

## FROM DATA TO DISCOVERY: EXPLORING & UNVEILING HEALTHCARE DATASET

**TASK:** Data Cleaning & Preprocessing.

**DATA SET:** Healthcare Data Report.

**DESCRIPTION:** Working with CSV file of healthcare data set that contains missing values, duplicates and inconsistent data formats.

**TOOLS:** Python, Pandas, VS Code, Jupyter Notebook & Openpyxl.

### **DISCOVERY:**

- Number of Columns in data set: 7.
- Number of Rows: 12,000.
- Number of Duplicate Values 0.
- Number of Empty Cells 0.

### **PROCESS:**

- Understanding the data set.
- Dataset Column header (Hospital Name, Country Name, Global Ranking, Hospital Size, Visibility, Rich Files, Scholarship Scores)
- Renaming all columns in the above dataset
  - Hospital Name (Before): Hospital\_Name (After)
  - Country Name (Before): Country\_Name (After)
  - Global Ranking (Before): Global\_Ranking (After)
  - Hospital Size (Before): Hospital Size (After)
  - Visibility (Before): Digital\_Presence (%) (After)
  - Rich Files (Before): Research\_Score (%) (After)
  - Scholarship Scores (Before): Education\_Support\_Score (%) (After)
- Addition of new columns (i.e. Overall\_Score & Hospital\_Performance) to the above dataset.
- Capitalizing each word of string data type in the above dataset.

**Output GitHub Link:** [https://github.com/AanDevAnalyst/Data\\_Cleaning\\_2/tree/main](https://github.com/AanDevAnalyst/Data_Cleaning_2/tree/main)