Overview

**Hospital Provider Cost Data Analysis:**

**Insights and Trends**

# Introduction

**Data Source:** This report provides an in-depth analysis of hospital provider enrollment and cost data obtained from the Provider Enrollment, Chain, and Ownership System (PECOS). 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 from the Centers for Medicare & Medicaid Services (CMS) at:  
[Hospital Enrollments - CMS Data Portal](https://data.cms.gov/provider-characteristics/hospitals-and-other-facilities/hospital-enrollments)

**Purpose and Audience:** The purpose of this analysis is to describe and explore trends within the hospital enrollment and cost data, uncovering patterns that may inform healthcare administrators, researchers, policy makers, and patient advocates. This analysis aims to enhance understanding of cost distribution, provider types, and service utilization, which can contribute to improved operational decisions and healthcare policies.

**Objective:** The objective of this study is to develop and showcase strong data analysis skills through a meaningful project addressing real-world healthcare data challenges. The insights derived from the analysis are intended to assist stakeholders in making informed decisions regarding hospital resource allocation, cost management, and regulatory oversight.

# Objectives

This section outlines the intended outputs and focus areas of the data analysis report, guiding the direction of the study.

The primary objective is to explore hospital provider cost report data to uncover detailed cost and utilization patterns across facilities. The analysis aims to:

* Summarize key financial health metrics for hospitals
* Benchmark cost efficiency across different provider types and regions
* Identify trends and potential areas for cost optimization
* Support healthcare researchers, administrators, and business teams in data-driven decision-making and strategic planning

I want to explore hospital provider cost reports to uncover detailed cost and utilization patterns across facilities, while also summarizing key financial health metrics and benchmarking cost efficiency to support both researchers and business teams in decision-making and strategic positioning.

**Goal Statement**The goal of this project is to perform a comprehensive analysis of hospital provider cost report data to identify patterns, trends, and potential areas for improvement. This work will serve as a learning exercise to strengthen my data analysis skills and will result in a polished portfolio piece showcasing my ability to transform raw data into actionable insights.

**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
* General public & patient advocates – to promote understanding of healthcare costs and trends
* Policy makers and government agencies – to guide regulation, funding allocation, and healthcare reform initiatives

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 Overview

Most of the data is stored in separate files, by year, in CSV format.

* Dataset Description: Number of records, variables/features, data types.
* Key Variables: Describe important columns (e.g., hospital ID, cost, service types, outcomes, dates).
* Data Refresh Info: Mention any data updates or special considerations.

# Data Quality Assessment

* Missing Data: Summary of missing or null values per column.
* Duplicates: Check for duplicate records.
* Outliers: Initial identification of unusual cost or service values.
* Data Consistency: Check for format or type inconsistencies.

# 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:

1. **Cost Trends:**
   1. How have hospital operating costs changed over the period 2011–2022?
   2. Which cost centers contribute most to total hospital expenses?
2. **Utilization & Efficiency:**
   1. How does patient utilization relate to hospital costs and efficiency?
   2. Do hospitals with higher service volumes achieve lower costs per patient?
3. **Geographic Insights:**
   1. How do hospital costs vary across states and regions?
   2. Are rural hospitals more or less cost-efficient than urban hospitals?
4. **Medicare Settlement & Reimbursement:**
   1. How do Medicare settlement amounts compare to reported costs?
   2. Are there trends in reimbursement gaps over time?
5. **Benchmarking & Best Practices:**
   1. Which hospitals demonstrate consistently strong cost efficiency?
   2. What patterns emerge when benchmarking cost per patient day or discharge across hospital types?

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.

**Data Collection:**

Indicate here the data required and the method of how they were collected.

**Guide Questions: What data was collected, and how were they collected?**

A survey was conducted last January 24-28, 2022. The questionnaire utilized a 5-point Likert scale to generate responses from the customers.

Responses are collected from 200 customers ages 15-75. They received the questionnaire through an email blast. Each respondent is given 24 hours to open and answer the survey.

