Project Title: Exploring the factors behind Hospital Readmission Rates

Team Members:

- Indirapriyadarshini Arunachalam (iarunachalam@umass.edu) - SPIRE ID: 34743317
- Athulya Anil (athulyaanil@umass.edu) - SPIRE ID: 34760586
- Rakshith Venkatesh (rakshith@umass.edu) - SPIRE ID: 34767593

Project Repository:
https://github.com/athulya-anil/hospital-readmission

Hospital Readmission Rates and Causes

Overview and Motivation:

Hospital readmissions are a persistent challenge in healthcare, carrying implications for both patient well-being and hospital performance. High readmission rates often point to underlying issues in care coordination, chronic disease management, or social determinants of health. Our project aims to shed light on these patterns through thoughtful and interactive visualizations that make hospital readmission data both accessible and insightful.

With experience in healthcare data engineering and software development, our team was motivated by a shared interest in bridging the gap between raw clinical datasets and meaningful insights. We set out to answer not just "what are the readmission rates?" but also "why are they happening?" and "where can we improve?" Through interactive exploration of trends by condition, geography, and demographics, we aspire to support better, data-informed decisions by hospital administrators, healthcare providers, and policy analysts alike.

Related Work:

Our work was guided by a blend of public health research, real-world hospital dashboards, and design patterns from previous COMPSCI 571 projects.

On the healthcare front, our primary dataset came from the **Centers for Medicare & Medicaid Services (CMS)**, which also maintains the **Hospital Compare** portal—a platform that visualizes hospital-level quality metrics. These dashboards, while informative, often require domain knowledge to interpret. We sought to build a more user-friendly interface that preserves the analytical richness of the CMS data while being accessible to a broader audience.

Academic research played a pivotal role in framing our analysis. Studies show that conditions such as **heart failure**, **chronic obstructive pulmonary disease (COPD)**, and **diabetes** are leading contributors to hospital readmission. Additionally, factors like **patient age**, **length of stay**, and **discharge disposition** are known to influence readmission likelihood. These findings helped shape the structure of our visualizations, with a strong emphasis on condition-based comparisons, demographic filters, and hospitalization metrics.

We were also inspired by well-designed data storytelling platforms like **Our World in Data** and dashboards from **HealthData.gov**, which showcase how large, complex datasets can be distilled into clear visual narratives. Their use of interactivity, progressive disclosure, and well-structured filtering mechanisms directly influenced our design.

In terms of past COMPSCI 571 work, the **“Pokedata”** project by Hoang and Siu served as an excellent example of how to structure interactive dashboards using linked views, scatterplots, and customized filtering. Their radar chart for stat breakdowns and coordinated brushing/filtering across views shaped how we thought about interactivity in our own tool—especially when letting users explore patterns by condition and demographic group. Another project — *“**Making Earthquake Data Accessible**”* by Walsh, Treviño, and Bett — reinforced the importance of maintaining consistency in visual grammar and layering data in a way that doesn’t overwhelm users. Their approach to integrating geographic, tabular, and comparative visualizations validated our decision to combine a choropleth map, condition-specific bar charts, and interactive filters within a unified dashboard.

