**Portfolio Project: Business Intelligence Solution for HCA Healthcare**

Rachael Herman

Colorado State University Global

MIS480: Capstone: Business Analytics and Information Systems

Professor Justin Bateh

October 9, 2022

**Portfolio Project: Business Intelligence Solution for HCA Healthcare**

The final capstone project for the Management Information Systems and Business Analytics program allows students to demonstrate knowledge and understanding of program materials. After choosing and researching a U.S.-based organization, a problem is identified, researched, and analyzed using a public data set for a business intelligence solution that would work for this organization. The organization of focus for this project is HCA Healthcare. As the business heavily emphasizes improving patient care, the intelligence solution will focus on publicly available CMS patient experience survey data from 2016 through 2020 to analyze overall patient satisfaction in the U.S. The goal is to identify areas where HCA Healthcare can focus resources to gain more knowledge of internal patient experiences. This paper describes the business problem, the public data set used, the business intelligence tools utilized, the code and functions performed for statistical analysis, data visualizations and interpretations, challenges, and recommendations of how HCA Healthcare can use the information.

**HCA Healthcare Background and Rationale**

HCA Healthcare is a U.S.-based organization that opened in 1968 and now has more than 180 hospitals and 2,000+ ambulatory care facilities around the country. While they have a few locations in the U.K., HCA is primarily in the United States, with headquarters in Nashville, TN (Our history,n.d.). The company’s vision and mission heavily emphasize patient satisfaction, using the data from more than 35 million annual visits to develop new technologies and best practices that improve patient care (Who we are, n.d.). HCA Healthcare was chosen for its size and for-profit status and because of my background in healthcare as a medical transcriptionist and then content marketer for small medical offices. Further, a HIPAA privacy and security certification offers me a rich understanding of the data in this industry.

**Business Problem to Address**

Patient satisfaction is the extent to which patients and their caregivers are happy with the care they receive. Patient experience is directly related to revenue, with positive experiences linked to greater profitability (Richter & Muhlestein, 20017). HCA Healthcare does not have any patient experience data available for public use. However, they have public reviews that would show areas for improvement. This project focuses on evaluating data from similar facilities to help HCA Healthcare identify areas for improvement within its organization. By evaluating the patient ratings on specific topics, HCA Healthcare can formulate an internal patient survey and aggregate reviews to compare to the national ratings presented here and identify areas for improvement at its facilities.

**The Data Set and Access**

Understanding patient care requires evaluating many variables, such as facility mortality rates, timeliness, the safety of care, and overall patient experience. The data set for analysis comes from CMS hospitals data archive and includes U.S. patient satisfaction survey data from 2016 to 2020 (*PQDC,* n.d.). This data is readily available on the CMS website. The five CSV files were combined into a single file and are stored in GitHub, Google Drive, and Microsoft Power BI. The merged set had 43 variables and more than 2 million rows of data. Cleansing was performed in Power Query, removing irrelevant columns and rows with unavailable or unimportant data. For example, patient rating cells marked "not available" were deemed unimportant and removed. After cleansing, the data set had 20 variables and 132,174 observations.

**Business Intelligence Tools**

R is an open-source programming language and statistical analysis tool that offers significant insight into qualitative and quantitative data. Microsoft Power BI is a robust visualization tool for analyzing and creating dynamic, interactive reports and dashboards. Therefore, these two tools were used for this analysis. RStudio is the integrated development environment (IDE) for R and is used to perform statistical computations to examine the data. Microsoft Power BI translates that data into a visual, interactive report where users can gain insights into national patient satisfaction data at proprietary (for-profit) facilities.

**Code and Functions Used to Evaluate the Data**

In RStudio, several packages were used for statistical computations and modeling. Once these packages were installed, the merged data file was read into the program and placed into a data frame object. The str() function gives an overarching view of the data in the file and helps determine the best modeling functions to utilize. This function shows a list of variables, the data types, and the first few entries for each variable.

Once the file was imported and the seed set, the topic was changed to a factor to help read in average ratings. The goal is to analyze average hospital and patient ratings at proprietary facilities to determine which was most significant. An exposition function, as well as the filter(), group\_by(), and summarise() functions found the average patient and hospital ratings by topic. The ungroup() function ensured that the grouping did not affect the rest of the analysis.

As patient ratings had variability, a scatter plot using the ggplot(), geom\_point(), geom\_abline(), and theme() functions provided a better look at the variability in the patient ratings per topic. Simple linear regression models using the lm() function found significant coefficients for general hospital and patient ratings based on the topics. A contingency table for patient ratings used the table() function and pulled in patient ratings for each topic, separated by proprietary facilities and all other ownership types. This table gives an idea of how many patient reviews (1-5) are given per topic and ensured an adequate sample size for proprietary facilities.

**Statistical Analysis Performed in RStudio**

The figures below demonstrate the code and functions used for analysis. Explanations and interpretations follow.

