Northeastern University
College of Professional Studies

Module 6 Project
Uncovering High-Cost Patterns in New York State Hospital

ALY6110 (CRN 80613): Big Data and Data Management

Professor: Olesya Agafontseva

28th June 2025

Nishtha Jatinbhai Patel

PRESENTATION ROADMAP

Comprehensive big data analysis approach to uncover high-cost patterns in New York State hospitals, from problem identification through actionable recommendations for healthcare cost optimization.

Collect data to Identify healthcare cost challenges and real world challenges.

Data collection and Real World Problem

Clean the dataset, handle missing values, and prepare data for comprehensive analysis.

NoSQL Databases - MongoDB Compass

Perform comprehensive statistical analysis.

Statistical Analysis & Modeling

Create compelling visualizations and interpret findings to extract actionable insights.

Visualization, Insights & dashboard

Develop strategic recommendations and implementation roadmap for stakeholders

Recommendations & Implementation

BUSINESS PROBLEM

The Health Cost Crisis: Healthcare costs in the United States have been rising dramatically, making it increasingly difficult for patients to afford necessary medical care.

Understanding cost drivers is essential for this 4 parameters:

PATIENTS

Financial planning and informed decision-making about healthcare options

HEALTHCARE PROVIDERS

Resource optimization and cost management strategies

POLICY MAKERS

Evidence-based healthcare policy development and regulation

INSURANCE COMPANIES

Risk assessment and premium calculation accuracy

"What factors most significantly influence hospital inpatient costs across facilities in New York and diagnoses?"

About Data & Reason for Data Selection

Why This Dataset?

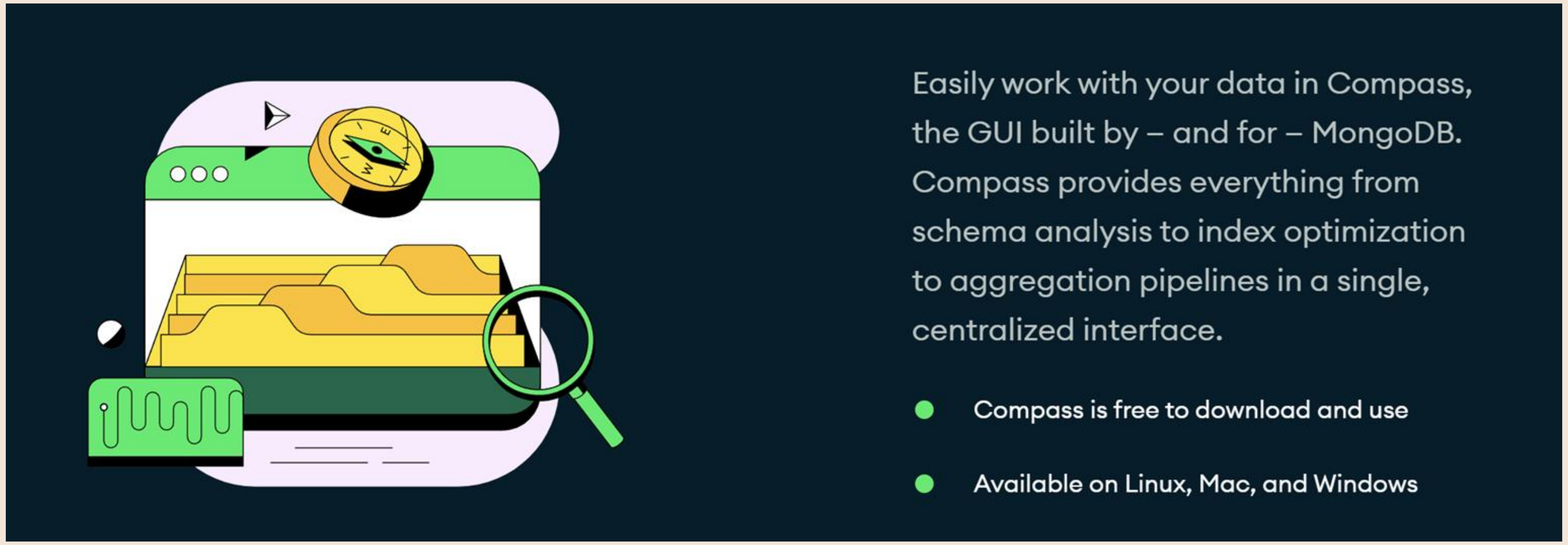
- **Scale & Scope:** Over 1M records representing diverse medical conditions and hospital types
- **Data Quality:** Government-maintained with standardized APR-DRG coding
- **Comprehensive Coverage:** Both charges (billed amounts) and costs (actual treatment costs)
- **Temporal Depth:** 9 years enabling trend analysis and pattern identification
- **Severity Classification:** Built-in illness severity coding (Minor, Moderate, Major, Extreme)

1M+ patient records (2009-2017) with complete financial data.

4 comprehensive variables support advanced analytics

1,081,672 comprehensive patient records.

NoSQL Databases – MongoDB Compass



NoSQL Databases are non-relational databases that store data in flexible formats like documents, key-value pairs, graphs, or wide-columns.

MongoDB Compass is a GUI for MongoDB, which helps to visually explore data, run queries, and optimize performance. It's widely used for interacting with NoSQL databases without using the command line.

Methodology used to Analyse the data
(Resources and Tools)

Data Cleaning & Preparation:

NoSQL Databases - MongoDB Compass

- Data Cleaning
- Checking Missing Values
- Handling missing Values
- Chose a Deletion Strategy
- Created a Clean Dataset Filter
- Applied the Deletion Method

Why NoSQL Databases - MongoDB Compass

- Reliable data cleansing
- Adaptable document-based structure perfect for managing the intricate information.

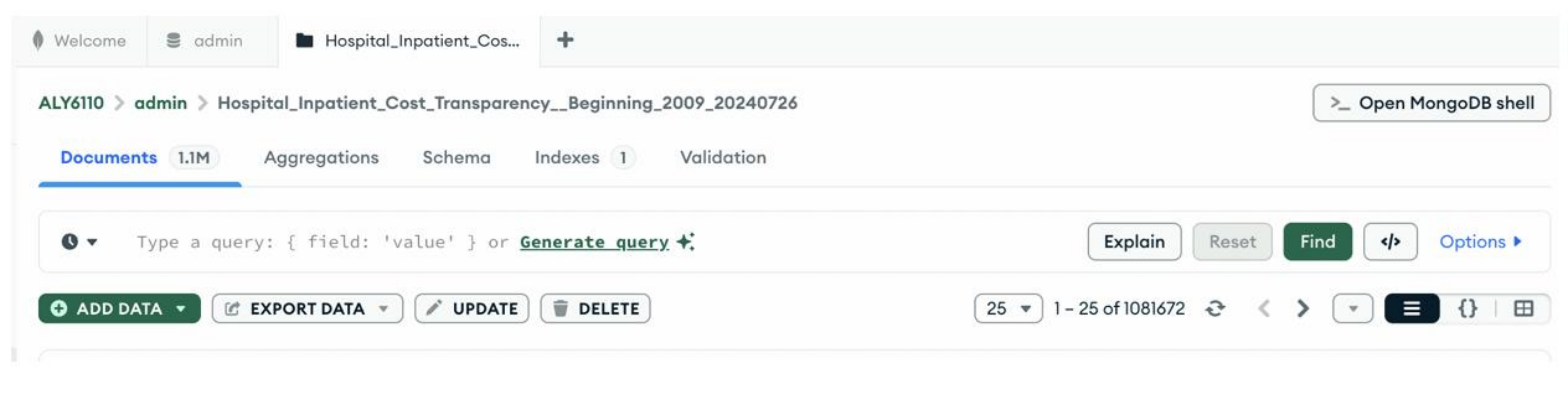
Data Visualization & Exploration: Tableau

- Built interactive dashboards
- Visualized trends by diagnosis, severity, facility, and discharges
- Used bar charts, scatter plots, treemaps, and boxplots
- Identified key cost drivers and outliers

Why Tableau?

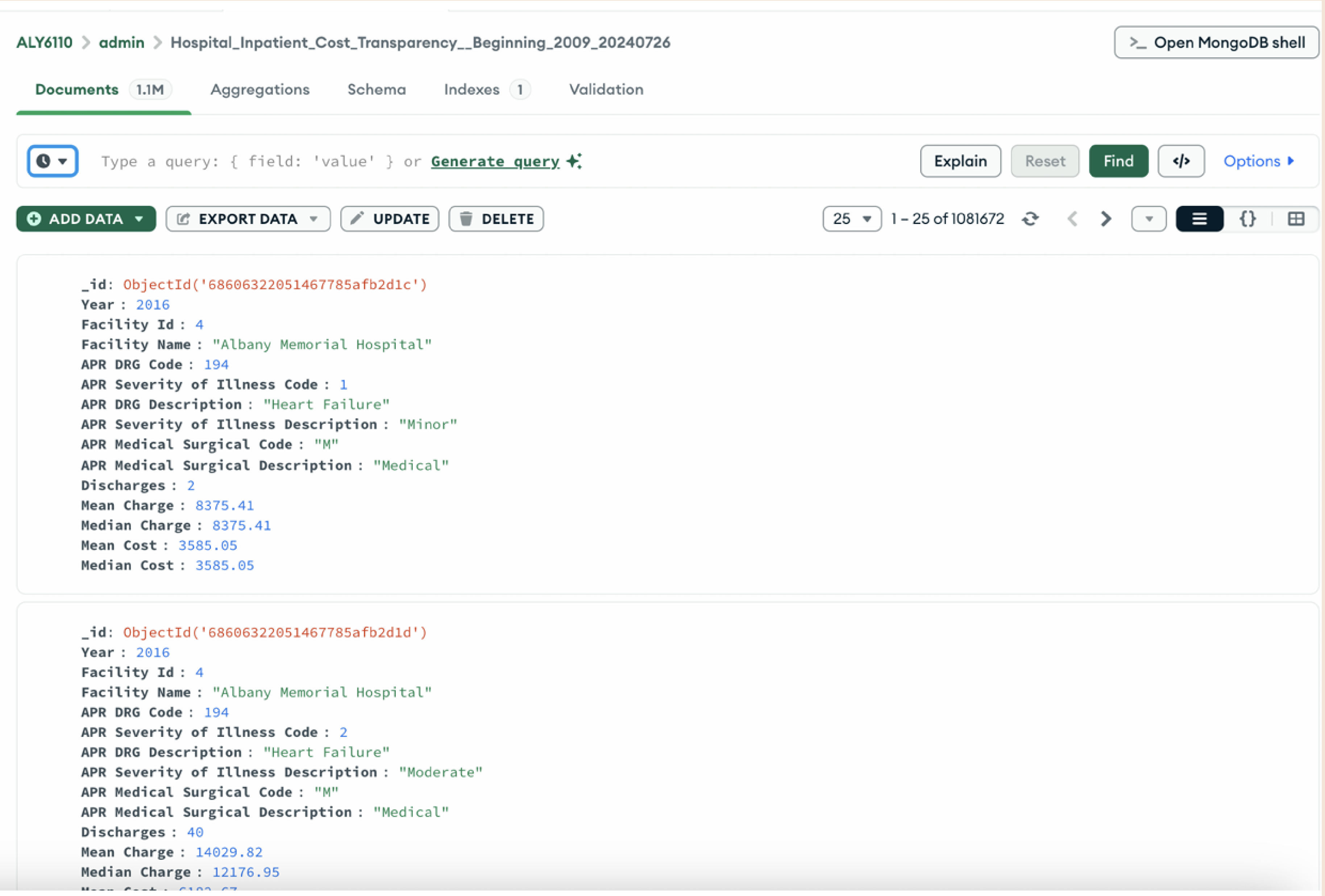
- Interactive Visual Explanation
- Intuitive Dashboard Building

