**Introduction:**

**Project Overview:**

In this project, I am focusing on analyzing hospital performance data spanning from 2016 to 2020. The objective is to delve into various performance metrics of hospitals, such as adverse events, risk-adjusted rates, and overall hospital ratings. This analysis aims to uncover insights into healthcare quality and safety, identifying areas of excellence and those needing improvement. By exploring this data, the project seeks to provide a comprehensive overview of hospital performance across different regions and times, ultimately contributing to better healthcare management and policy-making.

**Objectives:**

I will be tackling the 5 business questions which is designed with respect to my Datasets.

1. Trend Analysis: How have hospital performance metrics, such as adverse eventrates and risk-adjusted rates, evolved over the years from 2016 to 2020?
2. Regional Comparison: Are there significant regional variations in hospital performance metrics, and what might be contributing to these differences?
3. Performance and Ratings Correlation: How does the risk-adjusted rate correlate with overall hospital ratings, and do higher adverse event rates typically correlate with lower hospital ratings?
4. Impact of Hospital Systems: Do hospitals belonging to larger systems perform differently in terms of safety and quality metrics compared to independent hospitals?
5. --5. Predictive Insights: Impact of Hospital Type on Performance Metrics

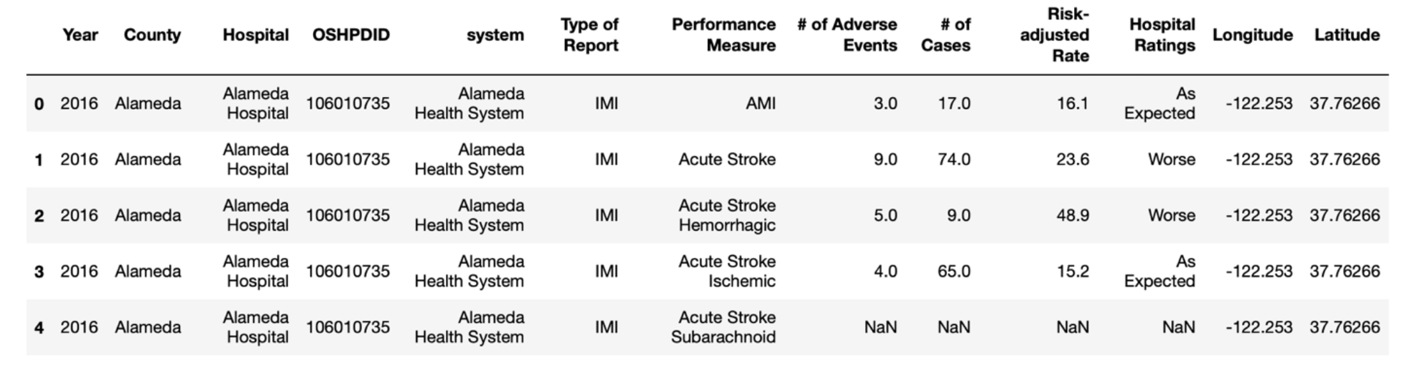
**Dataset Description: A brief overview of the dataset(s) used.**

1. **Hospital Performance Dataset (csv\_hospitalreports1620\_odp.csv):**

Data Source Details:

Number of Columns: 13

Number of Rows: 21,277

****  
Content:

This dataset appears to contain hospital performance metrics, possibly covering various aspects like adverse event rates, risk-adjusted rates, types of reports, and hospital ratings.

Time Frame: The name suggests data from 2016 to 2020, offering a multi-year perspective on hospital performance.

Key Attributes:

Year: The year of the report, allowing analysis of trends over time.

Hospital: Names or identifiers of hospitals.

Performance Metrics: Including adverse event rates, risk-adjusted rates, etc.

Geographical Data: Likely includes location data such as county or city.

