Data Cleaning Report – Barbados Health Registry

# 1. Overview

This report summarizes the data cleaning and preprocessing steps applied to the synthetic Barbados health facility registry dataset as part of the Amini Data Science Intern – Barbados assignment.  
The cleaning process aimed to make the dataset analysis-ready by correcting inconsistencies, standardizing formats, and removing noise.

# 2. Summary of Issues Found

- Inconsistent `facility\_id` formatting (case, spaces)  
- Mixed or unclear formats in `capacity` field (e.g., '12 beds', '15+')  
- Inconsistent date formats in `licence\_issue\_date` and `inspection\_date`  
- Geolocation stored as unstructured text (e.g., POINT or comma-separated)  
- Mixed capitalization and spacing in `region` values  
- Duplicate records present in the dataset

# 3. Cleaning Steps Taken

- \*\*Facility ID\*\*: Trimmed whitespace and converted all entries to uppercase.  
- \*\*Capacity\*\*: Extracted numeric digits from text values (e.g., '20 beds' → 20). Stored in a new column `capacity\_clean`.  
- \*\*Dates\*\*: Standardized `licence\_issue\_date` and `inspection\_date` using `pd.to\_datetime` with error handling.  
- \*\*GPS Location\*\*: Extracted latitude and longitude from the `gps\_location` column, handling both `POINT(x y)` and `lat,long` formats.  
- \*\*Region\*\*: Standardized the `region` column to title case and removed leading/trailing spaces.  
- \*\*Duplicates\*\*: Removed exact duplicate rows to avoid inflated analysis.

# 4. Assumptions Made

-Duplicates are dropped using df.drop\_duplicates()

-capacity is cleaned into a numeric column. Extracted numeric values only.

-region is standardized using .str.title() and cleaned to remove variations of "Parish"

-Reversed or corrupted region names (like "WERDNA .TS") are detected and fixed

-GPS coordinates are extracted from multiple formats, including:

* -POINT(long lat)
* -lat,long
* -DMS format like 13°9′5″N 58°58′44″W

-All dates (licence\_issue\_date, inspection\_date) are parsed and standardized to datetime format