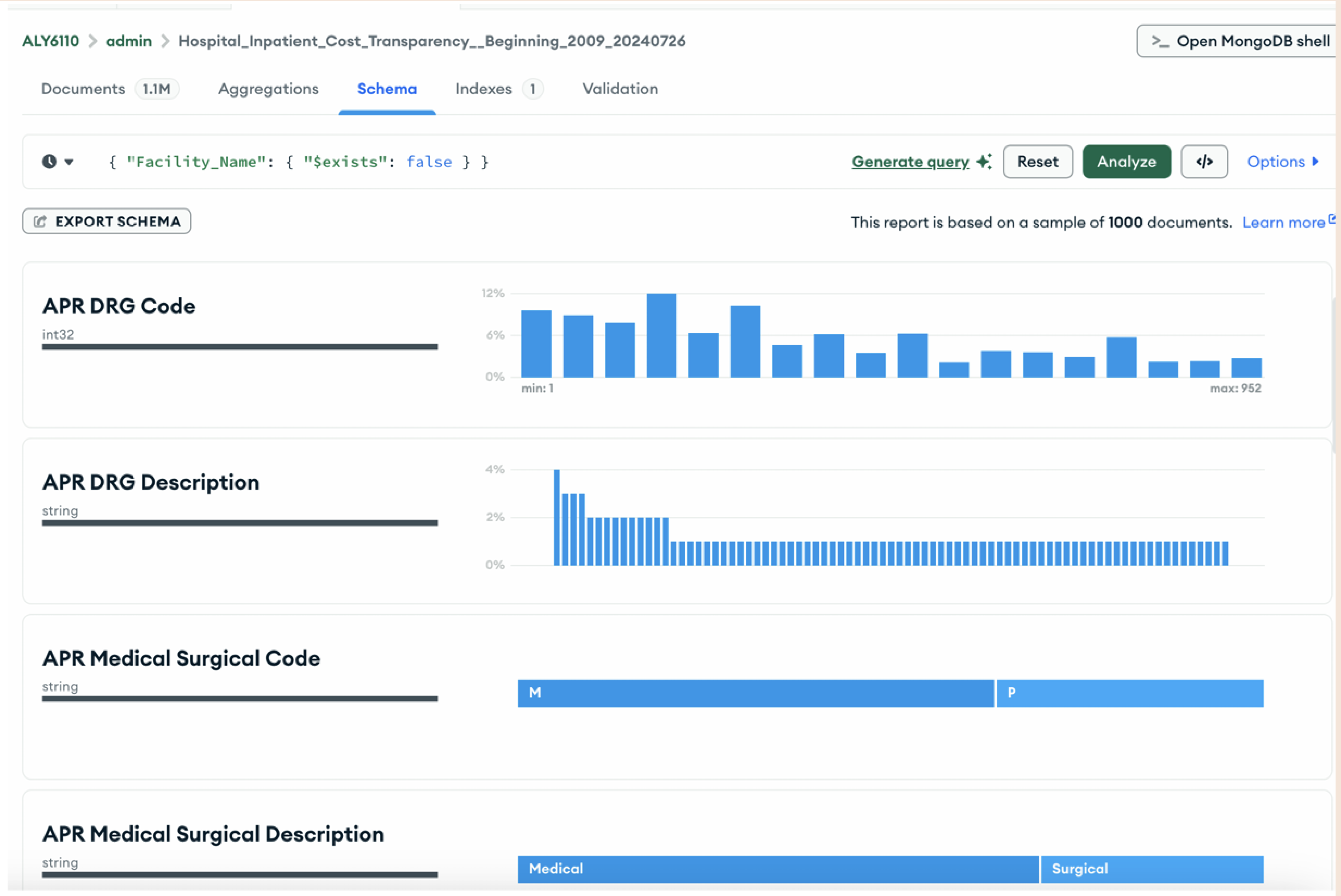
Step 1: Data Cleaning in MongoDB Compass



- Installed **MongoDB Community Server** and **Compass (GUI)**
- Connected to local MongoDB instance and Navigated to the collection - Hospital_Inpatient_Cost_Transparency__Beginning_2009_20240726
- Clicked **"Add Data"** to upload the dataset
- Successfully loaded **1.08 million records**
- Ready to explore data using **Documents**, **Schema**, and **Aggregations** tabs

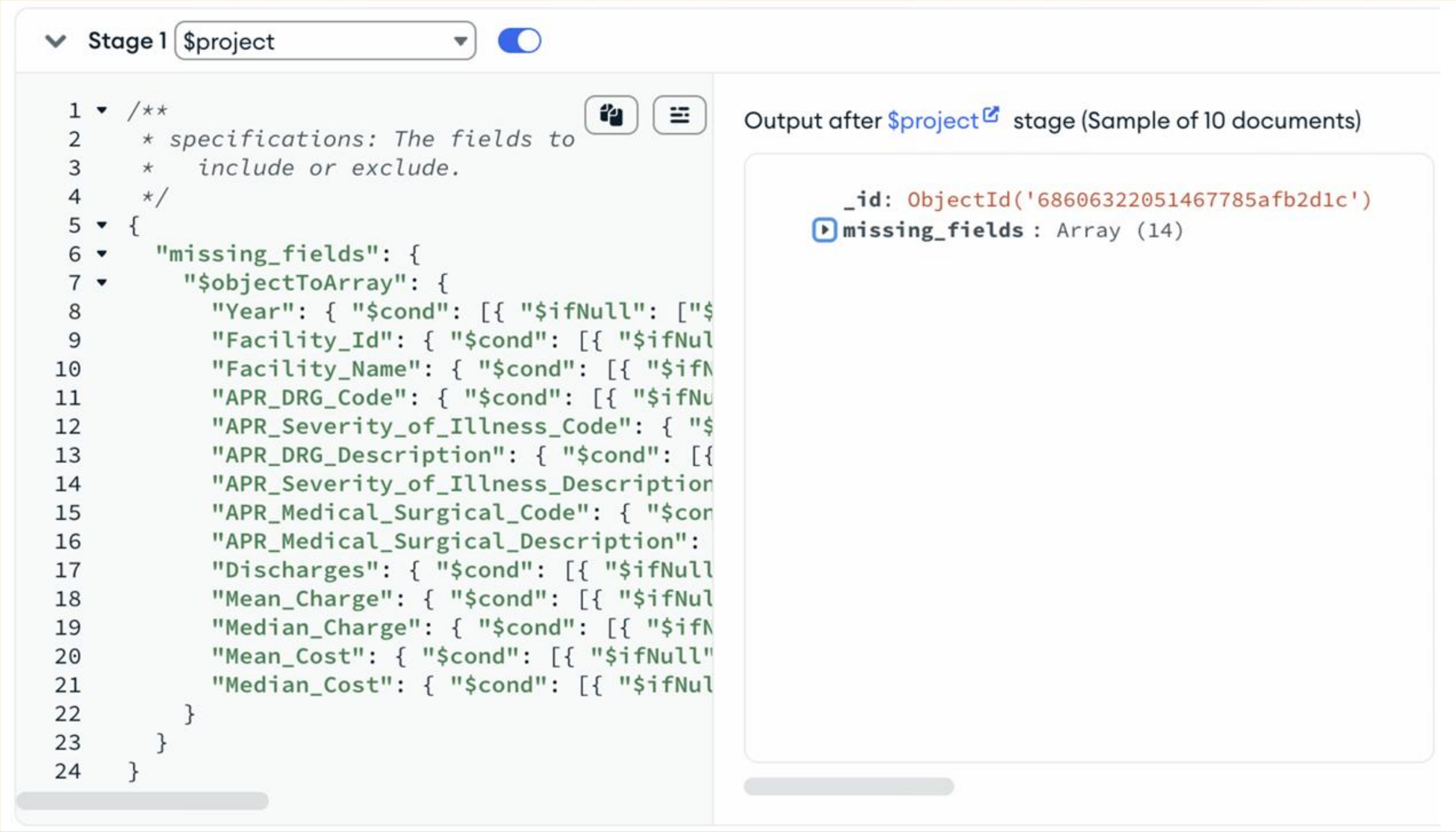
Step 2: Exploring and Auditing the Dataset in MongoDB Compass

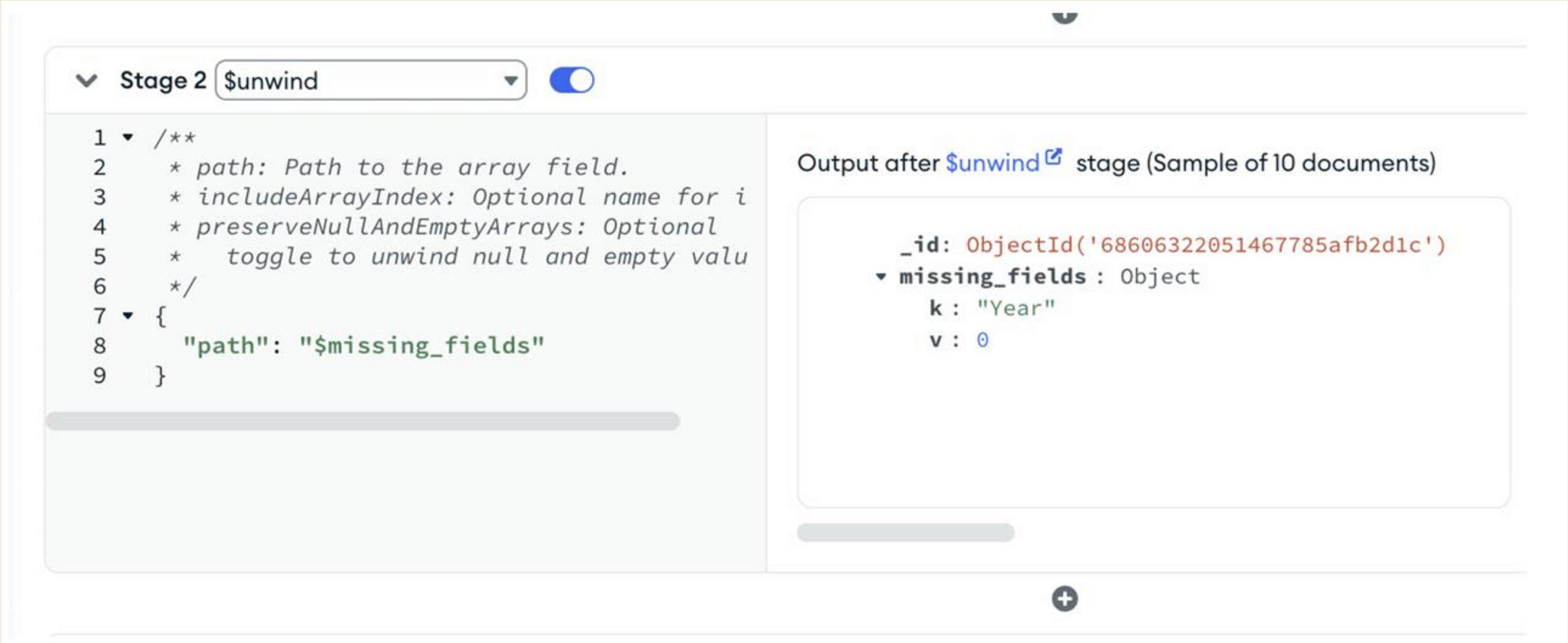




- Browsed hospital records in **Documents tab** to understand structure and key fields.
- Used **Schema tab** to check data types, value distributions, and spot missing values.
- Identified issues like **null entries**, **imbalanced categories**, and **format inconsistencies**.

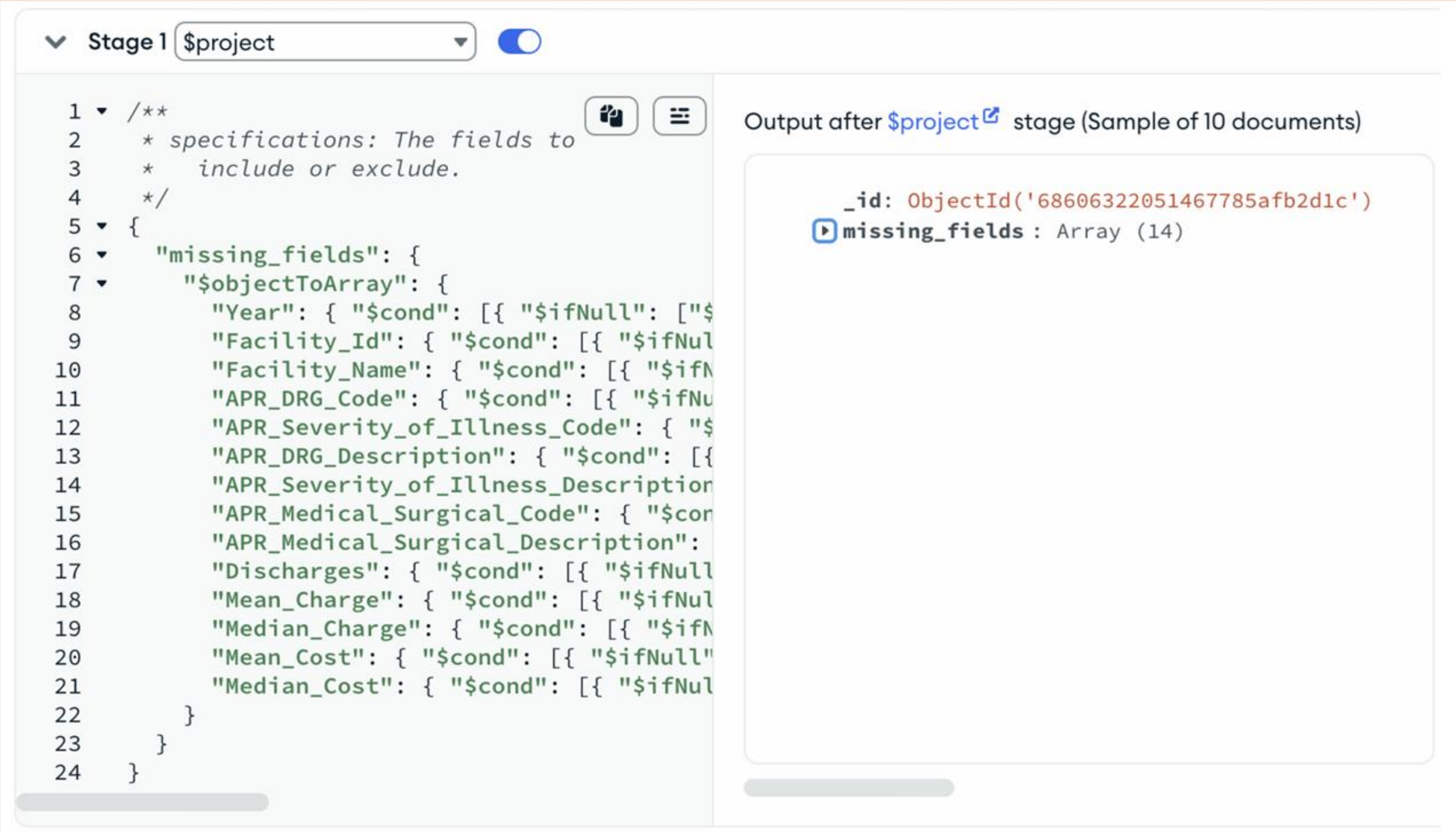
Step 3: Identifying Missing Values Using Aggregation Pipeline

