Table of Contents

[1 Summary 3](#_Toc152497447)

[2 Introduction 3](#_Toc152497448)

[2.1 Background: 3](#_Toc152497449)

[2.2 Importance: 4](#_Toc152497450)

[2.3 Role in Healthcare Evolution: 4](#_Toc152497451)

[2.4 About CMS: 5](#_Toc152497452)

[2.4.1 Regulatory Oversight: 5](#_Toc152497453)

[2.4.2 Data Collection and Dissemination: 5](#_Toc152497454)

[2.4.3 Consumer Empowerment: 5](#_Toc152497455)

[2.4.4 Quality Improvement Initiatives: 5](#_Toc152497456)

[2.4.5 Key Partner in Healthcare Transformation: 6](#_Toc152497457)

[2.5 Involved Technologies: 6](#_Toc152497458)

[2.6 Objectives: 7](#_Toc152497459)

[3 Data Loading and Overview: 7](#_Toc152497460)

[3.1 Data Loading: 7](#_Toc152497461)

[3.2 Data Merging: 8](#_Toc152497462)

[3.3 First Look of Data: 8](#_Toc152497463)

[4 Summary and Understanding of Data 9](#_Toc152497464)

[4.1 Exploratory Data Analysis 9](#_Toc152497465)

[4.2 Data Types: 10](#_Toc152497466)

[4.3 Missing Values: 10](#_Toc152497467)

[4.4 Unique Identifiers: 11](#_Toc152497468)

[5 Data Preprocessing: 11](#_Toc152497469)

[5.1 Handling Missing Values 11](#_Toc152497470)

[5.1.1 County Name Imputation: 11](#_Toc152497471)

[5.1.2 Refinement of County Name Data: 11](#_Toc152497472)

[5.2 Analyzing the Categorical Columns for "Not Available" / "Not Applicable" values: 11](#_Toc152497473)

[5.3 Filtering Rows and Imputation 13](#_Toc152497474)

[6 Data Analysis 13](#_Toc152497475)

[6.1 Facility Information 13](#_Toc152497476)

[6.1.1 Question 1: What are the most common facility names, cities, or states 13](#_Toc152497477)

[6.1.2 Question 2: How many completed surveys are available for each facility? 15](#_Toc152497478)

[6.2 Patient Experience Ratings 16](#_Toc152497479)

[6.2.1 Question 1: What is the overall distribution of patient experience ratings? 16](#_Toc152497480)

[6.2.2 Question 2: How does the distribution of patient experience ratings vary across different hospital types? 16](#_Toc152497481)

[6.2.3 Question 3: What is the trend of patient experience ratings over the years? Are there any significant changes observed? 17](#_Toc152497482)

[6.2.4 Question 4: Is there a notable difference in patient experience ratings among hospitals with different ownership types? 18](#_Toc152497483)

[6.3 Geographic Distribution 18](#_Toc152497484)

[6.3.1 Question 1: What is the geographic distribution of facilities by states? 18](#_Toc152497485)

[6.3.2 Question 2: What is the geographic distribution of facilities by cities? 19](#_Toc152497486)

[6.3.3 Question 3: What is the geographic distribution of facilities by counties? 19](#_Toc152497487)

[6.4 Hospital Information 20](#_Toc152497488)

[6.4.1 Question 1: What is the distribution of Hospital overall ratings? 20](#_Toc152497489)

[6.4.2 Question 2: What is the distribution of hospital ownership types? 20](#_Toc152497490)

[6.5 Rating Comparisons 21](#_Toc152497491)

[6.5.1 Question 1: What are the different rating comparisons available for each category? 21](#_Toc152497492)

[6.6 Time Analysis 22](#_Toc152497493)

[6.6.1 Question 1: How many unique years are covered in the dataset and survey count in each year? 22](#_Toc152497494)

[7 Feature Engineering 23](#_Toc152497495)

[7.1 Selecting Columns 23](#_Toc152497496)

[7.1.1 Dropping Irrelevant Features: 23](#_Toc152497497)

[7.1.2 Handling Missing Values: 23](#_Toc152497498)

[7.2 Feature Selection: 23](#_Toc152497499)

[7.3 Encoding the Categorical Values 23](#_Toc152497500)

[7.3.1 Separating Features X and Target Variable Y 23](#_Toc152497501)

[7.3.2 Categorizing Columns Based on Data 23](#_Toc152497502)

[7.4 Final Dataset Overview 24](#_Toc152497503)

[8 Machine Learning Classifier Implementation 25](#_Toc152497504)

[8.1 Performance Metrics: 25](#_Toc152497505)

[8.2 Summary of the Results 26](#_Toc152497506)

[9 Recommendations 27](#_Toc152497507)

[10 Limitations and Future Directions 28](#_Toc152497508)

[10.1 Limitations: 28](#_Toc152497509)

[10.2 Future Directions: 28](#_Toc152497510)

[11 Conclusion 28](#_Toc152497511)

[12 Acknowledgement 29](#_Toc152497512)

# Summary

The field of healthcare is changing, with a focus on using data to take wellinformed decisions which will improve patient outcomes and quality of care. The project delves into U.S. hospital customer satisfaction data from 2016 to 2020, aiming to empower healthcare decisionmakers with actionable insights. By analyzing diverse survey data, it uncovers trends in patient satisfaction, facility performance, and geographical distributions. Utilizing advanced statistical techniques and machine learning models like Decision Tree and Logistic Regression, the project predicts patient experience national comparison. The goal is to aid healthcare stakeholders in making informed decisions, improving resource allocation, and ultimately enhancing the quality of care provided by hospitals across the nation.

# Introduction

In the dynamic landscape of healthcare, the pursuit of quality and patientcentric care stands as a paramount objective for healthcare organizations. The convergence of technological advancements, burgeoning data availability, and the imperatives of evidencebased decisionmaking has ushered in an era where data analytics plays a pivotal role in revolutionizing healthcare management. This project embarks on a journey through the realms of U.S. hospital customer satisfaction data, spanning the years 2016 to 2020, with an overarching aim to revolutionize decisionmaking paradigms in the healthcare domain. The Centers for Medicare & Medicaid Services (CMS) offers extensive datasets, providing insights into hospitals beyond clinical data, allowing a profound understanding of healthcare quality components.

## Background:

The healthcare sector operates within a complex ecosystem where patient satisfaction, quality of care, and operational efficiency stand as cardinal pillars defining success. Over the years, healthcare providers have embraced a datacentric approach to gauge performance, with patient satisfaction emerging as a pivotal metric reflecting the efficacy of healthcare delivery.

The project draws upon extensive datasets procured from The Centers for Medicare & Medicaid Services (CMS), encompassing over four thousand Medicarecertified hospitals across the United States. These datasets, collected through CMS' Hospital Compare program, serve as a treasure trove of information, aggregating patient survey data, hospital ratings, and comparative analytics.

## Importance:

In an era characterized by the convergence of technology and healthcare, the significance of leveraging data analytics cannot be overstated. This project represents a paradigm shift in healthcare management, unlocking the latent potential within voluminous survey data to derive actionable insights. The endeavor to develop a predictive model rooted in analytics is not merely an academic pursuit but an imperative in steering healthcare organizations towards enhanced performance and patientcentric care.

At its core, this project stands as a testament to the transformative power of analytics in healthcare decisionmaking. By mining intricate datasets and employing advanced statistical models, the project seeks to unravel patterns, trends, and correlations inherent within patient experience national comparisons. Such insights serve as compass points for healthcare administrators, guiding them in optimizing resources, improving care delivery, and fostering a culture of continuous enhancement.

## Role in Healthcare Evolution:

The project's role transcends the realms of data analysis; it becomes a catalyst for change in the healthcare landscape. Its outcomes have a profound impact on strategic planning, resource allocation, and policy formulation within healthcare organizations. By amalgamating healthcare expertise with cuttingedge analytics, this project propels the evolution of decisionmaking paradigms, elevating patient satisfaction and quality of care as central tenets of healthcare delivery.

The burgeoning relevance of datadriven decisionmaking in healthcare underscores the timeliness and importance of this project. Its outcomes pave the way for informed strategies, efficient resource allocation, and a holistic approach towards fostering patientcentric care.

## About CMS:

The Centers for Medicare & Medicaid Services (CMS) plays a pivotal role in shaping the healthcare landscape in the United States. Its significance within this project and the broader healthcare domain is profound:

### Regulatory Oversight:

CMS serves as the governing body overseeing Medicare and Medicaid programs, wielding considerable influence in setting standards and regulations for healthcare providers. Through mandates and guidelines, CMS ensures adherence to quality standards, transparency in healthcare delivery, and data reporting requirements.

