**Project:** Analysis of Global Internet Usage 2000-2023

**Date:** 2025-04-01

**Analyst:** Peter Obikpe

**Data Source:** “Global Internet Usage by Country (2000-2023)” Downloaded from <https://www.kaggle.com/datasets/meleknur/global-internet-usage-by-country-2000-2023>

**Original Data Description:**

* The original dataset contains internet usage data for 217 countries
* It has 218 rows and 26 columns
* Columns include:
* “Country Name” (Text)
* “Country Code” (Text)
* “Years (2000-2023)” (Numeric)

**Data Cleaning Steps**

1. Handling rows with excessive missing data
* Action: Rows that contained missing data greater than 49% of the total data were deleted.
* Rationale: These rows contained insufficient information and could introduce significant data bias.
* Implementation: Used excel conditional format option to highlight cells with missing data greater than 49%, then deleted them.
* Result: 13 rows deleted. 205 rows remain.
1. Handling missing values in “2000 to 2022” column
* Action: Missing values in the “2000 to 2022” columns were imputed using the median of the non-missing values of each column.
* Rational: The missing values in the “2000 to 2022” columns were less than 20% and the column was crucial for analysis, so imputation was chosen over deletion.
* Implementation: Imputed missing values using the excel formula: =IF(ISBLANK(“Year\_Column”$2), MEDIAN(“Year\_Column”$2: “Year\_Column”$205), “Year\_Column”2) (applied to the entire “2000-2022” column.
* Result: All missing values in the “2000-2022” columns were replaced by the median.
1. Handling missing values in “2023” column
* Action: The “2023” column was deleted.
* Rational: The column contained insufficient data as it had 72% missing data and could introduce significant bias, so deletion was chosen over implementation.
* Implementation: Deleted the “2023” column.
* Result: 1 column deleted, 25 columns remain.

**Final Cleaned Dataset:**

* File name: “Internet\_Usage\_Cleaning.xlsx”
* Location: C:\Users\USER\Pictures\Data Analysis project\Internet Usage Cleaning.xlsx
* Changes: Rows with excessive missing values were deleted. Columns with excessive missing values were deleted. Missing values in columns with missing values less than 20% were replaced with the median of non-missing values in each column