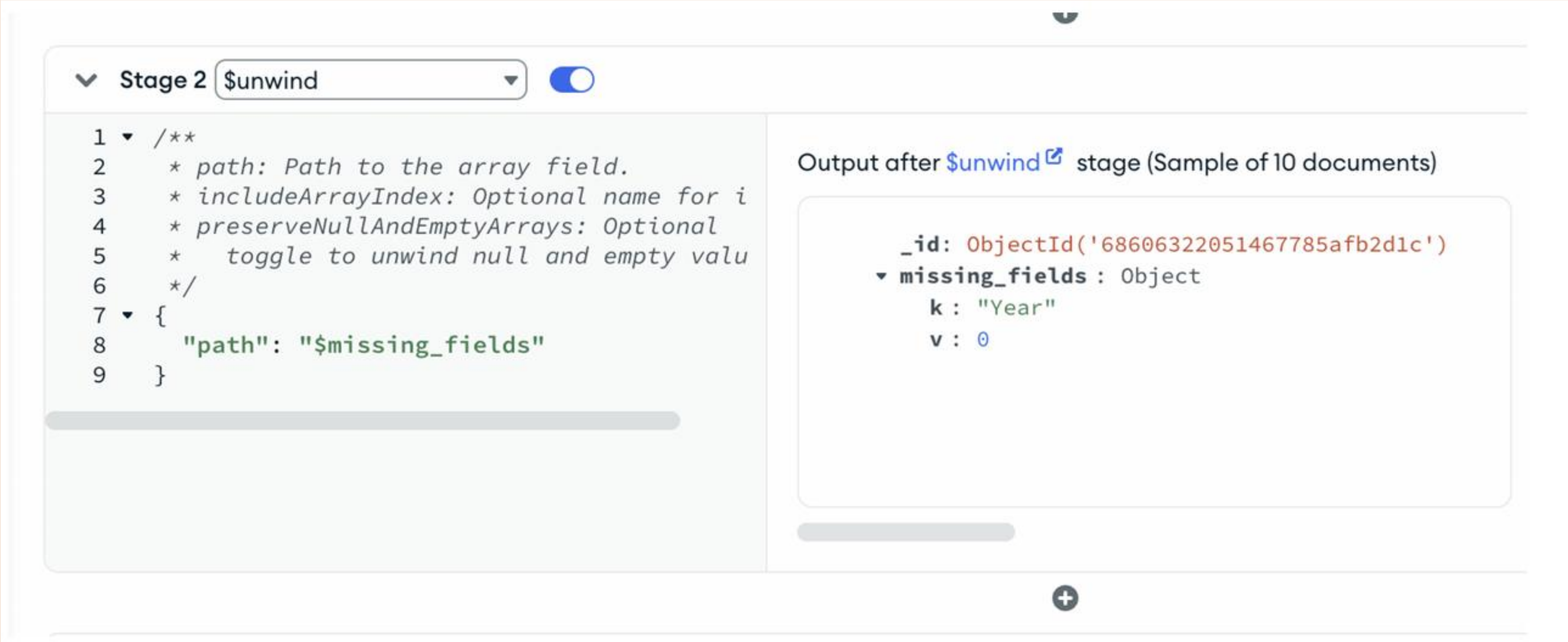


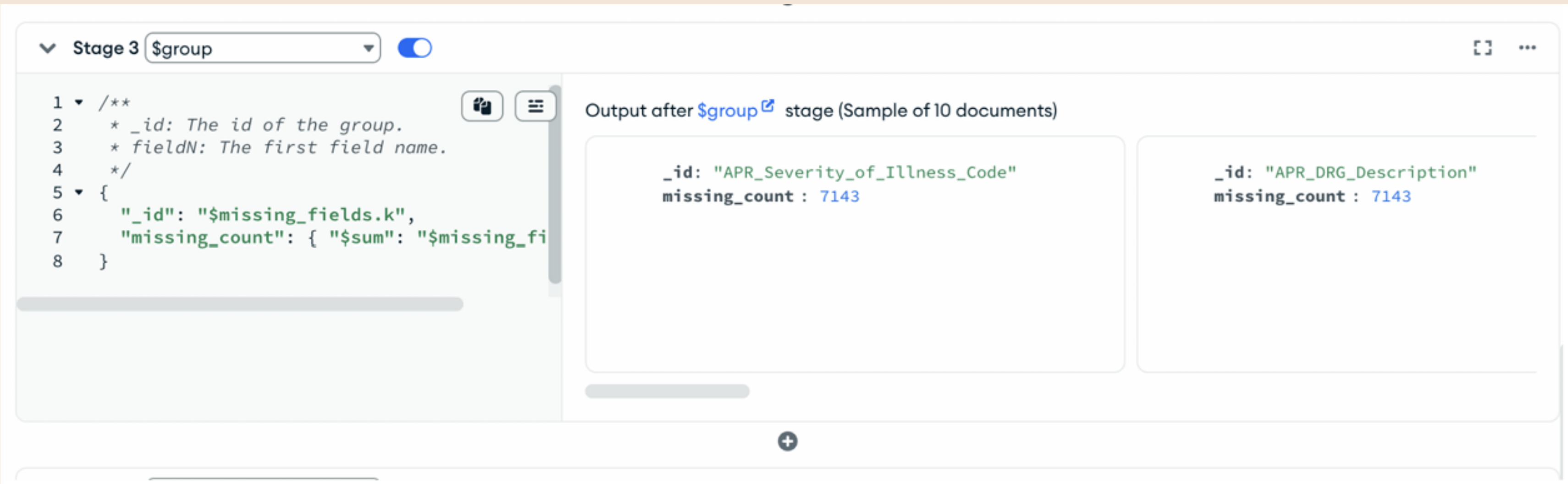


- Used **$project stage** to check for missing values across 14 key fields.
- Created a new array called missing_fields containing null-check results.
- **$unwind stage** broke down that array into individual field-check documents for easy analysis.

Step 3: Identifying Missing Values Using Aggregation Pipeline







- Used **$project stage** to check for missing values across 14 key fields.
- Created a new array called missing_fields containing null-check results.
- **$unwind stage** broke down that array into individual field-check documents for easy analysis.

Step 4: Handling Missing Values with $match

Stage 1 $match

```
{
  Year: { $exists: true, $ne: null },
  "Facility Id": { $exists: true, $ne: null },
  "Facility Name": {
    $exists: true,
    $ne: null,
    $ne: ""
  },
  "APR DRG Code": { $exists: true, $ne: null },
  "APR Severity of Illness Code": {
    $exists: true,
    $ne: null
  },
  "APR DRG Description": {
    $exists: true,
    $ne: null,
    $ne: ""
  },
  "APR Severity of Illness Description": {
    $exists: true,
    $ne: null,
    $ne: ""
  },
```

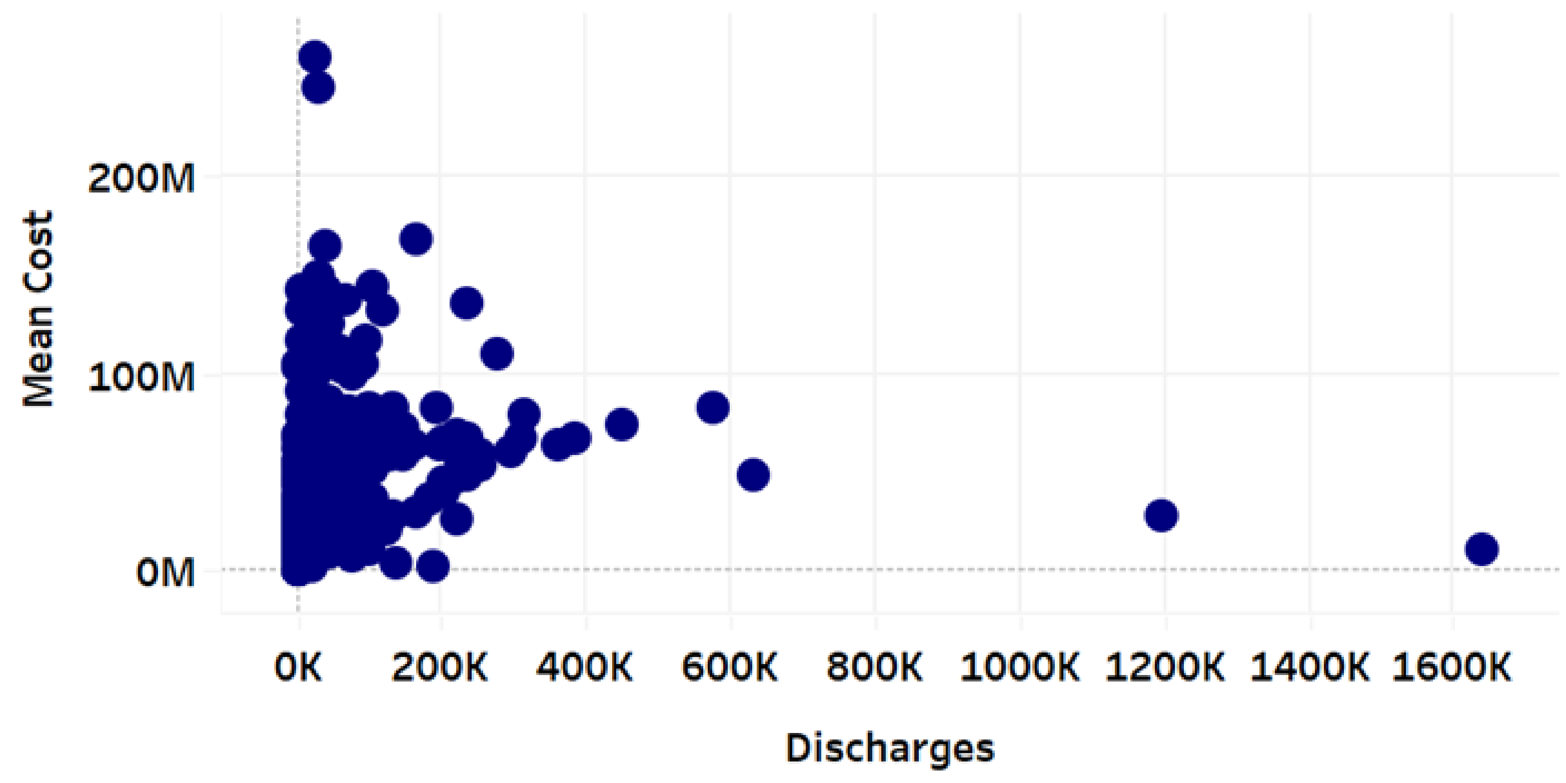
- Used a **$match filter** to keep only records where all essential fields are present, not null, and not empty.
- Focused on critical fields like Facility_Id, APR DRG Description, and Mean_Charge.
- Exported a clean dataset (~1.09M records) for reliable downstream analysis in Tableau.

ALL RESULTS

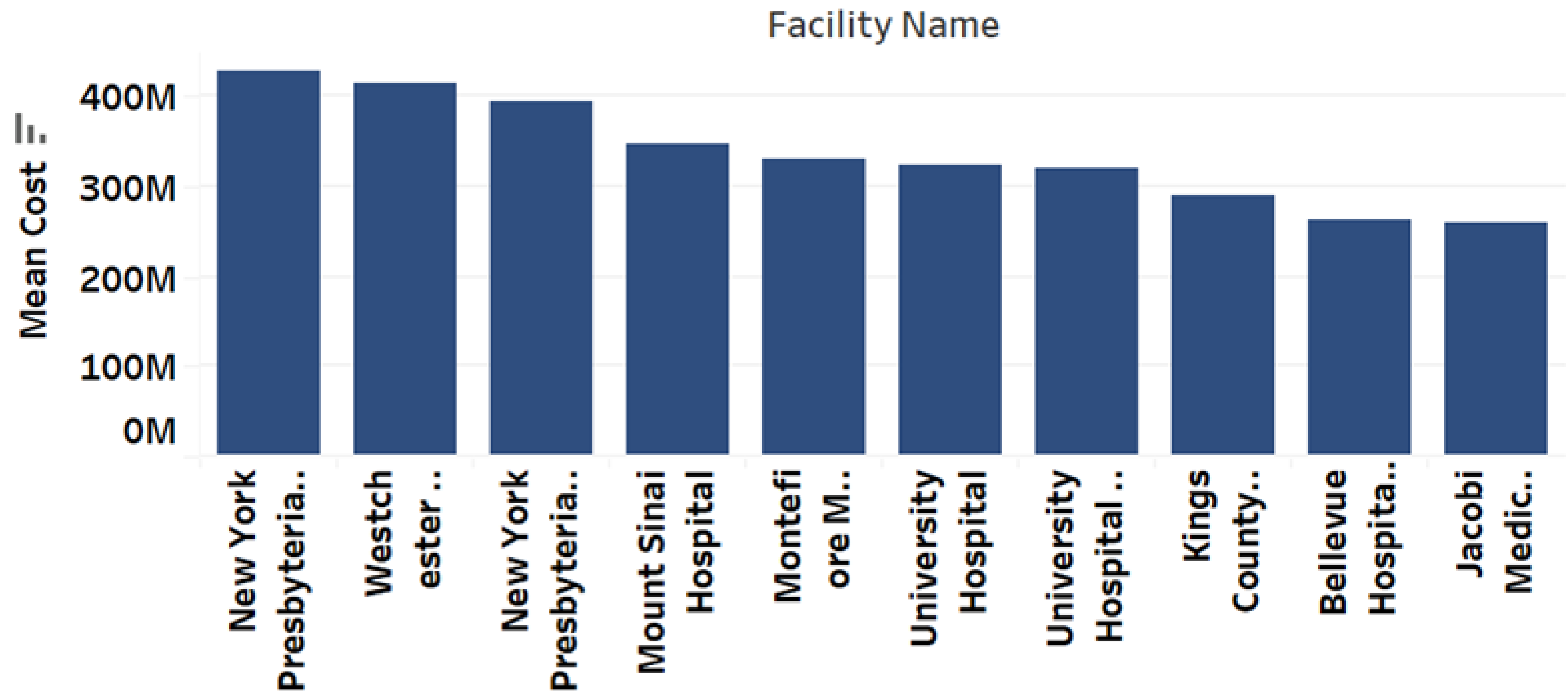
```
_id: null
total_docs : 1081672
Year_missing : 0
Facility_Id_missing : 0
Facility_Name_missing : 0
Mean_Charge_missing : 0
Mean_Cost_missing : 0
```