### Data Collection and Dissemination:

A cornerstone of CMS' function lies in collecting, curating, and disseminating extensive healthcare data. The Hospital Compare program, managed by CMS, collates comprehensive datasets encompassing patient survey ratings, hospital performance metrics, and comparative analytics. This vast repository empowers stakeholders with valuable insights into healthcare provider performance and patient experiences.

### Consumer Empowerment:

CMS, via initiatives like Hospital Compare, endeavors to empower healthcare consumers by providing access to transparent and comprehensible healthcare quality data. By disseminating this information through userfriendly platforms, CMS enables patients to make informed decisions about their healthcare choices, fostering a culture of accountability and consumer empowerment.

### Quality Improvement Initiatives:

The datadriven approach advocated by CMS catalyzes quality improvement initiatives within healthcare organizations. By leveraging benchmarking and comparative analytics, CMS encourages hospitals to strive for continual enhancement in patient care, operational efficiency, and overall quality performance.

### Key Partner in Healthcare Transformation:

In the context of this project, CMS stands as a crucial partner, providing the extensive datasets that underpin the analysis. The datasets sourced from CMS' Hospital Compare program serve as the backbone for predictive modeling, exploratory analysis, and deriving actionable insights. CMS' commitment to data transparency and quality aligns seamlessly with the project's objective of harnessing data analytics for informed decisionmaking and improved patient satisfaction.

CMS assumes a multifaceted role, serving as a regulatory authority, a custodian of extensive healthcare data, an advocate for consumer empowerment, and a catalyst for quality improvement initiatives. Its contributions are instrumental in shaping the trajectory of healthcare, underlining the pivotal role it plays in this project's endeavors to enhance patient satisfaction and elevate the quality of care.

## Involved Technologies:

The project extensively utilized Python 3 as the primary programming language within a Jupyter Notebook environment. Leveraging the power of Python, the project employed several key modules such as pandas for data manipulation, numpy for numerical computations, and matplotlib, seaborn, and plotly.express for data visualization. These modules facilitated comprehensive data analysis, visualization of trends, patterns, and relationships within the dataset, and the creation of informative visual representations for stakeholders.

The project heavily relied on scikitlearn, a prominent machine learning library in Python, encompassing a range of algorithms and tools for predictive modeling. It employed various classification algorithms such as Logistic Regression, Decision Trees, and Random Forests for the creation and evaluation of predictive models.

Furthermore, the project utilized preprocessing techniques such as LabelEncoder, OneHotEncoder, StandardScaler, and ColumnTransformer to preprocess and transform the dataset, ensuring compatibility with machine learning models. To evaluate the performance of these models, the project employed metrics including classification reports, confusion matrices, mean squared error, and r2 score to assess the models' accuracy and predictive capabilities.The environment settings were optimized for enhanced data display, and future warnings were muted to streamline the analysis process.

In summary, the project capitalized on Python's versatility, leveraging a suite of libraries and machine learning tools to conduct robust data preprocessing, exploratory analysis, predictive modeling, and performance evaluation. This tech stack provided the foundation for comprehensive datadriven decisionmaking in the healthcare domain, empowering stakeholders with actionable insights derived from extensive survey data.

## Objectives:

In light of the voluminous CMS dataset, our project was dedicated to comprehensively dissecting hospital performance dynamics. Our primary objectives encompassed two critical facets: firstly, to conduct an extensive evaluation of the evolution of hospital performance metrics over time, discerning patterns, bottlenecks, and areas ripe for enhancement. Secondly, we aimed to untangle the intricate matrix of variables impacting overall hospital ratings. Our ambition was to construct a predictive model capable of foreseeing healthcare quality, allowing proactive management strategies. The paramount significance of these objectives lies in their potential to revolutionize the healthcare domain. By deciphering the elements underpinning positive patient experiences, robust safety standards, and overall exceptional care, we aimed to catalyze informed decisionmaking, drive improved healthcare outcomes, and foster heightened public confidence in the healthcare infrastructure.

# Data Loading and Overview:

## Data Loading:

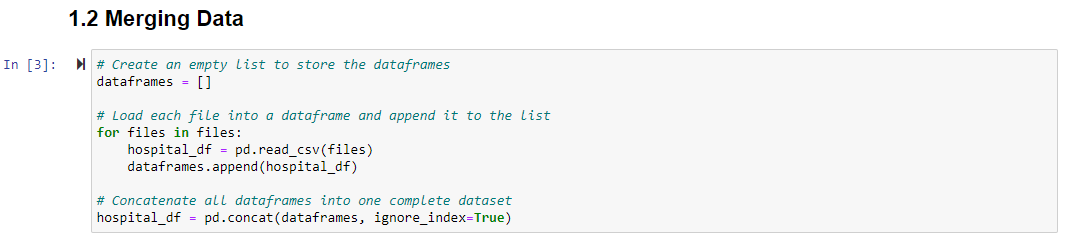
The initial phase of our project involved a meticulous process of data acquisition and loading. Leveraging Python's Pandas library, we imported the U.S. Hospital Customer Satisfaction dataset, encompassing a vast array of hospitalrelated attributes. The dataset, spanning from 2016 to 2020, contained a rich compilation of information from CMS, detailing various facets of hospital performance. Utilizing Pandas' efficient functionalities, we loaded this extensive dataset into a structured DataFrame format, enabling seamless manipulation and analysis. This meticulous data loading process laid the groundwork for our subsequent exploratory analysis and predictive modeling tasks.



**Figure 1**

## Data Merging:

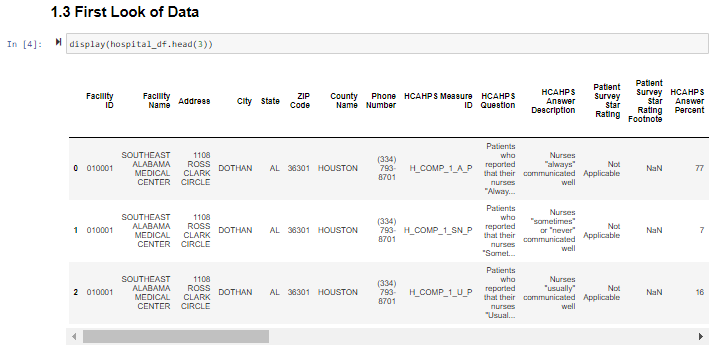
The data merging stage was pivotal in consolidating disparate datasets into a unified, comprehensive structure. Leveraging unique identifiers like Facility ID and Date, we merged multiple datasets seamlessly, ensuring data integrity and consistency throughout the process. This amalgamation allowed us to synthesize diverse sources of information, enriching our analytical capabilities. By merging survey data, hospital information, and performance metrics, we created a cohesive dataset that facilitated holistic analyses, empowering us to derive nuanced insights into hospital performance and patient experiences across various dimensions.



**Figure 2**

## First Look of Data:

The initial exploration of the dataset revealed a rich and expansive landscape encapsulating diverse facets of hospital performance and patient experiences. The comprehensive dataset encompassed various attributes such as facility information, survey metrics, geographical distributions, and hospital ratings across different categories. This multifaceted compilation provided a robust foundation for indepth analyses and predictive modeling.



**Figure 3**

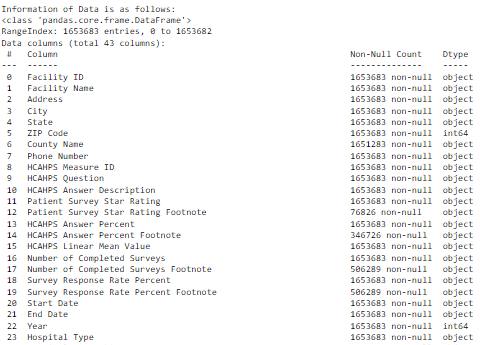
# Summary and Understanding of Data

## Exploratory Data Analysis

The EDA of the U.S. Hospital Customer Satisfaction dataset revealed a comprehensive view of the information it encapsulated. The dataset contained 43 columns and over 71 million entries, with 1.65 million records. It provided a first glimpse into various aspects, such as facility information, patient experience ratings, geographic distributions, hospital types, and comparisons across different quality metrics. The dataset comprised a mix of numerical and categorical columns. Notably, it contained numerical entries such as ZIP codes and years, while categorical data included facility names, addresses, cities, states, and various ratings and comparisons related to patient experiences and hospital quality.

The dataset covered the period from 2016 to 2020 and included a wide range of ratings and comparisons, denoting diverse hospital types, ownership structures, and emergency services. The exploratory analysis laid the foundation for a deeper dive into the data, shedding light on the intricacies of hospital performance, patient satisfaction, and various quality comparisons across healthcare providers.

