Hospital Quality Patterns in CMS Timely and Effective Care Data

Suhani Singh
INST 447: Data Sources and Manipulation
Instructor: Wei Ai
University of Maryland
December 19, 2025

1. Introduction

Background and Significance

Public reporting of hospital quality metrics plays an important role in enabling patients, policymakers, and healthcare organizations to compare performance and identify opportunities for improvement. The Centers for Medicare & Medicaid Services (CMS) published recent provider-level datasets that capture both hospital characteristics and performance on standardized "timely and effective care" measures. These data are widely used for transparency, benchmarking, and quality improvement initiatives, making them a valuable resource for exploratory analysis of hospital performance patterns.

Research Questions

This project examines how hospital performance scores vary across geographic regions and institutional characteristics using two CMS provider datasets:

- Timely and Effective Care – Hospital: provider-level performance measures across multiple clinical domains
- Hospital General Information: hospital attributes such as ownership type, hospital classification, availability of emergency services, and overall CMS hospital rating

The primary research question guiding this analysis is:
How do timely and effective care performance scores vary across states and hospital characteristics (ownership, type, emergency services, and overall rating)?

To support this question, the following sub-questions are explored:

- Which states exhibit higher or lower average performance scores for common conditions such as Emergency Department care and Sepsis Care?
- Do performance scores differ systematically by hospital ownership or hospital type?
- Is a hospital's overall CMS rating associated with higher performance on timely and effective care measures?

Approach Overview

Using Python and the pandas library, the two datasets were cleaned and integrated using a common facility identifier. Mixed-type score fields were converted to numeric values where possible, and missing data patterns were examined. Exploratory data analysis was then conducted through descriptive statistics and five visualizations to summarize variability, geographic patterns, hospital-level differences, and relationships between overall ratings and individual performance measures.

2. Dataset Description

Datasets and Sources

This project uses two publicly available datasets from the CMS Provider Data platform. The first dataset, Timely and Effective Care – Hospital, contains provider-level performance measures across multiple clinical domains and includes approximately 138,182 rows and 16 columns, with each row representing a hospital–measure–reporting period combination. The second dataset, Hospital General Information, provides hospital-level characteristics such as ownership type, hospital classification, emergency services availability, and overall CMS rating. This dataset contains 5,421 rows and 38 columns, with one row per hospital.

Volume

The size of the timely and effective care dataset is sufficient to support exploratory comparisons across multiple conditions, measures, and geographic regions. The hospital general information dataset provides comprehensive coverage of Medicare-certified hospitals included in the performance data and supports meaningful grouping and enrichment of performance measures by hospital characteristics.

Velocity

The timely and effective care dataset includes performance measures reported over defined reporting windows, captured using Start Date and End Date fields. The data used in this analysis span multiple reporting periods, primarily covering measures reported during 2023 and 2024, depending on the clinical measure. This project uses a static snapshot of the CMS Provider Data files rather than real-time updates, and results reflect the reporting periods included in that snapshot.

Variety

Together, the datasets include a mix of data types relevant to hospital performance analysis. These include categorical variables such as condition, measure identifiers, hospital ownership, hospital type, and state; numeric variables such as performance scores and overall hospital ratings (where available); and date fields representing measure reporting periods. This variety enables analysis across institutional, geographic, and clinical dimensions.

Veracity

Data completeness varies across measures and hospitals. The performance score field contains both numeric values and non-numeric entries such as "Not Available," meaning that not all measures provide comparable numeric metrics across hospitals. After conversion, approximately 38.8% of rows in the timely and effective care dataset contain usable numeric scores. Similarly, overall hospital ratings are missing for a substantial portion of hospitals, with approximately 52.9% of hospitals having a numeric rating available. These data quality limitations introduce potential bias and limit direct comparisons, and results should therefore be interpreted as exploratory rather than definitive assessments of hospital quality.

Value

By merging performance measures with hospital-level characteristics, the combined dataset enables exploratory insights into how reported quality metrics vary across states and different types of hospitals. This integration supports analysis of patterns related to geography, ownership structure, and overall ratings, providing context for understanding variation in publicly reported hospital performance.

3. Data Preparation and Exploration Process

Loading and Combining Data Sources

Both datasets were loaded into pandas from CSV files. To prepare for integration, the facility identifier (Facility ID) was standardized across datasets, and the timely and effective care dataset was merged with hospital-level attributes using a many-to-one left join. This structure reflects the underlying data generation process, where multiple performance measures are reported for each hospital, while hospital characteristics are recorded once per facility. The merged dataset therefore represents hospital performance across multiple CMS reporting periods, primarily from 2023–2024, rather than a single point in time.