Dashboard 1: Drivers of High Inpatient Costs

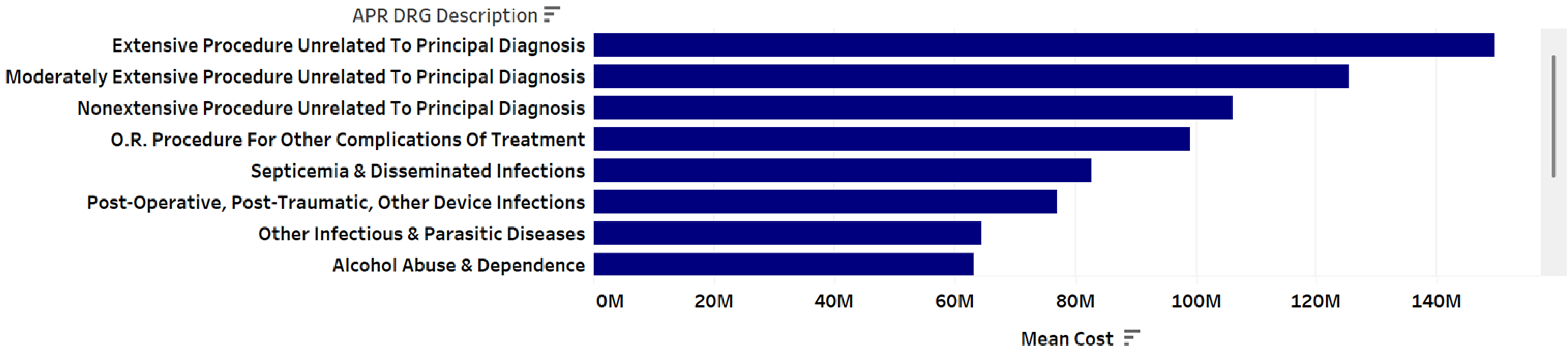
Discharges vs. Total Cost



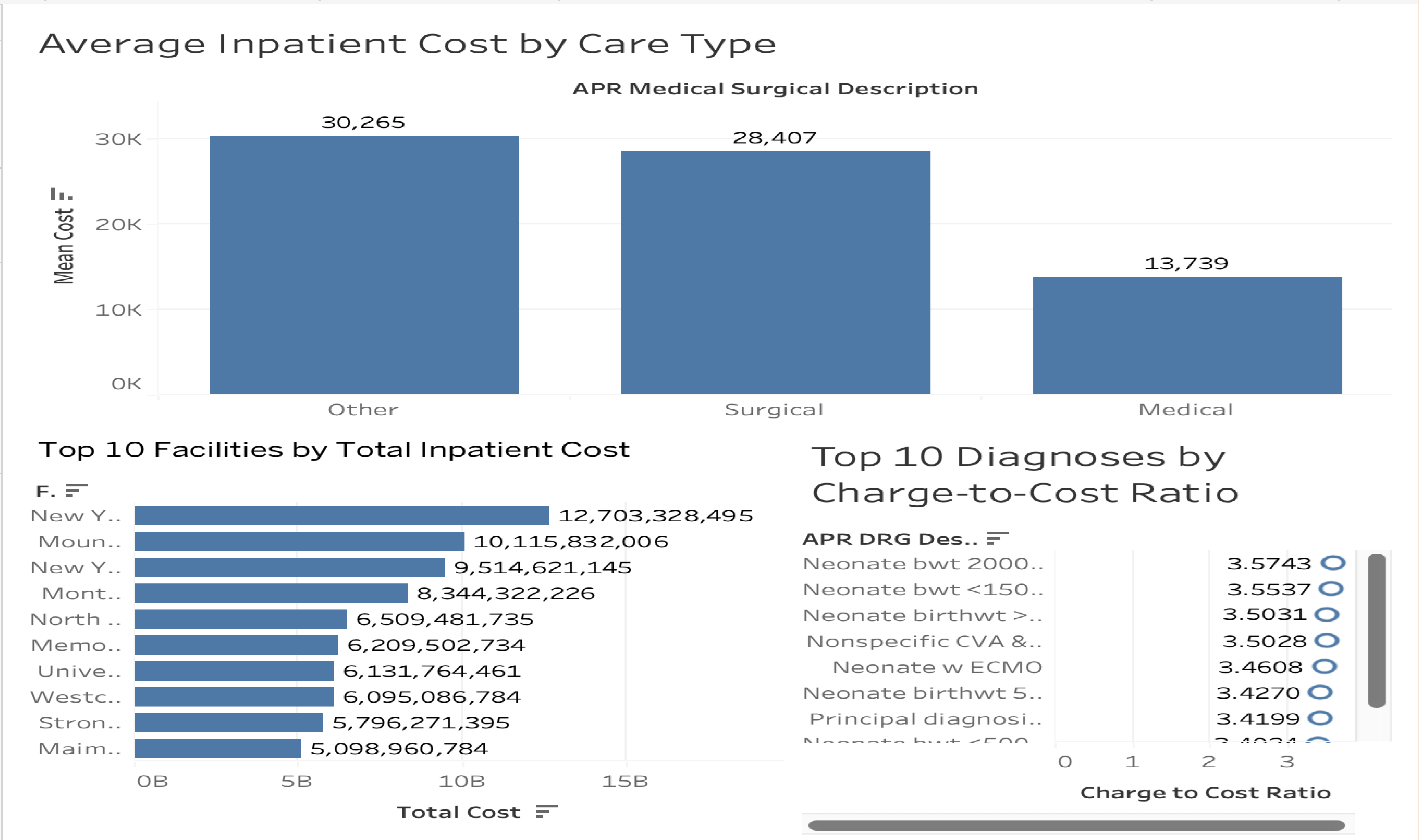
Top 10 Costliest Facilities



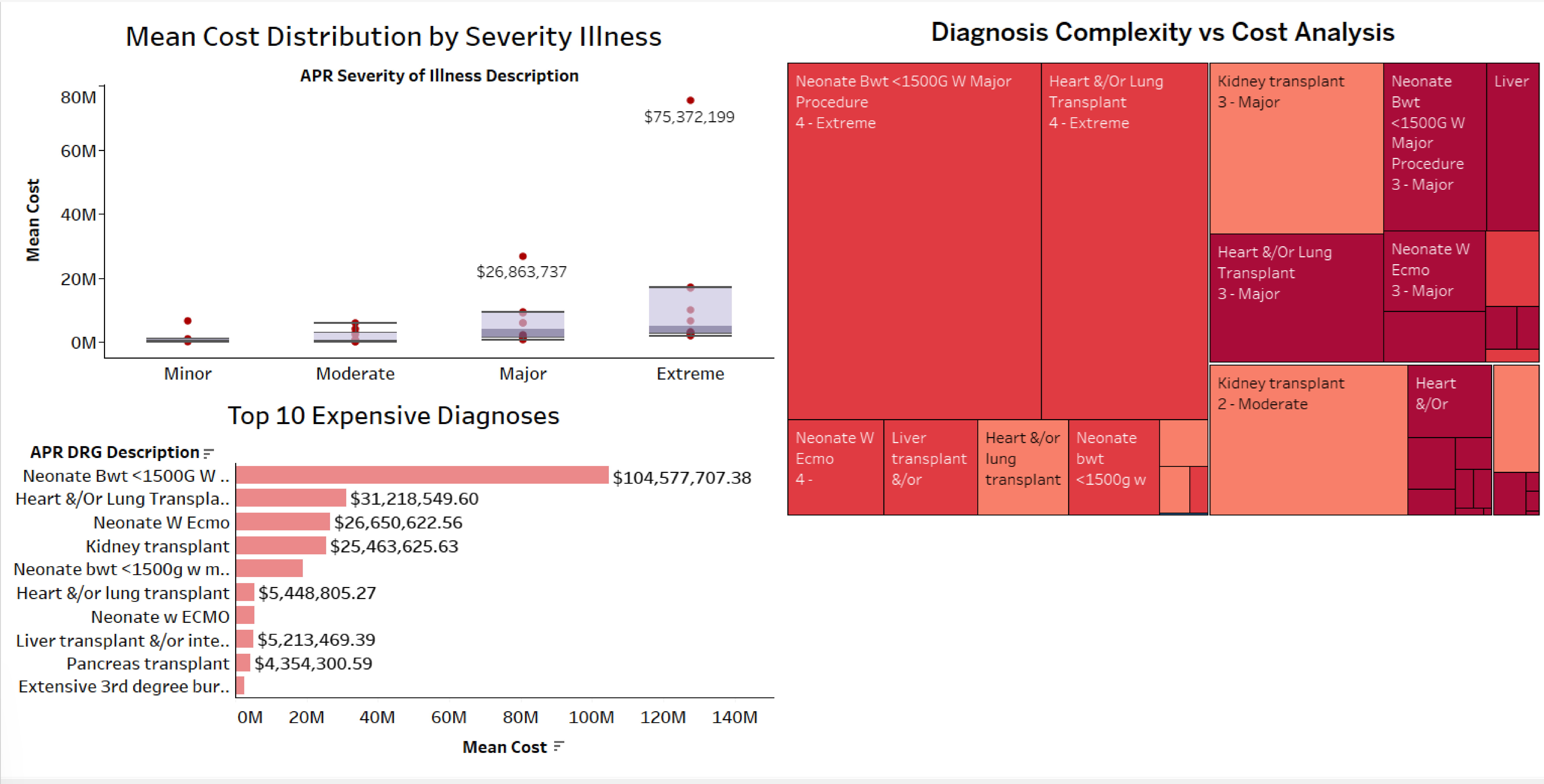
Top 10 Diagnoses by Average Cost



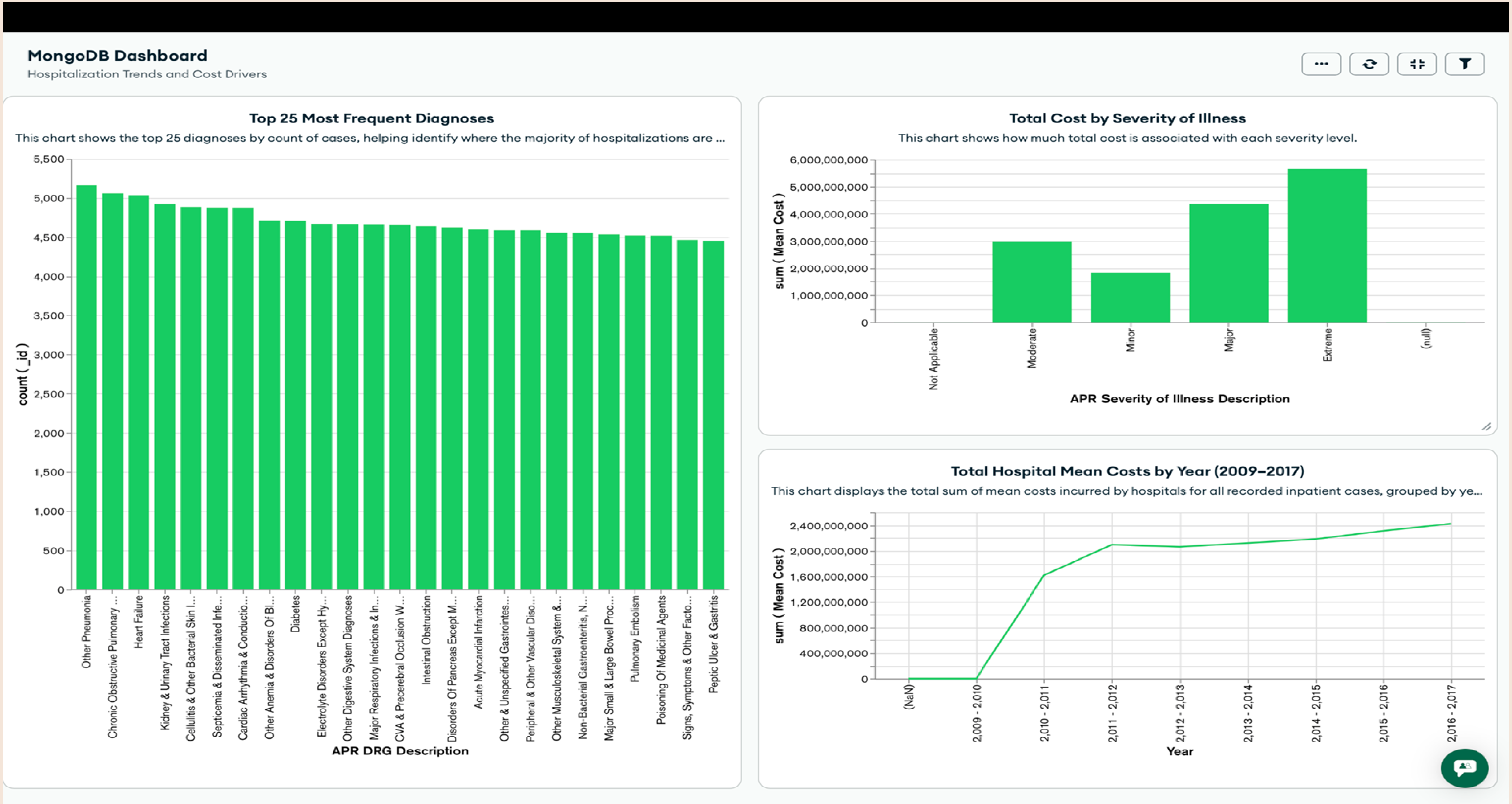
Dashboard 2: System-Level Cost Drivers and Transparency Gaps



