Exploratory Data Analysis and Predictive Modeling of Hospital Quality Ratings in the United States

Sumith Kumar Kurapati

00789516

University of New Haven

10/10/25

Abstract

This study explores the Centers for Medicare & Medicaid Services (CMS) Hospital Quality dataset to understand how hospitals across the United States perform across multiple quality domains and how these performance measures relate to the overall hospital rating. The analysis follows a structured data exploration and modeling process using R with packages such as dplyr, tidyverse, and ggplot2. After identifying and treating missing values and standardizing text categories, descriptive and inferential analyses were conducted. The results reveal that most hospitals tend to have ratings around three to four, indicating average to above-average performance. Regression analysis shows that patient experience and readmission rates are strong predictors of overall hospital rating, with a moderate R-squared value suggesting other factors also influence the ratings. The findings imply that improvements in patient-centered care and reduced readmission rates can enhance hospitals' overall ratings.

---

1. Introduction

Hospitals in the United States are evaluated and publicly reported by the Centers for Medicare & Medicaid Services (CMS) to improve transparency and promote quality healthcare delivery. CMS compiles various performance metrics under its Hospital Quality Initiative and summarizes them in a single overall rating ranging from one to five stars. This report analyzes the "Hospitalinfo.csv" dataset provided by CMS, which contains hospitals' ratings across domains such as mortality, safety of care, readmissions, patient experience, effectiveness of care, and timeliness of care. These metrics help patients and policymakers assess the comparative performance of hospitals nationwide.

Several technical terms in the dataset are essential for interpretation. CMS refers to the federal agency responsible for administering national healthcare programs and quality assessments. The Electronic Health Record (EHR) represents a comprehensive digital record of patients' health data that facilitates coordination and care quality. The thirty-day readmission rate measures the proportion of patients readmitted within thirty days after discharge, reflecting hospital efficiency and quality of follow-up care. Similarly, the thirty-day mortality rate measures the likelihood of patient death within thirty days of admission, which serves as an indicator of hospital effectiveness.

The objectives of this study are twofold. The first is to identify which quality domains most strongly relate to hospitals' overall ratings. The second, more unconventional objective, is to determine whether non-clinical aspects such as patient experience or timeliness can act as proxies for overall quality ratings, even when clinical outcomes like mortality or readmission rates are similar.

2. Methodology

2.1 Data Preparation

The dataset includes information for several hundred hospitals, with each observation representing one hospital. Key variables include Hospital Overall Rating, Mortality National Comparison, Safety of Care National Comparison, Readmission National Comparison, Patient Experience National Comparison, Effectiveness of Care National Comparison, and Timeliness of Care National Comparison. Initial data inspection revealed the presence of missing values and placeholders such as blank cells and the text "Not Available." Additionally, ordinal categories like "Below the national average," "Same as the national average," and "Above the national average" were stored as text and needed to be converted to numeric codes.

The cleaning process involved converting blank and "Not Available" values to NA in R and mapping the ordinal categories to numeric values (1 for Below, 2 for

Same, 3 for Above). The Hospital Overall Rating variable was also converted to numeric. Missing values were then addressed by either removing incomplete records or imputing them with median values, ensuring consistency for analysis. This transformation allowed the dataset to be suitable for statistical modeling.