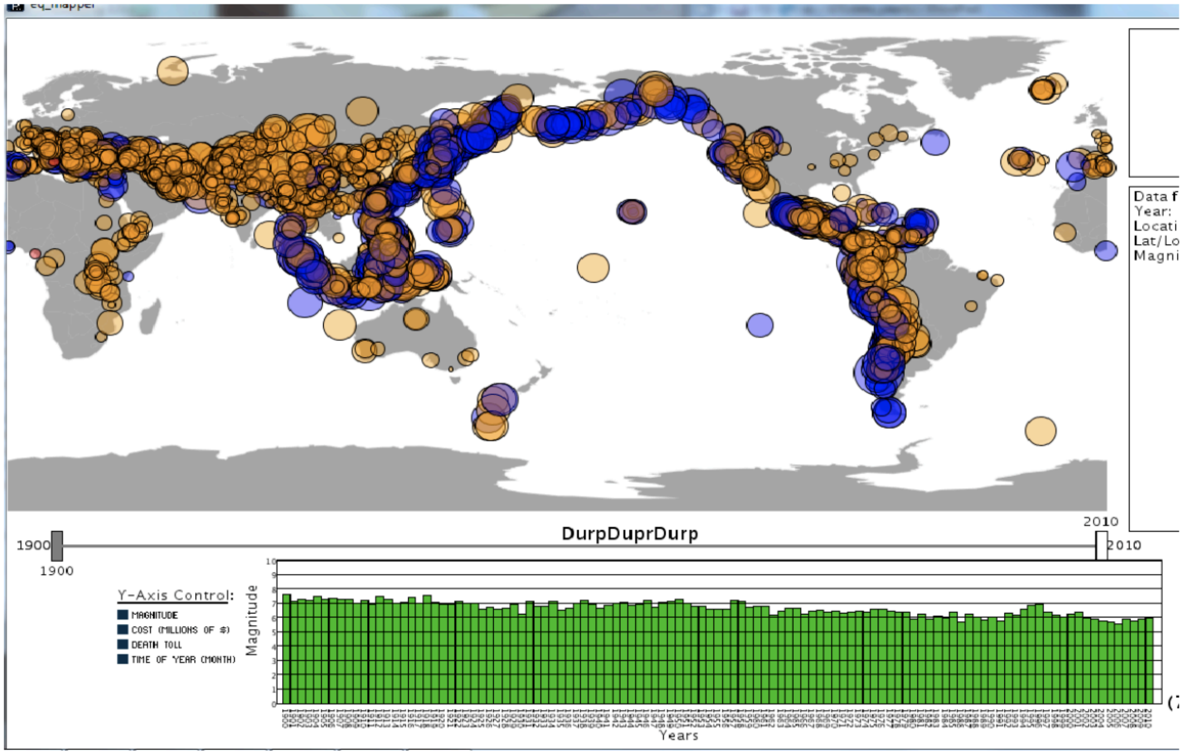
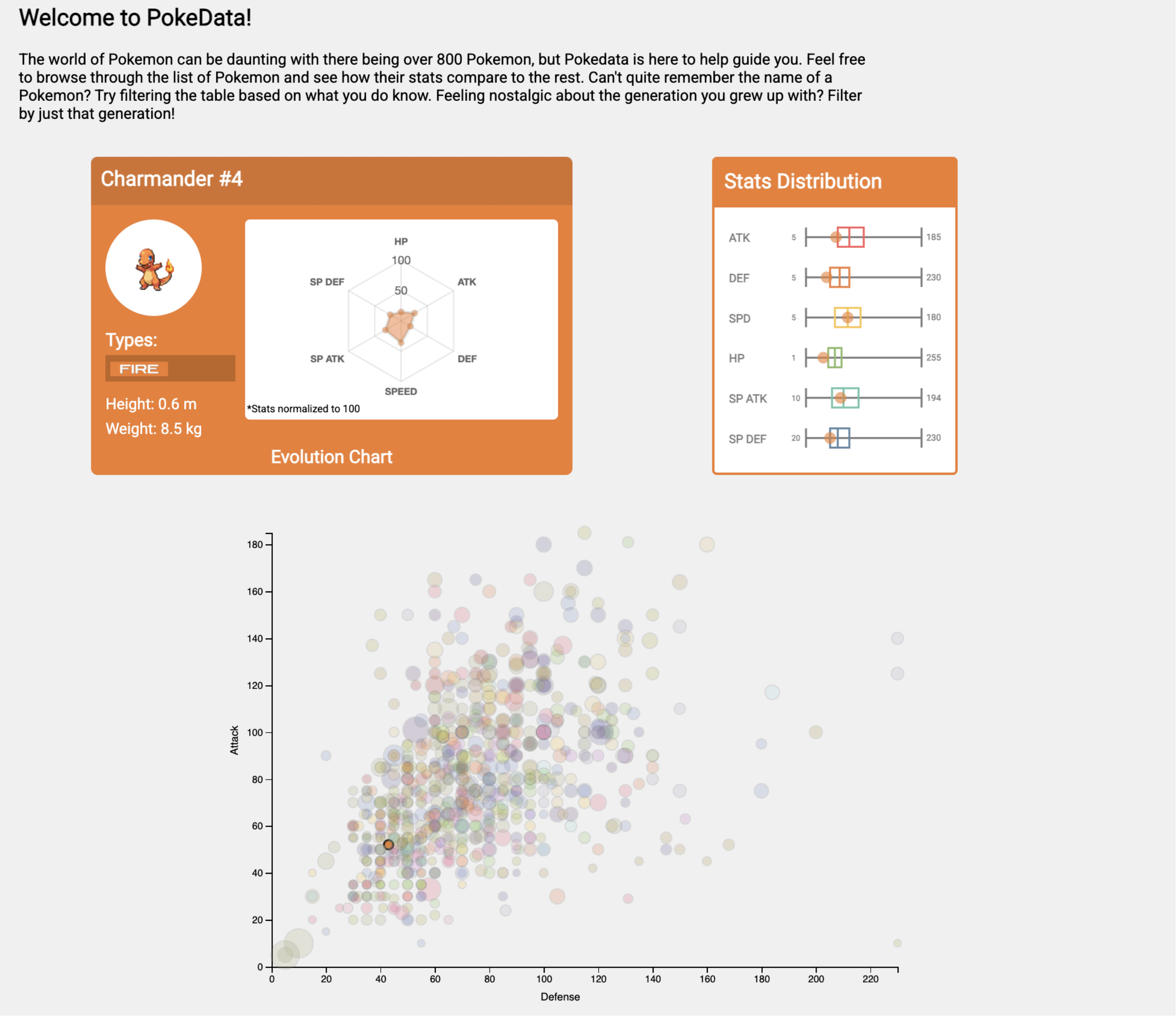


Figure 1 : Choropleth map inspiration from *“**Making Earthquake Data Accessible**”*



Welcome to PokeData!

The world of Pokemon can be daunting with there being over 800 Pokemon, but Pokedata is here to help guide you. Feel free to browse through the list of Pokemon and see how their stats compare to the rest. Can't quite remember the name of a Pokemon? Try filtering the table based on what you do know. Feeling nostalgic about the generation you grew up with? Filter by just that generation!