EDA of Customer Satisfaction of US Hospital Dataset is........

Size: 71108369 | Columns: 43 | Records: 1653683

**Figure 4**

## Data Types:

The dataset comprises a mix of data types, encompassing both numerical and categorical entries. Among the numerical data are columns like ZIP codes and years, while categorical information includes facility names, addresses, cities, states, and various ratings and comparisons related to patient experiences and hospital quality. These diverse data types enable a comprehensive analysis, providing insights into both quantitative and qualitative aspects of hospital performance.

## Missing Values:

Throughout the dataset, several columns exhibit missing values across various categories. Entries such as 'Patient Survey Star Rating Footnote,' 'HCAHPS Answer Percent Footnote,' and 'Hospital overall rating footnote' contain significant instances of missing data. Similarly, other columns related to national comparisons and specific survey footnotes also display considerable proportions of missing information. Addressing these missing values is crucial for accurate analysis and modeling.

## Unique Identifiers:

The dataset contains unique identifiers essential for distinguishing and categorizing different elements. The 'Facility ID' serves as a primary identifier for individual facilities, aiding in tracking and analyzing performance trends across specific hospitals. Additionally, other categorical columns, like 'HCAHPS Measure ID' and 'HCAHPS Question,' act as distinct identifiers for different survey measures and questions, facilitating a granular examination of patient experiences and care quality. These unique identifiers enable a more detailed and specific analysis of the dataset.

# Data Preprocessing:

In the data preprocessing phase, several steps were taken to ensure the data's quality and integrity.

## Handling Missing Values

To address missing values, a systematic approach was adopted:

Dropping Columns: Columns with a significant percentage (80%) of missing values were removed. This strategic elimination helps maintain data quality without compromising the analysis.

### County Name Imputation:

As the 'County Name' column contained missing data, an imputation method was employed. Missing values in this column were filled based on corresponding 'City' and 'State' values, ensuring accuracy and relevance in the geographical representation.

### Refinement of County Name Data:

Following the imputation process, any remaining missing values in the 'County Name' column were handled using appropriate techniques, ensuring the completeness of this critical geographical identifier.

## Analyzing the Categorical Columns for "Not Available" / "Not Applicable" values:

The second phase of data preprocessing involved an indepth analysis of categorical columns containing "Not Available" or "Not Applicable" values. This process was crucial, given the prevalence of these values across multiple columns: Not Available / Not Applicable Values Analysis

The following columns had "Not Available" or "Not Applicable" values, along with their respective percentages:

'Patient Survey Star Rating': 88.15%

'HCAHPS Answer Percent': 42.77%

'HCAHPS Linear Mean Value': 89.19%

'Number of Completed Surveys': 14.17%

'Survey Response Rate Percent': 14.17%

'Hospital overall rating': 24.22%

'Mortality national comparison': 28.74%

'Safety of care national comparison': 44.97%

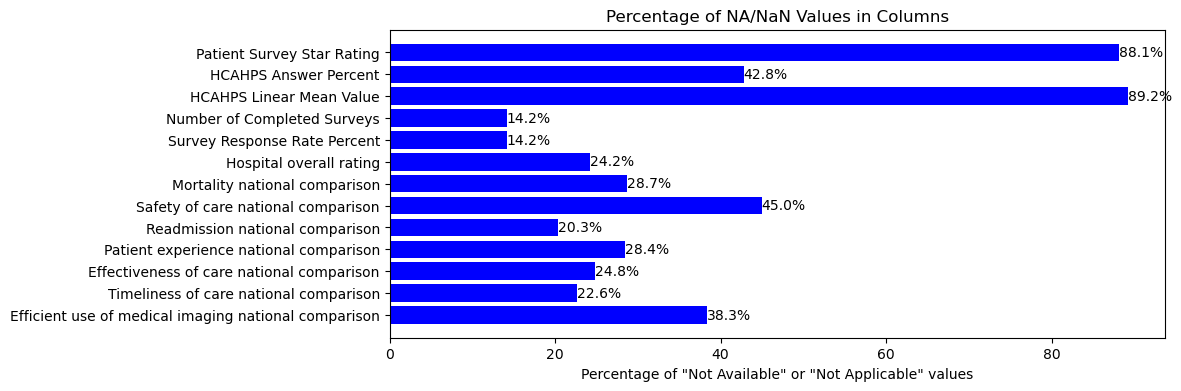
'Readmission national comparison': 20.32%

'Patient experience national comparison': 28.43%

'Effectiveness of care national comparison': 24.80%

'Timeliness of care national comparison': 22.59%

'Efficient use of medical imaging national comparison': 38.32%

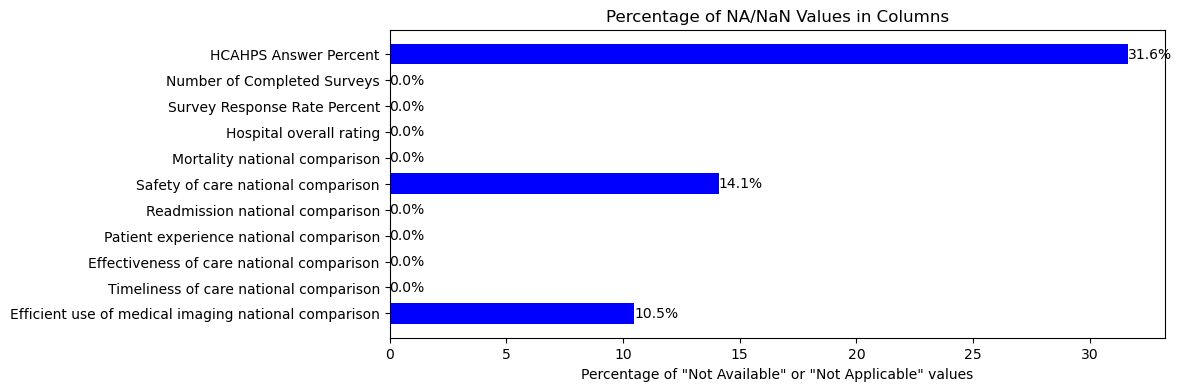


**Figure 5**

Handling "Not Available" Values

Given the significance of 'Patient Experience National Comparison' as the target variable, rows with "Not Available" values were dropped from this specific column.

Furthermore, columns such as 'Patient Survey Star Rating' and 'HCAHPS Linear Mean Value', which contained over 60% "Not Available" values, were deemed less informative and were subsequently dropped from the dataset.



**Figure 6**

## Filtering Rows and Imputation

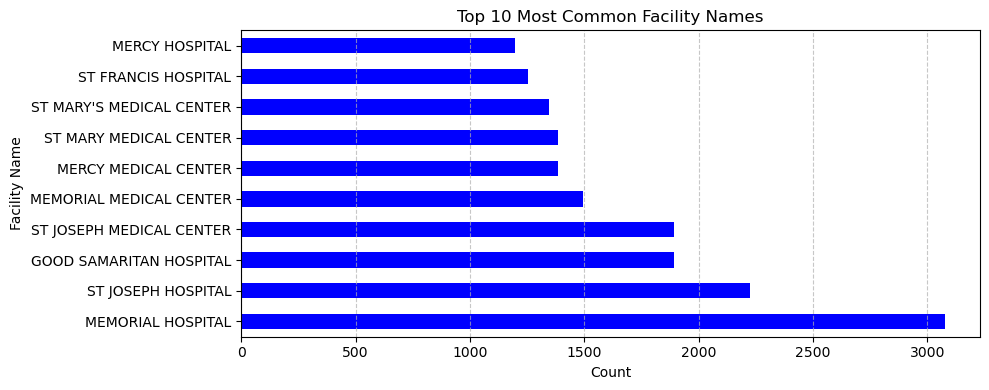
Rows containing less than 10% "Not Available" values were retained, ensuring data integrity. Additionally, missing values in columns such as 'HCAHPS Answer Percent', 'Safety of care national comparison', and 'Efficient use of medical imaging national comparison' were imputed with mode values to maintain data consistency and completeness.

These meticulous data preprocessing steps were pivotal in preparing a refined dataset, ensuring the quality and relevance of information used for subsequent analyses and model development.

