Cost Analysis for Inpatient Prospective Payment System

DA 5020 – Collecting, Storing and Retrieving Data

NORTHEASTERN UNIVERSITY

FALL 2017

Supervised By

Kathleen Durant

YALIM DEMIRKESEN

Contents

1 BUSINESS UNDERSTANDING

In the term project, I am analyzing a healthcare data on data.gov. The data is related to the health costs in the US states. It gives charges that are created by more than 3000 hospitals all around US that receive Medicare Inpatient Prospective Payment System (IPPS). The costs are related to top 100 most frequently billed discharges, which are paid under Medicare. With the provided data, I will be able to analyze the charges and their correlation with states.

This data set will make it possible for me to analyze the healthcare charges from each state. Since it is also a political issue, the healthcare costs are a popular subject. Throughout this research, the readers will have a chance to examine each hospital that is enrolled to the Inpatient Prospective Payment System. The data belongs to each and every contributor to IPPS from each city and each state. So, it gives a nice locational and updated information. Depending on the records, I will be able to provide answer to questions that are related to cost of each hospital and also their discharge numbers and how much they benefit from Medicare funds.

2 DATA UNDERSTANDING

The source of the data is Centers for Medicare & Medicaid Services' (CMS) inpatient database. The dataset that we are interested in is taken from October 1st, 2011 to September 30th, 2012. The database includes IPPS short term care, long term care, critical access hospital, non-medical, rehabilitation and psychiatric discharges.

There are in total 12 columns. Descriptions of each column can be found below:

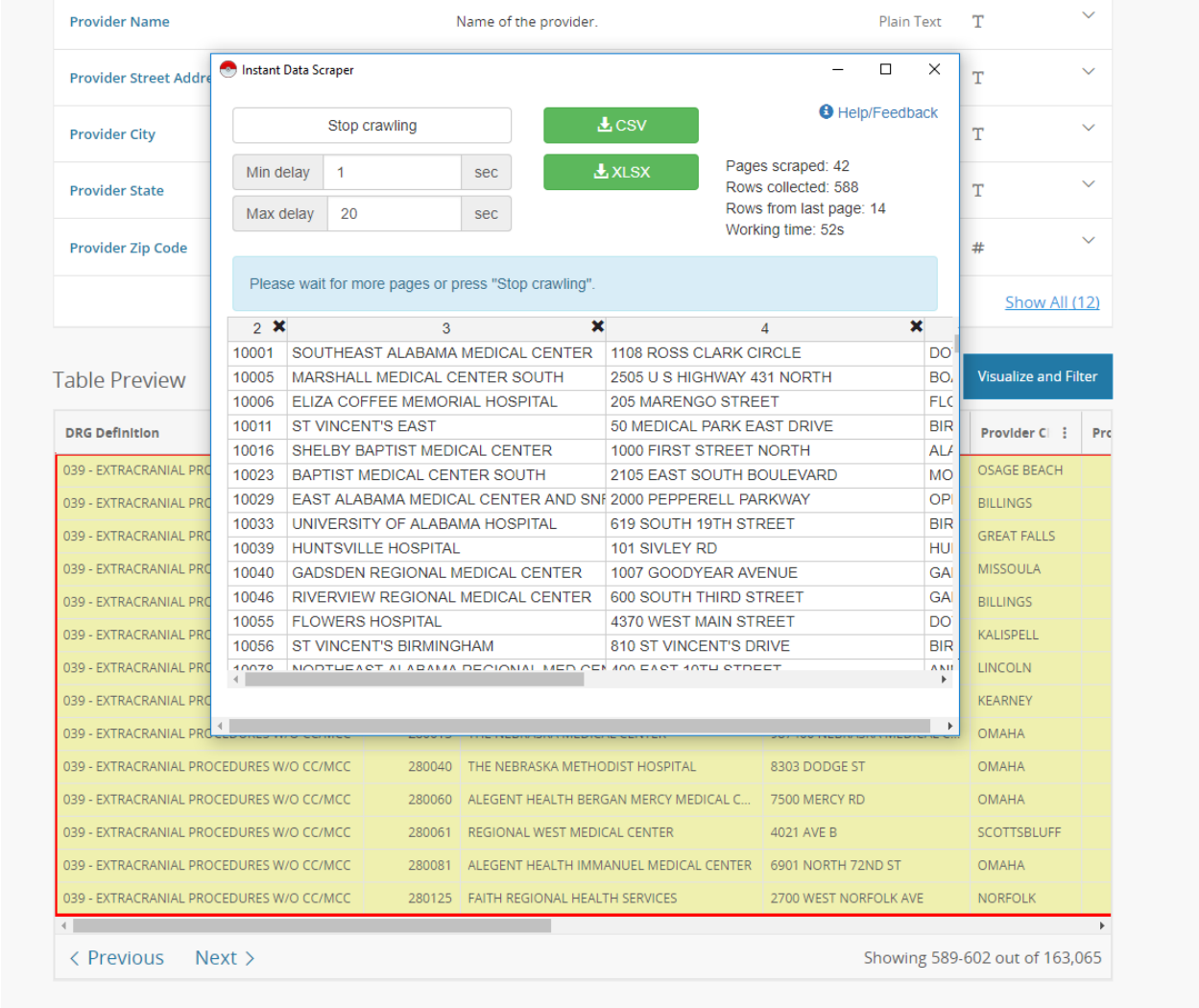
1. **DRG Definition**: The definition of the discharge that is stored in the CMS database about inpatient discharges
2. **Provider ID**: Number that identifies the provider, which is also qualified for Medicare Certification