Charmander #4

Types: FIRE

Height: 0.6 m

Weight: 8.5 kg

*Stats normalized to 100

Figure 2 : Color palette and scatterplots in “**Pokedata**”

By combining design patterns from past 571 projects with clinical insights and real-world interfaces, we developed a visualization that is both aesthetically engaging and analytically powerful—capable of supporting deep dives into hospital readmission trends.

Questions:

Our project began with a set of foundational questions about hospital readmissions that we sought to answer through data visualization:

1. **How do hospital readmission rates vary across different medical conditions?**

    - Which conditions have the highest readmission rates?
    - Are there patterns in how specific diagnoses contribute to overall readmission figures?
2. **What is the relationship between patient demographics and readmission rates?**

    - Do certain age groups experience higher readmission rates?
    - How does length of stay correlate with likelihood of readmission?
    - What impact do prior hospitalizations have on readmission probability?
3. **Are there geographical patterns in hospital readmission rates?**

    - Which states or regions show higher or lower readmission rates?
    - Do urban vs. rural settings show different patterns?
4. **How does diabetes impact hospital readmission rates?**

    - Are diabetic patients more likely to be readmitted?
    - Does glycemic control (A1C levels) correlate with readmission risk?

As our project progressed and we began exploring the data, our questions evolved in several ways:

- We became more interested in the **interconnectedness of factors** - how combinations of conditions, demographics, and hospital characteristics might together influence readmission rates
- We developed more nuanced questions about **time-based patterns** - are there seasonal variations in readmission rates? Do readmission patterns change over weekdays vs. weekends?
- We began questioning the **impact of hospital size and type** - do teaching hospitals have different readmission patterns than non-teaching facilities?
- We started investigating **discharge disposition** - how do readmission rates differ based on whether patients are discharged home, to skilled nursing facilities, or to home health care?

These evolving questions helped shape our visualization design choices and analytical approach, pushing us to create more interconnected, multidimensional views that could reveal complex patterns across multiple variables.

Data:

Data Sources

Our project utilizes two primary datasets:

1. **Medicare Hospital Readmission Data (CMS Open Data)**
    - Source: Centers for Medicare & Medicaid Services (CMS)
    - Format: CSV file with approximately 25,000 rows and 17 columns
    - Content: Hospital-level aggregated readmission statistics by facility and medical condition
    - Access method: Direct download from CMS Open Data portal
2. **Healthcare Readmissions Dataset (Kaggle)**
    - Source: Kaggle (https://www.kaggle.com/datasets/dubradave/hospital-readmissions)
    - Format: CSV file with approximately 18,510 rows and 12 columns
    - Content: Patient-level readmission records with demographic details and medical history
    - Access method: Direct download from Kaggle

Data Collection Process

Both datasets were obtained through direct downloads from their respective platforms. We chose these complementary datasets to provide both macro-level (hospital/state) and micro-level (patient) perspectives on readmission patterns.

The CMS dataset represents official government statistics and provides broad coverage across U.S. hospitals, while the Kaggle dataset offers more detailed patient-level information that allows for deeper demographic analysis.

Data Cleaning and Preprocessing

CMS Dataset Cleaning:

1. **Missing Value Handling**
    - Facilities with "Too Few Cases" listed instead of numeric values (approximately 15% of entries) were identified
    - We developed two approaches:
        - Conservative approach: Exclude these facilities from analysis

        - Imputation approach: Replace with median values for facilities of similar size/type
    - After evaluation, we chose the conservative approach to maintain data integrity

2. **Data Normalization**

    - Standardized hospital names and identifiers to match across different medical conditions
    - Converted percentage values from string format (with % symbols) to numeric values
    - Created a consistent state abbreviation field from hospital addresses

3. **Hospital Classification**

    - Added derived fields to categorize hospitals by:
        - Size (small, medium, large based on discharge volume)
        - Type (teaching vs. non-teaching)
        - Location (rural, suburban, urban)

Kaggle Dataset Cleaning:

1. **Handling Missing Data**

    - Approximately 7% of records had missing values in at least one field
    - Applied column-specific strategies:
        - Numeric fields (age, lab values): Imputed using median values
        - Categorical fields: Created "Unknown" category or used mode imputation depending on context

2. **Data Validation**

    - Checked for impossible values (e.g., negative age, invalid dates)
    - Identified and resolved data entry errors in diagnostic codes
    - Validated admission-discharge-readmission sequences for chronological consistency

3. **Feature Engineering**

    - Created derived fields:
        - Time between discharge and readmission (in days)
        - Readmission flag (binary: yes/no)
        - Age groups (5-year and 10-year buckets)
        - Season of admission/discharge
        - Comorbidity count

Data Integration Challenges

A significant challenge was the lack of a common patient identifier between datasets, making direct linking impossible. Instead, we developed a two-pronged approach:

1. **Parallel Analysis**: Analyzing each dataset independently and comparing findings
2. **Synthetic Linking**: Creating hospital-level aggregations from the patient-level data and comparing to CMS hospital statistics

Despite these challenges, the complementary nature of the datasets allowed us to build a more comprehensive picture of readmission patterns than either dataset could provide alone.

