****

**Hospital Provider Cost Reports**

**(2011 - 2022)**

***Exploratory Data Analysis: Financial & Operational Insights***

**Prepared by:** Jason Cao

**Date:** July 2025

**Data Source:**

Centers for Medicare & Medicaid Services (CMS) - Hospital Provider Cost Reports

# Introduction

**Data Source**

This report provides an in-depth analysis of hospital provider cost and enrollment data from the Centers for Medicare & Medicaid Services (CMS). The dataset originates from the Hospital Provider Cost Reports, with annual submissions spanning 2011–2022. Each report includes measures such as facility characteristics, utilization metrics, cost and charge data (both total and Medicare-specific), Medicare settlement figures, and financial statement information organized by CMS Certification Number (CCN).

Following the July 17, 2023 data refresh, the dataset was expanded to include additional provider types such as Critical Access Hospitals (CAH) and Rural Emergency Hospitals (REH), broadening the scope of the analysis.

The data is publicly available via the: [Hospital Provider Cost Report - CMS Data Portal](https://data.cms.gov/provider-compliance/cost-report/hospital-provider-cost-report)

**Purpose**

The purpose of this analysis is to explore financial and operational trends within U.S. hospitals, uncovering patterns that may support decision-making by healthcare administrators, researchers, policy makers, and patient advocates. By analyzing hospital costs, revenues, utilization, and capacity metrics, this work aims to contribute to a clearer understanding of resource allocation, financial performance, and service distribution across diverse hospital types.

# Objectives

The primary objective of this analysis is to examine hospital financial and operational performance trends using Medicare Cost Report data from 2011–2022.

**Business Questions**

**Financial Performance**

* How has hospital net income changed over time, both nationally and by state?
* What is the average cost-to-charge ratio across hospitals, and which hospitals are outliers?
* How has the cost-to-charge ratio evolved over the years?
* How do urban vs. rural hospitals differ in financial performance trends?
* What are the long-term trends in assets, liabilities, and fund balances?

**Capacity & Operations**

* Which states or counties have the highest hospital capacity in terms of beds and patient days?
* How has staffing (FTE employees) changed relative to patient volume?

**Care Delivery & Equity**

* What are the trends in charity care and uncompensated care over time?
* How much uncompensated care is provided by region or hospital type?
* How do inpatient vs. outpatient revenues vary by hospital size and type?

**Goal Statement**  
The goal of this project is to perform a comprehensive analysis of hospital provider cost report data to identify patterns, trends, and areas for improvement, while demonstrating my ability to transform raw data into actionable insights for my professional portfolio.

**Intended Audience**

The insights will be tailored to meet the needs of several stakeholder groups:

* Hospital administrators and executives – to inform operational and financial decision-making
* Researchers & analysts – to serve as a reference point for further studies
* Policy makers and government agencies – to guide regulation, funding allocation, and healthcare reform initiatives

# Data Overview

**Dataset Description**

The dataset is derived from annual Hospital Provider Cost Reports published by CMS between 2011 and 2022. Each year’s data was originally provided as a separate CSV file (such as CostReport\_2011\_Final.csv, CostReport\_2012\_Final.csv, CostReport\_2022\_Final.csv).

* **Number of Records:** 73,974 rows (excluding header)
* **Number of Columns:** 118 variables/features
* **Data Types:** A mix of categorical (such as hospital type, ownership, state), numerical (such as beds, employees, costs, revenues), and date/time fields (such as fiscal year, reporting period). Data types are consistent with the values provided.

**Key Variables**

* **Provider CCN (Hospital ID)**: Unique identifier for each hospital, consistent across years
* **Financial Metrics**: Net income, total costs, charges, revenues, assets, liabilities, charity care, uncompensated care
* **Utilization Metrics**: Patient days, discharges, inpatient vs. outpatient services
* **Capacity Metrics**: Number of beds, number of employees, interns/residents
* **Geographical & Organizational Attributes**: State, county, rural vs. urban status, ownership type, teaching status
* **Time Variables**: Fiscal year, reporting period start and end dates

**Data Refresh Information**

* **Update Frequency:** Reports are typically updated once every 12 months
* **Coverage:** Includes historical data spanning multiple fiscal years (2011–2022)
* **Special Considerations:**
  + The same hospital may appear across multiple years (trackable via CCN)
  + Some hospitals may have incomplete or missing submissions
  + Financial figures are reported in nominal dollars and are not inflation-adjusted
  + Updates reflect finalized reporting for the most recent fiscal year

# Scope

This report covers the analysis of the data spanning 11 years from 2011 through 2022.