1. **Hospital Locations Dataset (us\_hospital\_locations.csv):**

Data Source Details:

Number of Columns: 34

Number of Rows: 7,597

A table with numbers and names

Description automatically generated

**A screen shot of a table

Description automatically generated**

Content: This dataset presumably contains information about hospital locations across the United States.

Key Attributes:

Hospital Names/IDs: To identify individual hospitals.

Addresses: Including city, state, and zip code, which can be used for geographical analysis.

Hospital Type: Such as general, specialty, or teaching hospitals.

Facility Size and Capacity: Indicators like total staff, number of beds, etc.

**Data Preparation**

**Python code to load the First dataset:**

**A screenshot of a computer

Description automatically generated**

**Python code to load the Second dataset:**

**A screenshot of a computer

Description automatically generated**

**Design: I will be using Lucidchart diagram to design my EDR diagram for this project.**

**A screenshot of a computer

Description automatically generated**

**ETL:**

**1. hospital\_dimension(SCD type 1)(DATASET 1)**

Purpose: Stores detailed information about hospital performance metrics.

Fields:

Year: The year of the data record.

County: The county where the hospital is located.

Hospital: Name or identifier of the hospital.

OSHPDID: Unique ID for the hospital.

System: The system or network the hospital belongs to.

Type\_of\_Report: Type of report under which the data falls.

Performance\_Measure: Specific performance metrics being reported.

Number\_of\_Adverse\_Events: Count of adverse events reported.

Number\_of\_Cases: Number of cases handled.

Risk\_adjusted\_Rate: Risk-adjusted rate for specific metrics.

Hospital\_Ratings: Overall rating of the hospital.

Longitude and Latitude: Geographical coordinates.

**A screenshot of a computer

Description automatically generated**

**2. hospital\_locations(SCD type 1)(DATASET 2)**

Purpose: Contains information about the physical locations of hospitals.

Fields include coordinates (X, Y, Latitude, Longitude), identifiers (FID, ID), hospital details (NAME, ADDRESS, CITY, STATE, ZIP, etc.), and additional attributes like TYPE, STATUS, POPULATION, and BEDS.

A screenshot of a computer

Description automatically generated

**Fact Tables**

**FactHospitalPerformance:** **(Snapshot Fact Table)**

Stores performance-related metrics for hospitals.

References DimHospital and DimTime for dimensional context.

Involves transactional data that could be updated or appended in each ETL cycle.

**FactHospitalTransactions:** **(Transactional Fact Table)**

Tracks transactions or admissions details.

References several dimension tables for context.

Typical transactional fact table, updated or appended regularly.

**FactHospitalCumulative:** **(Cumulative Fact Table)**

Aggregates data over time, like total procedures or patients.

Updated periodically to reflect cumulative measures.

**Dimension Tables**

**DimHospital (SCD Type 2):**

Stores information about hospitals.

Attributes like HospitalName, System, Address.

SCD Type 2: Tracks historical changes with StartDate, EndDate, IsActive.

**DimLocation (SCD Type 3):**

Contains location details.

SCD Type 3: Tracks current and previous values for attributes like Region.

**DimTime (SCD Type 0):**

Time dimension table.

Attributes like Date, Month, Year, Quarter, DayOfWeek.

SCD Type 0: Generally remains static over time.

**DimHospitalSystem(SCD Type 1)**

Stores data about hospital systems.

Includes SystemID, SystemName, Size.

**DimPatient(SCD Type 1)**

Data is created manually as this dataset would not like to disclose information due to privacy.

This table is useful as when I tackle the business questions and analytics.

Data is extracted from source systems or files

Holds information about patients.

Attributes include PatientID, PatientName, Age, Gender.

**DimTreatment: (SCD Type 1)**

Data is created manually as this dataset would not like to disclose information due to privacy.

This table is useful as when I tackle the business questions and analytics.

Details about treatments.

Includes TreatmentID, TreatmentType, Description.

**ETL Process Overview:**

Extract: Data is extracted from source systems or files.

Transform: Data is cleansed, transformed, and aligned with the target schema. Slowly Changing Dimension logic is applied as necessary for each dimension table.