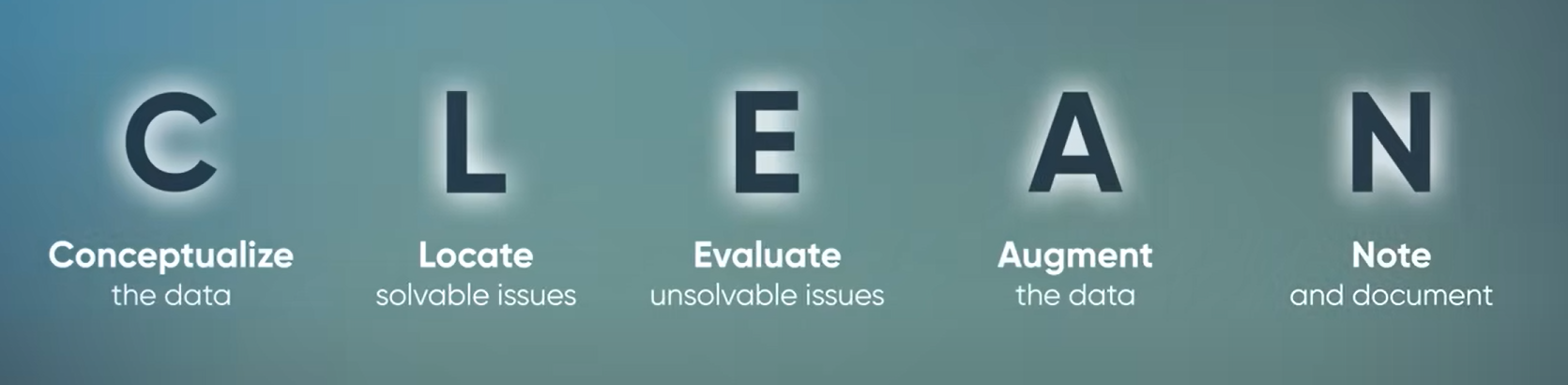
Data collected are stored and analyzed through Google Sheets.

Statistical Treatment:

Indicate the statistical methods applied to the data collected by filling in the table below.

**Guide Question: What statistical computations are applied to the data collected?**

# 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

* Identify grain, measures, and dimensions
  + What’s the grain of the table, aka what does each row in this dataset represent?
  + What are the quantitative values and the qualitative values and which ones have the most business relevance for this task?
  + Measures, or quantitative values absolutely need proper number formatting for calculations, and qualitative values need to have consistent categories for segmentation
  + Key metrics:
    - Number of Beds
    - Total Bed Days Available
    - Hospital Number of Beds For Adults & Peds
    - Hospital Total Bed Days Available For Adults & Peds
    - Cost of Charity Care
    - Total Bad Debt Expense
    - Cost of Uncompensated Care
    - Total Unreimbursed and Uncompensated Care
    - Total Salaries From Worksheet A
    - Overhead Non‐Salary Costs
    - Depreciation Cost
    - Total Costs
    - Inpatient Total Charges
    - Outpatient Total Charges
    - Combined Outpatient + Inpatient Total Charges
    - Wage‐Related Costs (Core)
    - Wage‐Related Costs (RHC/FQHC)
    - Total Salaries (Adjusted)
    - Contract Labor: Direct Patient Care
    - Wage Related Costs for Part ‐ A Teaching Physicians
    - Wage Related Costs for Interns and Residents
    - Cash on Hand and in Banks
    - Temporary Investments
    - Notes Receivable
    - Accounts Receivable
    - Inventory
    - Prepaid Expenses
    - Other Current Assets
    - Total Current Assets
    - Land
    - Land Improvements
    - Buildings
    - Leasehold Improvements
    - Fixed Equipment
    - Major Movable Equipment
    - Minor Equipment Depreciable
    - Health Information Technology Designated Assets
    - Total Fixed Assets
    - Investments
    - Other Assets
    - Total Other Assets
    - Total Assets
    - Accounts Payable
    - Salaries, Wages, and Fees Payable
    - Payroll Taxes Payable
    - Notes and Loans Payable (Short Term)
    - Deferred Income
    - Other Current Liabilities
    - Total Current Liabilities
    - Mortgage Payable
    - Notes Payable
    - Unsecured Loans
    - Other Long Term Liabilities
    - Total Long Term Liabilities
    - Total Liabilities
    - General Fund Balance
    - Total Fund Balances
    - Total Liabilities and Fund Balances
    - DRG Amounts Other Than Outlier Payments
    - DRG Amounts Before October 1
    - DRG Amounts After October 1
  + Key dimensions:
  + Conceptualizing and understanding the dataset is the first step to identifying what you actually need to care about and prioritize in your data cleaning process
* Identify critical vs. non-critical columns
  + Given this, which columns need to be high quality (aka ~80% complete or accurate) for this analysis?
  + Which columns are less important or not important at all?
* Understand definitions
  + For the columns that are important, what does each one mean?

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. The data spans the years 2011 through 2022.

### L - Locate and address solvable issues

* Formatting
  + State lookups
  + Using numeric code instead of name
* Consistency
* Duplicates

Make the most important columns usable.

Make sure to have a raw copy of data of this before cleaning.

### E - Evaluate 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

* 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

Create a change log.

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

This subpage presents the results of the data analysis.

*Present here the results of the study through the columns below.*

**Guide Questions: What are the findings of the methodology executed? What do these findings imply?**

## Findings

# 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