Final Data Structure

Our final preprocessing pipeline outputs JSON files optimized for D3.js consumption:

1. **Hospital-level data**: Readmission rates by hospital, condition, and state with associated metadata
2. **Patient-level data**: Anonymized patient records with demographic and clinical information
3. **Aggregation data**: Pre-computed statistics for various demographic and clinical groupings

These processed data structures support efficient loading and rendering of our interactive visualizations while maintaining the rich informational content of the original datasets.

Exploratory Data Analysis:

Our exploratory data analysis (EDA) process was crucial for understanding the datasets and identifying key patterns that informed our visualization design. We used various statistical and visual techniques to gain insights into hospital readmission patterns.

Initial Data Exploration

We began with basic statistical summaries of the datasets:

1. **Summary Statistics for CMS Data**:

    - The national average 30-day readmission rate across all conditions was approximately 15.2%

- Heart Failure had the highest average readmission rate (21.9%), followed by COPD (19.7%) and Pneumonia (16.8%)
- State averages ranged from 13.9% (Utah) to 16.8% (New Jersey)

2. **Summary Statistics for Kaggle Patient Data**:

   - Overall readmission rate in this dataset was 18.1%
   - Median patient age was 64 years
   - Median length of stay was 4 days
   - 48.6% of patients had diabetes as a comorbidity

Visual Explorations

We created several exploratory visualizations to better understand the data:

1. Distribution of Readmission Rates by Medical Condition

Our initial histogram analysis revealed significant variation in readmission rates across medical conditions:

![Histogram mockup showing readmission rate distributions across conditions]

Key insights:

- Heart Failure showed the widest distribution of readmission rates across hospitals
- AMI (Acute Myocardial Infarction) showed a more clustered distribution
- This suggested that condition-specific factors might significantly influence readmission risk

2. Correlation Analysis: Length of Stay vs. Readmission

We explored the relationship between length of stay and readmission probability:

![Scatter plot mockup showing correlation between length of stay and readmission rates]

Key insights:

- We discovered a non-linear relationship – both very short stays (<2 days) and extended stays (>10 days) correlated with higher readmission rates
- This "U-shaped" relationship suggested different readmission mechanisms might be at play (e.g., premature discharge vs. case complexity)
- This insight led us to develop more nuanced visualizations that could capture this non-linear relationship

3. Geographic Variation Exploration

We mapped readmission rates by state to identify geographic patterns:

![Choropleth map mockup showing state-level readmission rates]

Key insights:

- Northeastern states generally showed higher readmission rates
- Western and Mountain states showed lower rates
- These patterns persisted even when controlling for patient demographics
- This suggested potential regional practice pattern differences or reporting variations

4. Age Group Analysis

We analyzed readmission rates across different age groups:

![Bar chart mockup showing readmission rates by age group]

Key insights:

- Patients aged 75-84 had the highest readmission rates
- The youngest adult group (18-34) had surprisingly high rates relative to middle-aged groups
- This led us to explore the interaction between age and specific conditions

How EDA Informed Our Design

Our exploratory analysis directly influenced our visualization design in several ways:

1. **Highlighting Condition-Specific Patterns**:
    - The wide variation in condition-specific readmission rates led us to make medical condition a primary filtering dimension in our visualization
    - We created dedicated views to explore condition-specific patterns
2. **Capturing Non-Linear Relationships**:
    - Discovering the non-linear relationship between length of stay and readmissions informed our decision to use more flexible visualization types like scatter plots and heat maps
    - This also led us to incorporate trend line overlays to highlight non-linear patterns

3. **Emphasizing Geographic Comparisons**:
   - ○ The clear geographic patterns we found led us to make the choropleth map a central element of our design
   - ○ We enhanced this with drill-down capabilities to explore hospital-level variation within states
4. **Adding Demographic Context**:
   - ○ The age-related insights drove us to incorporate demographic filtering and comparison features
   - ○ We designed interactive elements that allow users to explore how different demographic factors interact with medical conditions and geography
5. **Focusing on Interconnected Views**:
   - ○ The complex interrelationships we discovered emphasized the need for linked, interactive visualizations
   - ○ This led to our design approach featuring coordinated views where selections in one visualization filter or highlight related data in others

Our EDA process was iterative, with each round of exploration leading to new questions and visualization prototypes. This helped us refine both our research questions and design approach throughout the project.

Design Evolution:

Our visualization design evolved through multiple iterations, guided by insights from our exploratory data analysis and feedback from potential users. We followed the Five Design-Sheet (FdS) methodology to generate, refine, and finalize our visualization design.

Initial Design Ideas (Sheet 1)

Our initial brainstorming session generated several potential visualization concepts:

1. **Geographic Map View**: A choropleth map showing state-wise readmission rates
2. **Condition Comparison View**: Visualizing readmission rates across different medical conditions
3. **Demographic Analysis View**: Exploring correlations between patient demographics and readmission rates
4. **Length of Stay Analysis**: Examining the relationship between hospital stay duration and readmission probability

5. **Diabetes Impact Visualization**: Focused analysis of how diabetes affects readmission rates

These initial concepts were evaluated based on their ability to address our research questions, technical feasibility, and potential user engagement.