3. **Provider Name**: Hospital facility name
4. **Provider Street Address**: Hospital facility address
5. **Provider City**: City where the hospital facility is located at
6. **Provider State**: State where the hospital facility is located at
7. **Provider Zip Code**: The hospital facility's zip code
8. **Provider HRR**: The Hospital Referral Region (HRR) where the provider is located.
9. **Total Discharges**: The number of discharges billed by the provider for inpatient hospital services
10. **Average Covered Charges**: The provider's average charge for services covered by Medicare for all discharges. There might be variations in this data because of the different hospital policies
11. **Average Total Payments**: The average total payments to all providers for all cases. Also included in average total payments are co-payment and deductible amounts that the patient is responsible for and any additional payments by third parties for coordination of benefits.
12. **Average Medicare Payments**: The average amount that Medicare pays to the provider for Medicare's share. Medicare payments don't include beneficiary co-payments and deductible amounts nor any additional payments from third parties for coordination of benefits.

3 DATA PREPARATION

3.1 Data Scraping

The first step of analysis is scraping the data from its source. For that purpose, I used the *Instant Data Scraper*. From the webpage that is indicated in the References, I scraped the data. It was a long process because of the amount of the rows that is in the data. Since there is an enormous number of rows in the records, I left the data scraper open more than several hours. In the data scraping process, I needed to indicate the correct table that needs to be extracted from the webpage. After scraping process, I downloaded all the scraped data to a csv file and it was ready to be uploaded to R. Below you can find the screenshots of data scraper on data.gov.



2	3	4	
10001	SOUTHEAST ALABAMA MEDICAL CENTER	1108 ROSS CLARK CIRCLE	DO
10005	MARSHALL MEDICAL CENTER SOUTH	2505 U S HIGHWAY 431 NORTH	BO
10006	ELIZA COFFEE MEMORIAL HOSPITAL	205 MARENGO STREET	FLO
10011	ST VINCENT'S EAST	50 MEDICAL PARK EAST DRIVE	BIR
10016	SHELBY BAPTIST MEDICAL CENTER	1000 FIRST STREET NORTH	AL
10023	BAPTIST MEDICAL CENTER SOUTH	2105 EAST SOUTH BOULEVARD	MO
10029	EAST ALABAMA MEDICAL CENTER AND SNF	2000 PEPPERELL PARKWAY	OP
10033	UNIVERSITY OF ALABAMA HOSPITAL	619 SOUTH 19TH STREET	BIR
10039	HUNTSVILLE HOSPITAL	101 SIVLEY RD	HU
10040	GADSDEN REGIONAL MEDICAL CENTER	1007 GOODYEAR AVENUE	GA
10046	RIVERVIEW REGIONAL MEDICAL CENTER	600 SOUTH THIRD STREET	GA
10055	FLOWERS HOSPITAL	4370 WEST MAIN STREET	DO
10056	ST VINCENT'S BIRMINGHAM	810 ST VINCENT'S DRIVE	BIR

DRG Definition				
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	280040	THE NEBRASKA METHODIST HOSPITAL	8303 DODGE ST	OMAHA
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	280060	ALEGENT HEALTH BERGAN MERCY MEDICAL C...	7500 MERCY RD	OMAHA
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	280061	REGIONAL WEST MEDICAL CENTER	4021 AVE B	SCOTTSBLUFF
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	280081	ALEGENT HEALTH IMMANUEL MEDICAL CENTER	6901 NORTH 72ND ST	OMAHA
039 - EXTRACRANIAL PROCEDURES W/O CC/MCC	280125	FAITH REGIONAL HEALTH SERVICES	2700 WEST NORFOLK AVE	NORFOLK

After picking the correct table, I located the “Next” button for data scraper to understand how to switch to the next page. Then it started gathering data from the source. An advantage of *Instant Data Scraper* was that it allowed me to download the data as csv file.