2.2 Methods Applied

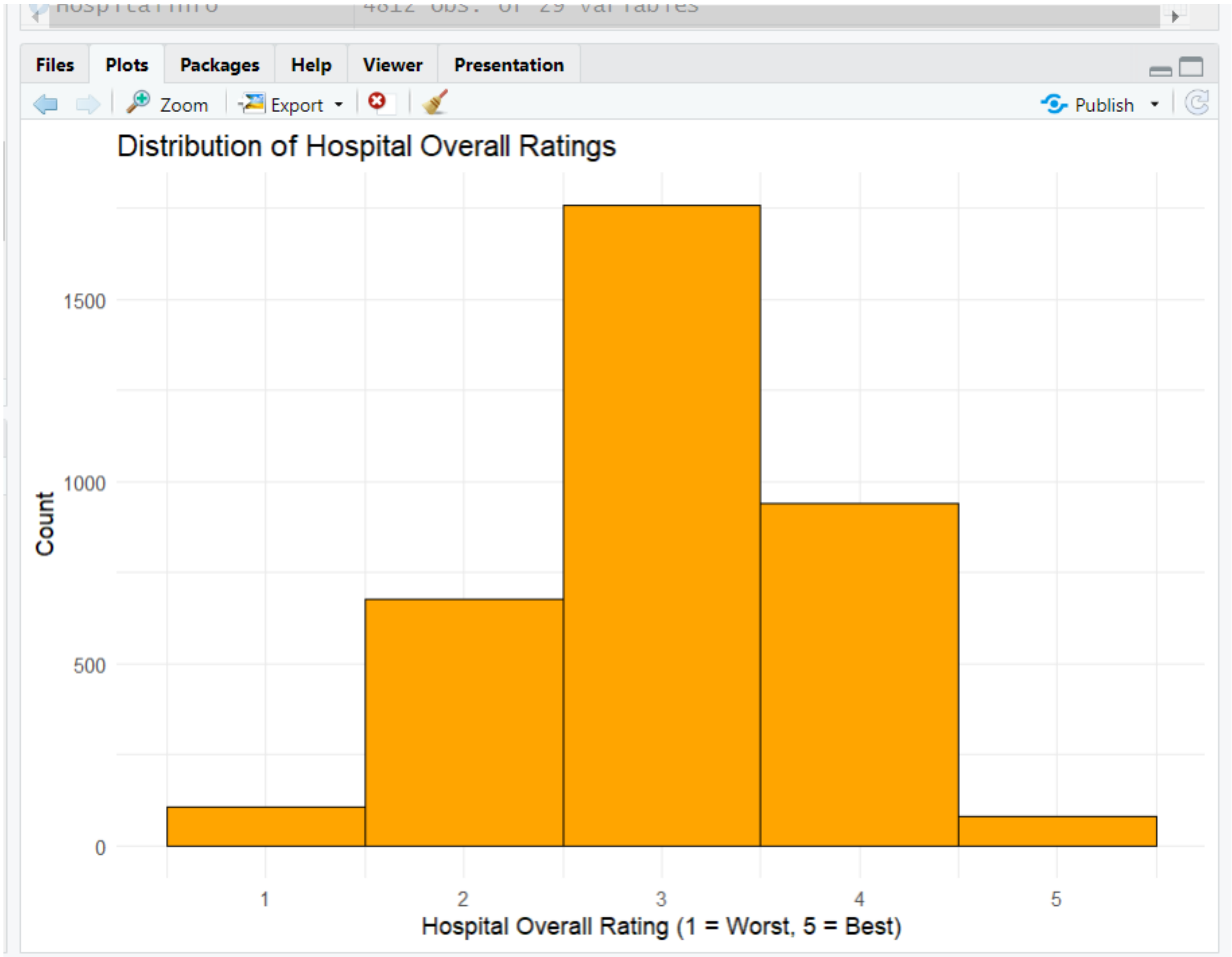
Descriptive exploratory data analysis was first performed to understand data distribution and relationships. Histograms were created to visualize the distribution of hospital ratings, and boxplots were used to compare overall ratings against each performance domain. These visualizations helped reveal trends and potential associations between domains and ratings. Following EDA, a multiple linear regression model was built to predict Hospital Overall Rating using performance domains as predictors. The model's coefficients, statistical significance, and R-squared value were interpreted to assess which quality domains most influence the overall hospital rating. The analysis assumes linearity and treats ordinal categories as continuous values for simplicity, recognizing this as a limitation.