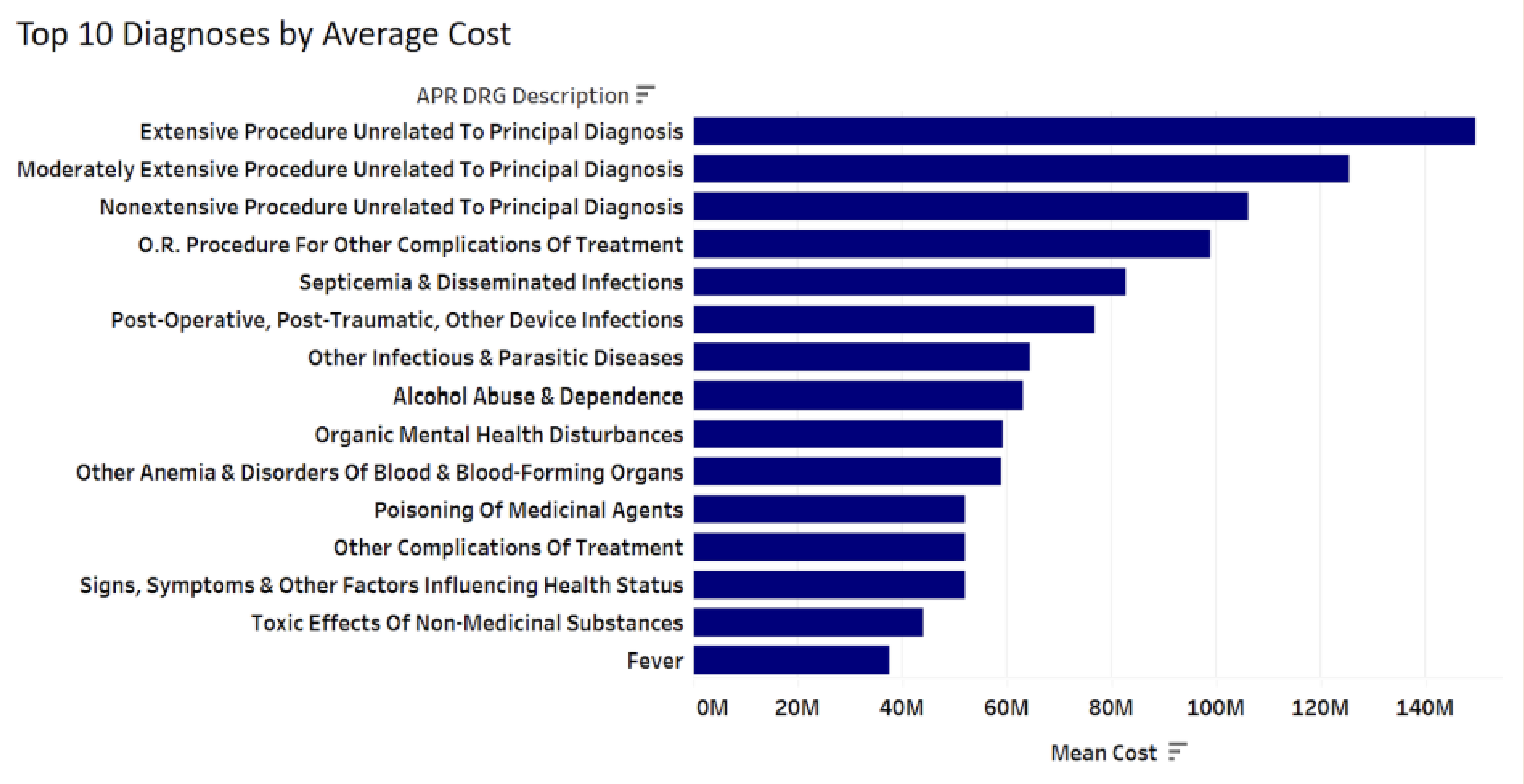
Dashboard 3: High-Impact Diagnoses and Cost Drivers



Dashboard 4: Hospital Cost Drivers: Diagnoses, Severity, and Time Trends

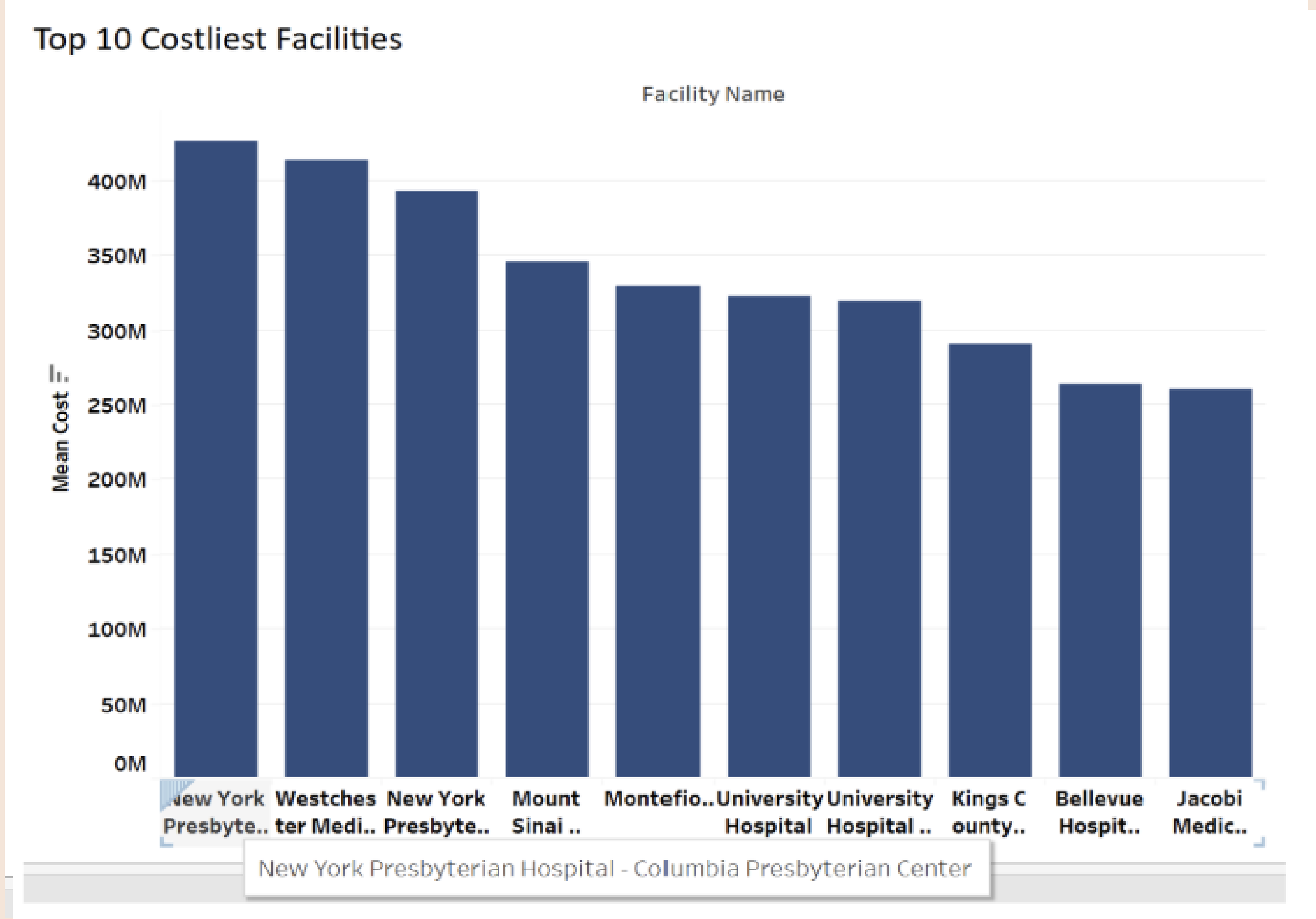


Unrelated and Post-Operative Procedures Drive the Highest Inpatient Costs Across Diagnoses



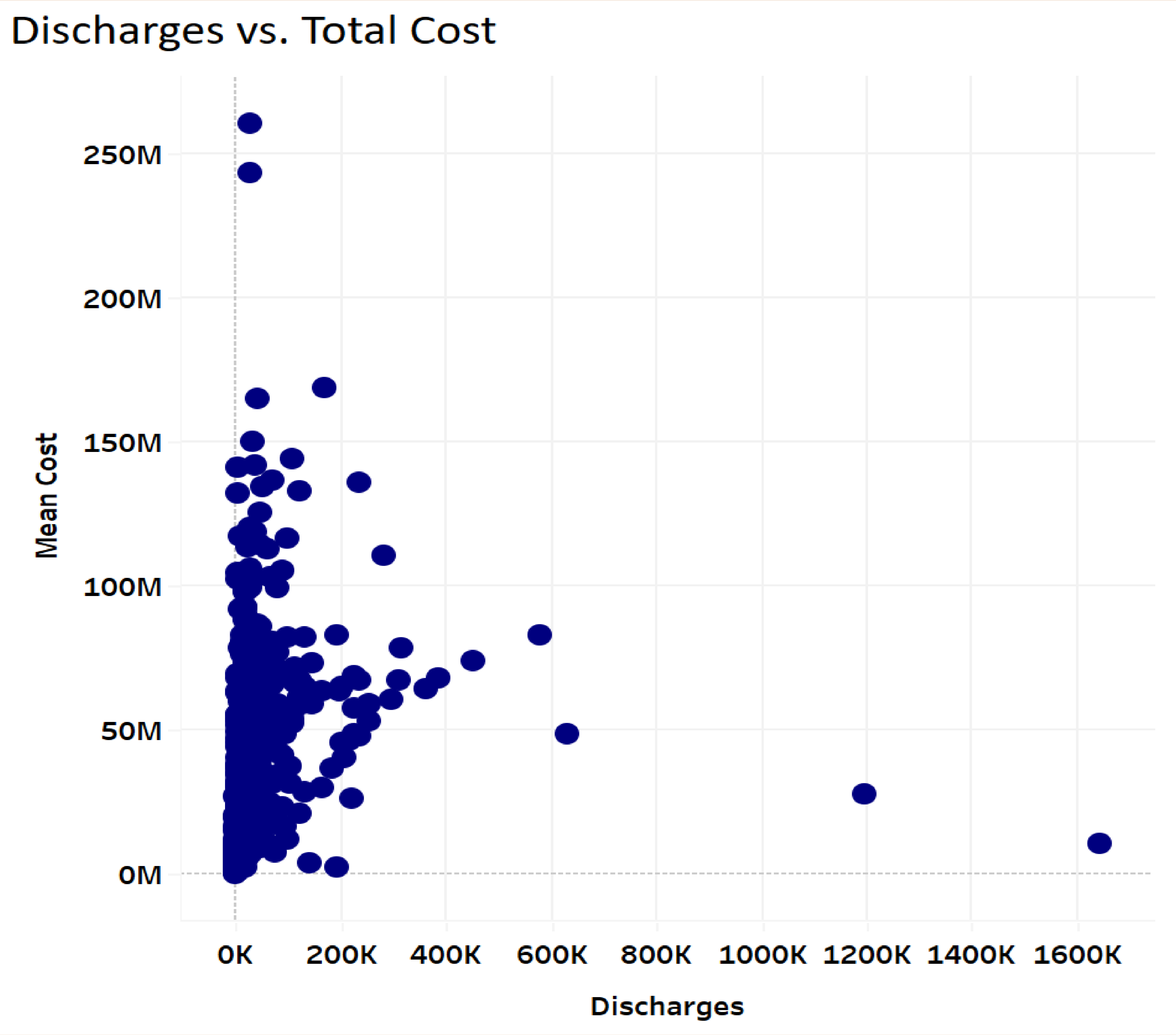
The visual highlights that procedures unrelated to the principal diagnosis and complex post-operative treatments account for the highest average inpatient costs, indicating these cases may involve complications, extended care, or resource-intensive interventions.