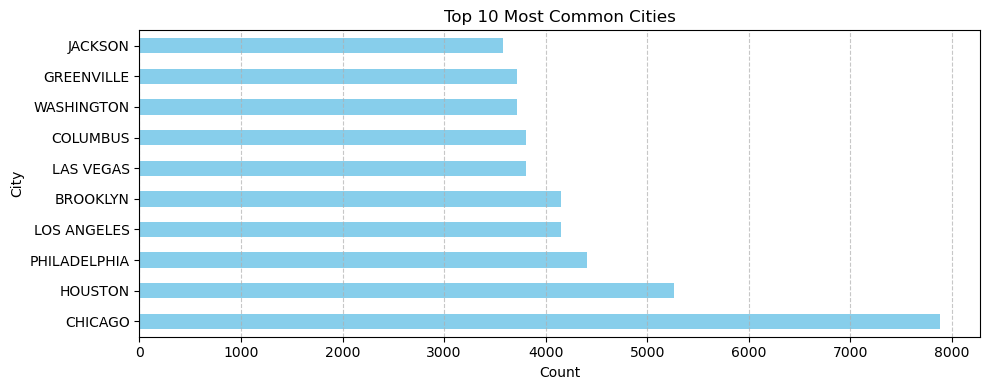
# Data Analysis

## Facility Information

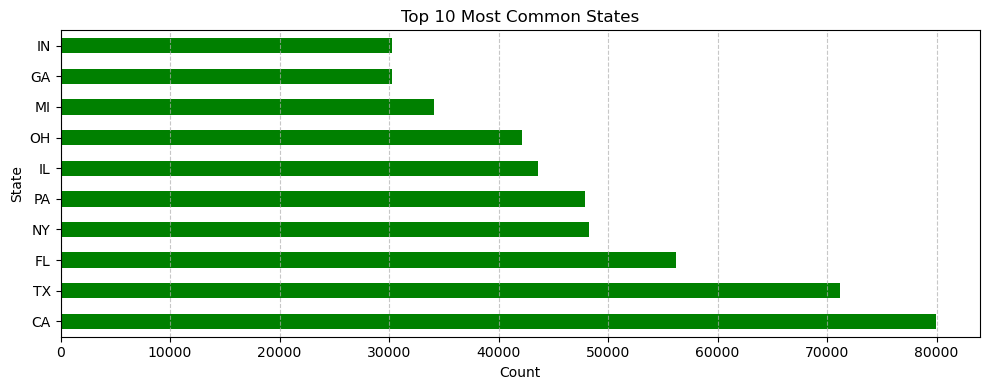
### Question 1: What are the most common facility names, cities, or states



**Figure 7**



**Figure 8**



**Figure 9**

Facility Names:

* Memorial Hospital appears as the most frequent facility name with 3076 occurrences. St Joseph Hospital and Good Samaritan Hospital follow with 2224 and 1890 occurrences, respectively. Several hospitals with similar names like St Joseph Medical Center, Memorial Medical Center, Mercy Medical Center, and St Mary Medical Center exhibit notable occurrences ranging from 1384 to 1890. Other hospitals such as St Mary's Medical Center, St Francis Hospital, and Mercy Hospital complete the top 10 list with occurrences ranging between 1198 to 1346.

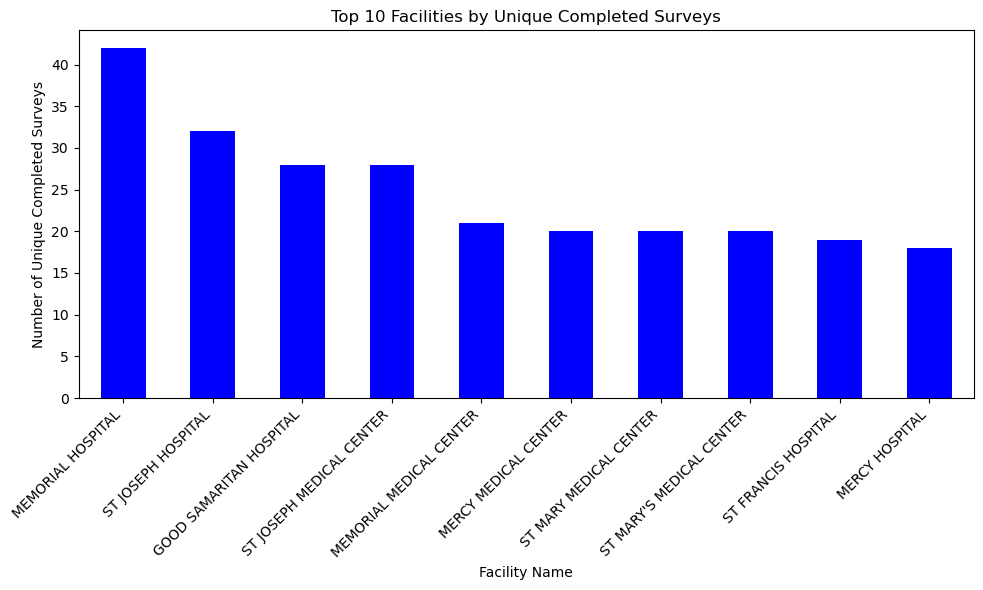
Cities:

* Chicago stands out significantly with 7882 occurrences, leading the list. Houston, Philadelphia, Los Angeles, and Brooklyn display a similar number of occurrences around 4152 to 5262. Las Vegas, Columbus, Washington, Greenville, and Jackson also appear in the top 10 list, each with 3577 to 3806 occurrences.

States:

* California (CA) tops the list with a substantial count of 79970. Texas (TX) follows closely with 71168 occurrences. Florida (FL), New York (NY), and Pennsylvania (PA) exhibit significant numbers, ranging from 47923 to 56196. Illinois (IL), Ohio (OH), Michigan (MI), Georgia (GA), and Indiana (IN) complete the top 10 list with occurrences between 30243 to 43571.

### Question 2: How many completed surveys are available for each facility?

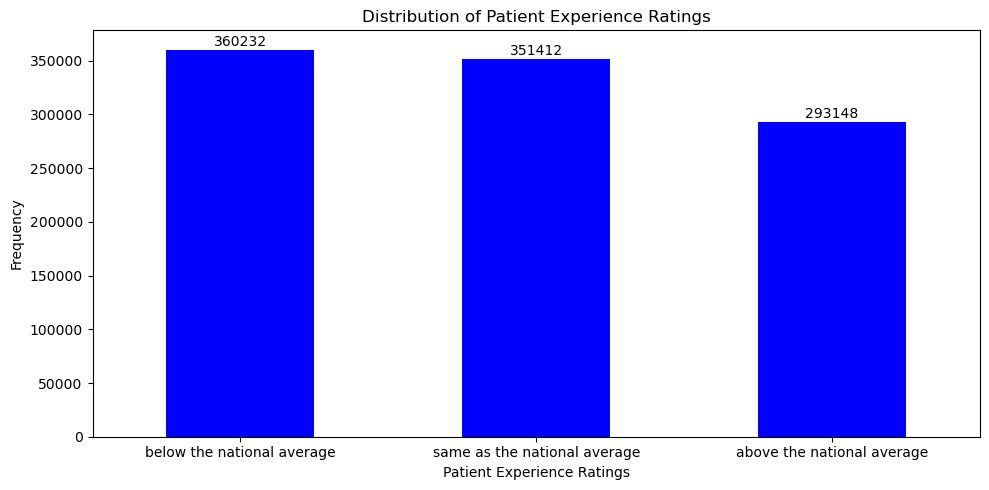


**Figure 10**

The top 10 facilities with the highest number of completed surveys are listed, revealing that Memorial Hospital holds the highest count with 42 completed surveys, followed closely by St Joseph Hospital with 32. Good Samaritan Hospital and St Joseph Medical Center share 28 completed surveys each, while several other facilities, including Memorial Medical Center, Mercy Medical Center, St Mary Medical Center, and St Mary's Medical Center, range between 20 to 21 completed surveys.

## Patient Experience Ratings

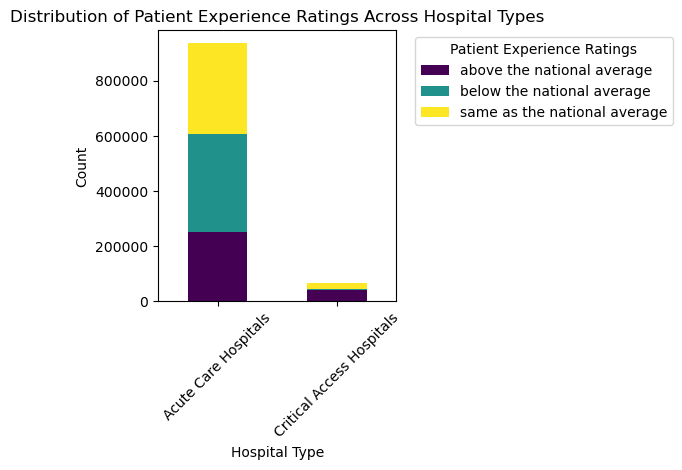
### Question 1: What is the overall distribution of patient experience ratings?



