

Enhancing Healthcare Experiences: A Deep Dive into Patient Satisfaction Metrics

Table of Contents

1. Executive summary	pg.3
2. Introduction	pg.4
3. Dataset Selection	pg.5
4. Data Management	pg.6
5. Data Exploration and Analysis	pg.7
6. Text Analysis	pg.9
7. Statistical Modeling	pg.10
8. Data Visualization	pg.11
9. Conclusion	pg.19
10. References	pg.20
11. Appendix	pg.21

Executive Summary

Purpose:

- The primary aim of this report is to offer a concise overview of the comprehensive project that assessed patient satisfaction metrics across healthcare facilities in the United States. This project sought to uncover trends and insights that could guide hospitals in enhancing the quality of patient care and operational efficiency.

Summary:

- The investigation entailed rigorous data processing techniques using R and Power Query, ensuring the accuracy and reliability of the dataset. Subsequently, I utilized Power BI for an exploratory data analysis (EDA), which facilitated a deep dive into the trends and patterns within the patient satisfaction scores. The analysis highlighted a concerning downward trajectory in overall hospital ratings, particularly in the domains of care transition, communication about medicines, and discharge information. Sentiment analysis, conducted through advanced natural language processing in R, identified key areas for improvement, most notably in the communication between patients and healthcare providers and the management of patient pain.

Major Findings:

- A continuous decline in patient satisfaction from 2015 to 2018 across multiple performance indicators.
- A significant correlation between the quality of communication and patient satisfaction levels.
- Pain management was frequently mentioned in patient feedback as an area needing improvement.

Introduction

- Project Scope:

The project focuses on data management, analysis, and visualization of the "Healthcare Patient Satisfaction Survey" dataset obtained from Kaggle. The dataset includes valuable feedback, ratings, and demographic information from patients, providing insights into their experiences with healthcare services.

- Data Management:

- The dataset will be cleaned using both R and Power Query to ensure accuracy and reliability in my analyses.

- Analysis:

- The primary objective is to utilize survey results to calculate patient satisfaction. This involves identifying key areas for service enhancement by analyzing patient feedback and ratings.

- Visualizations:


- Historical results of summary star ratings and linear mean scores for various survey metrics will be tracked through visualizations. This will offer a comprehensive view of trends and performance over time.

- Objectives:

- Calculate the best and worst hospitals for doctor and patient insights based on the survey data.
 - Determining the percentage of survey questions answered to gauge the completeness of the dataset.
 - Determine the sentiment of survey questions, providing a qualitative understanding of patient experiences.

Dataset Selection

- Dataset Name
 - Healthcare Patient Satisfaction – Data Collection 2016-2020
- Reason for Selection
 - In the ever-evolving landscape of healthcare, patient feedback is paramount. The "Healthcare Patient Satisfaction - Data Collection" dataset from Kaggle was chosen due to its comprehensive coverage of patient feedback, service ratings, and demographic data. Its richness allows for in-depth analysis and visualization of patient satisfaction trends, helping to identify key areas for improvement in healthcare services. The dataset offers potential for meaningful insights into patient experiences, enabling us to show what can be done to better meet and exceed patient expectations. This choice aligns perfectly with my objective of enhancing patient care quality through data-driven insights.
- Confirmation of Dataset Claim



Samhitha Reddy Annem

Nov 2, 2023

Group 1 - 18. Healthcare Patient Satisfaction - Data Collection: <https://www.kaggle.com/datasets/kaggleprollc/healthcare-patient-satisfaction-data-collection>

👍 (1 like)

↩ Reply

- Confirmation of Unique Choice
 - This dataset was selected after thorough consideration and is unique to my project, as confirmed on the course's discussion board. I have ensured that this dataset is not being used by any other group, maintaining the integrity of the project, and avoiding any duplication.

Data Management

Process Detailing

- Data Cleaning:
 - Data Import and Concatenation: I had five raw datasets of 5 different years. Utilized Power Query to import and concatenate the five raw datasets into a single dataset, this formed a comprehensive dataset by combining information from multiple datasets.
 - Column Datatype Adjustment: Modified the column data types to align with the appropriate formats for improving consistency and providing accurate analysis results.
 - Column Deletion: Identified and removed columns with high percentage of missing values (96%) as they do not contribute to the analysis and help in dimensional reduction. Also, removed columns which contain specific error messages and footnotes which do not add considerable value to the analysis.
- Data Preparation:
 - Handling Textual Errors: Addressed errors in the "Number of Completed Surveys" column, which initially contained text data in an integer column. Replaced occurrences of "FEWER THAN 50" with the value 50 for immediate resolution.
 - Managed errors in the "Hospital overall rating" column by creating a custom column that replaced errors with null values, as those errors were caused because of having text ("Not Available") in an integer data type column.
 - For specific columns ("Patient Survey Star Rating," "HCAHPS Answer Percent," and "HCAHPS Linear Mean Value"), modified cell values from "Not Applicable" to null and "Not available" to zero. Adjusted data types to integer for improved analysis.
 - Derived Column Removal: Eliminated the "Year" column as it was derived from the "End Date" column, streamlining the dataset, and reducing redundancy.
 - Standardization: Standardized the term "National" across all relevant columns, ensuring consistency, enhancing clarity and for removing redundancy across the rows.

