

ACTION PLAN

**Hospital Quality of Care Analysis Using Data
Analytics, Data Science, and AI**

NITYA SHUKLA 31/12/25

1. Project Overview

Objective

The objective of this project is to analyze hospital quality of care using publicly available healthcare data from the Centers for Medicare & Medicaid Services (CMS). The analysis aims to identify patterns and differences in hospital performance—specifically readmission outcomes—and to examine how hospital characteristics such as service volume and geographic location relate to quality of care.

Why this objective is valid (critical check):

You are not claiming causation. You are analyzing **associations**, which is appropriate for observational healthcare data.

2. Project Goals

1. To explore hospital readmission-related metrics using real-world healthcare data.
 2. To perform exploratory data analysis (EDA) to understand distributions and trends.
 3. To conduct statistical hypothesis testing to compare hospital performance groups.
 4. To integrate multiple datasets to enrich analysis with hospital characteristics.
 5. To optionally apply simple predictive or explanatory modeling.
 6. To communicate findings clearly through visualizations and documentation.
-

3. Data Acquisition Plan

Primary Dataset

Source: CMS (Centers for Medicare & Medicaid Services)

- **Dataset:** FY 2025 Hospital Readmissions Reduction Program (HRRP)
- **Format:** CSV
- **Purpose:** Provides excess readmission ratios (ERR), predicted and expected readmission rates, and discharge volumes.

Secondary Dataset

- **Source:** CMS Hospital General Information

- **Format:** CSV
- **Purpose:** Provides hospital characteristics such as state, ownership, hospital type, and overall rating.

Rationale for dataset choice:

The HRRP dataset provides outcome-based quality metrics, while the Hospital General Information dataset provides contextual variables needed for comparison and hypothesis testing.

4. Environment Setup

Software & Tools

- Python (via Jupyter Notebook)
 - Pandas & NumPy for data manipulation
 - Matplotlib & Seaborn for visualization
 - SciPy for statistical testing
 - (Optional) Power BI for dashboard creation
 - GitHub for version control and documentation
-

5. Data Preparation Plan

1. Load CSV files into Pandas DataFrames.
2. Standardize column names for consistency.
3. Identify and handle missing values in key variables.
4. Aggregate HRRP data so that each hospital is represented by a single row.
5. Merge HRRP data with hospital general information using facility_id.
6. Validate merges and check for data type mismatches.

Critical note:

Aggregation is necessary because HRRP data contains multiple rows per hospital due to multiple measures.

6. Exploratory Data Analysis (EDA) Plan

EDA will include:

- Summary statistics for ERR, discharges, and readmission rates.
 - Distribution plots (histograms and boxplots) for ERR.
 - Comparisons of ERR across:
 - Volume groups (low vs high discharge hospitals)
 - Geographic regions (states)
 - Correlation analysis between discharge volume and ERR.
 - Visual inspection of outliers and missing data patterns.
-

7. Feature Engineering Plan

- Create a **volume group** variable (Low vs High volume) based on discharge counts.
- Create indicator variables for data availability (e.g., valid ERR present).
- Standardize categorical variables where necessary.

Why this matters:

Feature engineering simplifies complex raw data into interpretable variables suitable for hypothesis testing.

8. Statistical Inference & Hypothesis Testing Plan

Planned hypotheses include:

1. Volume vs Readmissions

- Compare ERR between low-volume and high-volume hospitals using an independent t-test.

2. Geographic Variation

- Compare ERR across states using ANOVA.

Statistical significance will be evaluated using p-values with a standard alpha level (0.05).

Important limitation acknowledged:

Results indicate statistical association, not causation.

9. Predictive Modeling Plan (Optional)

If included:

- Use a simple linear regression model to examine whether hospital volume and characteristics can explain variation in ERR.
- Emphasis will be placed on interpretability rather than prediction accuracy.

If excluded:

- The project remains complete and valid through EDA and inference alone.
-

10. Visualization & Reporting Plan

- Use Python visualizations for analysis.
 - Create clear, labeled charts suitable for academic presentation.
 - Optionally export cleaned datasets for Power BI dashboards.
 - Summarize findings in written reports with interpretation.
-

11. Documentation & Deliverables

Deliverables

1. Jupyter Notebook with:
 - EDA
 - Statistical tests
 - Visualizations
 2. Action Plan document (this file)
 3. Research/Theory Report (separate document)
 4. (Optional) Power BI dashboard
 5. (Optional) GitHub repository with README
-

12. Timeline

Phase	Task	Estimated Duration
Phase 1	Data acquisition & setup	1 day

Phase	Task	Estimated Duration
Phase 2	Data cleaning & aggregation	1–2 days
Phase 3	EDA & visualization	2 days
Phase 4	Hypothesis testing	1 day
Phase 5	Optional modeling	1 day
Phase 6	Documentation & reporting	2 days