Guiding Questions:

As part of this project, the analysis will explore questions such as:

* Financial Performance
  + How has hospital net income changed over time, both nationally and by state?
  + What is the average cost-to-charge ratio across hospitals, and which hospitals are outliers?
  + How has the cost-to-charge ratio evolved over the years?
  + How do urban vs. rural hospitals differ in financial performance trends?
  + What are the long-term trends in assets, liabilities, and fund balances?
* Capacity & Operations
  + Which states or counties have the highest hospital capacity in terms of beds and patient days?
  + How has staffing (FTE employees) changed relative to patient volume?
* Care Delivery & Equity
  + What are the trends in charity care and uncompensated care over time?
  + How much uncompensated care is provided by region or hospital type?
  + How do inpatient vs. outpatient revenues vary by hospital size and type?

These questions will help define the scope of the analysis and ensure that the findings provide meaningful insights for hospital administrators, researchers, policy makers, and patient advocates.

# Methodology

This section presents the methodology used in the development of the data analysis.

**Type of Analyses**

**Descriptive analysis:**

* Visualize cost distribution by department and facility characteristics
* Analyze Medicare settlement trends and utilization data
* Explore patterns that might indicate cost drivers or inefficiencies

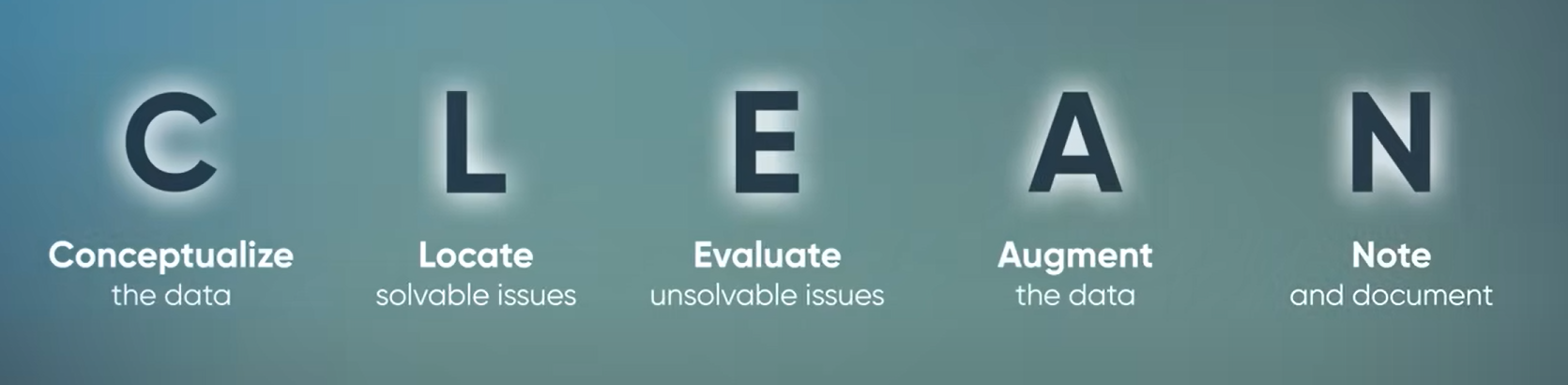
**Financial insights:**

* Highlight key KPIs like cost-to-reimbursement ratios
* Detect and flag unusual cost spikes or trends in expenses
* Provide easy-to-understand summaries for quick decision-making

**Marketing/strategy angle:**

* Benchmark facilities against each other on cost efficiency
* Show utilization strengths in high-demand services
* Present geographic and service-line heatmaps for competitive insights

# Data Cleaning



Framework used (CLEAN):

* **C - Conceptualize**
* **L - Locate solvable issues**
* **E - Evaluate unsolvable issues**
* **A - Augment and improve the dataset**
* **N - Note and document**

**C - Conceptualize the Data**

This step is about **understanding the dataset** in context before making any changes. You explore the purpose of the data, its structure, and what questions it can realistically answer. By conceptualizing, you set expectations, identify key variables, and recognize limitations upfront.