Transformation and Integration

- Data Transformation Techniques:
 - No transformation was needed to normalize the data.
- Data Integration Techniques:
 - We did not integrate and merge datasets from other sources.
- Rationale for Techniques Used:
 - The steps I took to clean the dataset to make sure that I got rid of unnecessary or error-filled columns. This helps in creating a dataset that is more dependable and precise when I analyze it.
 - Our cleaning process was always focused on making sure the dataset fits perfectly with what I want to achieve in the project. I aimed to maintain the quality of the data, ensuring it's reliable and ready for meaningful analysis.

Following these careful data cleaning and preparation procedures, the dataset is now well-prepared for in-depth changes and analysis. This sets a strong base for extracting valuable insights that align perfectly with what I want to achieve in the project.

Data Exploration and Analysis

Identify Trends/Patterns

- The PowerBI provided insights into hospital distribution by state, with Texas and California having the highest count of hospital overall compared. It details trends in hospital ratings over time, which show improvements except for 2016. This upward trend indicates a positive development in hospital ratings over the observed period.
 - The mortality rate is compared to national averages, with most facilities aligning with the norm the majority, 73.28% (3.27K facilities), have the same mortality rates as the national average.
 - Emergency services are widely available, with over 90% of facilities equipped to handle emergencies.
 - Additionally, HCAHPS linear mean scores and star ratings are examined, revealing a general decline in patient experience categories over the years. The report concludes with a need for improvement in hospital services, particularly in care transition, medication communication, and discharge information.
 - Linear mean scores and Star ratings which are used to quantify patient experiences at hospitals by HCAHPS (Hospital Consumer Assessment of Healthcare Providers and Systems). Linear mean scores are derived from individual survey responses, averaged, adjusted for patient mix and survey mode, rescaled to a 0-100 scale, and rounded. Star ratings, ranging from 1 to 5 stars, are assigned based on these scores using a clustering algorithm to ensure similarity within, and differences between, categories. These measures help consumers understand and compare hospital performance.

Statistical Descriptions

- Fitting distributions to survey answer percentages

```
set.seed(1)
#Sample Data
data <- sample(df$HCAHPS.Answer.Percent, length(df$HCAHPS.Answer.Percent)/2)

# Create Sequence
min <- min(data)
max <- max(data)
seq <- seq(min, max, by = .01)

# Calculate Parameters
beta <- mean(data)/var(data)
alpha <- beta * mean(data)
lambda <- 1/mean(data)

# Fit gamma and exponential distributions
gamma_dist <- dgamma(seq, alpha, beta)
exp_dist <- dexp(seq, lambda)

# Plot
hist(data, prob = TRUE, breaks = 100, main = "Histogram of Survey Answer %", xlab = "Survey Answer %")
lines(seq, gamma_dist, col = "red", lw = 3)
lines(seq, exp_dist, col = "blue", lwd = 3)
```

- Set sequence and calculated parameters of exponential and gamma distributions to fit a line to the histogram to better understand the distribution of the variable.

- Bootstrapping

```

set.seed(1)
boot = function(B, data){
  n=length(data)
  resample.boot=matrix(sample(data, size = B*n, replace=TRUE), B, n)
  mean.boot=apply(resample.boot, 1, mean)
  se=sd(mean.boot)
  list(se=se, mean=mean(mean.boot))
}

resample.boot=matrix(sample(data, size = 1000*length(data), replace=TRUE), 1000, length(data))
mean.boot=apply(resample.boot, 1, mean)

hist_data <- data.frame(x = mean.boot)

ggplot(hist_data, aes(x = x)) +
  geom_histogram(binwidth = 0.01, fill = "lightblue", color = "black", alpha = 0.7) +
  labs(title = "Bootstrapped Survey Answer %",
       x = "Mean Survey Answer %") +
  theme_minimal()
boot(1000, data)

Normt.CI=function(x, alpha){
  n=length(x)
  df=n-1
  ME=qt(1-alpha/2, df)*sd(x)/sqrt(n)
  c(mean(x)-ME, mean(x)+ME)
}

ggplot(hist_data, aes(x = x)) +
  geom_histogram(binwidth = 0.01, fill = "lightblue", color = "black", alpha = 0.7) +
  geom_vline(xintercept = c(Normt.CI(mean.boot, 0.05)[1], Normt.CI(mean.boot, 0.05)[2]),
            linetype = "dashed", color = "red", linewidth = 1) +
  labs(title = "Bootstrapped Survey Answer %",
       x = "Mean Survey Answer %") +
  theme_minimal()
print(paste("95% CI Lower Bound:", Normt.CI(mean.boot, 0.05)[1], "95% CI Upper Bound:", Normt.CI(mean.boot, 0.05)[2]))

