Tutorial 3 - Exploring Data Preprocessing with a Real-World Dataset

Objective

To understand the basics of data preprocessing by applying techniques to a real-world dataset, specifically from your selected **SDG-related project dataset**.

Task

Based on the dataset you selected in **Tutorial 2**, identify and list at least **two data quality issues** you observed during your dataset analysis.

After watching the video **Data Preprocessing for Data Scientists**, reflect on the following four key areas of data preprocessing. https://www.youtube.com/watch?v=VCx0ZI-XcSQ

1. **Data Cleaning** – Did the dataset contain missing values, duplicate records, or inconsistent entries?
2. **Data Integration** – Was there a need to combine data from multiple sources? If yes, what inconsistencies or challenges did you encounter during the merge process?
3. **Data Transformation** – Were there unscaled numerical features, unencoded categorical variables, or inconsistent data formats that needed transformation?
4. **Data Reduction** – Did the dataset include unnecessary columns or high-dimensional data that could be reduced for better efficiency?

Show Your Work

Attach a link to your **Google Colab notebook** that demonstrates the practical steps you took to address **any one or more** of the issues identified above.

Make sure your notebook is **shared publicly or set to "Anyone with the link can view."**

Tip

Your submission may include short explanations, code snippets, or visual outputs (e.g., plots, dataframes) that clearly support your analysis and solution.

Submission link

https://forms.office.com/r/AbP47gBCp7