Analyzing and Predicting the hospitalization Costs using Data Science - DSC520_FinalProject_Part1

Guruprasad Velikadu Krishnamoorthy

2023-02-12

Contents

1 Introduction

Problem Statement: With the tremendous advancement that has been made in technology in the last few decades, things that used to be considered luxuries such as smartphones or 50-inch television are not very expensive anymore. However, even with all the technological progress one thing that has not changed much is the cost of medical treatment. In fact, many medical procedures have gotten so expensive over the years that some are not affordable, even after having Health Insurance and patients have to go for external sources such as GoFundMe to collect funds. When the doctor suggests a medical treatment and when we call the hospital to ask how much the treatment might cost, there is a good chance the hospital may not be willing to tell the actual amount, considering how complicated the health care system in the US is. Many people get sucked into this mess without completely understanding all the nuances of the healthcare system and end up with a massive bill that they may not be able to pay. An article from CNBC states that two-thirds of all bankruptcies in the US are tied to issues with Medical Bills. With such a huge problem in hand, there is no better solution to get an idea of how much an essential but non-emergency procedure would cost in a hospital. (Konish n.d.)

Why would someone be interested in the research? Medical bills can be very expensive, especially if you don't have insurance or if you get treatment from a hospital that is out of the health insurance network. Most hospitals are not transparent about sharing the treatment costs and they redirect to reach out to health insurance instead. However, not many people know that medical bills can be negotiated with hospitals. If we have the right stats and numbers, the hospitals may be willing to negotiate the bills and it can save thousands of dollars and sometimes can even save bankruptcies for some families. This research focuses on studying the pattern behind the medical costs across various states and cities in the US and identifying a pattern in the costs. Patients can use the data to choose care wisely and avoid expensive hospitals if they have a choice. Also, these numbers can help negotiate the bills with the hospitals.

How Data Science can help with the research? Data science can be used to study patterns and derive statistics and compare prices over a period. Also, data from different datasets can be combined to derive the relationship between the variables and a model can be built to predict the prices in the coming years.

2 Research Questions

1. What is the average cost for treatment, say, for example, Chemotherapy, in each state across the US?
2. What is the total number of inpatient stays across the states for a specific medical procedure during a

given period?

3. How much has the average cost of a medical procedure (for Chemotherapy, as an example) varied over the years across a state, Texas for instance?
4. What is the average price of Chemotherapy treatment across various cities in Texas?
5. Given a zip code in Texas, if multiple hospitals offer the same treatment for Chemotherapy, how different are the costs?
6. Are the hospitals operating for Profit or Non-Profit? Is there a substantial difference in the costs for the same medical procedures between the For-Profit and Non-Profit hospitals? By what percentage?
7. How many hospital beds are available across a city or a state? What are their inpatient occupancy rates? Do "For-Profit" hospitals provide more beds in the state than "Non-Profit"? By how much percent?
8. Is there insight into the Profit margin and other financials of the "for-profit" and "non-profit" hospitals for a given period? How profitable are the "For-Profit" hospitals?
9. Is there any correlation between the number of uninsured people across the state/ region to the average cost of medical treatments across the region?
10. Is there a pattern in the variation in the cost of a medical procedure over a period? Can it predict the prices of treatment in the future?

3 Approach

The problem that this paper address is to study the disparity in hospitalization costs across the country. Though the costs may differ from state to state, many times they are so much different even within the same city or county. Also, hospitals are not transparent about sharing the medical costs, so it can be very frustrating to undergo a medical procedure not knowing how much it costs and if the hospital is overcharging for the procedure or not. The solution addresses this problem by studying the trends in medical costs (for a medical procedure as an example) across the state or region and over a period and provides an overall idea about the differences in the cost. The results will be plotted in graphs and plots for visualization. The study also provides results of costs that are grouped based on a state or city or a region for a medical procedure. This can also be used by the patients to compare the costs with the explanation of benefits to getting a picture of whether the costs look normal or inflated.

4 How the approach addresses (fully or partially) the problem

The approach includes importing the data from different periods and comparing the trends in the costs of medical treatment. We use the average costs in the state as a benchmark to compare the costs and identify how the hospital's costs compare with others in the same region. We study the occupancy rates of hospitals and compare them with the state's average and discover the pattern between inpatient occupancy and their costs. Also, the Tables and plots in R will be used to visualize the results for easy understanding. Also, the costs of treatments may vary between counties and if the difference is substantial, one can choose to undergo treatment from a cheaper hospital from a neighboring county or a neighboring state in some cases.