```

- Sampled with replacement 1000 times to return a distribution of samples means which provides a normal distribution where I calculated standard error and mean.

```

$se
0.0900444349161937
$mean
34.7202404432979

```

■
Calculated the 95% confidence interval for the mean which returned these values:

```
[1] "95% CI Lower Bound: 34.7151851097236 95% CI Upper Bound: 34.7263375450307"
```

Contextual Implications

- This analysis helps a patient determine what hospital to seek treatment and which to avoid. This analysis helps hospitals understand what areas they are lacking in and where they can allocate more resources to improve results.
- This analysis helps determine how many survey takers answer a given question to determine the effectiveness of the survey at collecting information and provides feedback to help improve future surveys.

Text Analysis

- Text Preprocessing
 - We chose to use the "bing" lexicon in conjunction with tokenization to determine the sentiment of the text.
 - The "bing" lexicon is a valuable resource because it provides a pre-defined list of words categorized as either positive or negative. This lexicon allows us to quickly assign sentiment scores to individual words in the text.
 - By using the "bing" lexicon, I can identify the overall sentiment of the text, whether it leans towards positivity or negativity. This helps us gain insights into how the language used in the dataset may affect responses and perceptions.
 - The steps include converting text to lowercase, removing punctuation and numbers, and excluding common words (stopwords) that don't add much meaning to the analysis. This cleaning helps make the data more uniform and ensures that the analysis focuses on the more meaningful words.
- Sentiment Analysis:
 - We used tidytext, dplyr, and tidyr in R along with the "bing" lexicon to analyze sentiment. These powerful tools enable us to efficiently process and analyze text data.
 - We applied sentiment analysis to understand the emotional tone of the text and gauge respondents' sentiments. In analysis, the sentiment is Mostly Positive Sentiment.
 - This Analysis helped us for uncovering trends and patterns in the data related to sentiment.
- Topic Modeling
 - We did not employ topic modeling in my analysis because it was not applicable to the dataset. This dataset primarily consisted of survey questions and responses, and there was no need to identify specific topics. Instead, the focus was on understanding sentiment, which was more relevant to my objectives.
- Interpretation
 - The sentiment analysis conducted on the dataset has predominance of positive sentiments over negative ones. This indicates that the feedback from the survey questions related to patient satisfaction is mostly positive. Understanding how I phrase the questions elicits accurate and unbiased responses.

Statistical Modeling

- Model Used
 - For the statistical analysis, I wanted to understand the survey answer percentage. To explain this dependent variable, I utilized an ANOVA model. This model was used to determine the survey answer percentage and explain if it is statistically significant.
- Assumptions Made
 - We initially assumed that the survey answer percentage average is statistically different when grouped by survey question. The Hypothesis test for this model is as follows:
 - H0: There is no significant difference between the means of questions.
 - Ha: At least one question's mean is different than the others
- Model Results

```
ANOVA <- aov(HCAHPS.Answer.Percent ~ factor(HCAHPS.Question),
             data = df)
```

```
summary(ANOVA)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(HCAHPS.Question)	71	218805345	3081765	106895	<2e-16 ***
Residuals	254145	7326956	29		

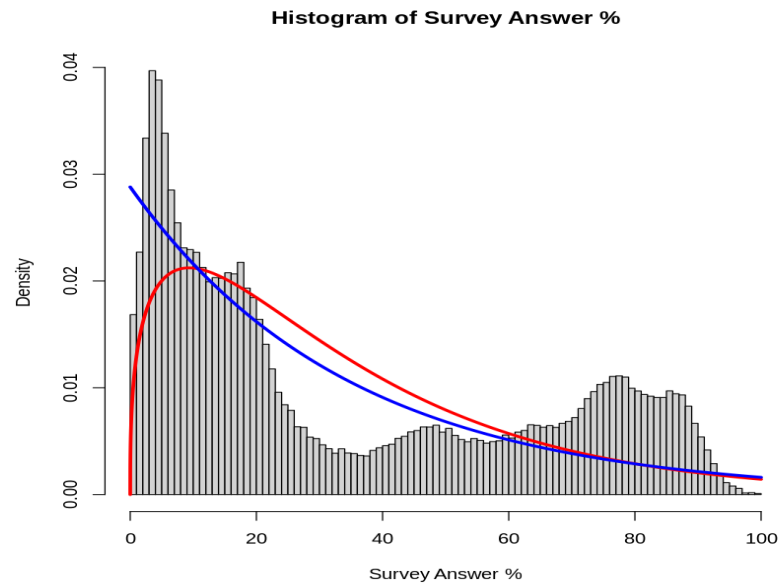
```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