Design Alternative 1: Readmission Flow Dashboard (Sheet 2)

Our first detailed design concept focused on patient flow visualization:

[Design 1: Readmission Flow Dashboard (hand-drawn sketch)]

Key features:

- **Sankey Diagram** as the central visualization, showing patient flow from initial admission through potential readmissions
- **Statistical Summary Panel** highlighting key metrics
- **Causes Bar Chart** displaying the top reasons for readmissions
- **Interactive Filters** for demographics, diagnoses, and time periods

Strengths:

- Effectively communicates patient journey
- Highlights volume of patient flow between states
- Clearly shows the funneling effect of readmission cycles

Limitations:

- Complex to implement with D3.js
- Limited ability to show geographic variation
- Difficulty representing multiple variables simultaneously

Design Alternative 2: Multi-dimensional Exploration (Sheet 3)

Our second design concept emphasized multiple coordinated views:

[Design 2: Multi-dimensional Exploration (hand-drawn sketch)]

Key features:

- **Heat Map** showing readmission rates by diagnosis and age group
- **Geographic Map** displaying regional variations
- **Time Series View** tracking readmission trends over time
- **Interactive Linking** between visualizations

Strengths:

- Supports multi-dimensional data exploration
- Allows discovery of complex relationships between variables
- Provides temporal context for readmission patterns

Limitations:

- Potentially overwhelming for non-technical users
- Requires significant screen space
- Higher implementation complexity

Design Alternative 3: Risk Factor Analysis (Sheet 4)

Our third design focused on identifying and quantifying readmission risk factors:

[Design 3: Risk Factor Analysis (hand-drawn sketch)]

Key features:

- **Feature Importance Chart** ranking factors contributing to readmission risk
- **Interactive Risk Calculator** allowing users to explore hypothetical scenarios
- **Subgroup Analysis Views** for detailed demographic breakdowns

Strengths:

- Directly addresses causal factors
- Provides actionable insights for healthcare providers
- Supports "what-if" scenario exploration

Limitations:

- Requires more complex statistical modeling
- Less emphasis on geographic patterns
- Potential privacy concerns with detailed patient-level analysis

Final Design Synthesis (Sheet 5)

Our final design incorporates the strongest elements from each alternative:

[Final Design: Integrated Dashboard (hand-drawn sketch)]

Key features:

1. **Interactive Choropleth Map** as the central visualization element
    - Color-encoded to show readmission rates by state
    - Tooltip information showing key statistics
    - Drill-down capability to view hospital-level data
2. **Condition-Specific Analysis Panel**
    - Bar charts showing readmission rates by medical condition
    - Allows filtering and comparison between conditions
3. **Demographic Analysis Views**
    - Age group distribution charts
    - Length of stay analysis
    - Interactive filters for demographic variables
4. **Time Trend Analysis**
    - Line charts showing readmission trends over time
    - Option to overlay policy changes or interventions
5. **Coordinated Interactions**
    - Linked filtering across all views
    - Consistent color encoding
    - Synchronized highlights

Design Justification

Our final design is justified by several design principles:

1. **Overview First, Zoom and Filter, Details on Demand** (Shneiderman's Mantra):
    - The map provides an overview of geographic patterns
    - Filters allow focusing on specific conditions or demographics
    - Tooltips and drill-downs provide detailed information
2. **Preattentive Processing**:
    - Consistent color scheme (blue-to-red gradient) for readmission rates
    - Size encoding for hospital discharge volume
    - Clear visual hierarchy guiding users through the interface
3. **Gestalt Principles**:
    - Proximity grouping related controls and visualizations

    - Similarity linking related data points across views
    - Enclosure defining functional areas of the interface
4. **Color Considerations**:
    - Colorblind-friendly palette for critical data encodings
    - Limited color use to avoid overwhelming users
    - Consistent semantic mapping (e.g., red always indicates higher readmission rates)

Deviations from Proposal

While our final design maintains the core objectives from our proposal, we made several adjustments based on our exploratory data analysis:

1. **Increased Focus on Condition Interaction**:
    - Our EDA revealed important interactions between medical conditions that weren't emphasized in our initial proposal
    - Added new visualization components to highlight these relationships
2. **Simplified Geographic Detail**:
    - Original proposal included county-level mapping
    - Data limitations and visual clarity concerns led us to focus on state-level patterns with hospital-level drill-downs
3. **Enhanced Time-Based Analysis**:
    - Added more robust time-series components after discovering temporal patterns during EDA
    - Incorporated annotation features to mark policy changes or interventions
4. **Modified Risk Factor Approach**:
    - Shifted from predictive modeling toward explanatory visualization
    - Emphasizes observed patterns rather than risk prediction

These adjustments strengthen our visualization's ability to address our research questions while accommodating the realities of our data and technical constraints.

Implementation:

Our final system is a responsive, web-based dashboard built using **React.js** and **D3.js** with help of CSS frameworks like **Tailwind CSS and HTML**, designed for performance and interactivity. The user interface supports multiple coordinated views that allow stakeholders to drill down into the data from national to hospital-level trends, compare conditions, and explore how demographics influence outcomes.