Identify the grain, key metrics, and dimensions of the dataset:

What’s the grain of the table? (or what does each row in this dataset represent?)

* Each record (row) represents a hospital’s annual Medicare cost report submission for a specific fiscal year.

**Key Metrics:**

| **Category** | **Key Metrics** | **Why It Matters** |
| --- | --- | --- |
| **Operational Capacity** | - Number of Beds  - FTE – Employees on Payroll  - Interns/Residents (FTE) | Measures hospital size, staffing levels, and training capacity |
| **Utilization Metrics** | - Total Patient Days (Medicare, Medicaid, Other)  - Total Discharges (by program)  - Total Bed Days Available | Shows how heavily resources are being used and demand for services |
| **Financial Performance** | - Total Costs  - Total Charges (Inpatient + Outpatient)  - Net Patient Revenue  - Net Income  - Cost-to-Charge Ratio | Captures efficiency, profitability, and pricing strategies |
| **Uncompensated & Charity Care** | - Cost of Charity Care  - Bad Debt Expense  - Total Unreimbursed & Uncompensated Care | Highlights the financial burden of unpaid or underpaid care |
| **Revenue Mix & Payer Dependence** | - Inpatient Revenue vs. Outpatient Revenue  - Net Revenue from Medicaid  - Medicaid Charges  - Medicare Utilization (Title XVIII Days)  - Disproportionate Share Adjustment (DSH) | Identifies reliance on public vs. private payers and revenue diversification |
| **Balance Sheet / Financial Health** | - Total Assets  - Total Liabilities  - Fund Balance | Assesses long-term financial stability and solvency |
| **Geographic & Organizational Context** | - State / County  - Rural vs. Urban  - Facility Type  - Type of Control | Enables segmentation by region and ownership type |

**Key Dimensions:**

| **Category** | **Key Dimensions** | **Why It Matters** |
| --- | --- | --- |
| **Time** | - Fiscal Year Begin Date  - Fiscal Year End Date  - Cost Report Year | Enables trend analysis across years and comparisons of hospital performance over time |
| **Geography** | - State Code  - County  - City  - Zip Code  - Rural vs. Urban | Provides regional segmentation and urban/rural comparisons |
| **Hospital Identity** | - Provider CCN (unique hospital ID)  - Hospital Name  - Street Address | Identifies each facility and allows linking across years or to external datasets |
| **Organizational Attributes** | - CCN Facility Type (e.g., acute, critical access)  - Provider Type  - Type of Control (ownership: nonprofit, government, for-profit) | Enables comparisons across different hospital structures and governance models |
| **Operational Characteristics** | - Number of Beds  - FTE – Employees on Payroll  - Interns/Residents (FTE) | Provides context for scale and staffing when analyzing outcomes |

Overall, this dataset contains provider information, with each record representing a unique healthcare provider. Key columns include provider details, utilization data, and financial costs and charges broken down by cost center. Additionally, the dataset includes supplementary information such as facility characteristics, Medicare settlement data, and financial statement details.

**L - Locate & Address Solvable Issues**

Here, I searched for **issues in the dataset that can be directly fixed**. These include missing values, duplicates, outliers, or inconsistent formats. The focus is on identifying problems that are within my control to resolve without external dependencies.

**Examples of Solvable Issues:**

1. Inconsistent data formats
2. Inconsistent categorizations
3. Nulls
4. Duplicates

**First Pass at Data Cleaning:**

1. Eyeball for obvious issues
2. Filter to check distinct values
3. Document any issues in the Issues Log

**Steps to start addressing solvable issues:**

* Formatting
  + State lookups
  + Using numeric code instead of name
* Finding inconsistent categorizations and documenting it in issues log
* Identifying Duplicates

**E - Evaluate Unsolvable Issues**

Not every problem has a quick fix. This step is about **assessing data issues that cannot be resolved with the information at hand**, such as missing context, incomplete external references, or biased sampling. Evaluating these helps me understand their potential impact on analysis and decide whether to adjust the scope or proceed with caution.