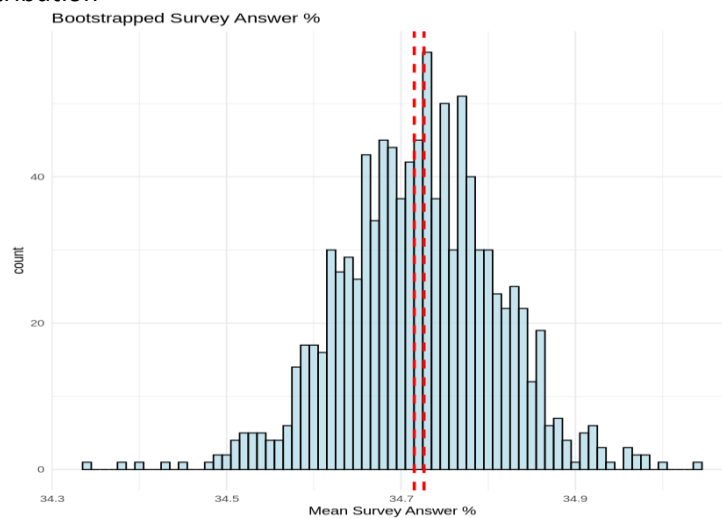
- After running the model, I observed a low p-value that is statistically significant between alpha levels 0 and 0.001 which means I can reject the null hypothesis and accept the alternative hypothesis that at least one question's mean is different than the others.
- Implication of Model Results
 - With the output of this model, I can conclude that I need to focus on identifying and fixing the lowest performing survey questions to address the problem of low survey answer percentages.

Data Visualization

R Visuals



This visualization shows the distribution of survey answer percentage being fit by both an exponential and gamma distribution. Both distributions underestimate on the left and right tails and overestimate in the center of the distribution.



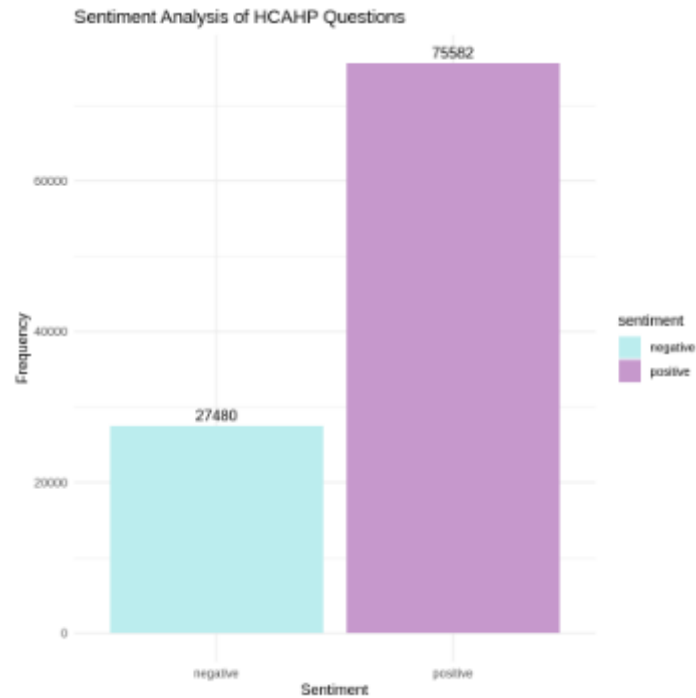
This visualization shows the survey answer percentage normally distributed after bootstrapping (resampling with replacement 1000 times and creating a distribution of the sample means). This visual also shows a 95% confidence interval for the mean:

Standard Error: .09

Mean: 34.72

95% CI Lower Bound: 34.72

95% CI Upper Bound: 34.73

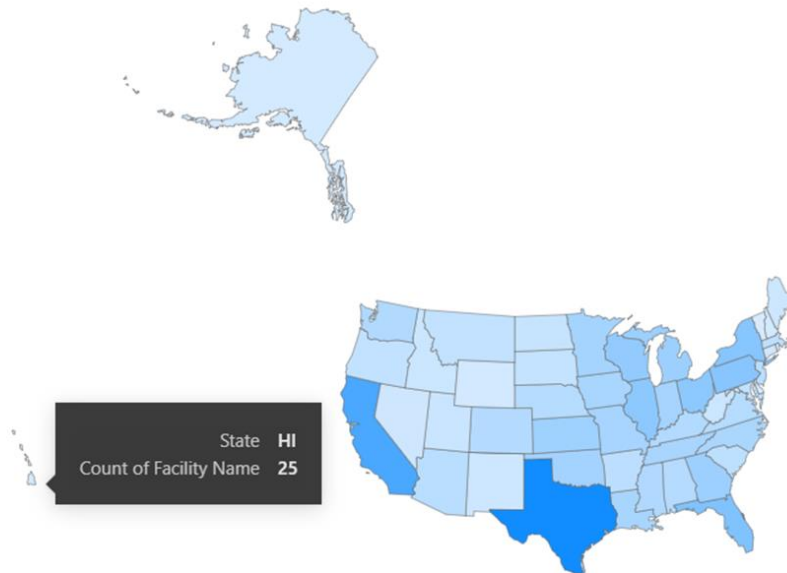


This barplot shows the total negative and positive sentiment scores after tokenizing the text and combining it with the “bing” lexicon to get a sentiment rating per word. From this graph I can clearly see that there is a far greater positive sentiment than negative

Power BI Visuals

Distribution of Hospitals Across States

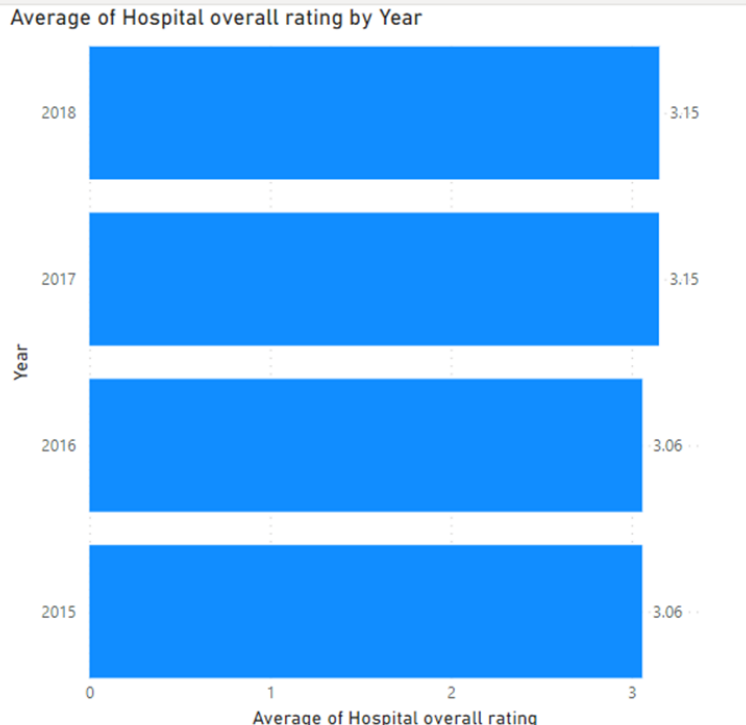
(Dark blue represents the highest and light blue shade represents low hospitals)



The visualization provided appears to be a choropleth map of the United States, which is commonly used to show statistical data by geographical regions through varying shades of color. In this case, the map is titled "Count of Facility Name by State," suggesting that it displays the distribution of a certain type of facility across different states. Most of the states are colored in a light shade, indicating a lower count, while two states, Texas (533 facility) and California (384 facility) are colored in a much darker shade. This implies that Texas and California have a significantly higher count of the facility compared to other states.

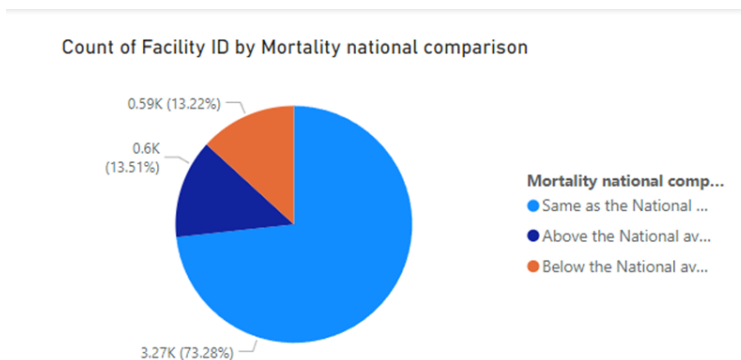
The Rating of hospitals over each year

The year shown is start year (example if its 2015 the rating is from year 2015-2016)



The bar chart shows the average overall hospital rating by year from 2015 to 2018. There has been a slight increase in the average rating from 2015 to 2016, remaining steady at 3.06. However, there's a noticeable improvement in the following years, with the average rating rising to 3.15 in both 2017 and 2018. This upward trend indicates a positive development in hospital ratings over the observed period.