This project may partially address the problem statement of identifying the affordable hospitals in the region as the study does not include the health insurance data in the research. Also this study uses Medicare data for research and sometimes hospitals may charge a different price for treatments than the medicare costs. Hence the study from the research can be used a baseline for further research. The rates are compared across hospitals with the assumption that they are all within the same network, which may not be true in all cases. This project is to identify the pattern of how hospitals charge different rates for the same treatments in the same region and to predict costs in the future. Also, visualization of the data using tools such as Tableau or Power BI can be a great enhancement for this research which is not included in this project. Visualization can add more features such as the ability to select multiple hospitals and select procedures in the hospitals and compare the costs.

Below is the high-level approach for addressing the research project:

Identifying the datasets : There are a variety of sources to find the datasets such as www.data.cms.gov, www.cdc.gov, www.who.int, www.eia.gov, Kaggle, etc. However, the real datasets on hospitals and the costs of procedures may be a challenge to find. So, the first step is to invest time to find the right dataset.

Exploring the dataset : Once the datasets are found, the second step is to study the data and understand each field in the dataset. We may not find all the relevant data in the same dataset. So, we may have to identify more than one dataset and join them to extract meaningful insights. Hence understanding the data is very important.

Draft a high-level plan : The next step is to come up with a high-level plan (algorithm) for the approach to address the problem. This step will require identifying the fields that we are planning to join across different datasets and analyzing the number of records in the original data to see if a subset of data is required for analysis.

Identifying the technical requirements : The next step is technical where we identify the list of R packages and plots required to address the problem. In this step, we will perform a very high-level analysis of the technical requirements as more packages can be added later during the development stage. Further steps of cleansing data and transforming the data will be done during the development phase.

Development and exploration : After the steps mentioned in the above question are complete, the next steps involve loading the data into R and cleansing and transforming the data. Depending upon the size of the dataset, a subset of data can be used for the research. In this stage, we can identify new fields required to be used in the plots to join the datasets.

5 Datasets

Seven datasets are used in this research and the details are given below.

i. **MUP_IHP_RY21_P02_V10_DY18_PrvSvc.csv**
ii. **MUP_IHP_RY21_P02_V10_DY19_PrvSvc.csv**
iii. **MUP_IHP_RY22_P02_V10_DY20_PrvSvc.csv**

("Medicare Inpatient Hospitals - by Provider and Service" 2022)

iv. **MUP_IHP_RY21_P02_V10_Dy18_Geo.csv**
v. **MUP_IHP_RY21_P02_V10_Dy19_Geo.csv**
vi. **MUP_IHP_RY22_P02_V10_Dy20_Geo.csv**

("Medicare Inpatient Hospitals - by Geography and Service" 2022)

- ***Description:*** The Datasets 1,2 and 3 contain data about the use, hospital charges, and payments for more than 3000 US hospitals that received IPPS payments (Inpatient Prospective Payment System). Each dataset is for the years 2020, 2019, and 2018 respectively. The Datasets 4,5, and 6 are like the ones above except the data is summarized at state and nation for each category of the medical procedure. The 3 datasets are for the years 2020, 2019, and 2018 respectively.
- ***Description of the fields:***
  - ***Rndrng_Prvdr_CCN*** : This describes the Rendering provider's CMS certification number, which is a unique Identifier to identify a hospital.

- ***Rndrng_Prvdr_Org_Name*** : Denotes the Rendering Provider organization name which is the name of the hospital.
- ***Rndrng_Prvdr_St, Rndrng_Prvdr_City, Rndrng_Prvdr_State_Abrvtn, and Rndrng_Prvdr_Zip5*** : These fields indicate the street address, City name, State name, and the Zip codes of the hospital.
- ***DRG_Cd***: Classification code for the Diagnosis Related Group that indicates unique code for each type of medical procedure.
- ***DRG_Desc*** : Description of the Diagnosis Related Group code. This indicates the description of the medical procedure.
- ***Tot_Dschrgs*** : This indicates the total number of all discharges billed by all the providers.
- ***Avg_Tot_Pymt_Amt*** : This indicates the average total payment made to all providers which include the copays, deductibles, etc.
- ***Rndrng_Prvdr_Geo_Lvl*** : This indicates if the data is summarized at the National level or State level.
- ***Rndrng_Prvdr_Geo_Desc*** : This indicates the name of the state.

- **Observations about the data**
  - The data in these datasets do not have any empty or missing values.
  - However, to distinguish the data in the datset from each other a new column for Year should be added.
  - The Filed CCN has some values that are of length 5 and some are of length 6. We may need to standardize the length by adding a leading 0 to join the datasets that are used in this project.
  - Some of the datasets are huge with up to 200,000 rows and have up to 15 columns each, though only the columns described above will be utilized in this project.