API Access and Data Acquisition

To demonstrate API-based data acquisition, a subset of records was retrieved programmatically via the CMS Open Data API and parsed from JSON into a pandas DataFrame, confirming schema consistency with the CSV-based data used for analysis.

Data Cleaning Decisions

Several cleaning steps were applied prior to analysis. Performance scores were converted to numeric values using coercion, which transforms non-numeric entries such as "Not Available" into missing values. Hospital overall ratings were converted to numeric values using the same

approach. To improve interpretability of visualizations, the analysis focused on the most common condition categories while still preserving broad coverage of the dataset.

In addition to standard cleaning and aggregation steps, ChatGPT, a large language model (LLM) was used to assist with data processing. Specifically, it was used to transform frequently occurring CMS "Measure Name" values into structured metadata by grouping long, heterogeneous measure names into a small set of clinically meaningful categories and identifying score directionality (higher-is-better versus lower-is-better). The resulting mapping was reviewed and merged back into the analysis dataset as additional variables, enabling grouped exploratory comparisons while preserving the original numeric score values.

Data Quality Issues and Impact

The primary limitation of the data is the presence of missing or non-numeric performance scores. Many measures are reported using different units, suppressed for reporting reasons, or marked as unavailable, which limits direct comparability across hospitals and measures. As a result, findings from this analysis should be interpreted as exploratory patterns rather than definitive rankings of hospital quality.

Exploratory Data Analysis Workflow

Exploratory data analysis focused on understanding score availability and variability across conditions and hospital characteristics. Descriptive summaries of score completeness, means, and medians were computed by condition, and the most frequently reported measures were identified to determine which portions of the dataset were most suitable for analysis. These steps informed the selection of conditions and groupings used in subsequent visualizations.