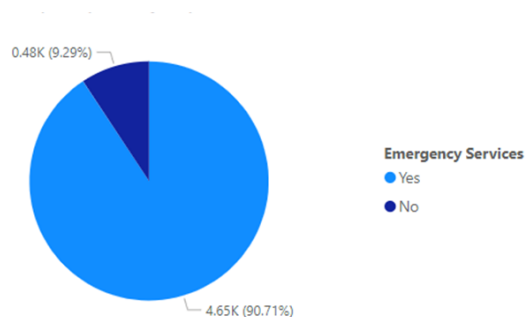
Mortality Rate Comparison with National Score Comparison



The pie chart presents data on the count of facilities by mortality national comparison. It shows that the majority, 73.28% (3.27K facilities), have the same mortality rates as the national average. A smaller, yet significant portion, 13.51% (0.6K facilities), are above the national average, while 13.22% (0.59K

facilities) are below the national average. This distribution highlights that most facilities align with the national mortality rates, indicating consistency in outcomes across the majority of these facilities.

Emergency Services Provided?

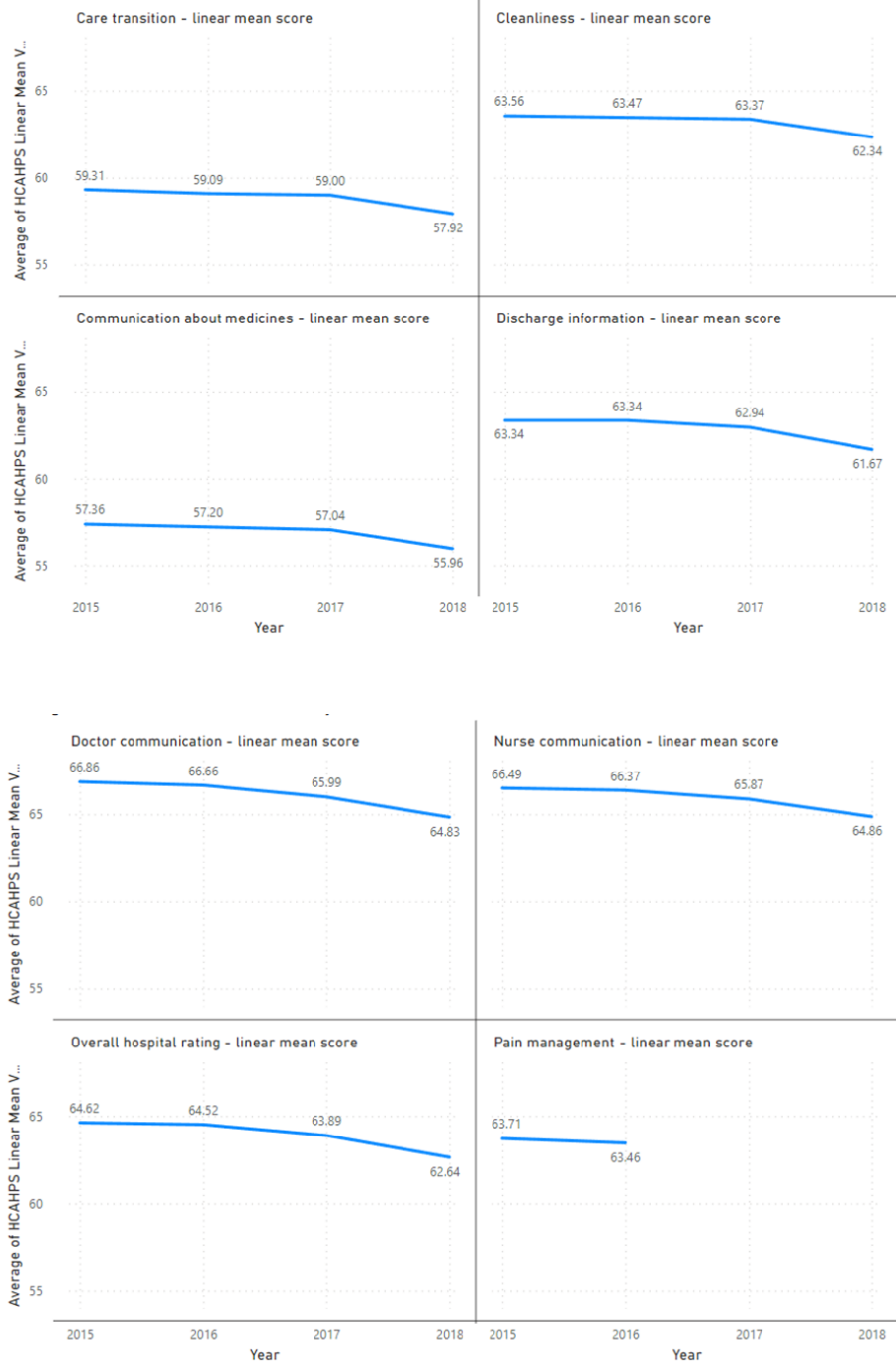


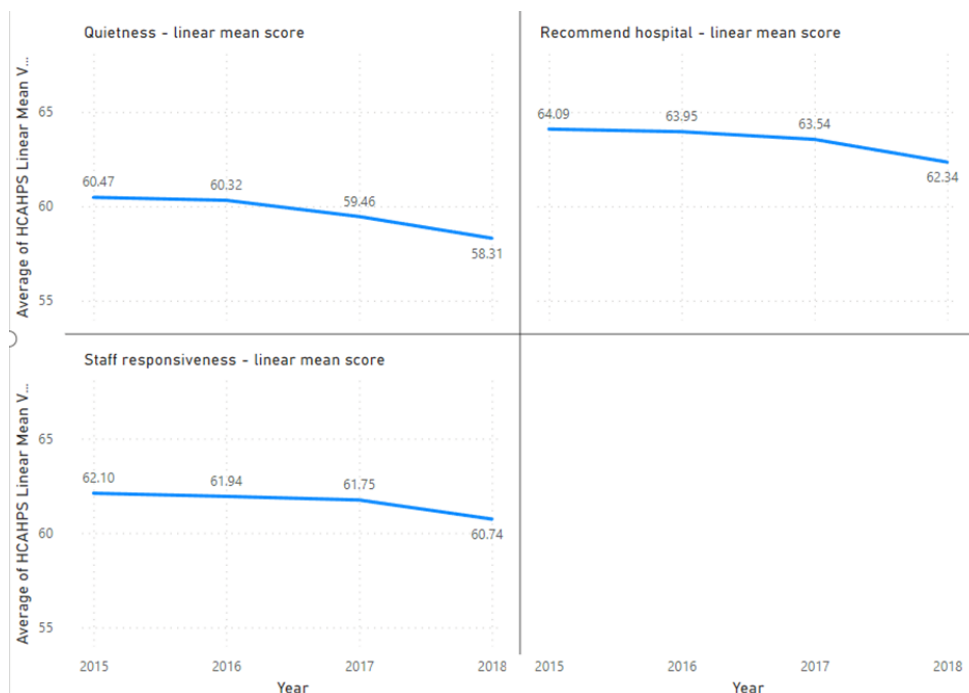
The pie chart illustrates the availability of emergency services in facilities, indicating that a vast majority, 90.71% (4.65K), offer emergency services. In