**Figure 11**

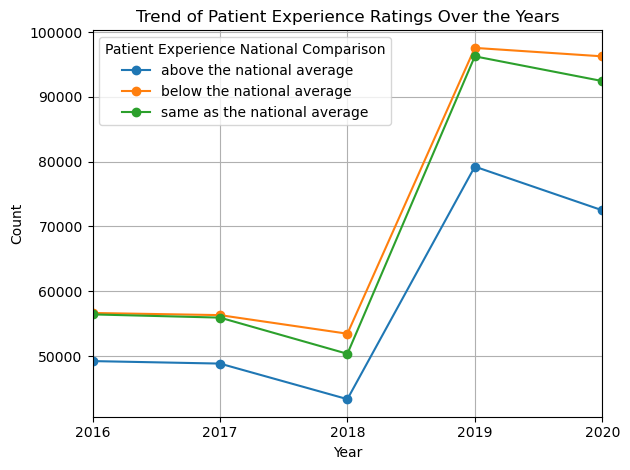
The distribution of patient experience ratings across different hospitals reveals an intriguing pattern. The majority of hospitals fall within the "Below the National Average," "Same as the National Average," and "Above the National Average" categories, with significant counts in each classification.

### Question 2: How does the distribution of patient experience ratings vary across different hospital types?

The analysis reveals distinct variations in patient experience ratings across different hospital types. Among Acute Care Hospitals, the count for ratings above the national average stands at 250,810, surpassing those below the national average, which amount to 357,512. Conversely, Critical Access Hospitals show a considerably lower count of ratings above the national average, totaling 42,338, with a minor count of 2,720 below the national average. Moreover, Acute Care Hospitals display a considerable number of ratings aligning with the national average, reaching 328,224, whereas Critical Access Hospitals demonstrate a notably smaller count of ratings at 23,188 that meet the national average criteria. These findings underscore notable disparities in patient experience ratings between Acute Care Hospitals and Critical Access Hospitals.

**Figure 12**

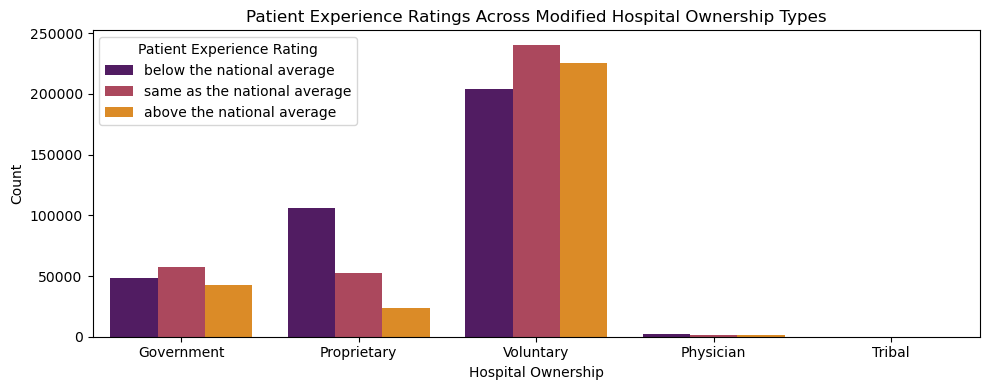
### Question 3: What is the trend of patient experience ratings over the years? Are there any significant changes observed?



**Figure 13**

The results depict a comprehensive overview of patient experience ratings categorized as above, below, and the same as the national average across different years. The data highlights fluctuations and patterns in patient experiences from 2016 to 2020. Across these years, there are noticeable variations in the counts of ratings, indicating potential shifts in patient perceptions. Specifically, while there are fluctuations in the counts of ratings above and below the national average, the counts of ratings similar to the national average showcase a decreasing trend in recent years. This insight suggests a potential change in patient experiences, warranting a closer examination of factors influencing these variations across different years.

### Question 4: Is there a notable difference in patient experience ratings among hospitals with different ownership types?

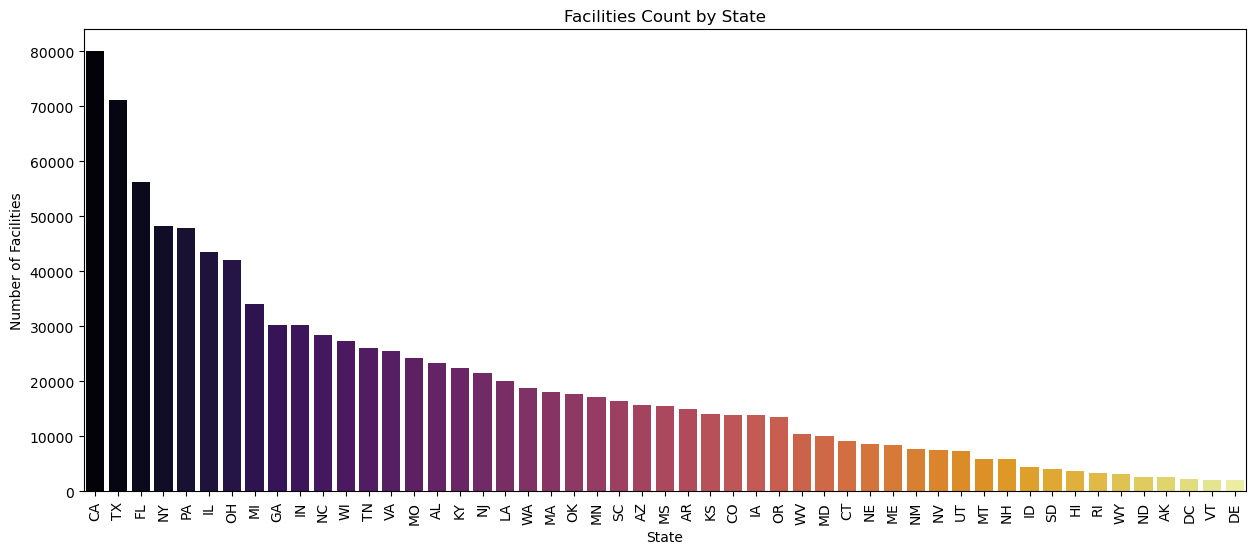


**Figure 14**

The distribution of patient experience ratings across the modified hospital ownership types demonstrates that, among the ownership categories, 'Voluntary' hospitals have the highest count of ratings above, below, and the same as the national average. 'Government' hospitals follow, with a sizable count of patient experience ratings across all three categories—above, below, and same as the national average. 'Proprietary' hospitals exhibit a notable count across all three categories as well, though relatively fewer compared to 'Government' and 'Voluntary' hospitals. 'Physician' and 'Tribal' categories show significantly lower counts of patient experience ratings across these three national comparison categories, with 'Tribal' hospitals having no ratings below or above the national average.

## Geographic Distribution

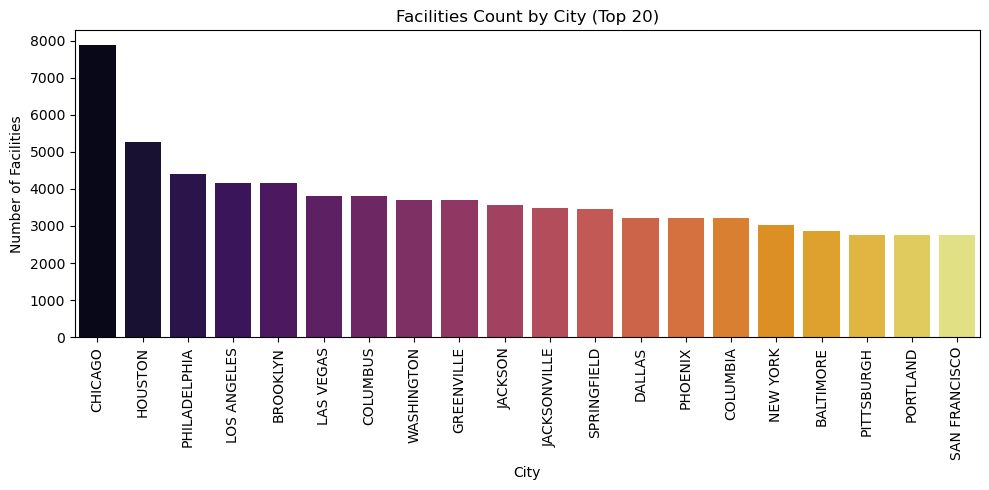
### Question 1: What is the geographic distribution of facilities by states?