**Figure 1**

*RStudio str() Function to Examine Variables and Convert Topic to Factor*

Graphical user interface, text, application, email

Description automatically generated

**Figure 2**

*Average Patient and Hospital Ratings at Proprietary Hospitals*

A picture containing text

Description automatically generated

**Figure 3**

*Scatter Plot to Look at Average Patient Ratings by Topic*

Chart, scatter chart

Description automatically generated

**Figure 4**

*Linear Regression Model for Hospital Rating Predicted by Topic*

Text

Description automatically generated with medium confidence

**Figure 5**

*Linear Regression Model for Patient Rating Predicted by Topic*

Table

Description automatically generated with medium confidence

**Figure 6**

*Contingency Tables for Patient Ratings Per Topic and Ownership Type*

1. *Table for All Ownership Types*

Table

Description automatically generated with medium confidence

1. *Table for Proprietary vs. All Other Ownership Types*

Table

Description automatically generated

**Interpretation of Statistical Analysis**

The analysis reveals that hospital ratings are relatively the same across all topics and will not provide much insight into possible improvements. This is demonstrated in the average summary in Figure 2 and the linear regression model shown in Figure 4. As patient ratings are the more important variable in understanding patient satisfaction, much of the analysis focuses on that variable rather than the overall hospital ratings. However, it is worth noting that “overall hospital rating” is a survey topic. As HCA Healthcare is a proprietary (for-profit) facility, the data is filtered to focus on this type of facility.

The content structure was examined to ensure the correct data types were used for essential variables, as seen in Figure 1. The average summaries per topic are shown in Figure 2 and demonstrate ratings for all five years in proprietary facilities. The data shows some variation between patient ratings per topic but not hospital ratings. Figure 3 demonstrates a weak correlation between topics in the scatter plot for patient ratings, with ratings scattered far apart and a relatively straight regression line.

Figures 4 and 5 show linear regression models for hospital and patient ratings as predicted by the topics. In Figure 4, none of the topics are predictors of hospital ratings. Figure 5 identifies several topics as predictors of patient rating. The R-squared and adjusted R-squared values are low, indicating weak predictors (Frost, 2017). However, the higher t-values and F-statistic in the linear model for patient ratings indicate a good fit model with higher confidence. The contingency tables in Figure 6(a-b) show how many patient reviews (1-5) are on each topic. Figure 6a is for all ownership types across all five years, and Figure 5b is for proprietary facilities across all five years vs. all other ownership types combined. The majority of ratings are three stars. However, there are many values under the 2 rating at proprietary facilities.

**Data Visualization in Power BI**

Power BI is used for data visualization in the form of a report. This is an excellent way to visualize data and help measure KPIs (Laursen & Thorlund, 2016). The Power BI report is interactive, so users can drill down into the data to get more insights. No special functions or codes were used to create the visualizations. Filters were utilized to help drill down into the data.

**Figure 7**

*Power BI Report Visualization*

Chart, treemap chart

Description automatically generated

**Figure 8**

*Drill Down Example of Lowest Rating Topic in Treemap*

Chart, treemap chart

Description automatically generated

**Figure 9**

*Drill Down Example of State Ratings by Topic*

Chart, treemap chart

Description automatically generated

**Interpretation of Data Visualizations**

Statistical analysis showed that hospital ratings were uniform across all survey topics. Therefore, they were not a big part of the visualizations; however, they are included as a card for reference. The report is focused on proprietary facilities similar to HCA Healthcare. Figure 7 shows the entirety of the report. The treemap is a telling picture of the data with average ratings per topic, as is the table. In Figure 8, the topic with the lowest rating, cleanliness, is chosen for analysis. The average patient rating year over year, the number of surveys and the average overall hospital rating are shown. To dig further, one could choose a state from the map to see how that state’s for-profit facilities fare in each topic, as seen in Figure 9.

The report allows users to choose views based on year, topic, or state. The drill-down opportunities offer in-depth insight into the selected items. The goal is to identify areas where patients have had the most unsatisfactory experiences in facilities similar to HCA Healthcare. The lowest patient ratings were cleanliness, care transition, and doctor communication. The states with the lightest gray have the lowest ratings. Pain management was only surveyed for two years, which needs to be considered when looking at the total ratings. HCA Healthcare would benefit from targeting the states with the lowest ratings in the topics that showed the most need for improvement. Tailoring surveys to these states on those topics could help identify critical need areas.

**Figure 10**

*GitHub Repository Showing Uploads*

Graphical user interface, text, application, email

Description automatically generated

**Challenges and Recommendations**

The biggest challenge was combining the files and cleaning the data set, which had more than 2 million rows of data and 43 variables. The cleaning and organization of the data took many hours to complete, along with some trial and error. Another challenge was in deciding which software to use. I started with SAS, Python, and Tableau. However, these programs proved challenging for my purposes, so I opted instead to use RStudio and Microsoft Power BI.

HCA Healthcare’s business intelligence solution centers around patient experience. The patient’s ratings on specific topics are critical to understanding where facilities should focus resources for improvement. Per the data above, these areas for improvement include cleanliness, doctor communication, and care transitions. HCA Healthcare can tailor internal surveys and target states with the lowest ratings to get the best insights.

**References**

Frost, J. (2017, April 16). *How to interpret R-squared in regression analysis.* Statistics by Jim. Retrieved September 27, 2022, from <https://statisticsbyjim.com/regression/interpret-r-squared-regression/>

Laursen, G.H. N., & Thorlund, J. (2016). *Business analytics for managers: Taking business intelligence beyond reporting* (2nd ed.). Wiley Professional Development (P&T). <https://mbsdirect.vitalsource.com/books/9781119302537>

Our history. (n.d.). *HCA Healthcare.* Retrieved September 17, 2022, from <https://hcahealthcare.com/about/our-history.dot>

*PQDC.* (n.d.). Retrieved September 8, 2022, from <https://data.cms.gov/provider-data/archived-data/hospitals>

Richter, J.P. & Muhlestein, D.B. (2017). Patient experience and hospital profitability: Is there a link? *Health Care Management Review, 42(*3), 247-257. doi: 10.1097/HMR.0000000000000105.

Who we are. (n.d.) *HCA Healthcare.* <https://hcahealthcare.com/about/>