Data Cleaning and Preprocessing Report

This report summarizes the data cleaning and preprocessing steps applied to the Economic Freedom Index dataset in this notebook.

1. Data Loading and Initial Inspection:

- The dataset was loaded from the provided CSV file (heritage-index-of-economic-freedom-20250825135744.csv).
- Initial inspection using .head() and .info() was performed to understand the data structure, column types, and identify initial missing values.
- Column names were standardized by converting to lowercase and replacing spaces with underscores.

2. Handling Missing Values:

- The extent of missing values for each column was assessed using .isnull().sum().
- Median imputation by country was applied to fill missing values in the independent variables within the training and test sets separately. This approach assumes that the median value for a country is a reasonable estimate for missing data points within that country's time series.
- Rows with missing values in the target variable (overall_score_without_monetary_freedom) were dropped from both the training and test sets.
- Rows with any remaining missing values in the independent variables after country-wise imputation were also dropped.

3. Text Standardization:

- The 'Country' column was standardized by converting all entries to lowercase and removing leading/trailing whitespace to ensure consistency.

4. Feature Engineering (Creation of new target variable):

- A new dependent variable, overall_score_without_monetary_freedom, was created by calculating the mean of all independent economic freedom indicators *excluding* 'Monetary Freedom'. This was done to address the potential tautology of predicting an overall score that includes the predictor variable itself.

5. Data Splitting:

- The dataset was split into training and testing sets based on a temporal split at the beginning of the year 2020. This ensures that the model is evaluated on data from a time period later than the data it was trained on, simulating a real-world forecasting scenario.

Summary of Impact:

These cleaning and preprocessing steps were crucial for preparing the data for regression modeling. Handling missing values allowed us to utilize more of the available data, while standardizing text and column headers ensured consistency. The creation of the new target variable enabled a more focused analysis of the impact of other economic freedom indicators on the overall score, independent of Monetary Freedom. The temporal split ensured a realistic evaluation of the model's predictive performance on unseen future data.