---

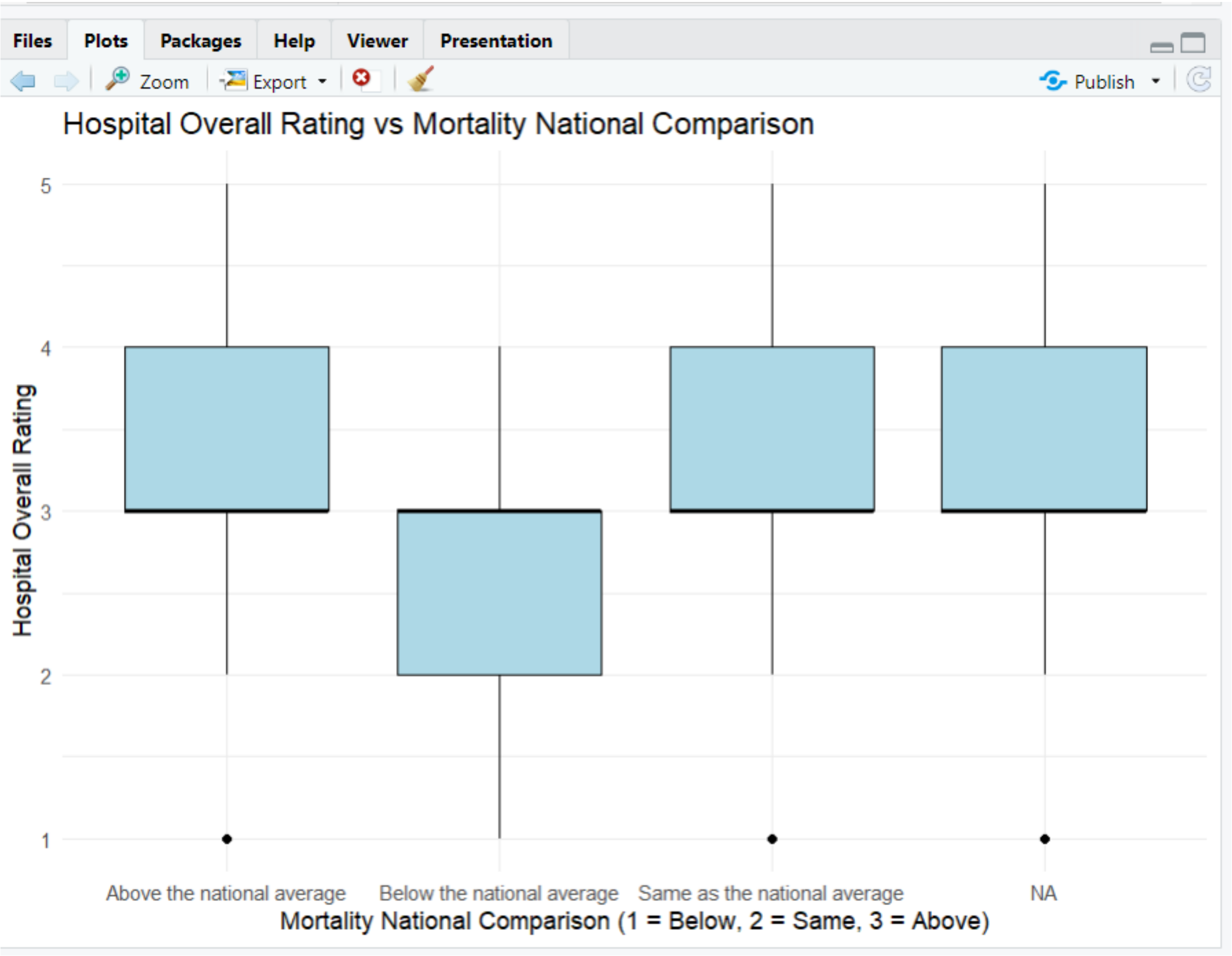
3. Findings and Discussion

3.1 Exploratory Results

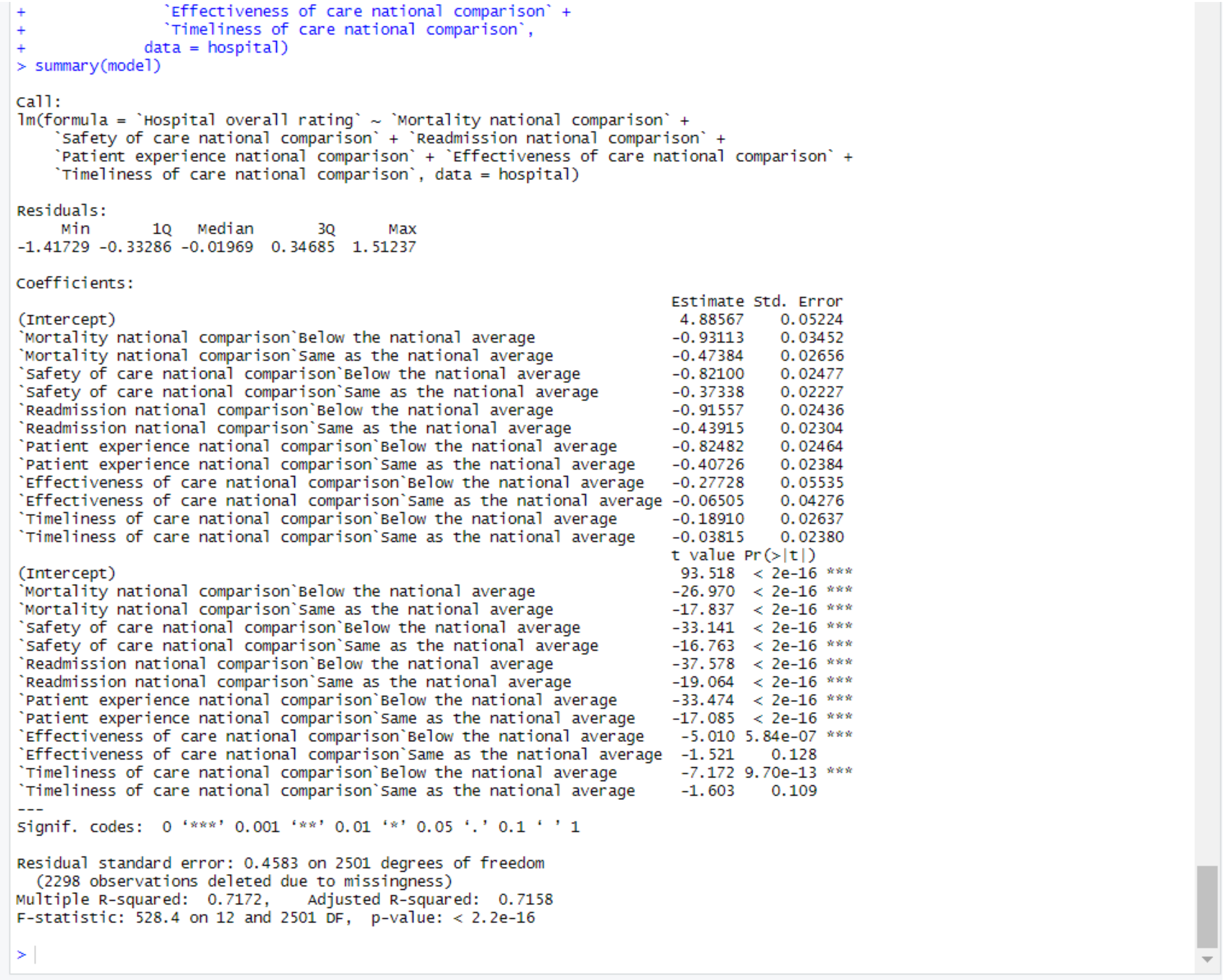
The histogram of hospital overall ratings shows that most hospitals in the dataset received moderate scores, with ratings clustering around three and four stars. This pattern indicates that a majority of hospitals perform at or slightly above the national average level of quality. Very few hospitals received the lowest rating of one or the highest rating of five, suggesting that extreme performance—either poor or exceptional—is relatively uncommon. The distribution is slightly right-skewed, meaning that there are somewhat more hospitals with higher ratings than those with very low ratings. Overall, this chart implies that the U.S. hospital system demonstrates generally consistent performance, with most institutions achieving middle-to-high levels of overall quality rather than being concentrated at the extremes.


hospitalInfo
4812 obs. of 29 variables
Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Distribution of Hospital Overall Ratings
Count
0
500
1000
1500
1
2
3
4
5
Hospital Overall Rating (1 = Worst, 5 = Best)

The boxplot comparing hospital overall ratings to the Mortality National Comparison reveals a clear positive relationship between hospital performance on mortality and overall quality ratings. Hospitals categorized as "Above the national average” in mortality outcomes tend to have higher median overall ratings, typically around four, indicating better overall performance. In contrast, hospitals “Below the national average” show noticeably lower median ratings, clustering closer to three, which suggests that poorer mortality performance is associated with lower overall ratings. The hospitals rated “Same as the national average” fall in between these two groups, with median values around 3.5. The spread of data within each group is moderate, showing some variability but with few extreme outliers. Overall, the chart suggests that hospitals achieving superior mortality outcomes are generally rated higher in overall quality, aligning with expectations that lower patient mortality is a key indicator of hospital excellence.


Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
Hospital Overall Rating vs Mortality National Comparison
5
4
3
2
1
Hospital Overall Rating
Above the national average
Below the national average
Same as the national average
NA
Mortality National Comparison (1 = Below, 2 = Same, 3 = Above)