3.2 Preparing for MongoDB

After uploading the data to R, I started to check data whether there are missing values. I wasn’t expecting any missing values but just to be sure, I checked the data and created a function that counts the missing values in each column. The list suggested that there are no missing values in the dataset.

The second step was preparing data for mongoDB. For this purpose, two datasets are created. One included all the provider information including their id numbers, names, street addresses, cities, states, zip codes and region description. Second dataset was about related to discharges. Discharge definition, provider id number, provider name, total charges, average covered charges, average total payments and average Medicare payments are attached to the second dataset.

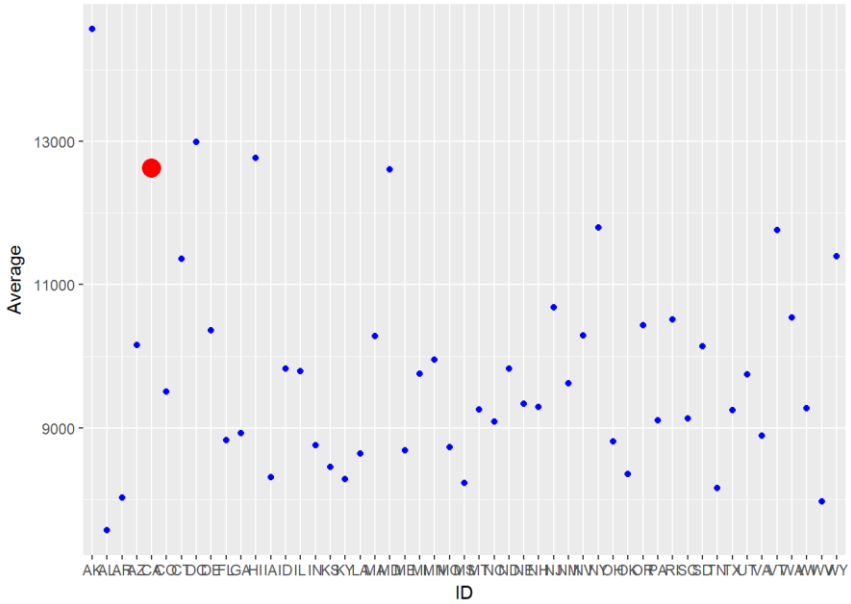
3.3 Inserting Data to MongoDB

First step here is data insertion to mongoDB. That can take place once the connection with the mongoDB is started. Using the mongo function, I created three different databases and three different tables. First two are the explained ones and as the third, I inserted my whole dataset so that I can use it for the queries that involves both numerical values and string information.

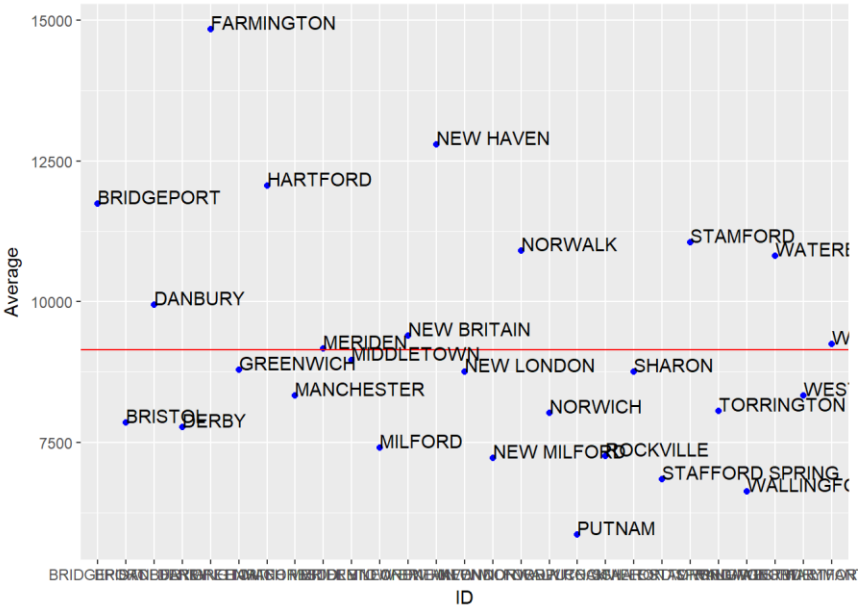
4 RESULT OF QUERIES

Below the results of the queries can be found:

- Number of Hospitals in Massachusetts: For this query, I generated count function from mongoDB library. As a result, I ended up with the solution of 240.
- The State with Most Hospitals: To find out the state with most healthcare locations, I grouped the data considering the state names and counted the values. As a result, R suggested that Texas is the state with the most number of hospitals. There are 1240 hospitals. California follows Texas with 1196 and then comes Florida with 664.
- Average Coverage Charges per State: Again, after grouping the data, I ended up with the average coverage charges per state. CA is the first state which is followed by New Jersey and Nevada.
- Average Total Payments: Visualizing this with a scatter plot is much easier and more effective.

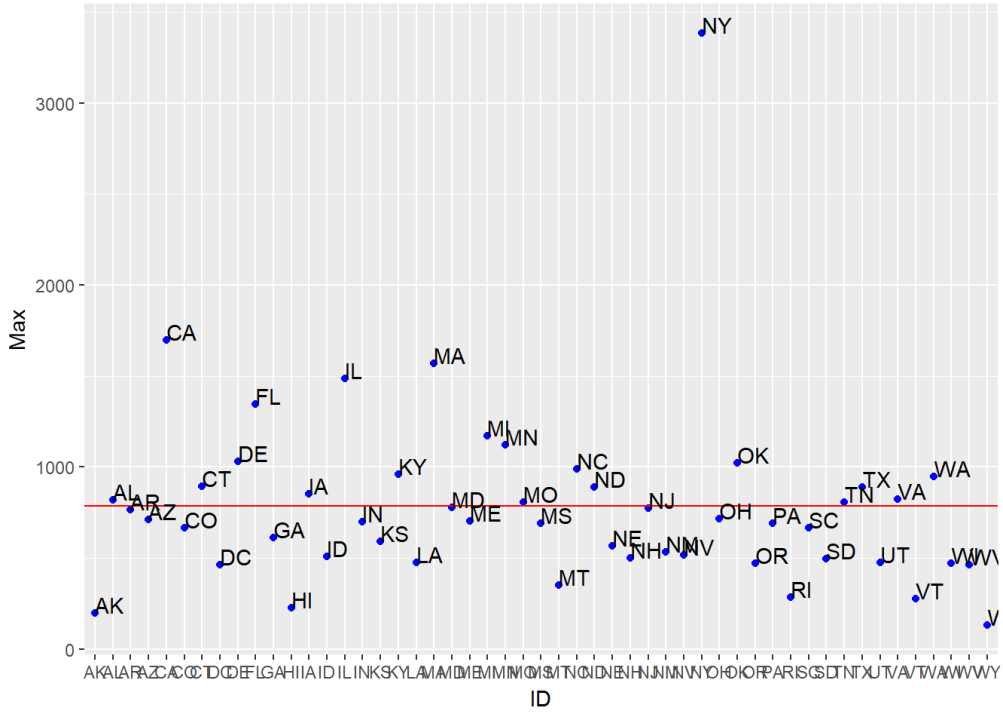


This graph suggests that CA, is one of the states with the highest average total payments, whereas Alaska is the state with lowest average total payments. Then I created a linear regression model to see what contributes to the total average payment. As a result, it was total discharges, average covered charges and average Medicare Payments.



- Average Medicare Payments in Connecticut: Above graph describes this query better. Farmington is the city with the highest Medicare payments, whereas the Putnam is the lowest.

- The Hospital with Most Discharges: The answer to this question is Hospital for Special Surgery which is in New York. It seems like an outlier but I didn't give a chance of having a wrong entry in the governmental dataset. That made me think that it might just be an outlier. Also, what is meant by special surgery is not clear. They might be small surgeries or daily operations.



When we make the same analysis also only for the hospitals, we get again that exact hospital as the outlier with a much higher y value that is much bigger than the other hospitals around.

5 CONCLUSION

As a solution, I benefited from mongoDB a lot. It helped to develop the models much easier than they would be performed by a regular R command. With the help of MongoDB , I generated several tables that helped me with the creation of relational database. With the aggregate, count and find functions I utilized the dataset. On the other hand I would be able to have the same outcome with more lines of R codes.

6 REFERENCES

- https://data.cms.gov/Medicare-Inpatient/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3
- https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Inpatient2011.html