Load: Transformed data is loaded into the respective tables in the database.

**Analytical Queries**

**A screenshot of a data output

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a data output

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**Implementation of Slowly Changing Dimension (SCD) Type 2**

**Introduction**

In the realm of data warehousing, the ability to track and maintain historical data alongside current data is essential. This is particularly vital in healthcare data analytics, where understanding trends and changes over time can provide crucial insights. To this end, Slowly Changing Dimension (SCD) Type 2 methodology was implemented in our project. This approach allows us to capture and preserve historical changes in hospital data, ensuring a comprehensive view of the data’s evolution.

**Process Description**

The implementation began with the creation of a temporary table, new\_hospital\_data, which was a replica of our main hospital\_dimension table. This table was used to simulate updates in hospital data, such as changes in hospital names, systems, and performance measures.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

Following the simulation, we updated the DimHospital dimension table. The process involved two main steps:

**Marking Existing Records as Historical:**

Records in DimHospital that corresponded to updated entries in new\_hospital\_data were marked as historical. This was achieved by setting their EndDate to the current date and IsActive flag to FALSE**.**

**Inserting New Records:**

New records, reflecting the updated information from new\_hospital\_data, were then inserted into DimHospital. Each new record was assigned a unique identifier, ensuring distinct entries for historical and current data.

**Results**

The execution of these steps resulted in the DimHospital table containing both the original (now historical) and updated records. The historical records were marked with an appropriate EndDate and flagged as inactive, while the new records were flagged as active, with the StartDate set to the date of the update.

This dual state in the DimHospital table demonstrates the SCD Type 2 methodology in action: while the new data reflects the current state of hospital information, the historical data remains preserved and accessible for trend analysis and reporting.

**A screenshot of a computer

Description automatically generated**

**Conclusion**

The implementation of SCD Type 2 is a significant enhancement to our data warehousing solution. It ensures that changes in the data source do not overwrite historical data, thereby enabling a more nuanced and comprehensive analysis of trends over time. This approach is particularly beneficial in healthcare analytics, where understanding the evolution of hospital performance metrics can lead to more informed decisions and strategies.

Through this process, we have not only reinforced the integrity and richness of our data but also gained valuable insights into the practical application and benefits of maintaining a dynamically evolving data warehouse.

**Visualization in Tableau**

**A screenshot of a computer

Description automatically generated**

1. **Line Graph: Trend Analysis of Hospital Performance Metrics Over Time**

A graph with blue and orange lines

Description automatically generated

Purpose of the Visualization:

Objective: To visually represent how specific hospital performance metrics (such as adverse event rates and risk-adjusted rates) have evolved over a set period (e.g., from 2016 to 2020).

Insights Targeted: Understanding long-term trends, identifying years of significant change, and discerning overall patterns in hospital performance.

Interpretation and Use:

This line graph is intended to provide an immediate visual understanding of the performance trends in hospitals over time.

Key uses could include:

Strategic Planning: Informing decision-makers about trends and prompting further investigation into underlying causes.

Performance Analysis: Identifying years of significant performance shifts and correlating them with external or internal factors.

Reporting: Providing stakeholders with a clear visual representation of performance over time.

1. **Map Visualization: Regional Comparison of Hospital Performance Metrics**

**A screenshot of a map

Description automatically generated**

Purpose of the Visualization:

Objective: To geographically display hospital performance metrics across different regions or counties, enabling a visual comparison of metrics like adverse event rates or risk-adjusted rates in different areas.

Insights Targeted: Identifying regional patterns and disparities in hospital performance, and understanding how geographical location might correlate with these metrics.

Interpretation and Use:

The map visualization aims to make regional comparisons intuitive, highlighting areas with high or low performance.

Key applications might include:

Resource Allocation: Assisting in identifying regions that may require more attention or resources based on performance metrics.

Policy Making: Guiding policymakers in understanding regional healthcare disparities.