New York-Based Hospitals Lead in Average Inpatient Costs Across Facilities



New York Presbyterian Hospital incurs the highest inpatient costs, exceeding $400 million, indicating significant resource utilization that may be driven by high patient volume, advanced treatments, or specialized care.

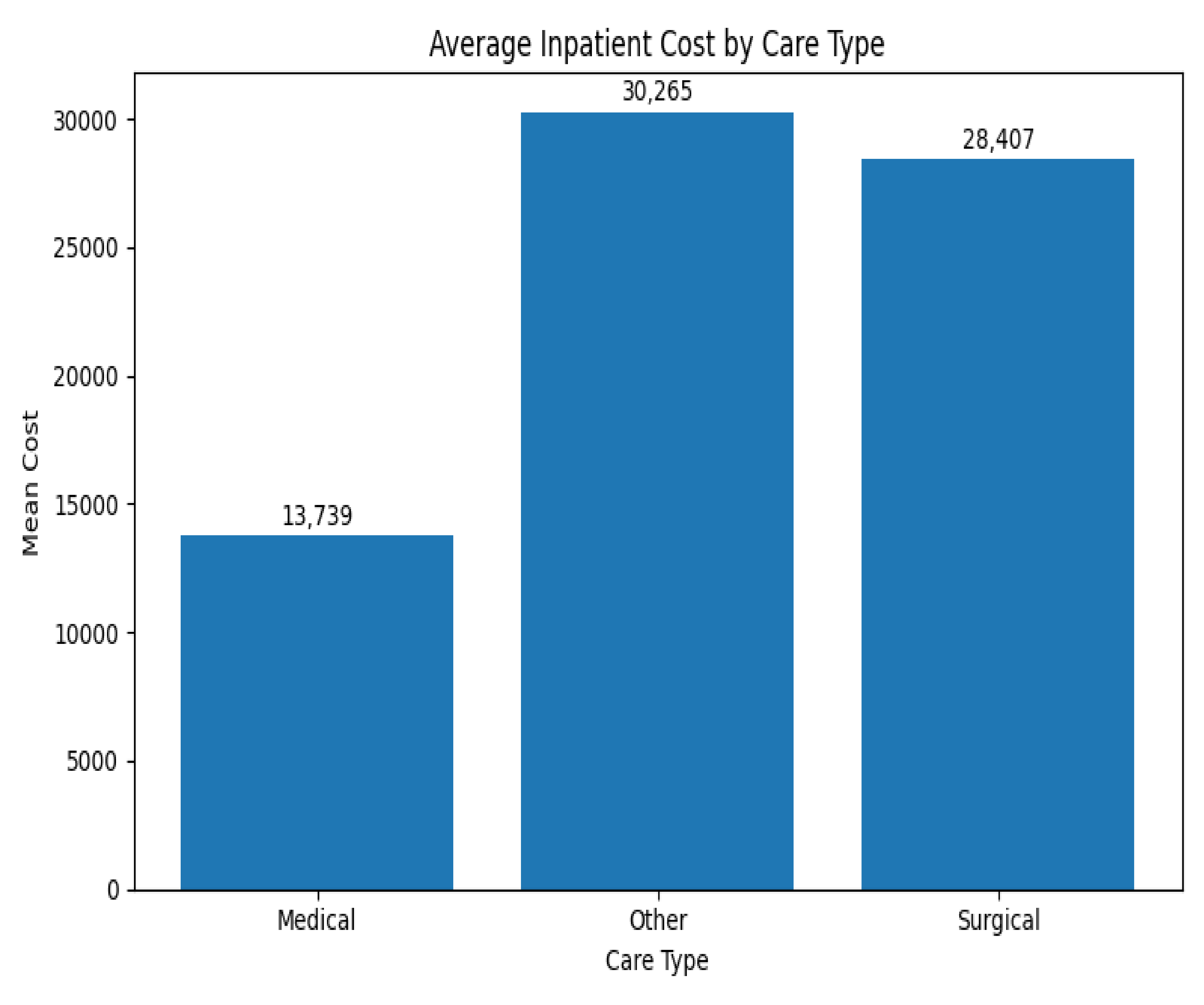
Costly Yet Rare Diagnoses Dominate Inpatient Spending



The scatter plot shows that while most diagnoses have moderate discharge volumes, some of the highest average costs are associated with relatively rare conditions indicating that low-frequency, high-complexity cases contribute disproportionately to overall hospital expenses.

Several outliers show high costs with low discharges, indicating rare but expensive treatments that may need further review.

Average Inpatient Cost by Care Type

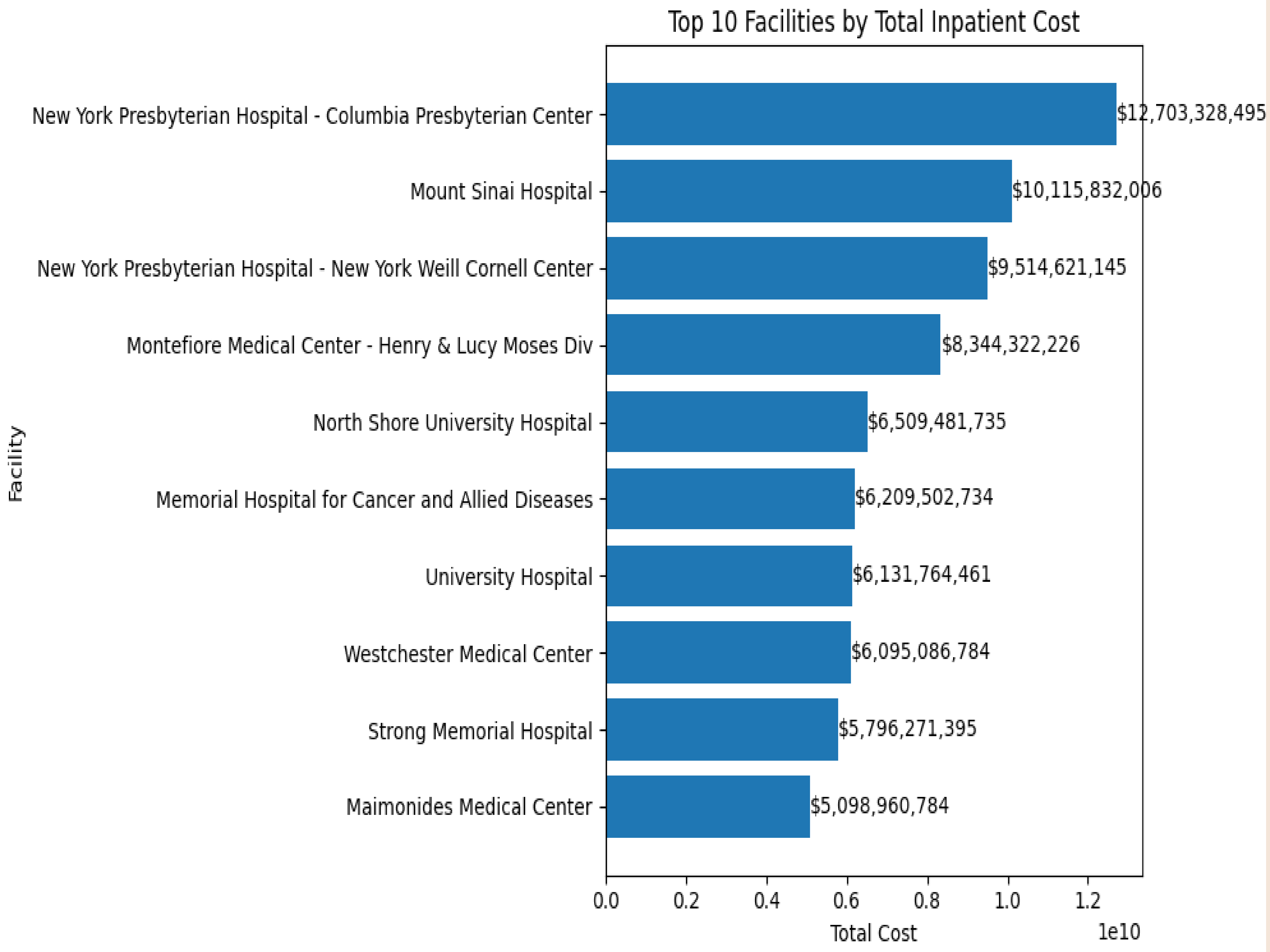


Insight:

- Surgical care is significantly pricier than Medical care.
- "Other" category appears to have the highest average cost overall.

Implication: Hospitals can evaluate whether high-cost surgical interventions are necessary or if non-surgical alternatives could reduce costs.

Top 10 Facilities by Total Inpatient Cost

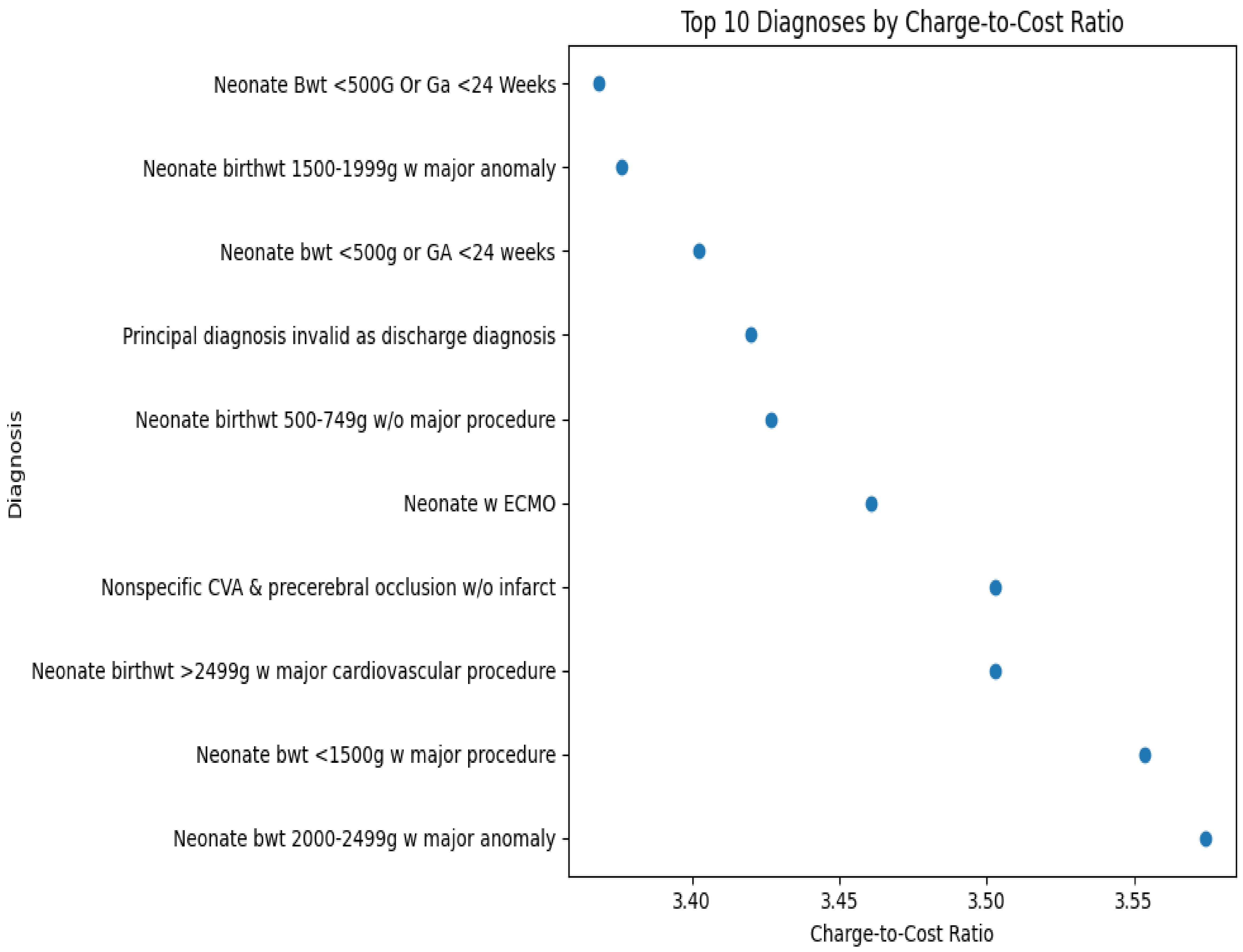


Insight:

- New York Presbyterian Hospital-Columbia Presbyterian Center is responsible for a massive share of inpatient costs.
- Total cost is not per patient -> It reflects the overall burden on the healthcare system.

Implication: These facilities should be focus points for cost containment, efficiency audits, or funding decisions.