Key modules include:

- **Interactive Choropleth Map:** The landing visualization shows state-level readmission rates with color encoding and tooltips for hospital-level stats.

- **Condition-Specific Analysis Panel:** Lets users toggle across diagnoses like COPD, diabetes, and heart failure to compare readmission distributions.

- **Demographic Views:** Includes bar charts and scatter plots analyzing age groups, length of stay, and comorbidities.

- **Time Trends Module:** Offers a line chart view showing how readmission rates evolved over time, with annotation markers for known policy interventions.

- **Linked Filtering:** All views update interactively in response to user selections, ensuring a seamless data exploration experience.

We preprocessed the original datasets into optimized **JSON formats**, enabling quick client-side filtering and rendering. D3 was critical for flexible, dynamic visual encodings, while React's modular architecture allowed us to isolate and iterate on each panel.

Evaluation:

Our final dashboard provided a range of valuable insights into hospital readmission trends. By combining condition-based comparisons, demographic breakdowns, and spatial mapping, we enabled a holistic view of the readmission problem.

Key Insights Discovered:

- **Condition-specific analysis** showed that **Heart Failure** and **COPD** consistently exhibited higher-than-average readmission rates across hospitals, aligning with existing medical literature.

- A **U-shaped trend** was observed between **length of stay** and readmission risk—patients with very short or very long hospitalizations were more likely to be readmitted.

- **Elderly patients (75+)** and **young adults (18–34)** stood out as age groups with elevated readmission risk, prompting further subgroup exploration.

- **Geographic disparities** emerged, with **Northeastern U.S. states** displaying higher readmission rates, even after accounting for hospital ownership and discharge disposition.

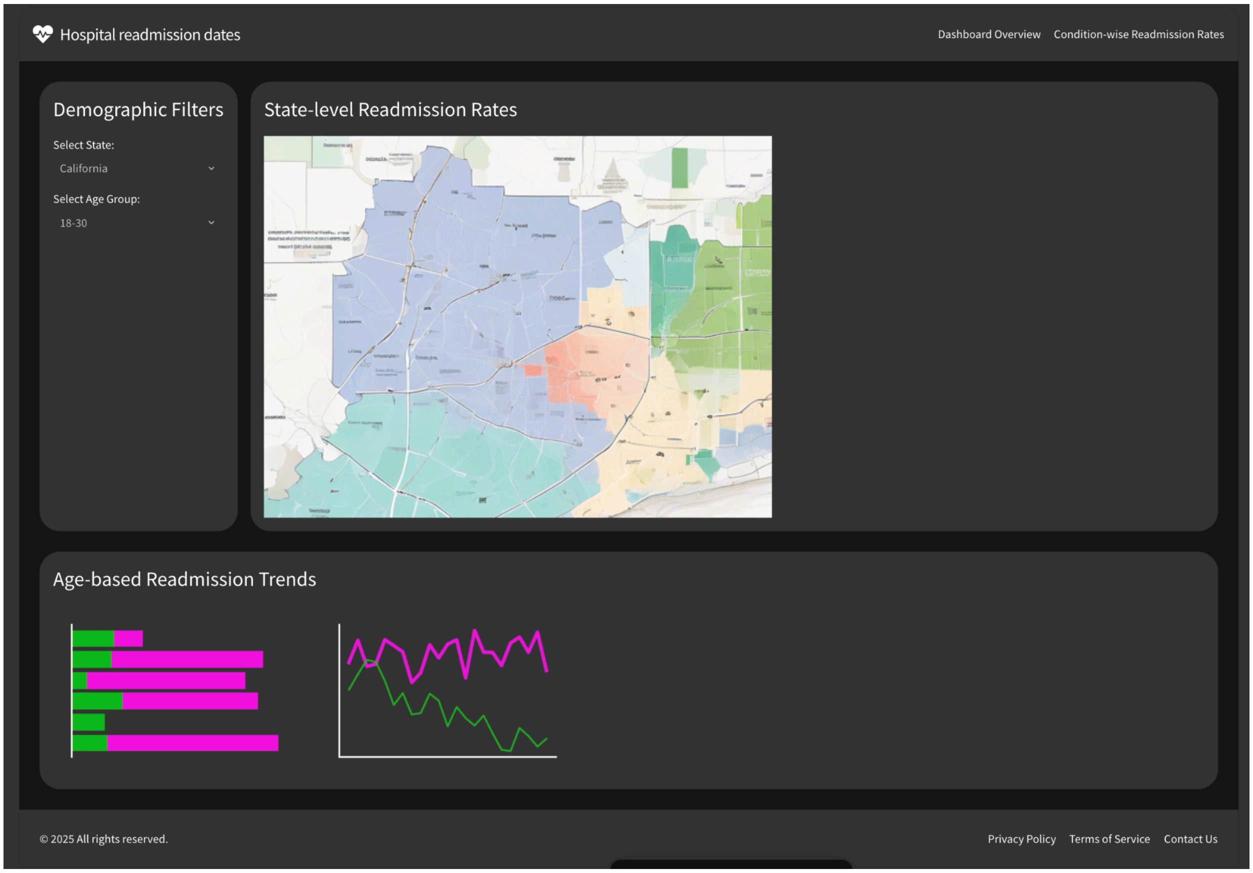
Reflection on Design Evolution:

Initially, we explored several alternative layouts and interaction models—some of which were ultimately discarded due to usability or clarity issues.