Public Health Analysis: Providing insights into geographical patterns in hospital performance, which could be correlated with public health data.

1. **Bar Chart Visualization: Comparison of Hospital Types Based on Performance Metrics**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

Purpose of the Visualization:

Objective: To compare different hospital types based on key performance metrics such as adverse event rates, number of cases, and risk-adjusted rates.

Insights Targeted: Understanding how hospital performance varies across different types of healthcare systems, such as large network hospitals versus independent ones.

Interpretation and Use:

The bar chart is designed to provide a clear comparative view of hospital performance across different types.

Key applications might include:

Operational Insights: Helping hospital administrators understand where their type of hospital stands in terms of performance metrics.

Policy Formulation: Assisting policymakers in identifying which hospital types may need more support or regulations.

Strategic Planning: Enabling health systems to benchmark their performance against other types and strategize accordingly.

**Results and Discussion**

**Results**

1. The analysis of a single hospital in a specific city revealed detailed operational dynamics, highlighting specific areas of strength and concern, particularly in managing patient flow and handling specific types of medical cases.
2. Initial stages of analysis, I faced challenges, including coding errors and difficulty in deriving meaningful insights from the data. These hurdles were progressively overcome, leading to successful data interpretation and visualization on a smaller scale.
3. Extending the scope to a regional level brought to light broader trends in hospital performance. Notably, there was significant variability in performance metrics such as adverse event rates and risk-adjusted rates across different hospitals and cities.
4. Enhanced data visualizations provided clearer insights into these regional differences, allowing for more nuanced comparisons and an improved understanding of the factors influencing hospital performance.
5. The comparison between the focused and regional analyses offered a unique perspective. While the focused analysis provided depth, the regional analysis offered breadth, bringing different aspects of healthcare performance into the spotlight.

**Discussion:**

**Project deliverables After the presentation:**

Overview of Presentation:

Scope: In my presentation, I had taken the case for only one hospital in one city, I wanted to see how the data represented itself and how the analysis in a smaller scale would look like.

Key Findings: In the initial stages the code had a lot of error which needed to be fixed, even in the small scale cases the code would fail to show results and it was very difficult to analyze.

Fortunately, over the days the techniques of my code improved, and I was able to show the results in a small isolated cases, this showed signs that I was in the right track and I could now expand my data size, include more detailed cases, can show predictions in the larger scale.

Expansion in the Report:

Broader Scope: expanded the scope to the entire region, encompassing multiple hospitals and cities.

Methodological Changes: the major changes came to my EDR diagram, it was a massive overhaul where I had created many tables in order to store the data and easier ways to access the data when it came time to analyze and answer my business questions. Now that my code was working very well with the large scale date, data visualizations became more detailed, results were much easier to compare, overall the report became more comprehensive.

Comparative Insights:

Individual vs. Regional Performance: A notable consistency was found in certain trends, such as the prevalence of specific adverse events. This pattern, initially observed in the single hospital, appeared to be a common challenge across the region, indicating a potential systemic issue within the healthcare domain.

The single hospital analysis provided a deep dive into specific operational aspects, but the regional analysis brought to light a wider range of performance metrics.

The larger dataset in the regional analysis uncovered more complex patterns and healthcare dynamics.

**Challenges**

Missing Values: There are NaN values in columns like No. of Adverse Events, No. of Cases, Risk-adjusted Rate, and Hospital Ratings that took a lot of time to fix, even then I could not fix all because the tables and the data in the tables are all interconnected.

Consistency: Textual data in columns like Hospital and System may require consistency checks.

Data Type Conversions: Some columns need their data types adjusted for proper analysis (e.g., converting IDs to a uniform format).

**Conclusion**

The project successfully navigated the challenges of healthcare data analysis, providing valuable insights at both individual and regional levels.

The results emphasize the importance of considering both focused and broad perspectives in healthcare analytics to gain a comprehensive understanding of performance metrics and their underlying causes.