4. Analysis and Findings

Figure 1. Distribution of Numeric Scores by Condition

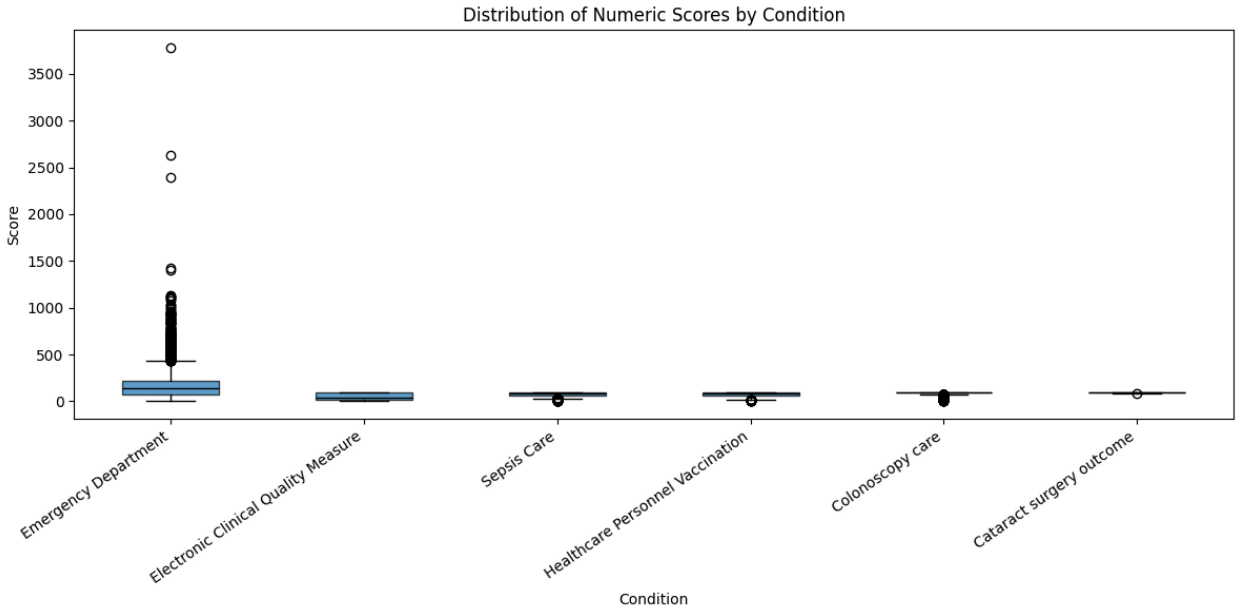

Figure 1 uses a boxplot to compare the distribution of numeric performance scores across major condition categories. This chart type was chosen because it highlights differences in medians, variability, and the presence of outliers across groups. Emergency Department measures display the widest spread and the most extreme outliers, indicating substantial variability in reported performance across hospitals. In contrast, conditions such as Cataract Surgery Outcome and Colonoscopy Care exhibit tighter distributions with fewer extreme values, suggesting more consistent reporting or narrower performance ranges. This variability reflects differences in how measures are defined and operationalized, particularly for time-based Emergency Department metrics that can vary widely by hospital context. These findings address the research question by identifying which condition areas exhibit the greatest variability and therefore require more cautious interpretation when comparing aggregated scores.

Figure 2. Top 15 States by Average Emergency Department Score

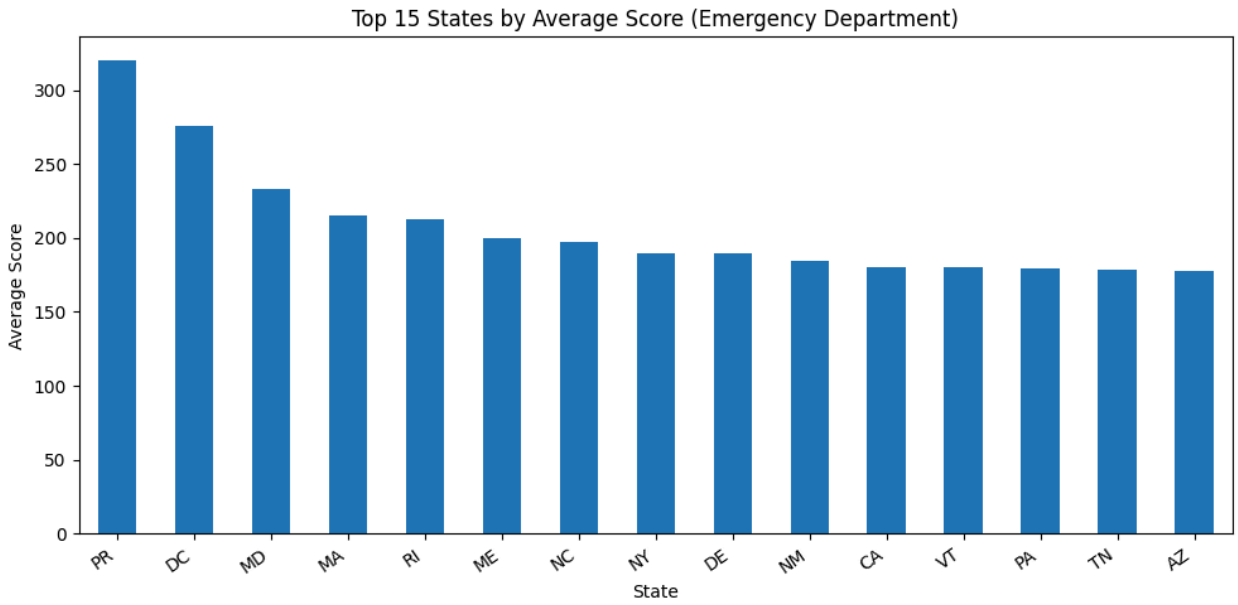

Figure 2 presents a ranked bar chart of the top 15 states based on average Emergency Department scores. A bar chart was selected to facilitate clear comparison and ranking across states. Puerto Rico and the District of Columbia appear at the top of the distribution, followed by several northeastern states such as Maryland, Massachusetts, and Rhode Island. While these differences suggest geographic variation in reported Emergency Department performance, they should be interpreted cautiously. State-level averages may be influenced by hospital composition, the number of reporting facilities, and the mix of Emergency Department measures included in each state.