- ***Period when data was collected:*** Dec, 2022

vii. **NASHP_HCT_Data_2022_November.csv**

("Hospital Provider Cost Report" 2022)

- ***Description:*** This contains the Healthcare Coset report details collected by the NASHP team (National Academy for State Health Policy) in collaboration with rice university. It contains the Healthcare cost report details such as hospital details, Bed size, Occupancy Rates, and lots of financials such as

net revenue, net income, etc. It contains data for approximately 5100 hospitals across the US for the period between 2010 to 2021.

- ***Description of the fields:***
  - ***CCN#*** : This indicates the Unique identifier for each hospital (CMS Certification Number). This denotes a Medicare Provider Number.
  - ***Facility Type*** : This indicates if the hospital is a Critical Access Hospital or a Short-Term Facility.
  - ***Year*** : Indicates the reporting year.
  - ***Hospital Name, Hospital Abbreviated Name, Address, Zip Code, City, State*** : Indicates the name and address of the hospital facility.
  - ***Hospital Ownership Type*** : This indicates the classification of the hospital if it is "For-Profit" or "Non-Profit" or "Governmental".
  - ***Bed Size*** : Indicates the number of beds available for Inpatient care and does not include the beds for emergency rooms or Outpatient care.
  - ***Inpatient Occupancy*** : Percentage of beds occupied by inpatient care during the given year.
  - ***Net Income (Loss) and Net Profit Margin*** : Indicates the financial metrics that will be used in this project.
- **Observations about the data** :
  - The dataset has 85 fields and 51894 rows of data, though in this project about 12 fields will be used.
  - This dataset requires quite a bit of cleansing. The field "Hospital Name" and "Hospital Abbreviated name" has invalid numeric values that should be excluded.
  - The Address field and City have invalid values such as ".", "X"," Y", etc. should be excluded.
  - The field's Bed Size, Occupancy rates, Net Revenue, and Net Profit margin all have invalid values such as "-99999" and "." Should be excluded or defaulted to 0.
- ***Period when data was collected:*** 2022

6 Required Packages

- ***ggplot2:*** This is one of the most important packages used for data visualization and is used to create Aesthetic visualization plots.

- ***lubridate:*** The dataset contains date fields and hence lubridate is used to extract individual date components if needed.
- ***dplyr and plyr:*** This is used for various type of data manipulation.
- ***stringr:*** This is used for String Manipulations.
- ***tidyr:*** This is used to tidy up the data.
- ***purrr:*** This is used to apply the function to a list and also to use the map functions.
- ***reshape2:*** This package is used for Data munging to convert the rows to columns or for data Melting operations.
- ***Magrittr :*** This package will be used to utilize the pipe functionality
- ***DT:*** This is for an interactive tabular experience through DataTables JavaScript library
- ***rmarkdown and knitr:*** These are required to knit the documents as pdf using RMarkdown.

7 Plots and Table needs

- ***Plots***
  - *Histograms*: This is to show the distribution of values for a variable.
  - *Scatterplots*: This is to show two variables in comparison with each other
  - *Boxplots*: This is to identify the outliers in the data.
  - *Density plots*: This is to show the density distribution
  - *Facet_wrap*: This takes data in one variable and splits data in it and creates separate panes for each set to fit them in the plot.
  - *Facet_grid*: This is like facet wrap but it applies all levels of a variable to a row and column and fits them in a plot.
  - *Violin Plot* : These are similar to the box plots but can explain the density of the data.
  - *line plot*: Line plot to show the relation between two variables.
  - *qq-plot* : These can be used to explain how close is the data to Normal distribution.
  - *Heat Map* : Heat map can be used to explain the density in the pattern of data based on the Heat color schemes.

- ***Tables:***
  - *kable* : Tabular data can be printed using kable
  - *Datatable*: From DT package

8 Questions for Future Steps

This study also involves studying the pattern of the data and creating a regression model to predict the medical costs. So far in this course we have learnt about the Linear regression models. Hence logistic regression and machine learning models should be studied to see if the Machine learning models can be useful for the Project. Also, there are many geom functions in ggplot that I have not used and many unknown packages which have to researched to see if any of them can be useful for the project.

References

"Hospital Provider Cost Report." 2022. Centers for Medicare & Medicaid Services. https://catalog.data.gov/dataset/hospital-provider-cost-report.

Konish, Lorie. n.d. "This Is the Real Reason Most Americans File for Bankruptcy." *CNBC*. Accessed February 11, 2023. https://www.cnbc.com/2019/02/11/this-is-the-real-reason-most-americans-file-for-bankruptcy.html.

"Medicare Inpatient Hospitals - by Geography and Service." 2022. Centers for Medicare & Medicaid Services. https://catalog.data.gov/dataset/medicare-inpatient-hospitals-by-geography-and-service-1bbc8.

"Medicare Inpatient Hospitals - by Provider and Service." 2022. Centers for Medicare & Medicaid Services. https://catalog.data.gov/dataset/medicare-inpatient-hospitals-by-provider-and-service-9af02.