**1. Overview**

The dataset analyzed contains information on digital service usage across different districts in Rwanda. The dataset has 1025 entries and 7 columns. The objective of the data cleaning process was to identify and correct data quality issues to prepare a reliable dataset for analysis.

**2. Key Data Issues Found**

**Missing Values:**

* Users\_Reported had **524 missing values**.
* Satisfaction\_Score\_(%) had **515 missing values.**
* Monthcolumn had **88 missing values.**

**Inconsistencies:**

* Variations in the text format of the **District column** (e.g: “Huyex ”, “Huye-”, etc).

**Duplicate Records:**

* **25 duplicate rows** were found in the dataset.

**3. Steps Taken to Clean the Data**

Handled Missing Values:

* Replaced missing values in Users\_Reported and Satisfaction\_Score\_(%) with their respective **median values**, ensuring robust handling against outliers.
* For **Month**, text values were standardized using .str.strip().str.title() and missing months were filled using the most frequent value (mode) within each district group.

**Standardized Text Formats:**

Ensured consistent capitalization and removal of unnecessary spaces in text fields (e.g., “Huyex” instead of “Huye”).

**Removed Duplicates:**

Detected and deleted all **25 duplicate records** to maintain data integrity.

**Validation:**

Rechecked for missing or duplicate values after cleaning to confirm data completeness.

**4. Assumptions Made**

* Median values were considered suitable replacements for continuous numeric variables, assuming the data distributions were not heavily skewed.
* The mode of each district’s month was used as a logical estimate for missing month entries.
* Duplicates were assumed to be accidental re-entries rather than legitimate repeated observations.

**5. Result**

After cleaning, the dataset contained no missing or duplicate values, standardized text formats, and consistent, reliable entries, so it is ready for further analysis and visualization.