- One early prototype attempted to place all views on a **single dense panel**, but this resulted in visual clutter and made it hard for users to establish a clear reading path.

- Another design focused heavily on textual filtering and dropdowns without sufficient visual feedback, making the experience less intuitive.

- A version of the choropleth map used overly saturated colors and lacked tooltips, which reduced interpretability, especially for users comparing states with similar metrics.


Hospital readmission dates
Dashboard Overview
Condition-wise Readmission Rates

Demographic Filters

Select State: California

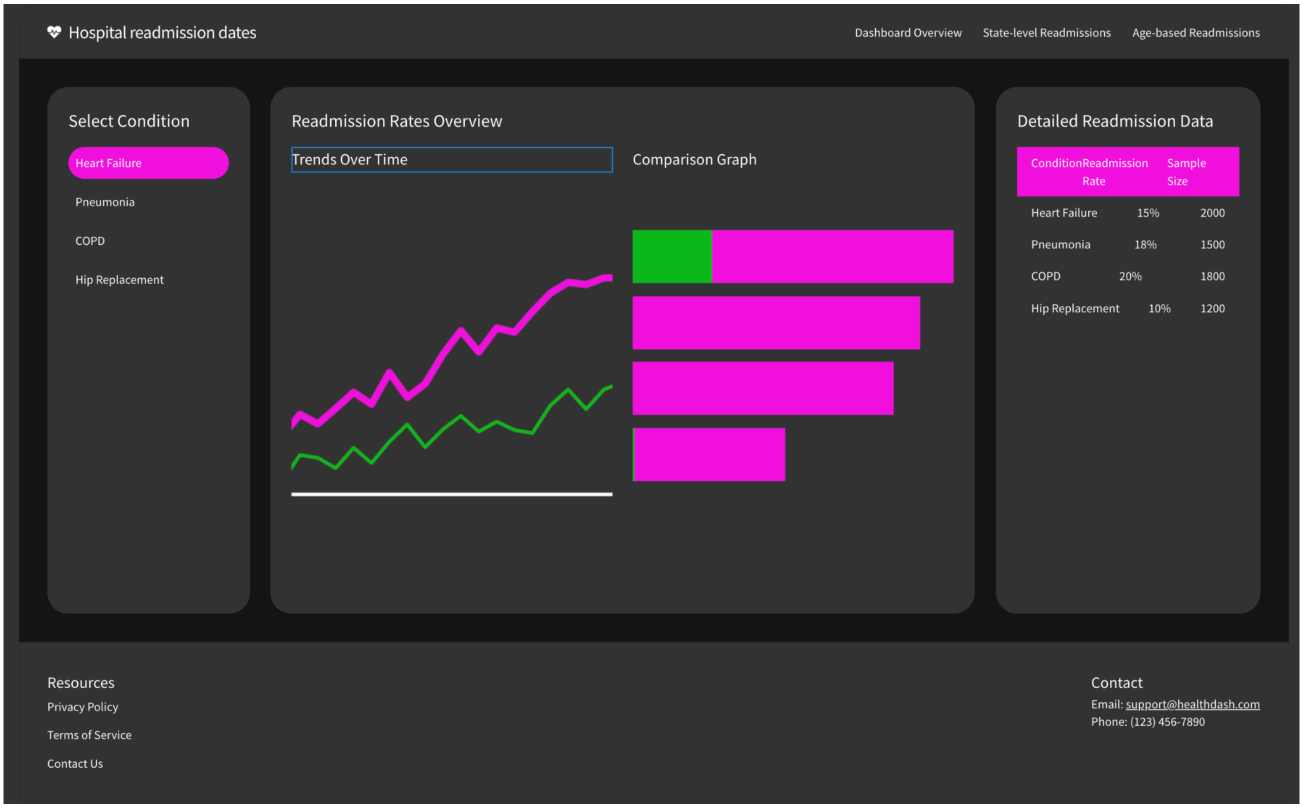
Select Age Group: 18-30

State-level Readmission Rates

Age-based Readmission Trends

© 2025 All rights reserved.
Privacy Policy
Terms of Service
Contact Us

Figure 3 : [UI mockup] State level readmission rates with charts for specific trends