Top 10 Diagnoses by Charge-to-Cost Ratio



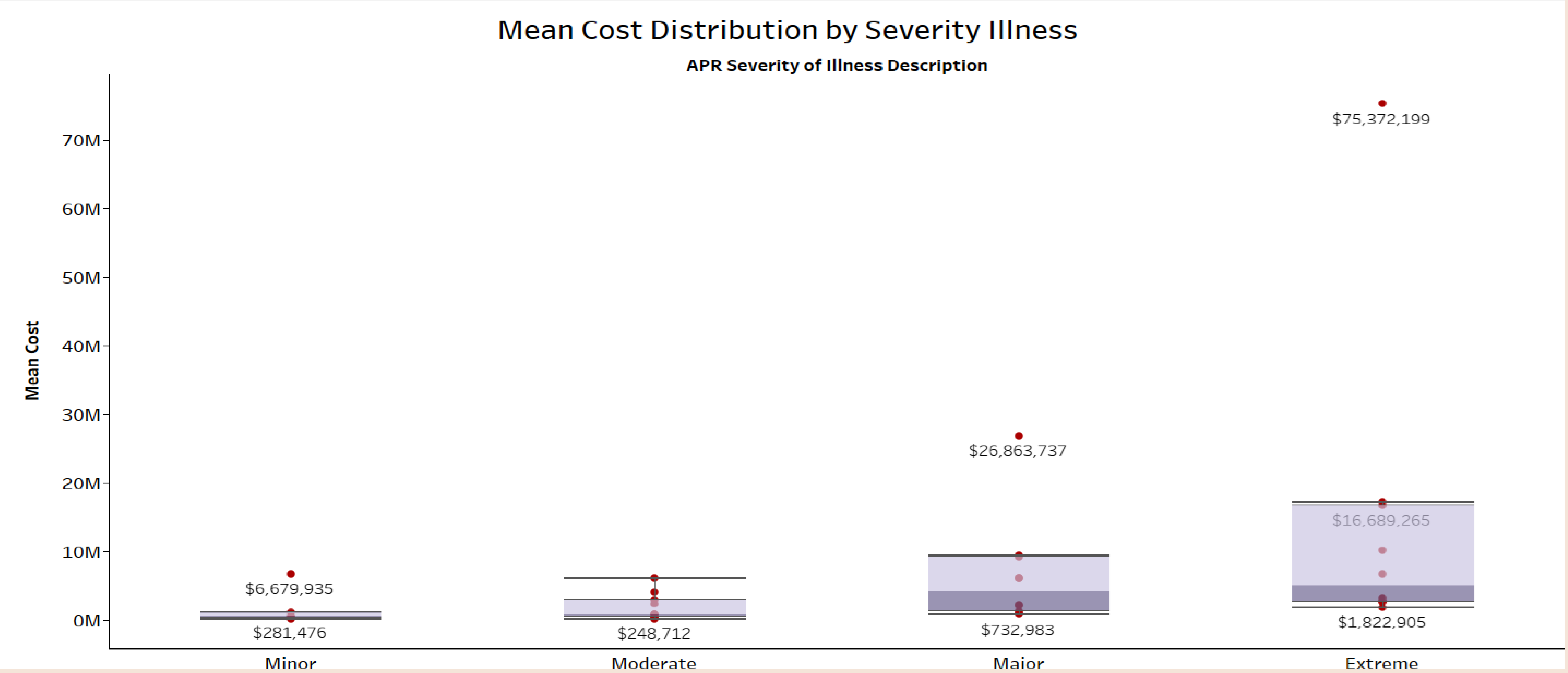
Insight:

- Neonate birthweight 2000–2499g w/ major anomaly has the highest charge-to-cost ratio, meaning hospitals charge over 3.5x what it costs them to treat.
- Most top diagnoses with high markups are neonatal-related, especially involving low birth weight or congenital anomalies.

Implication:

These diagnoses may require cost transparency reviews or billing audits to ensure ethical pricing, especially for vulnerable groups like newborns.

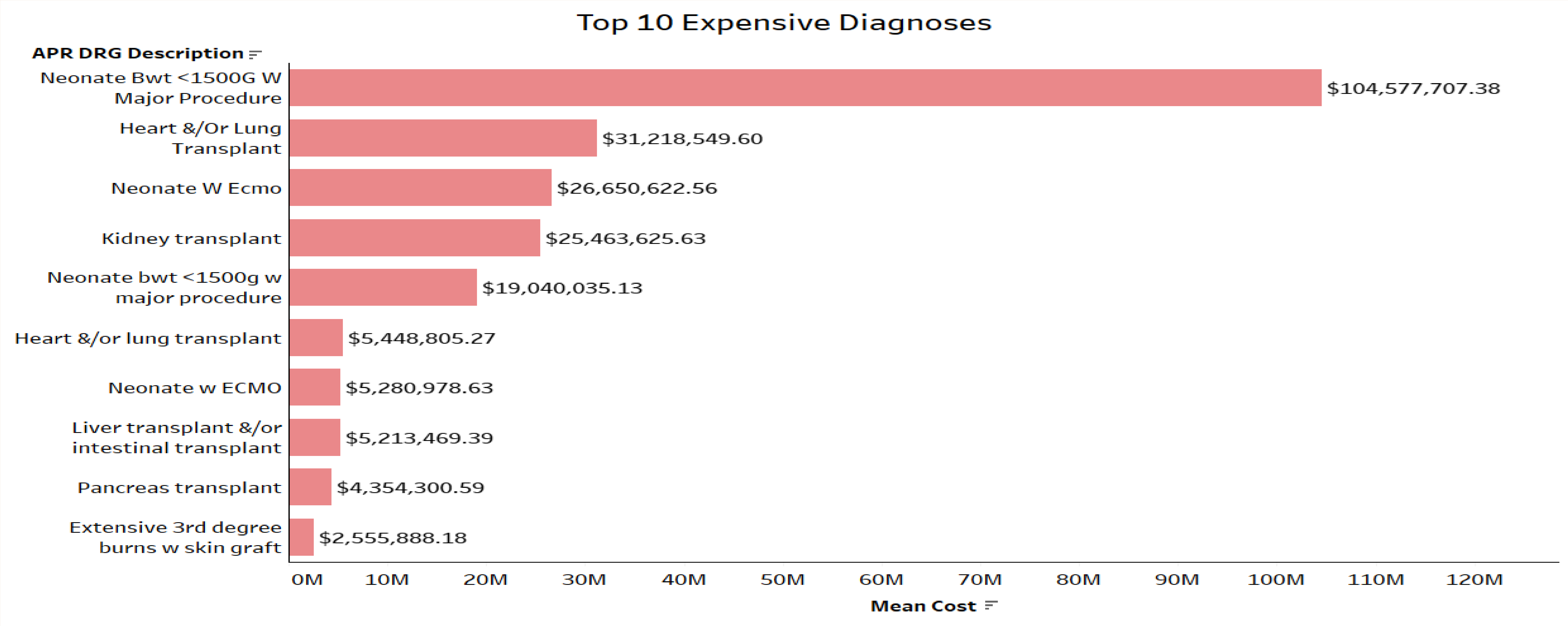
How Illness Severity Impacts Inpatient Treatment Costs



Insights: Patients classified under "Extreme" and "Major" severity levels have significantly higher mean costs compared to "Minor" and "Moderate" cases, indicating a direct relationship between severity and resource utilization.

The "Extreme" severity group shows the widest range and largest outliers, with costs exceeding $75 million, suggesting inconsistent treatment costs across facilities or complex care requirements.

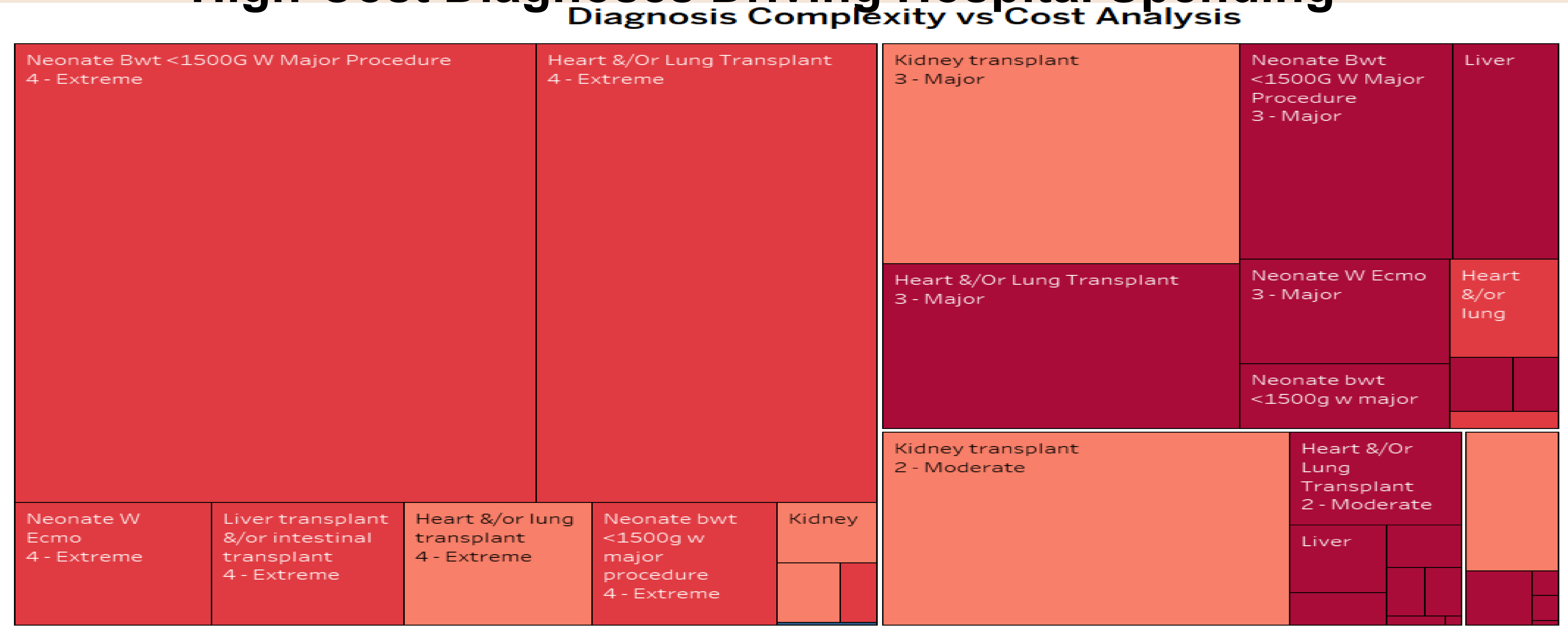
High-Cost Diagnoses Driving Hospital Spending



Insights: Tracheostomy procedures dominate as the two most expensive categories, exceeding $240M in costs.

Top 3 diagnoses account for significantly higher costs than remaining 7, suggesting focused management on these high-impact procedures.

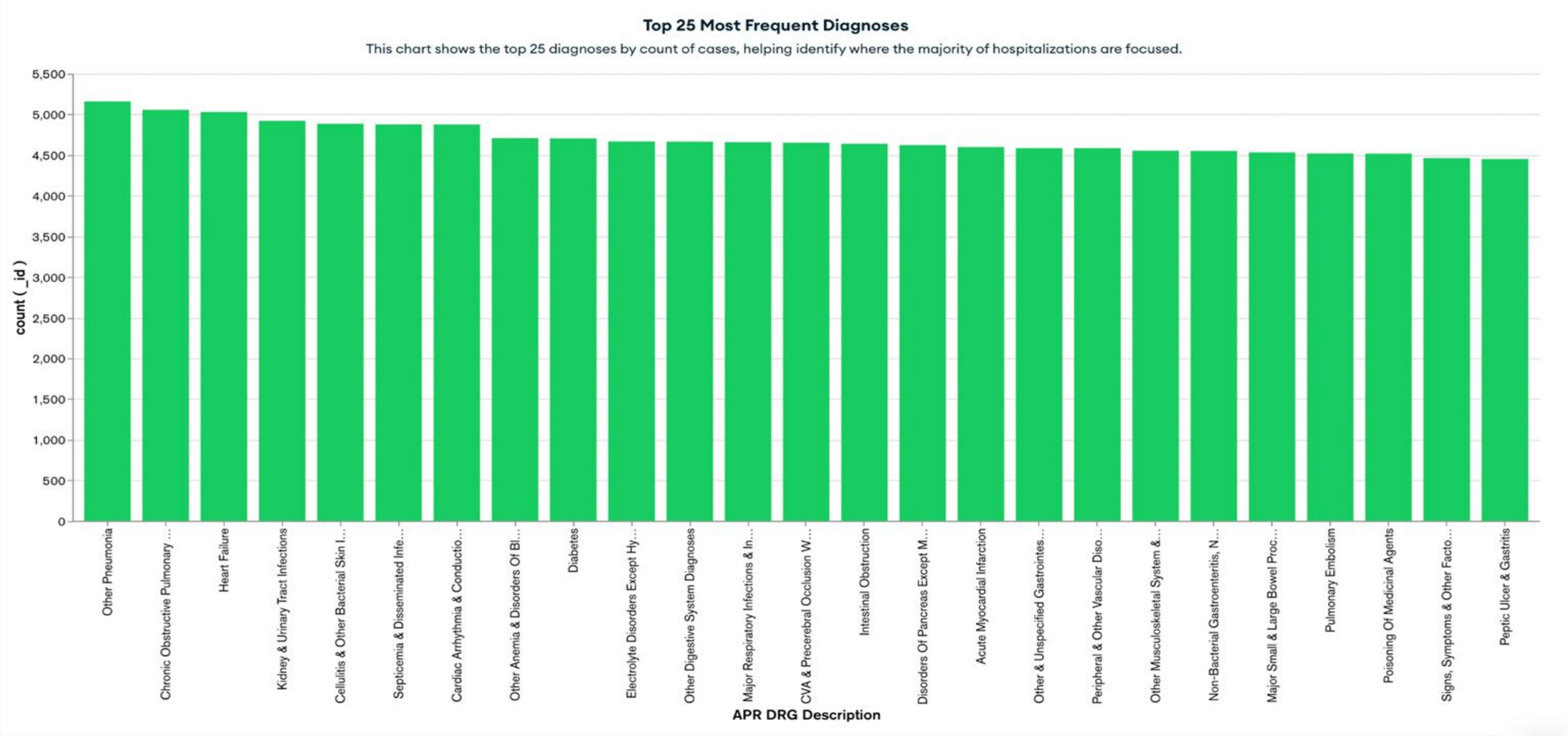
High-Cost Diagnoses Driving Hospital Spending



Insights: Neonate <1500G procedures represent the perfect storm of high volume (largest rectangles) and extreme costs (darkest red), making them priority #1 for cost reduction efforts.