**Figure 15**

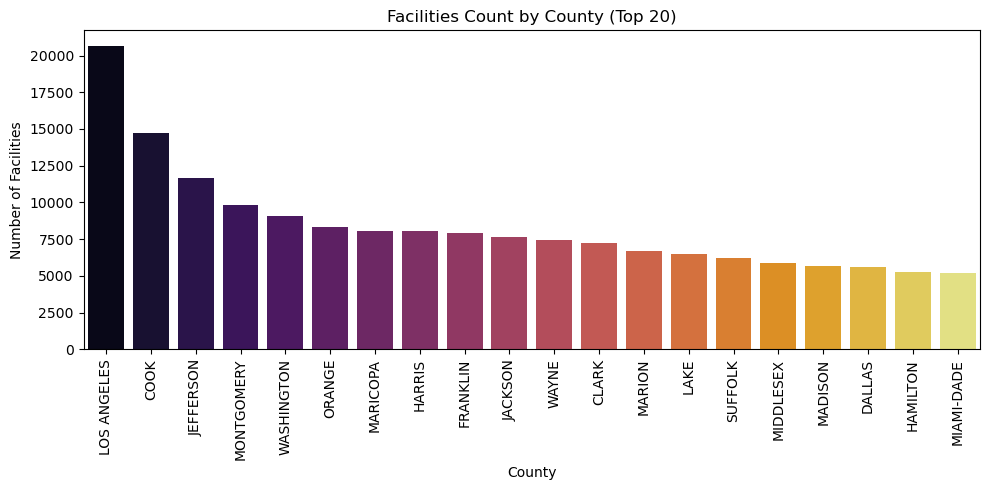
The distribution of healthcare facilities across different states shows significant disparities in the dataset. California (CA) has the highest count of facilities with approximately 79,970, followed closely by Texas (TX) with 71,168 facilities and Florida (FL) with 56,196 facilities. Conversely, the District of Columbia (DC) and Vermont (VT) have the lowest count, each having only 2,329 and 2,076 facilities, respectively.

### Question 2: What is the geographic distribution of facilities by cities?



**Figure 16**

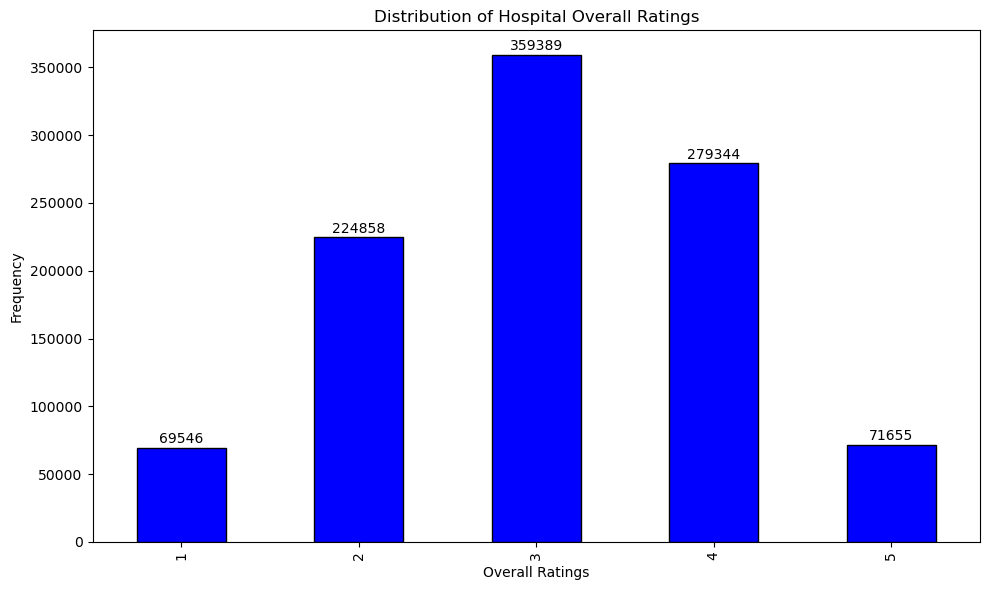
### Question 3: What is the geographic distribution of facilities by counties?



**Figure 17**

## Hospital Information

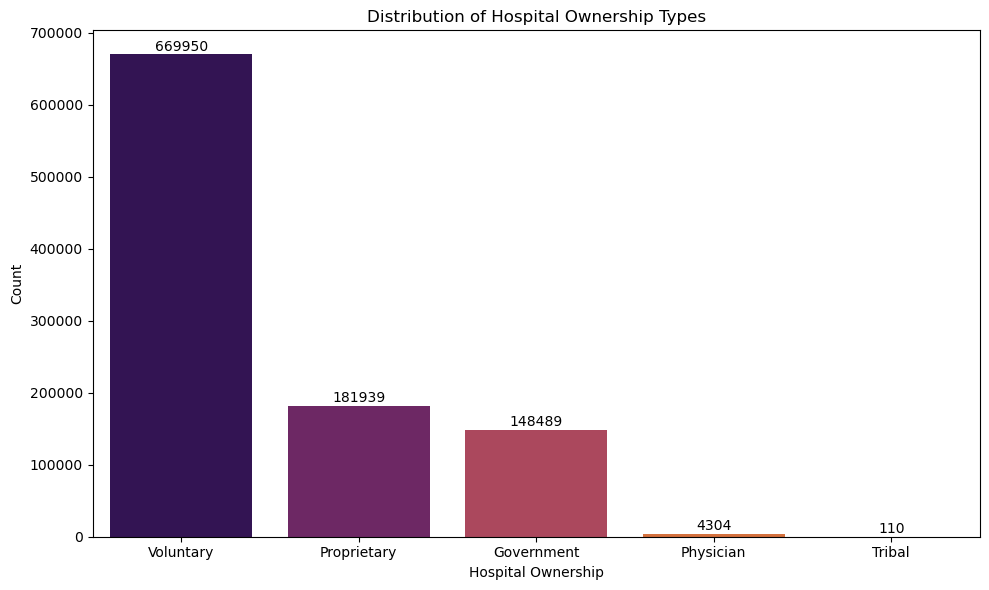
### Question 1: What is the distribution of Hospital overall ratings?



**Figure 18**

The distribution illustrates that Overall Ratings of Hospital has a significant difference. The frequency for “3” is on top, with following the ratings “4” and “2”.