Figure 3. Average Score by Hospital Ownership

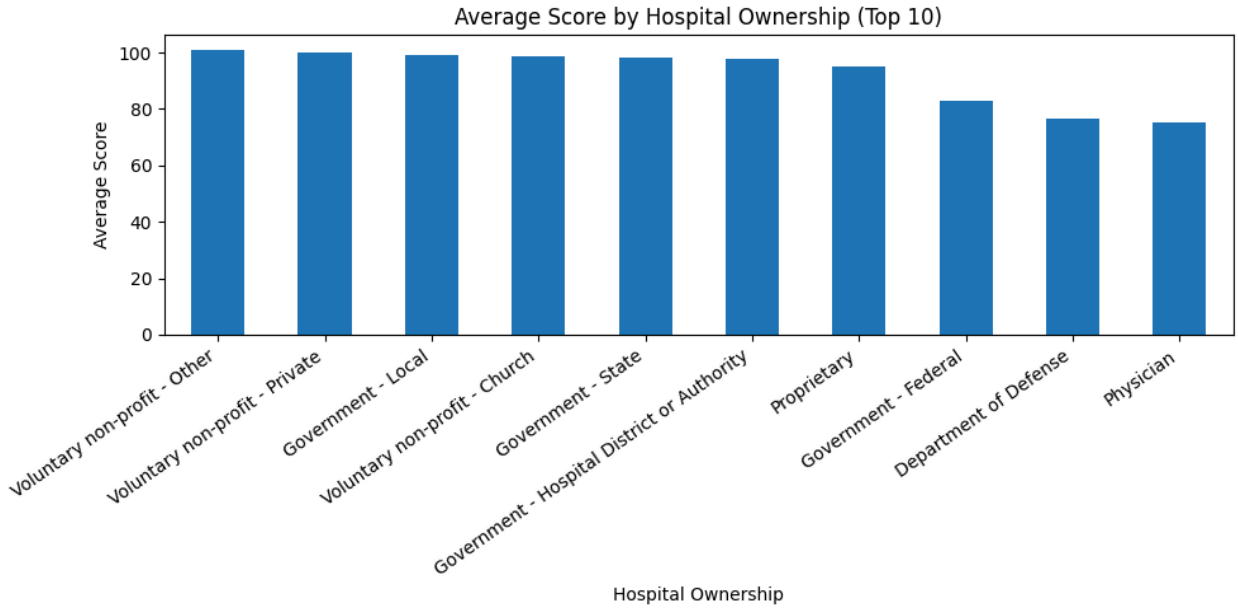

Figure 3 compares average numeric scores across hospital ownership categories using a bar chart, which is well suited for highlighting group-level differences across discrete classifications. Voluntary non-profit hospitals, both private and church-affiliated, show slightly higher average scores relative to government-operated and physician-owned hospitals. Federal and Department of Defense hospitals appear toward the lower end of the displayed averages. These patterns are exploratory and do not imply causal relationships, as ownership categories differ in mission, size, and service scope. However, this figure contributes to the research question by suggesting that organizational characteristics may be associated with systematic differences in reported performance measures.