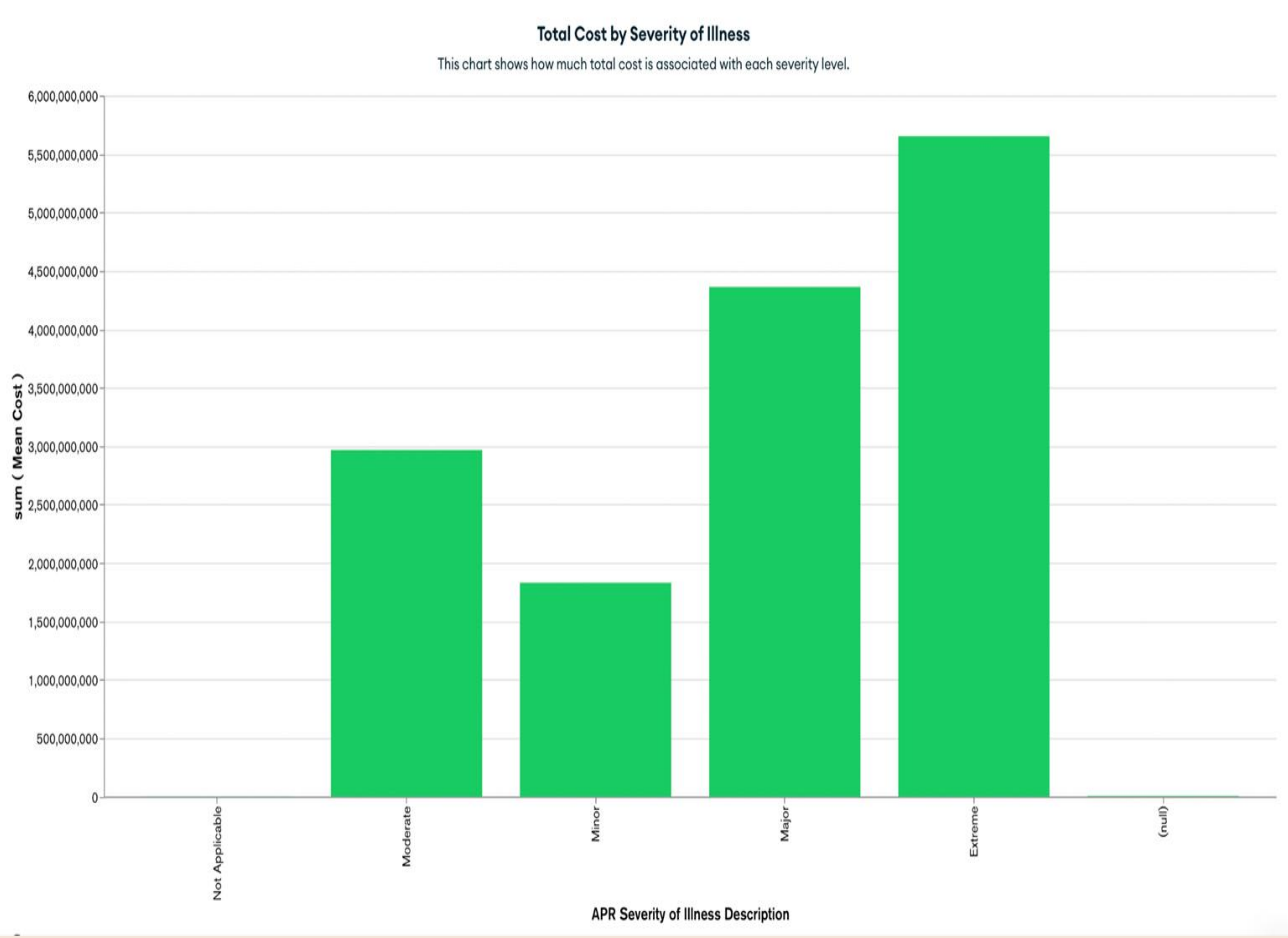
Clear color progression from orange (moderate severity) to deep red (extreme severity) confirms that illness complexity directly drives costs, supporting severity-based pricing and resource allocation strategies.

Top 25 Most Frequent Diagnoses



We began our **MongoDB dashboard** by identifying the **top 25 diagnoses** leading to **hospital admissions**. Conditions like **pneumonia**, **COPD**, and **heart failure** were most frequent, highlighting **key pressure points** in the **healthcare system**. This helped us understand where hospitals face the **highest demand** and set the stage for further **cost** and **severity analysis**.

Total Cost by Severity of Illness

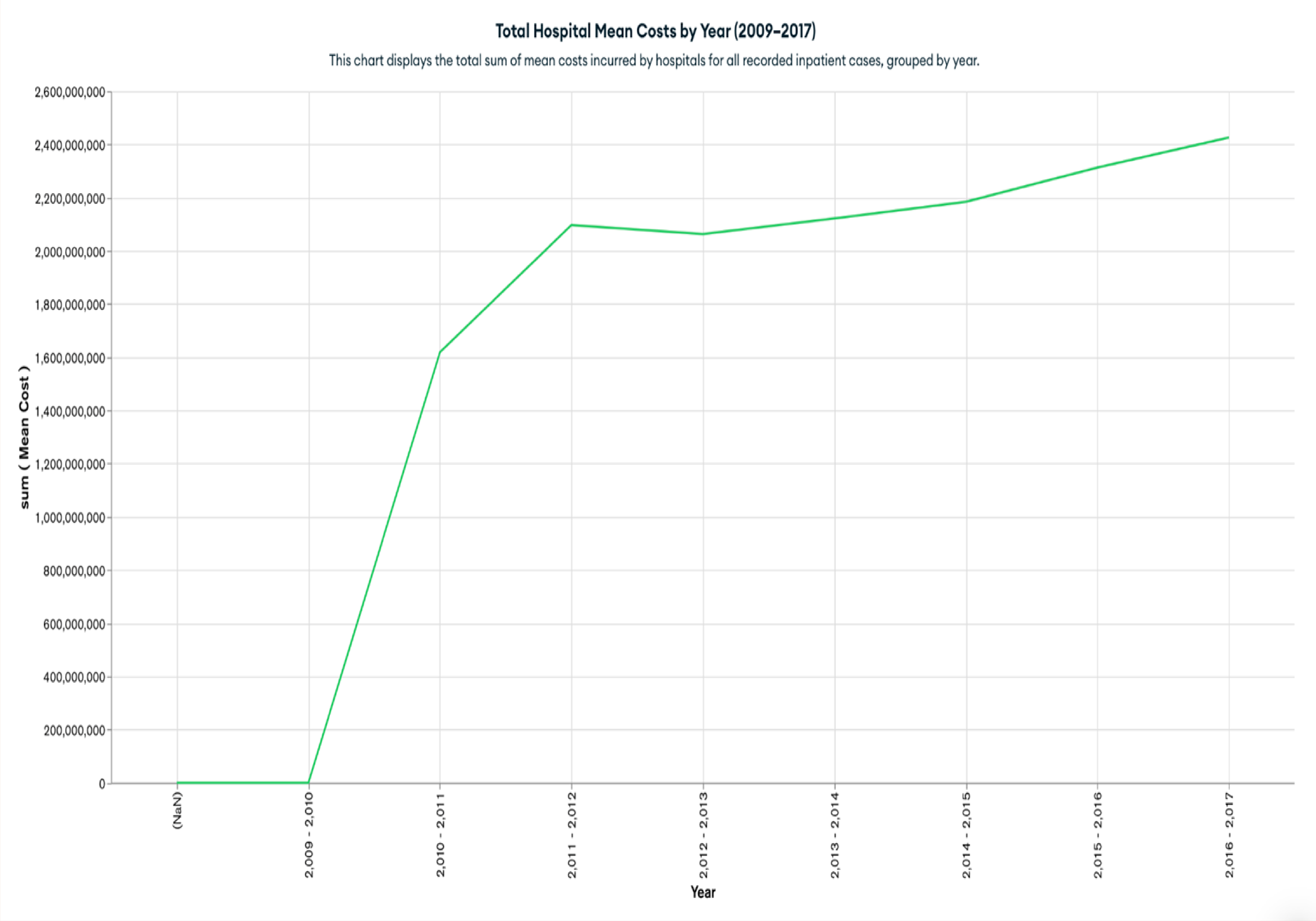


We explored how **hospital costs** vary with the **severity of illness** using the **APR Severity of Illness Description** field. This **column chart** visualizes the **total mean cost** across severity levels, showing how expenses rise with **clinical complexity**.

The findings show that **Extreme** and **Major** cases account for the **largest share** of costs, despite being **less frequent**. This highlights a steep increase in **financial burden** with higher severity.

These insights suggest **opportunities for cost control**, especially by improving **care pathways** or investing in **preventive strategies** for **high-severity** patients.

Total hospital mean costs from 2009 to 2017



This **line chart** shows the **total hospital mean costs** from **2009 to 2017**, helping us observe how **inpatient expenses** have changed over the years.

There was a noticeable **spike between 2010 and 2011**, followed by a gradual increase through 2017. This trend reflects the **rising cost of care** over time.

These findings highlight the need for **long-term cost monitoring** and can support better **financial planning** for hospitals and policymakers.

Key Findings

Major discoveries

- **Cost Distribution:** Hospital costs follow a log-normal pattern with 80-90% between $1,000-$100,000.
- **Charge-Cost Correlation:** Extremely strong relationship (0.98) between charges and actual costs
- **Severity Impact:** Clear correlation (0.35) between illness severity and treatment costs.
- **Volume-Cost Relationship:** Inverse relationship - low-volume treatments are significantly more expensive

Critical Insights from Visualizations

- **Charge-to-Cost Ratios:** Consistent across diagnosis types, validating pricing transparency
- **Facility Performance:** Large hospitals achieve economies of scale with lower per-patient costs
- **Complexity vs. Cost:** Clear treemap visualization shows relationship between diagnosis complexity and expenses

Final Conclusion and Recommendations

Conclusion:

Our big data analysis of 1M+ hospital records reveals that **healthcare costs are largely predictable and follow logical patterns** based on illness severity, treatment complexity, and hospital specialization. The strong correlation between charges and costs (0.98) indicates transparent pricing practices in most cases.

Recommendations

- **Implement Tiered Pricing Models:** Develop severity-based pricing that reflects actual resource requirements and promotes transparency
- **Audit High-Cost Outliers:** Investigate cases exceeding $7M to identify root causes and implement cost controls
- **Optimize Specialty Care:** Focus resources on high-volume hospitals to achieve economies of scale
- **Enhance Cost Transparency:** Provide patients with predictive cost estimates based on DRG codes and severity levels

Strategic Outcomes and Future Steps

Strategic Impacts & Benefits

- **For Hospitals:** Data-driven resource allocation and identification of cost optimization opportunities
- **For Patients:** Better cost predictability with 90% of procedures falling in $10K-$50K range
- **For Policymakers:** Evidence-based insights for preventive care investments in high-volume conditions
- **For Healthcare System:** Improved efficiency through centralization of complex, low-volume treatments

Future Directions

- **Preventive Care Focus:** Target common conditions like heart failure and pneumonia with early intervention programs
- **Specialty Hospital Optimization:** Develop regional centers of excellence for rare, high-cost treatments
- **Real-time Cost Monitoring:** Implement dashboard systems for continuous cost tracking and management
- **Patient Financial Planning:** Create tools using our findings to help patients prepare for healthcare expenses

References

Primary Dataset: New York State Hospital Inpatient Discharges (SPARCS) 2009-2017

Source: Kaggle Dataset - "Hospital Charges in US Exploratory Analysis". https://www.kaggle.com/datasets/wajahat1064/hospital-inpatient-cost-data-by-new-york-state

NoSQL Databases: MongoDB. (n.d.-c). What is NoSQL? NoSQL databases explained. https://www.mongodb.com/resources/basics/databases/nosql-explained

Original Data Provider: New York State Department of Health

Weinstein, M. C., et al. (2010). *Comparative effectiveness research what it is and what it can offer. New England Journal of Medicine*, 362(5), 460–465. https://www.nejm.org/doi/full/10.1056/NEJMsb0902534

Luo, J., et al. (2016).*Big data application in biomedical research and health care: a literature review.Biomedical Informatics Insights*, 8, 1–10. https://doi.org/10.4137/BII.S31559

Kumar, A., & Khajuria, A. (2020). *Exploring cost determinants in inpatient care: A systematic review. Health Economics Review*, 10(1), 12. https://healtheconomicsreview.biomedcentral.com/

Healthcare Cost and Utilization Project (HCUP). *Overview of the APR-DRG System.* https://www.hcup-us.ahrq.gov/toolssoftware/aprdrg/aprdrg.jsp

Thank You