### Question 2: What is the distribution of hospital ownership types?

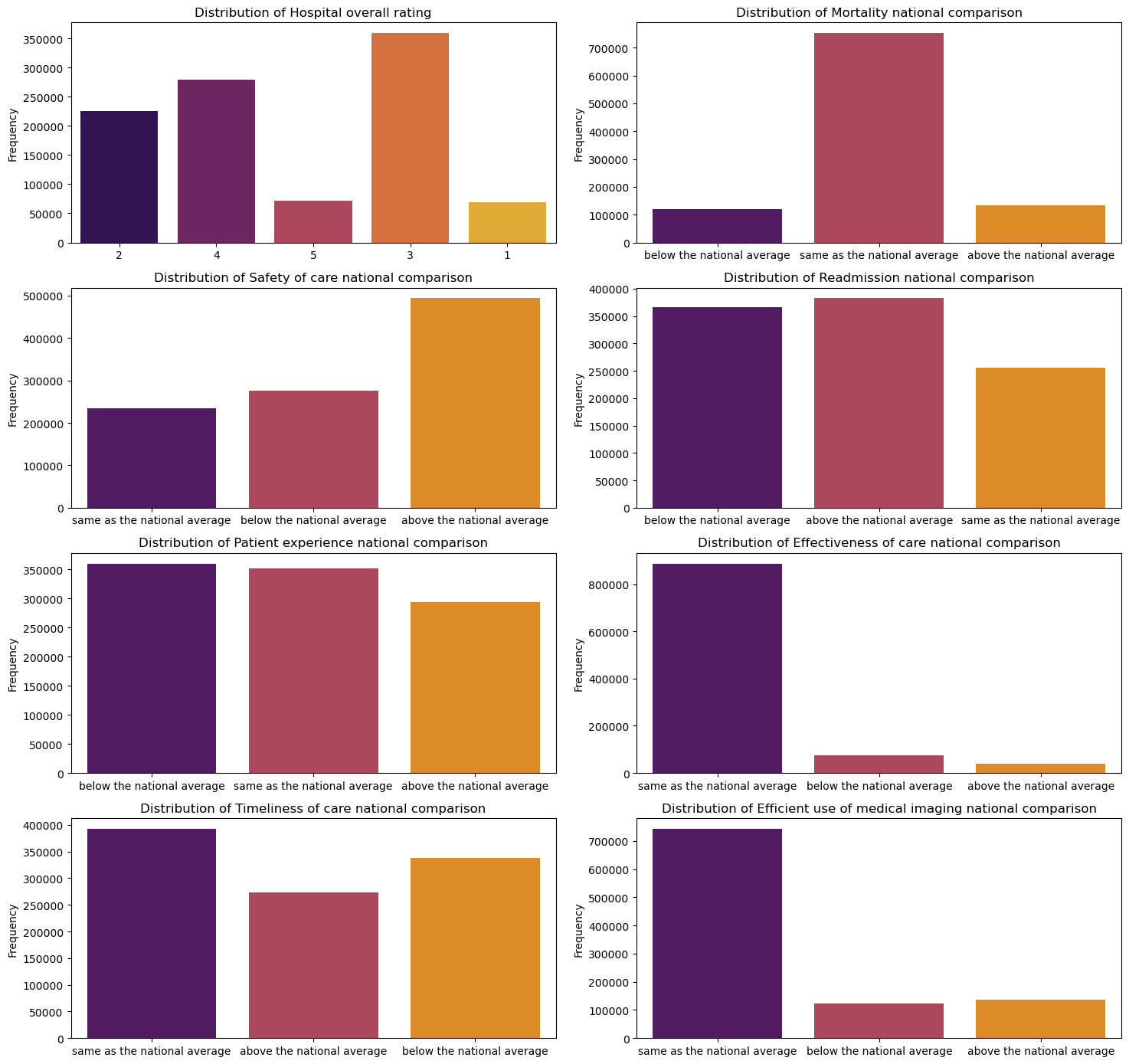


**Figure 19**

The distribution of hospital ownership types in the dataset portrays a varied landscape. "Voluntary" hospitals dominate the dataset, accounting for 669,950 instances, representing a substantial portion. Following this, "Proprietary" hospitals make up around 181,939 instances, establishing a significantly smaller presence. "Government" hospitals emerge as another noteworthy category, comprising approximately 148,489 instances. In contrast, "Physician" and "Tribal" hospital ownership types are less prevalent, with 4,304 and 110 instances, respectively. Overall, the dataset showcases a diverse range of hospital ownership types, with "Voluntary" hospitals occupying the most substantial proportion.

## Rating Comparisons

### Question 1: What are the different rating comparisons available for each category?

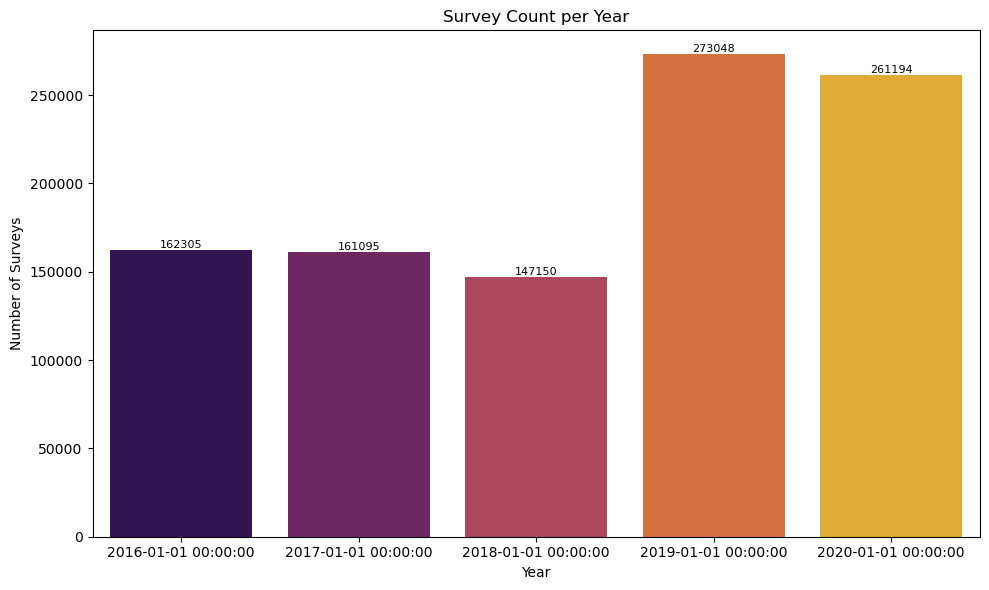


**Figure 20**

Hospital Overall Rating, Mortality National Comparison, Safety of Care National Comparison, Readmission National Comparison, Patient Experience National Comparison, Effectiveness of Care National Comparison, Timeliness of Care National Comparison and Efficient Use of Medical Imaging National Comparison reflecting the results as per Above the National Average, Same as the National Average, Below the National Average.

## Time Analysis

### Question 1: How many unique years are covered in the dataset and survey count in each year?



**Figure 21**

The year 2019 amassed the highest number of surveys, totaling 273,048, closely followed by 2020 with 261,194 surveys. The years 2016 and 2017 recorded relatively similar counts, each comprising approximately 162,305 and 161,095 surveys, respectively, while 2018 accounted for 147,150 surveys. The distribution illustrates a gradual increase in survey participation from 2016 onwards, reaching its peak in 2019 before a relatively consistent count in the subsequent years, with a slight decline in 2020.

# Feature Engineering

## Selecting Columns

### Dropping Irrelevant Features:

The 'Patient Survey Star Rating' column exhibited a high percentage of missing values, exceeding 80%. Due to this, it was deemed unreliable for analysis and was consequently removed from the dataset.

### Handling Missing Values:

In the case of the 'Number of Completed Surveys' column, instances with missing values were identified. As the number of missing values in this particular column was deemed minimal, the decision was made to remove the rows containing these missing values. This ensures a robust dataset for the subsequent analysis and model development.

## Feature Selection:

A crucial step in the feature engineering process was the selection of columns vital for the development of the machine learning model. This involved carefully choosing specific attributes that could significantly influence the predictive power of the model. The selected columns were determined based on their relevance and impact on the target variable.

## Encoding the Categorical Values

### Separating Features X and Target Variable Y

To facilitate further data processing and modeling, the dataset was divided into features (X) and the target variable (y).

**Features (X):** The feature set (X) consists of all columns from the dataset except for the 'Patient experience national comparison' column.

**Target Variable (y):** The target variable (y) specifically represents the 'Patient experience national comparison' column, which is the focal point of prediction or analysis.

### Categorizing Columns Based on Data

The columns in the dataset were categorized based on their data types into two distinct groups:

**Categorical Columns:**

Columns such as 'HCAHPS Question,' 'HCAHPS Answer Description,' 'Hospital Type,' 'Hospital Ownership,' 'Emergency Services,' 'Mortality national comparison,' 'Safety of care national comparison,' 'Readmission national comparison,' 'Effectiveness of care national comparison,' 'Timeliness of care national comparison,' and 'Efficient use of medical imaging national comparison' were identified as categorical features. These columns contain qualitative information or categorical data.

**Numeric Columns (excluding 'Year' and 'Hospital overall rating'):**

Other columns, excluding the categorical columns mentioned above, 'Year,' and 'Hospital overall rating,' were identified as numerical features. These columns are considered to contain numerical or continuous data and will be treated accordingly during further preprocessing and analysis.

This categorization helps structure the data, allowing for specific processing strategies tailored to the different data types present in the dataset.

## Final Dataset Overview

The final dataset has undergone a series of preprocessing steps and feature engineering techniques, resulting in a refined structure ready for machine learning analysis.

**Dataset Size and Composition:**

**Size:** The dataset comprises a refined set of features and a target variable for machine learning analysis.

**Columns:** It includes a subset of columns that are deemed significant for predicting the 'Patient experience national comparison' target variable.

**Records:** The number of records in the final dataset reflects the processed and cleaned observations used for model development.

**Preprocessed Features:**

**Selected Features:** The dataset includes relevant features chosen based on their importance and relevance to predict the patient experience.

**Categorical Columns:** Categorical columns have been appropriately encoded or prepared for model ingestion, enhancing their suitability for machine learning algorithms.

**Numeric Columns:** Numeric columns, excluding those designated as categorical and specific columns like 'Year' and 'Hospital overall rating,' are structured for numerical analysis and model training.

**Target Variable (y):** The 'Patient experience national comparison' column serves as the target variable for the predictive modeling task.

**Processed Target:** Preprocessing steps have been applied to ensure the target variable is ready for training models, removing 'Not Available' or 'Not Applicable' instances.

**Readiness for Machine Learning:**

**ModelReady Format:** The dataset is prepared in a format optimized for machine learning tasks, with cleaned, structured features, and the target variable in a form suitable for training models.

