Title: Data Cleaning Report

Prepared by:

FATOKI GBEMISOLA

Date: 10th January, 2025

INTRODUCTION

This dataset set is a healthcare data from Green Data Solution, the organizer of the data cleaning challenge. And it contains 1660 rows and 12 columns. The columns are ID, Name, Age, Gender, City, Blood Type, Education, Employment Status, Salary, Health Conditions, Credit Score and Date of admission.

Objectives: the objective of this exercise is to learn how to improve the quality and reliability of the given dataset by identifying and correcting inconsistencies, errors, missing values and duplicates.

DATA CLEANING PROCESS

Methodology

- Missing Value Treatment (Removal)
- Duplicates (Removal)
- Sorting
- Data Type Conversion
- Standardizing capitalization

Tool Used: Excel

FINDINGS

Summary of Issues

- Duplicates (3 rows)
- Missing Values (Gender = 98%, Education = 13%, Health Condition = 3%)
- Unsorted Data
- Inconsistent data entries

Actions Taken

- Corrections Made:

  Duplicates: Conditional formatting was used on the ID column with the format set to unique or duplicates value, after which find and replace (CTRL + H) was used to determine the true duplicate before their removal. **3 rows** were found to be true duplicate.

  Missing Values: the dataset was explored using **=CountA(A9:A1669)** on each column to count the total value in them, the ID was found to be complete, and the missing value was calculated by using the formula **=1 – B2/A2**.

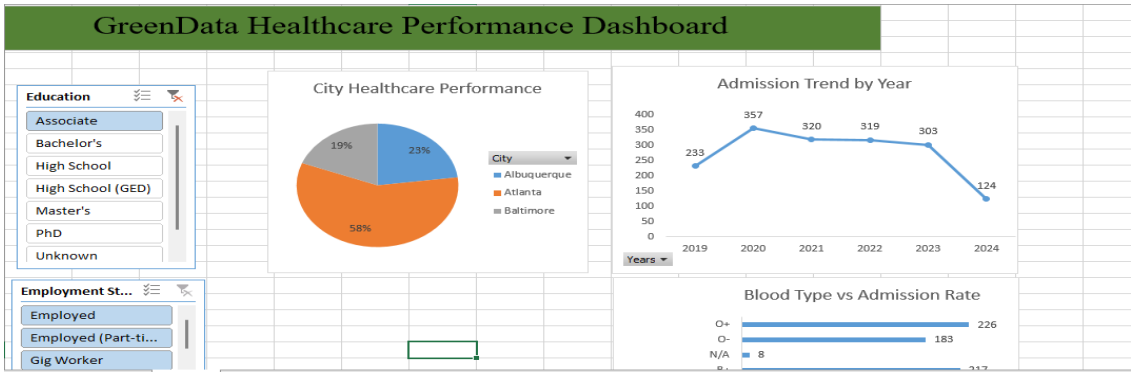
  Inconsistent Data Entries: the dataset has a lot of inconsistencies in its rows and columns, the Name column was treated using **=Proper(B2)** to standardize capitalization all through the column. Age, Salary, Credit Scores which are numbers have text in some rows which was treated using **find and replace**, while the Date of Admission column was treated using the **text-to-column tab** and **=Date(year, month, day).**

- Before the cleaning, statistical analysis could not be done to the Age, Salary, and Credit Score column but after cleaning the columns statistical analysis useful and applicable.

RESULTS

- Cleaned Dataset Overview: The dataset has been properly cleaned with no inconsistencies, error and properly sorted in ascending order, names are properly written, city has no abbreviations and it has been filled to have 3 unique values. All data are in correct format.
    a. The dataset now contains 1657 rows and 11 columns.
    b. with Statistical summary

Sum of Salary	Average of Salary	Average of Age	Sum of Credit Score	Count of ID
69552180	42000.11	32.97886473	782028	1656



RECOMMENDATION

The ID's are not unique to the patients which I believe it was because the dataset was generated from different cities and hospitals, so for a comprehensive analysis, I will suggest that the dataset should include the Hospital where the patients were admitted and the gender columns should be filled.

CONCLUSION

Cleaning data is essential in data analysis as it prevent bias in result and it increases the analyst confidence level, proper cleaning of data improve the quality of analysis.