**Report on the Greendata Solutions Challenge**

**By Lauretta Ogbenna**

**Project Overview**

**Objective**: To clean and standardize the given dataset to meet industry standards.  
**Dataset Description**: The original dataset contained 1660 rows and 12 columns before cleaning.

**Data Cleaning Steps**

**1. Removing Duplicate Records**

Using the "Remove Duplicates" tool under the Data ribbon in Excel, I identified and removed 3 duplicate records to ensure data accuracy and integrity.

**2. Resolving Column-Wise Inconsistencies**

**Names Column**:

* Corrected inconsistencies in letter casing by applying the PROPER formula.
* Removed unnecessary whitespaces using the TRIM function.
* Used FIND AND REPLACE to remove “Title” (e.g Ms, Mr, Dr)
* Replaced the cleaned version of the column with the original data using COPY AND PASTE (value).

**Age Column**:

* Filled empty age values with the mean age applying AVERAGE function to maintain data completeness.
* Converted all decimal values to whole numbers for uniformity.

**Gender Column**:

* Identified inconsistencies in gender classification (e.g., "M" and "male" for males, "F" and "female" for females).
* Standardized the data to use only "Male" and "Female" using IFS statement function.
* For blanks, I filled them with "Others," as per the guidelines which allowed three gender categories: Male, Female, and Others.

**City Column**:

* I addressed inconsistencies in upper and lower casing using the same method applied to the Names column.
* Standardized abbreviations and full names (e.g short forms and long forms) to ensure uniformity.

**Blood Type Column**:

* Since this is sensitive information, I avoided deleting incomplete data. Instead, I marked the 8 missing values as "Not Specified."

**Educational Degree Column**:

* Deleted this column entirely as it was irrelevant to the medical records being analyzed.

**Employment Status Column**:

* Identified inconsistencies in employment classification (e.g., “Retired-Early”, “Student-Part Time”, "Employed-Gig Work”).
* Standardized the data to use only “Employed, Unemployed and Student" using IFS statement function.
* For “Retired” Individuals, I filled them with “Unemployed” because they are not actively working or contributing to the labor force.

**Salary Column**:

* Cleaned extraneous information, such as employment type, by using the TEXT TO COLUMN feature.
* Filled missing values for unemployed individuals with $0, as they typically have no salary, by using the FIND AND REPLACE feature.

**Health Condition Column**:

* Standardized inconsistencies in health status descriptions (e.g., "Excellent" and "excellent").
* Rows with blank health condition entries was replaced with “Not Specified”, as this column was critical for the analysis.

**Credit Score Column**:

* Removed this column entirely, as it was not relevant to the project’s objectives.

**Date of Admission Column**:

* Filtered the data, removed all “Monday – Sunday”, restructured the column to enhance clarity and usability using the SHORT DATE feature.

**Final Outcome**

After cleaning and processing the dataset, the final version contained 1643 rows and 10 columns. This streamlined version of the data is accurate, consistent, and ready for further analysis.

**Key Insight**

Atlanta is the city in need of higher healthcare based on health condition and admission dates.