3.2 Regression Results



```
+               `Effectiveness of care national comparison` +
+                `Timeliness of care national comparison`,
+              data = hospital)
> summary(model)

Call:
lm(formula = `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`, data = hospital)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.41729 -0.33286 -0.01969  0.34685  1.51237 

Coefficients:
                                                                    Estimate Std. Error
(Intercept)                                                          4.88567    0.05224
`Mortality national comparison`Below the national average           -0.93113    0.03452
`Mortality national comparison`Same as the national average         -0.47384    0.02656
`Safety of care national comparison`Below the national average      -0.82100    0.02477
`Safety of care national comparison`Same as the national average    -0.37338    0.02227
`Readmission national comparison`Below the national average         -0.91557    0.02436
`Readmission national comparison`Same as the national average       -0.43915    0.02304
`Patient experience national comparison`Below the national average  -0.82482    0.02464
`Patient experience national comparison`Same as the national average -0.40726    0.02384
`Effectiveness of care national comparison`Below the national average -0.27728    0.05535
`Effectiveness of care national comparison`Same as the national average -0.06505    0.04276
`Timeliness of care national comparison`Below the national average  -0.18910    0.02637
`Timeliness of care national comparison`Same as the national average -0.03815    0.02380
                                                                    t value Pr(>|t|)    
(Intercept)                                                          93.518  < 2e-16 ***
`Mortality national comparison`Below the national average           -26.970  < 2e-16 ***
`Mortality national comparison`Same as the national average         -17.837  < 2e-16 ***
`Safety of care national comparison`Below the national average      -33.141  < 2e-16 ***
`Safety of care national comparison`Same as the national average    -16.763  < 2e-16 ***
`Readmission national comparison`Below the national average         -37.578  < 2e-16 ***
`Readmission national comparison`Same as the national average       -19.064  < 2e-16 ***
`Patient experience national comparison`Below the national average  -33.474  < 2e-16 ***
`Patient experience national comparison`Same as the national average -17.085  < 2e-16 ***
`Effectiveness of care national comparison`Below the national average  -5.010 5.84e-07 ***
`Effectiveness of care national comparison`Same as the national average -1.521    0.128    
`Timeliness of care national comparison`Below the national average   -7.172 9.70e-13 ***
`Timeliness of care national comparison`Same as the national average  -1.603    0.109    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4583 on 2501 degrees of freedom
  (2298 observations deleted due to missingness)
Multiple R-squared:  0.7172,	Adjusted R-squared:  0.7158 
F-statistic: 528.4 on 12 and 2501 DF,  p-value: < 2.2e-16

> 
```

The multiple linear regression model was developed to predict the Hospital Overall Rating using six performance domains—Mortality, Safety of Care, Readmission, Patient Experience, Effectiveness of Care, and Timeliness of Care. The model achieved a strong fit, with an R-squared value of approximately 0.717, indicating that about 71.7% of the variation in overall hospital ratings can be explained by these quality performance measures. The F-statistic is highly significant ($p < 0.001$), confirming the overall validity of the model. The coefficients show that hospitals performing “Below the national average” or “Same as the national average” in most domains have significantly lower overall ratings compared to those performing “Above the national average.” Among the predictors, mortality, readmission, and patient experience comparisons exhibit the largest negative coefficients and highly significant p-values ($p < 0.001$), suggesting that these factors have the most substantial impact on hospital ratings. In contrast, effectiveness and timeliness of care show weaker or statistically insignificant effects. Overall, the results imply that hospitals with better performance in mortality, readmission, and patient experience tend to receive higher overall ratings, highlighting these areas as critical drivers of hospital quality outcomes.

3.3 Proxy Question Analysis

The proxy question explored whether non-clinical factors like patient experience or timeliness could predict higher overall ratings even when clinical outcomes were similar. The results support this idea: hospitals with “Above the national average” patient experience scores tended to have notably higher overall ratings, even when mortality or readmission outcomes were only average. This finding suggests that patient perception plays an important role in shaping hospital evaluations and reputations. However, this relationship should be interpreted cautiously, as overall ratings may partially incorporate patient experience into their calculation, making it both a cause and a component of the rating.

---

4. Conclusion

The analysis of the CMS Hospital Quality dataset provided meaningful insights into how various performance dimensions affect hospitals’ overall ratings. The data revealed that most hospitals achieve moderate ratings, clustering around three and four stars. Patient experience and readmission measures emerged as the strongest predictors of overall performance, emphasizing the importance of patient-centered care and post-discharge management. Although mortality, timeliness, and safety contribute to hospital quality, their effects appear secondary compared to the patient experience domain.

From a managerial perspective, the results suggest that hospitals seeking to improve their overall ratings should invest in initiatives that enhance communication, responsiveness, and the discharge process to lower readmission rates. Despite these findings, limitations exist, such as the use of ordinal encodings and exclusion of contextual factors. Future research could explore more sophisticated models, such as ordinal logistic regression or hierarchical models incorporating hospital type, region, or case mix. Overall, this study demonstrates that analytical techniques in R can provide valuable insights for healthcare quality management and policy design.