Figure 4. Overall Hospital Rating vs. Timely and Effective Care Scores

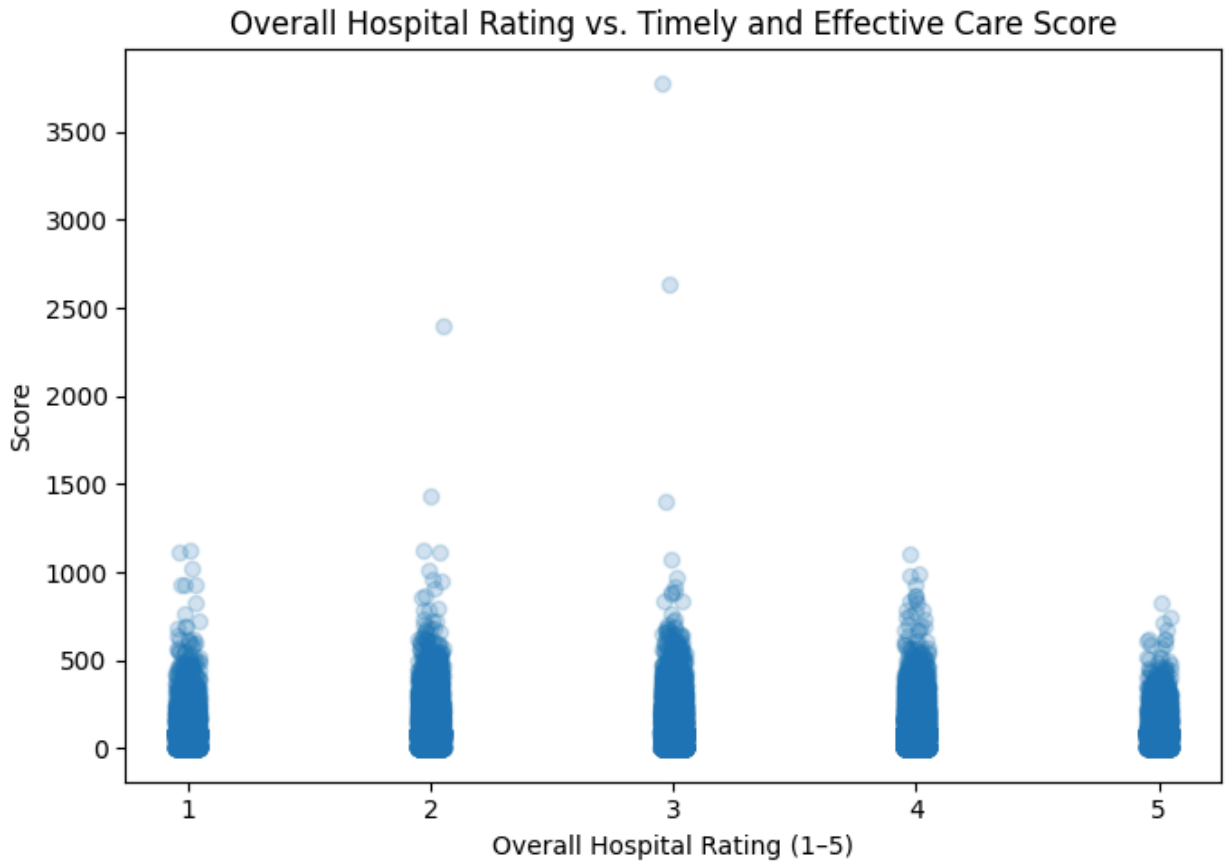

Figure 4 uses a scatter plot to examine the relationship between a hospital's overall CMS rating (on a 1–5 scale) and its timely and effective care measure scores. Scatter plots are well suited for evaluating potential associations between two numeric variables. The visualization shows substantial variability in performance scores within each rating level and no strong linear relationship between overall rating and individual measure scores. Although higher-rated hospitals appear slightly more concentrated within certain score ranges, extreme values are present across nearly all rating levels. This suggests that overall hospital ratings aggregate

multiple dimensions of quality that may not align directly with specific process-level performance measures. As a result, this figure addresses the research question by showing that higher overall ratings do not necessarily correspond to consistently higher timely and effective care scores.

