Dataset Overview

The National Health Facility Registry dataset is a synthetic compilation of five years of administrative data derived from licensing spreadsheets, inspection forms, and OCR-scanned documents. The dataset is intended to represent all hospitals, clinics, and community health centres within the country.

Due to the varied nature of its source systems, the dataset exhibited multiple data quality issues, necessitating systematic profiling and cleaning before it could be considered suitable for analytics.

Exploratory Data Analysis (Profiling)

Initial exploratory data analysis focused on understanding the dataset's structure, completeness, and reliability.

Dataset Characteristics

- The dataset contains approximately 13,000 records across 9 attributes.
- Most columns are stored as text (object type), including fields that logically represent numeric or date values.
- Key identifying fields such as facility_name and region have high completeness, while others (e.g. faculty_id, capacity, gps_location, remarks) show substantial missingness.

Duplicate Records

- 1672 exact duplicate rows were identified.
- 2034 logical duplicates were detected when grouping by facility_name and region, indicating multiple records referring to the same real-world facility.
- Variations across duplicate records suggest updates from inspections, licensing renewals, or data entry inconsistencies.

Major Data Quality Issues Identified

The following key issues were identified during profiling:

1. Duplicate facility records arising from multiple administrative sources.
2. Inconsistent text formatting, including casing and whitespace issues.
3. Non-numeric values in numeric fields, particularly in the capacity column.
4. Missing or incomplete metadata, especially for location and inspection-related fields.
5. Lack of a fully reliable unique identifier, as facility_id was missing for a significant portion of records.

Cleaning Decisions & Rationale

To address these issues, a reproducible cleaning pipeline was implemented with the following decisions:

- Capacity values were converted to numeric where possible, with non-numeric values retained as missing rather than replaced with misleading defaults.
- Exact duplicates were removed entirely.
- Logical duplicates were resolved using facility_name and region as composite identifiers, retaining the most recent record.
- Missing categorical values (e.g. facility_type) were filled with "Unknown" where appropriate to preserve row integrity.

Final Output

The result of this process is a single tidy dataset (cleaned_health_registry.csv) that:

- Represents one row per facility
- Is free from duplicate records
- Uses standardized, consistent formatting

- Is suitable for analytics, visualization, and AI-based applications

Relevance for Knowledge Graphs & RAG

A cleaned and deduplicated facility registry enables reliable entity resolution, which is essential for constructing health infrastructure knowledge graphs. Standardized facility names, locations, and attributes allow facilities to be represented as nodes with well-defined relationships.

For Retrieval-Augmented Generation (RAG) systems, the cleaned dataset supports accurate retrieval of facility-level information in response to natural language queries such as:

> "Which licensed clinics operate in St. Lucy?"

This demonstrates the dataset's readiness for downstream AI applications.