contrast, a small fraction, 9.29% (0.48K), do not provide these services. This suggests that most of the facilities analyzed have the capability to handle emergency cases.

Linear mean scores and Star ratings which are used to quantify patient experiences at hospitals by HCAHPS (Hospital Consumer Assessment of Healthcare Providers and Systems). Linear mean scores are derived from individual survey responses, averaged, adjusted for patient mix and survey mode, rescaled to a 0-100 scale, and rounded. Star ratings, ranging from 1 to 5 stars, are assigned based on these scores using a clustering algorithm to ensure similarity within, and differences between, categories. These measures help consumers understand and compare hospital performance.

The Linear mean score values

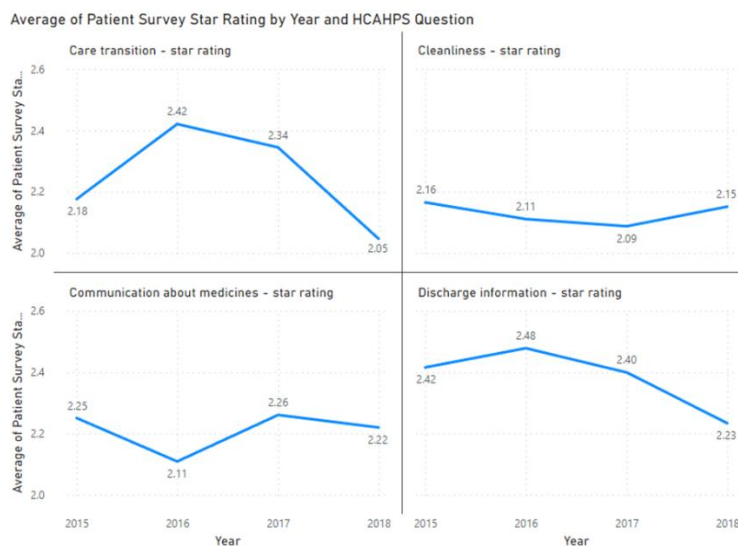
Average of HCAHPS Linear Mean Value by Year and HCAHPS Question