Figure 5. Average Score by LLM-Derived Measure Category

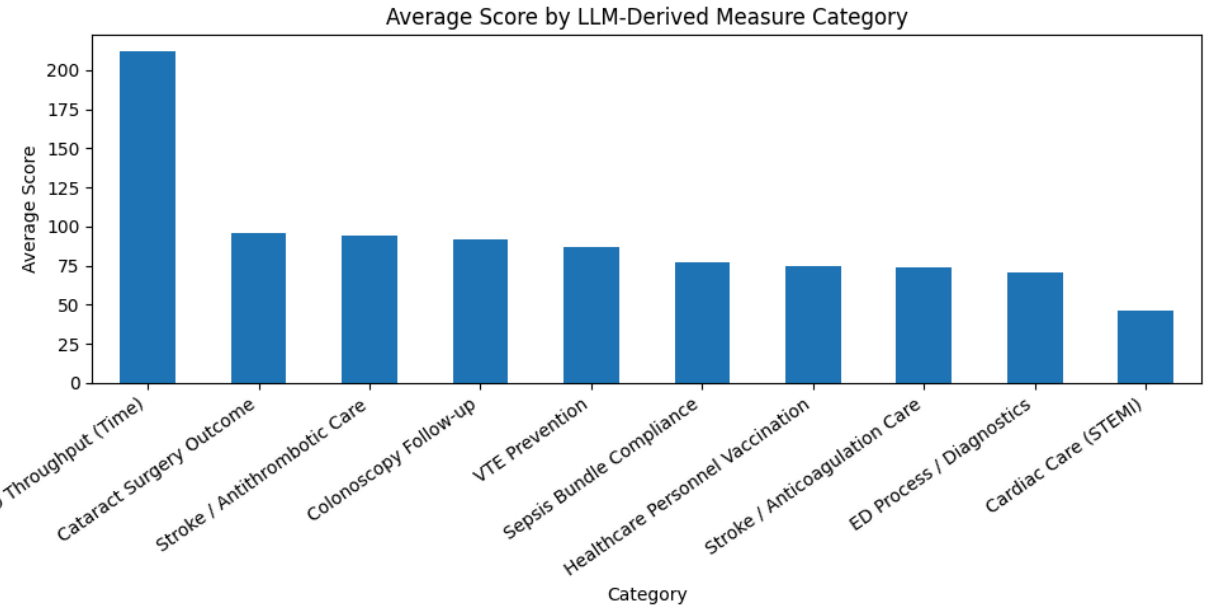

Figure 5 uses a bar chart to compare average numeric scores across measure categories derived using ChatGPT to group heterogeneous CMS performance measures into clinically meaningful domains. Emergency Department Throughput (Time) exhibits the highest average scores, reflecting the inclusion of time-based measures reported in minutes, which naturally produce larger numeric values. In contrast, categories such as Cardiac Care (STEMI), ED Process and Diagnostics, and Healthcare Personnel Vaccination show lower average scores, consistent with measures reported as rates or percentages. While these differences should not be interpreted as direct indicators of relative quality, this figure demonstrates how LLM-assisted categorization improves interpretability by organizing diverse performance measures into coherent clinical groupings. This analysis directly supports the research question by enabling category-level comparisons that would not be feasible using raw measure names alone.

5. Conclusion

This analysis explored patterns in hospital performance using CMS timely and effective care measures integrated with hospital-level characteristics. Across conditions, Emergency Department measures exhibited the greatest variability and the most extreme outliers, reflecting the heterogeneous and time-sensitive nature of emergency care metrics. In contrast, measures

related to procedures such as cataract surgery and colonoscopy care showed tighter score distributions, suggesting more consistent reporting or narrower performance ranges.

State-level comparisons revealed noticeable geographic variation in average Emergency Department scores. However, these differences are likely influenced by factors such as hospital composition, the number of reporting facilities, and variation in measure availability rather than performance alone. As a result, state-level rankings should be interpreted cautiously and viewed as descriptive patterns rather than definitive indicators of quality.

Differences were also observed across hospital ownership categories, with voluntary non-profit hospitals generally exhibiting higher average scores than government-operated or physician-owned hospitals. These associations are exploratory and do not imply causal relationships, as the analysis does not control for hospital size, patient mix, or service scope. Similarly, overall CMS hospital ratings showed only a weak association with individual timely and effective care measure scores, suggesting that aggregate ratings capture broader dimensions of quality that do not necessarily align with specific process-level measures.

Overall, these findings highlight the importance of contextualizing publicly reported hospital quality metrics and acknowledging data limitations when comparing performance across institutions. While the descriptive patterns identified in this analysis are informative, they should not be interpreted as definitive rankings of hospital quality.

Confidence in these findings is moderate, given the substantial presence of missing or non-numeric scores and the heterogeneous nature of reported measures. Future work could extend this analysis by standardizing scores within measure types, focusing on individual measures rather than pooled scores, and incorporating additional CMS datasets such as readmission outcomes or patient experience measures (HCAHPS: Hospital Consumer Assessment of Healthcare Providers and Systems) to provide a more comprehensive view of hospital quality.

6. Reflection on Tools and Process

AI tools were used selectively to support interpretation and preprocessing rather than to automate core analytical decisions. In particular, ChatGPT was used to transform unstructured CMS measure names into structured categories and to identify score directionality (higher-is-better versus lower-is-better). This step improved interpretability and enabled category-level comparisons that would not have been feasible using raw measure names alone. The LLM outputs were validated by spot-checking measure descriptions and reviewing category

assignments before integration into the dataset, ensuring that all quantitative analysis remained transparent and reproducible.

AI was most helpful for organizing and summarizing complex, domain-specific terminology, while it was less reliable when measure names lacked sufficient contextual detail. This required human judgment to verify and refine outputs, reinforcing the importance of treating AI tools as assistive rather than authoritative. The project also highlighted the value of choosing appropriate data access methods based on analytical goals. While APIs are well suited for targeted queries and automated workflows, CSV snapshots proved more practical for large-scale exploratory analysis and reproducibility.

Overall, this project strengthened my understanding of exploratory data analysis, data integration, and responsible AI use in data science workflows. It reinforced the importance of balancing technical efficiency with interpretability, validating AI-assisted outputs, and clearly distinguishing exploratory analysis from causal inference when communicating results.