Examples of unsolvable issues:

* Missing data
* Nonsensical data
  + Bad state
* Calculate the magnitude (% impacted of the issue)
  + If most records are (>70%) are missing or nonsensical, column is likely unusable
  + If few records (<10%) are missing or nonsensical, can likely keep data as is
  + If in the middle, make a judgment depending on how critical that column is

Document your thought process and record the severity.

About 10% of delivery timestamps were missing, and also 5% of the currency information was missing. However, these are not critical to this analysis. So they were left as is. For the 3$% of refund dates that showed up as being before the sales day, they were actually excluded from the analysis, so as to not bias the data.

**A - Augment and Improve the Dataset**

Once solvable and unsolvable issues are clarified, you work to **enrich the dataset**. This might mean adding external data sources, deriving new variables, or improving data quality through feature engineering. The goal is to make the dataset more useful, insightful, and aligned with your analysis objectives.

* Create additional columns through calculations
  + Column for number of days between Fiscal year begin date - Fiscal year end date
* Add supplementary information from another source
  + Add region column for more in-depth geographic analysis

I augmented the dataset by adding in the time to deliver, the time to ship, and the time to refund, as well as regional information so that we could better segment the sales and refunds trends and also understand the data in another geographic dimension.

**N - Note and Document Issues**

The final step is **documenting what I did, why, and how**. Good documentation ensures transparency and reproducibility. It provides a clear record of data issues found, decisions made, and transformations applied, helping both myself and others trust the analysis.

Create a change log.

Document the issues that you found, the magnitude, and the severity of the issues, and if the issues were resolved.

# North Star Metrics and Dimensions

List out attributes that are metrics and dimensions

# Summary of Insights

**Plan Type:**

* At the start of the pandemic: Enterprise plans had a significant spike and record-high at $226K bookings in one week. This spike then significantly decreased and slowed in both summer of 2020 and summer of 2021.
* Business plans have completely dropped off at the end of 2021 - investigate whether there is an issue with the product or a new competitor for this market.

**Plan Period:**

* Yearly plans make up over 90% of bookings, but have exhibited a significant drop-off towards the end of 2021 that was not seen in monthly plans.

**Plan Region:**

* During the peak-COVID months in early 2020, North America made up almost two thirds of weekly bookings, and most of the spike in enterprise was contained to this region.

# Recommendations & Next Steps

Here, we revisit our business questions and share with you our high-level business recommendations.

**Business Questions**

**Financial Performance**

* How has hospital net income changed over time, both nationally and by state?
* What is the average cost-to-charge ratio across hospitals, and which hospitals are outliers?
* How has the cost-to-charge ratio evolved over the years?
* How do urban vs. rural hospitals differ in financial performance trends?
* What are the long-term trends in assets, liabilities, and fund balances?

**Capacity & Operations**

* Which states or counties have the highest hospital capacity in terms of beds and patient days?
* How has staffing (FTE employees) changed relative to patient volume?

**Care Delivery & Equity**

* What are the trends in charity care and uncompensated care over time?
* How much uncompensated care is provided by region or hospital type?
* How do inpatient vs. outpatient revenues vary by hospital size and type?

# Strategic Recommendation

* Focus on Technology sub-category and Phones and Chairs as they are highest selling and most profitable. Bundle them with the less profitable products such as Bookcases, Table and Chairs to offset the losses.
* Selling Bookcases and Tables result in huge losses, so Super Store has to consider to bundle them together with High Selling or Profitable sub-category such as Chairs, Copiers, Phones and Office Supplies products.
* For Home Offices customers, these people might be busy with work and less likely to spend time selecting individual products, so create a Home Office package with products used for offices such as table, chairs, phone, copiers, storage, label, fasteners, bookcases.
* For loss-making products like Supplies, Bookcases, Tables, consider to either drop these from the catalogue or change suppliers and bargain for cheaper price.
* Consumer and Corporate Segment make up more than 70% of customerbase. Target them, especially customers from the East and West region in the Top 10 cities with Highest Sales by introducing special promotions and bundles for mass Consumer and Home Offices and send promotional emails or flyers.