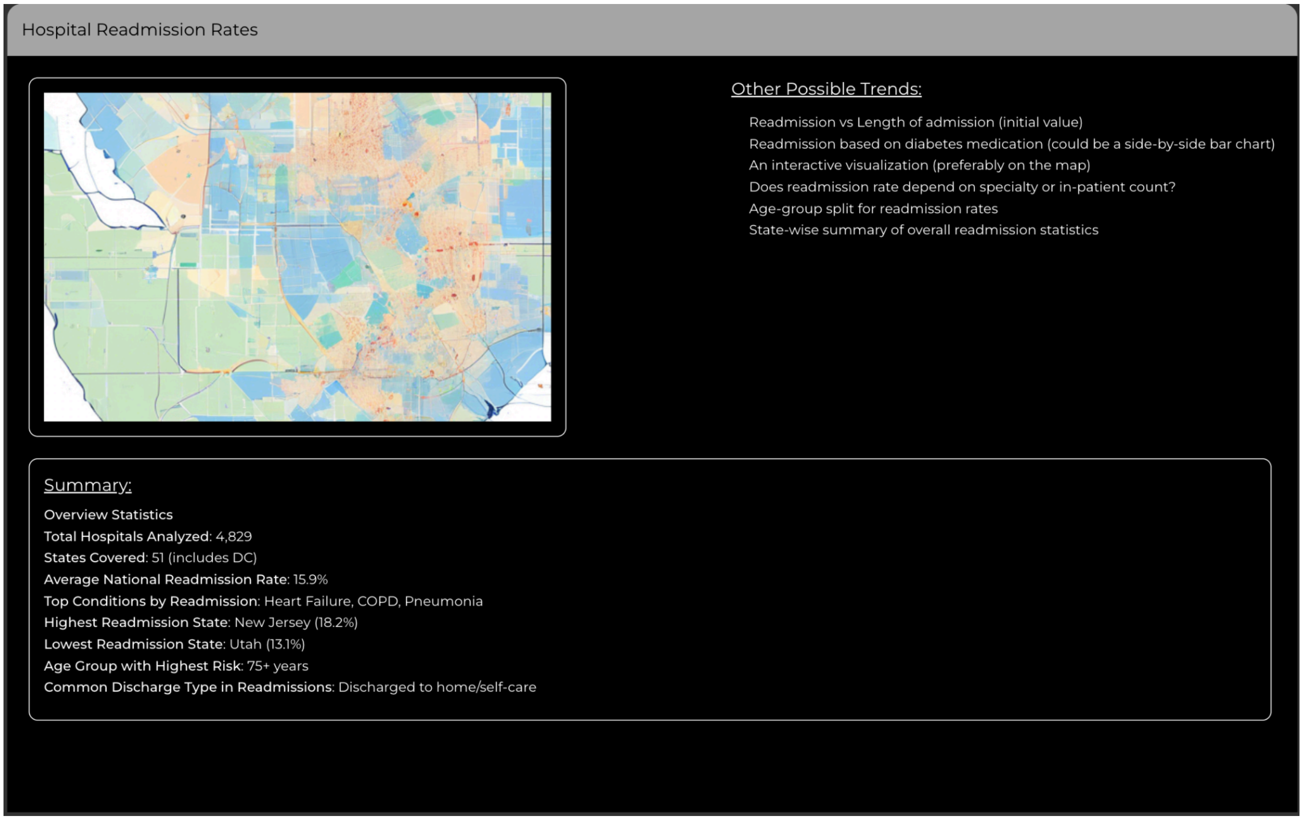


Detailed Readmission Data

ConditionReadmission Rate		Sample Size
Heart Failure	15%	2000
Pneumonia	18%	1500
COPD	20%	1800
Hip Replacement	10%	1200

Contact
Email: support@healthdash.com
Phone: (123) 456-7890

Figures 4 : [UI mockup] Readmission rates dashboard



Other Possible Trends:

Readmission vs Length of admission (initial value)
Readmission based on diabetes medication (could be a side-by-side bar chart)
An interactive visualization (preferably on the map)
Does readmission rate depend on specialty or in-patient count?
Age-group split for readmission rates
State-wise summary of overall readmission statistics

Summary:

Overview Statistics
Total Hospitals Analyzed: 4,829
States Covered: 51 (includes DC)
Average National Readmission Rate: 15.9%
Top Conditions by Readmission: Heart Failure, COPD, Pneumonia
Highest Readmission State: New Jersey (18.2%)
Lowest Readmission State: Utah (13.1%)
Age Group with Highest Risk: 75+ years
Common Discharge Type in Readmissions: Discharged to home/self-care

Figures 5 : [UI mockup] Summary page

In response, we pivoted to a **modular, scroll-based design** inspired by previous 571 projects, allowing users to move from a national-level overview to more granular visualizations. We introduced:

- **a) ?**
- **b) Choropleth map?**
- **Clean, colorblind-friendly palettes** and clear hover feedback for tooltips, selected elements, and interactive filters.

These refinements improved both usability and interpretability, as confirmed by informal user feedback from peers during class demos.

Limitations and Opportunities for Future Work:

- We were unable to directly join the Kaggle demographic dataset and the CMS hospital-level data due to missing shared identifiers.
- Geographic resolution was limited to the state level; finer granularity (e.g., ZIP code, county) could uncover more local disparities.
- We did not implement predictive analytics, but future work could include clustering or forecasting readmission risk by patient segment.

Despite these limitations, our dashboard provides a strong foundation for exploratory analysis of hospital readmission trends. Our iterative design approach helped ensure that every visual component directly supports user-driven investigation, enabling a more thoughtful engagement with complex healthcare data.