**Processed and Refined:** Through extensive preprocessing, missing value handling, feature selection, and categorization, the dataset is refined and primed for training machine learning models to predict patient experience national comparison effectively.

# Machine Learning Classifier Implementation

The data journey began with meticulous preprocessing, handling missing values, and refining the dataset. Feature engineering streamlined essential columns and categorized hospital ownership types for better interpretation. Categorical values were encoded, and the dataset was transformed for machine learning compatibility.

Machine learning classifiers, including Decision Tree and Logistic Regression, were implemented and rigorously evaluated. The Decision Tree model demonstrated approximately 67.37% accuracy, while Logistic Regression yielded around 56.04%. Each model showcased unique strengths and weaknesses across metrics, underscoring the importance of choosing the right model for healthcare predictive analytics. This comprehensive process encompassed preprocessing, feature engineering, encoding, and rigorous model assessment for healthcare performance prediction.

## Performance Metrics:

**Accuracy:** Decision Tree achieved approximately 67.37% accuracy, while Logistic Regression yielded around 56.04% accuracy.

**Precision:** Decision Tree displayed precision scores of roughly 65.74% for the first class, 74.50% for the second, and 60.86% for the third. Logistic Regression exhibited precision scores of about 55.54% for the first class, 62.89% for the second, and 47.48% for the third.

**Recall:** Decision Tree showed recall scores of approximately 71.32% for the first class, 75.75% for the second, and 55.41% for the third. Logistic Regression had recall scores of about 60.90% for the first class, 68.22% for the second, and 39.44% for the third.

**F1 Score:** The F1 scores for Decision Tree were roughly 68.41% for the first class, 75.12% for the second, and 58.03% for the third. For Logistic Regression, the F1 scores were around 58.09% for the first class, 65.45% for the second, and 43.01% for the third.

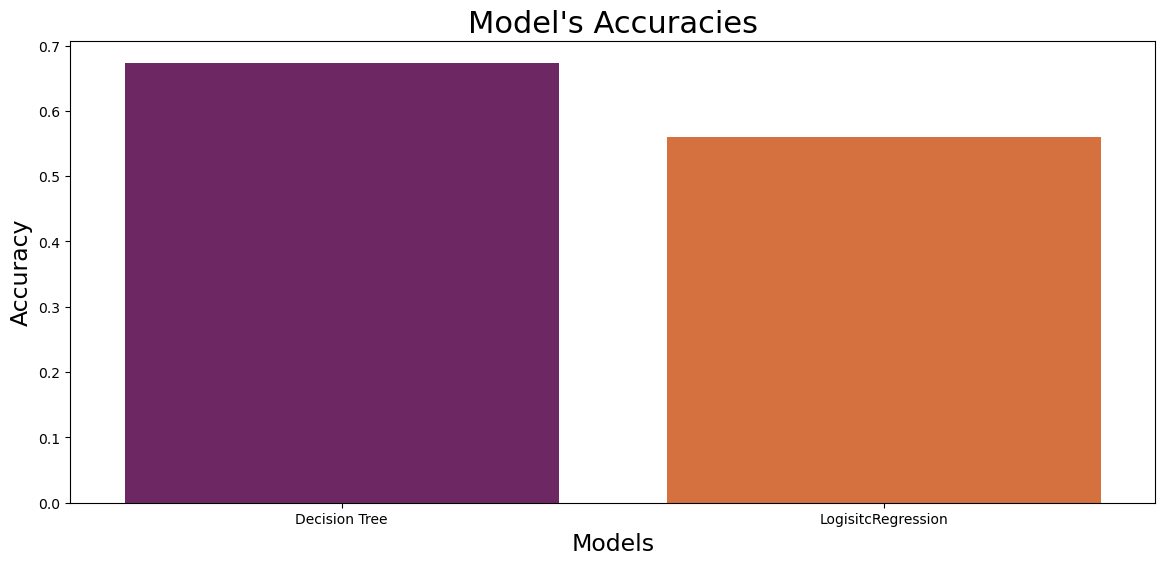
**Confusion Matrix:** Both models' confusion matrices detailed the true positives, true negatives, false positives, and false negatives across the three classes, providing an insightful breakdown of classification performance.

**Model Performance Summary:**

**Decision Tree:** Demonstrated relatively higher performance across all metrics compared to Logistic Regression. It showcased better accuracy, precision, recall, and F1 scores for all three classes.

**Logistic Regression:** While yielding lower performance metrics compared to Decision Tree, it still provided valuable insights into the dataset. It showcased a tradeoff between precision and recall, particularly between classes, which could be significant for specific healthcare prediction applications.

## Summary of the Results

The comparison between the Decision Tree and Logistic Regression models revealed notable differences in performance. The Decision Tree model demonstrated superior overall accuracy at around 67.37% compared to Logistic Regression's approximately 56.04%. Across precision, recall, and F1 scores for each class, the Decision Tree consistently outperformed Logistic Regression, showcasing better performance in predicting different categories within the dataset. Although Logistic Regression showed lower accuracy, its tradeoff between precision and recall could be valuable for specific healthcare prediction scenarios. In summary, the Decision Tree model emerged as the more effective choice due to its higher performance across various evaluation metrics, making it better suited for this particular predictive task.

**Figure 22**

# Recommendations

Based on the analyses and model performances, several recommendations can enhance the predictive outcomes and insights derived from the healthcare dataset. Firstly, given the Decision Tree model's superior performance in accurately predicting different categories within the dataset, it stands as the preferred model for this specific prediction task. However, to further refine and improve predictions, several steps can be taken.

**Feature Engineering Refinement:** Continuous refinement of feature selection and engineering could contribute significantly to model enhancement. Exploring additional relevant features or crafting more complex features might offer a deeper understanding of the data, potentially improving predictive capabilities.

**Further Data Exploration:** Delving deeper into data exploration could uncover hidden patterns or relationships that might have been overlooked initially. Investigating correlations between features, outlier detection, or exploring alternative data transformations may yield insights valuable for model improvement.

**Model Tuning and Optimization:** For both the Decision Tree and Logistic Regression models, hyperparameter tuning and optimization could be beneficial. Finetuning parameters using techniques like crossvalidation might enhance model generalization and performance.

**Addressing Class Imbalance:** If present, class imbalances can significantly affect model performance. Techniques such as oversampling minority classes or using weighted loss functions might mitigate this issue and improve model predictions for underrepresented classes.

**Continued Evaluation and Monitoring:** Continuous monitoring of model performance is crucial, especially in a healthcare context where predictions impact decisions. Implementing a robust monitoring system ensures the model's reliability and effectiveness over time, allowing for prompt updates or modifications as needed.

In conclusion, a combination of ongoing model refinement, and feature enhancement, could further elevate the predictive accuracy and realworld applicability of the healthcare prediction models. Continued exploration and refinement are key to developing robust, reliable, and impactful predictive models in healthcare settings.

# Limitations and Future Directions

## Limitations:

The limitations primarily revolve around the data quality, including missing values and limited granularity in certain features, which may have impacted model performance. The selected models might not capture all intricacies in patient experience determinants. More complex models could potentially improve predictions. The model's applicability might be restricted due to the dataset's geographical or temporal constraints.

## Future Directions:

In order to enhance the current dataset's depth and bolster model robustness, it's crucial to pursue additional datasets or more detailed information. Exploring advanced modeling techniques like deep learning or ensemble methods could better capture intricate patterns within the data. Validating the model's performance using external datasets or in real-world healthcare environments will ascertain its applicability beyond the current scope. This continuous improvement cycle, coupled with a focus on ethical considerations and fairness metrics, will pave the way for a more comprehensive and adaptable healthcare prediction system.

# Conclusion

In conclusion, this comprehensive analysis of the healthcare dataset has shed light on crucial factors influencing patient behavior and the drivers behind patient experience. Identifying and addressing these factors is paramount for healthcare providers aiming to enhance patient satisfaction and overall healthcare quality. The exploration of various features, such as survey ratings and national comparisons, has allowed us to discern patterns and trends that contribute to the patient experience. Furthermore, the analysis has spotlighted hospitals performing below average, prompting the need for closer scrutiny and potential interventions. By leveraging these insights, healthcare institutions can strategically focus on areas that require improvement, ultimately fostering a patient-centric approach and elevating the overall quality of healthcare services. Continued monitoring and adaptation based on emerging trends will be key to sustaining positive patient experiences and fostering a culture of continuous improvement in healthcare delivery.

# Acknowledgement

We extend our thanks to the Centers for Medicare & Medicaid Services for furnishing the invaluable data utilized in this analysis.