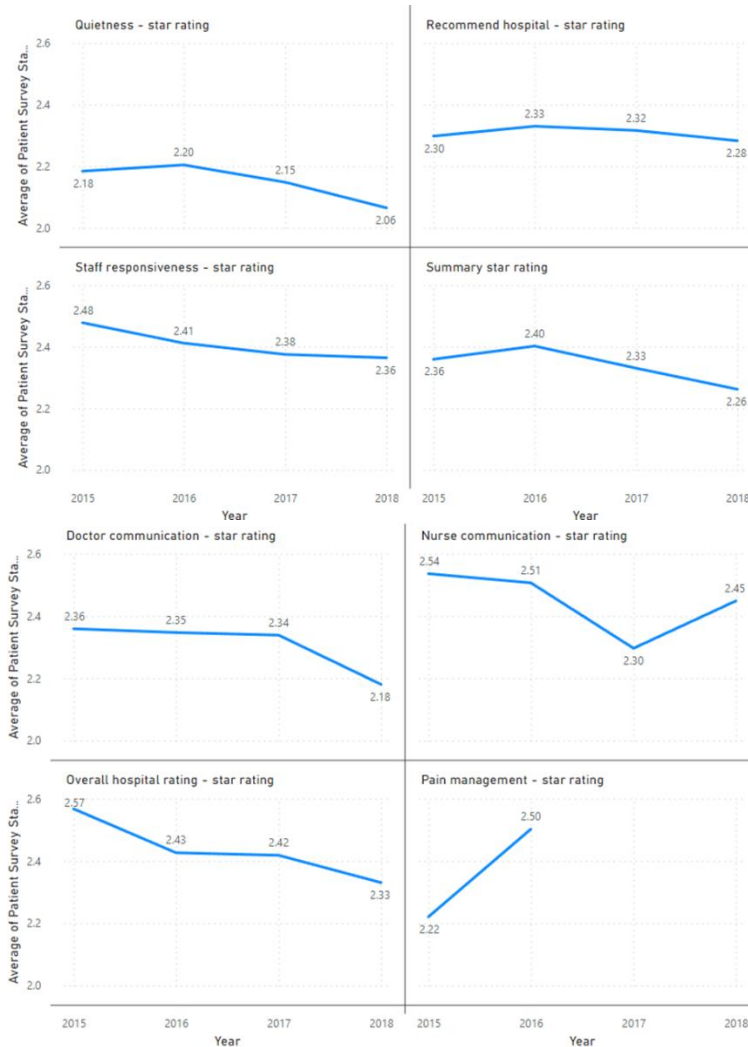




The series of images depict line graphs tracking the average HCAHPS linear mean scores across various categories from 2015 to 2018. There is a general downward trend in scores over the years in categories such as care transition, communication about medicines, cleanliness, and discharge information. The scores for doctor and nurse communication remain relatively stable with slight variations. Overall hospital ratings and staff responsiveness show a consistent decrease, while pain management and hospital recommendations also decline. The data suggests a need for hospitals to address these areas to improve patient experiences.

Average Star Rating for Each Question.





The images show a downward trend in patient survey star ratings across various HCAHPS questions from 2015 to 2018. Care transition, cleanliness, communication about medicines, and discharge information ratings declined. Doctor communication and overall hospital ratings decreased, with a slight dip in nurse communication and staff responsiveness. Pain management saw an initial drop followed by an increase, while the quietness of the hospital environment continued to decrease. Recommendations for hospitals and summary star ratings also showed a downward trend, suggesting areas for potential improvement.

- Adherence to Best Practices
 - Our visuals follow best practices, because I used the right chart to display my data and used relevant and accurate labels and titles.

Conclusion

- Summary
 - This comprehensive study, utilizing the U.S. Hospital Customer Satisfaction dataset from 2016 to 2020, revealed critical insights into hospital operations and patient satisfaction. I employed advanced data management techniques, including data cleaning and preparation in R and Power Query. This exploration and analysis highlighted trends in hospital ratings and patient satisfaction metrics. The sentiment analysis, primarily positive, offered a deeper understanding of patient feedback, while the statistical modeling (ANOVA) validated significant differences in survey responses.
- Significance
 - This research has significant implications for healthcare providers. By analyzing patient feedback and satisfaction metrics, hospitals can identify areas requiring improvement, thereby enhancing patient care. The findings also provide a robust foundation for hospitals to strategize resource allocation and improve operational efficiency.
- Limitations and Recommendations:
 - Our study, though extensive, faced limitations due to the dataset's scope and potential biases in self-reported data. Future research should expand the dataset range and explore alternative methodologies to counter these biases. Additionally, exploring more granular data, such as demographic-specific patient feedback, could yield further insights. Implementing advanced statistical techniques or machine learning models may also refine the analysis, providing a more nuanced understanding of patient satisfaction determinants.

References

ABeyer. "U.S. Hospital Customer Satisfaction 2016-2020." *Kaggle*, 1 June 2021,
www.kaggle.com/datasets/abrambeyer/us-hospital-customer-satisfaction-20162020.

Appendix

Presentation: [Final Presentation.pptx](#)

Power BI Dashboard: <https://app.powerbi.com/groups/f1fd5f87-61eb-4b85-b971-2652153f95f7/reports/5cc368e2-1edd-4b4e-a9ed-b1a32d3969e2/ReportSection?experience=power-bi>