Emissions Dataset:**

This dataset contains annual CO₂ emissions for multiple countries. Key columns include:

* country – country name
* year – the year of emission measurement
* CO2\_emissions – emission value in metric tons

**Observed Issues:**

* Duplicate records for the same country-year combination
* Missing or blank country entries, particularly for smaller countries
* Inconsistent or missing data affecting longitudinal analysis.

**3. Data Cleaning Methodology**

The cleaning process ensures the datasets are accurate, reliable, and ready for further analysis. The following steps were applied theoretically:

1. **Removal of Duplicate Records**
   * Duplicate entries can distort analysis results.
   * For the general dataset, duplicates were identified and removed based on country names.
   * For CO₂ emissions, duplicates were identified based on **country-year combinations**, ensuring each record represents a unique measurement.
2. **Handling Missing Values**
   * Missing values, especially for smaller countries, were addressed by removing incomplete rows where country information was absent.
   * This ensures that analyses do not include ambiguous or undefined records.
3. **Standardization of Column Names and Data Types**
   * Column names were standardized for consistency (e.g., country, year, CO2\_emissions).
   * Numeric columns, such as emission values and area measurements, were confirmed as numeric types to support calculations.
4. **Data Transformation for Analysis**
   * For time series or longitudinal analysis, CO₂ emissions data can be aggregated or summarized by year.
   * General dataset metrics were organized to allow trend analysis and comparison across countries.

**4. Challenges and Considerations:**

* **File Accessibility**: Ensuring the dataset could be loaded without path errors.
* **Data Quality**: Dealing with missing country names or incomplete measurements for smaller countries.
* **Duplicate Records**: Identifying true duplicates versus legitimate repeated measures.
* **Large Datasets**: Ensuring that transformations were efficient to handle large records without performance issues.

These challenges required careful attention to avoid introducing errors during cleaning.

**5. Outcomes:**

After data cleaning and transformation:

* **General Dataset**:
  + Duplicate country records removed
  + Missing or blank entries for smaller countries addressed
  + Dataset ready for visualization and further statistical analysis
* **CO₂ Emissions Dataset**:
  + Unique country-year records retained
  + Missing or blank countries removed
  + Dataset ready for trend analysis, regional comparison, and climate research

Both datasets are now **consistent, accurate, and ready for analysis**.