Healthcare Data Application Process Book

Kirsten Schulz, Pamela Peng, Trevor Levey

STOR 320 Final Project

TABLE OF CONTENTS

Overview and motivation 2

Related work 3

Initial Questions 3

Exploratory Data analysis 4

Final analysis 6

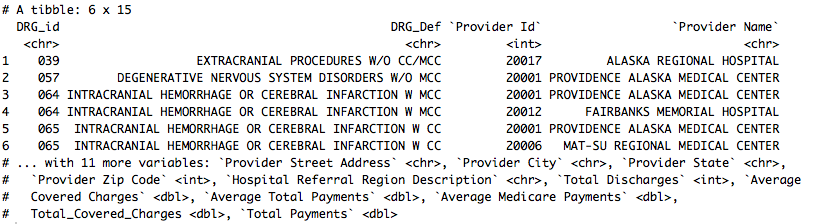
presentation 7

Overview and Motivation

Our project aimed to make use of R’s shiny package to make a data analysis application. In our search for a data set, we looked for something that could be tidied and aggregated into several groups that could be interactively selected and analyzed within the app. The data set we chose, from data.gov, examines average health care costs and amount of this cost that is covered for different diseases, healthcare providers, cities, and states.

Using this data, we hoped to make an app that could help a user find statistically significant differences in healthcare coverage between states or regions, and also explore and analyze the differences between the costs of different treatments. Our initial hope was that the user could select any state or region and view aggregated averages for these metrics (costs and coverage amount) for whatever they selected. This plan would allow this data to be easily presented and used in an argument by a non-technical user requiring statistical analysis for some argument or proposal concerning healthcare and health insurance.

The data structure of the tided used for the app development is summarized here:



Related Work

Our group was influenced by the presentation of the shiny package by the guest lecturer, Dr. Frances Tong. This lecture showed how intuitive and powerful shiny could be to create a variety of applications for data analysis. Because of this, our group was in complete agreement that we would find a data set that would fit well in a shiny application.

Initial Questions

As our project plan was geared towards an application, our initial questions were not based on any single combination of states, regions, or disease costs, but rather on what our application could help a user to answer. We aimed to get as much information as we could from the data, just from using our application. As such, we thought over the different aggregations we could apply to the data (regional, state, local, healthcare provider, or disease), and figured out useful questions for each:

1. How is our data set organized and what information does it include?
2. What is the average, lowest and highest inpatient medical cost in each state and region?
3. How the cost of inpatient medical care varies across states and regions?
4. What is the relationship between Average Covered Charges and Average Total Payments across states and regions?
5. Are there certain patterns about costs in big/small cities (urban-rural difference) within the same state?
6. Are there any relationships between the hospital size (by the number of hospital’s total discharges) and the inpatient medical costs?

The data organization question led to slight tidying of the data. We used regular expressions to remove dollar signs from the money values, and then made them integers. We also split the disease column into names and disease ID’s. The first few questions, after the data organization question, dealt with healthcare costs and coverage for each state and region. The fourth question looked at more specific analysis possibilities for our app, grouping by local regions and cities. The fifth looks at a possible analysis based on healthcare provider aggregation.

As we began to develop the app, we realized some of these questions would not be practical, and some questions required rethinking. Through exploratory analysis, we found that aggregating by city would have made the app more cluttered and the analysis would have been limited anyway. We thought about other ways to represent this aggregation, but in the end, decided the application would not support that scope. For number 5, we found that this would not work well in the app, because we found no notable results from this branch of analysis. The questions requiring averages to answer required some extra thought. Whenever we calculated an aggregated average, we realized that since the data was already aggregated, we would have to add weight the groups by size. We took this step repeatedly throughout the project.

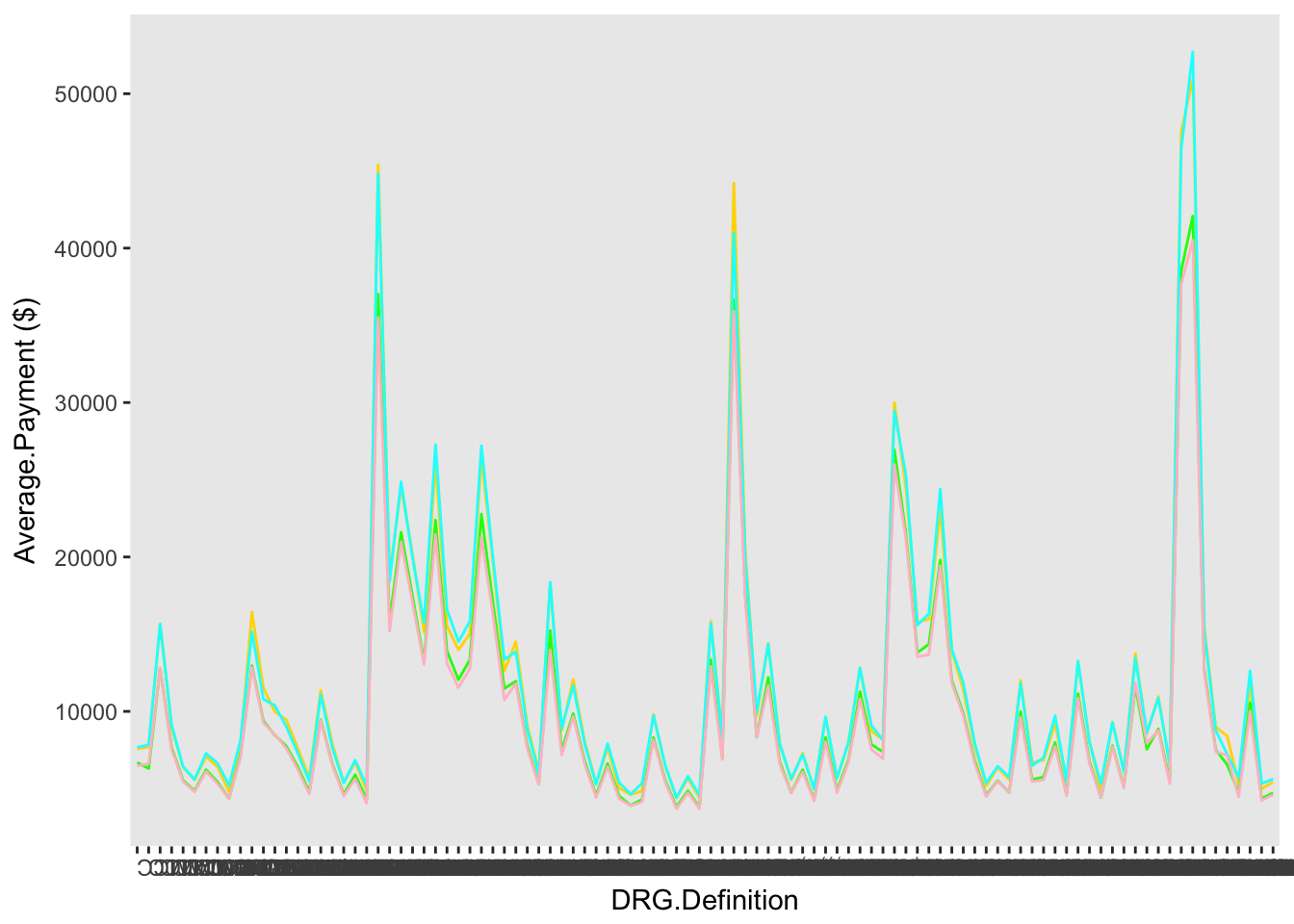
Exploratory Data Analysis

Please let me know if I left out anything major here

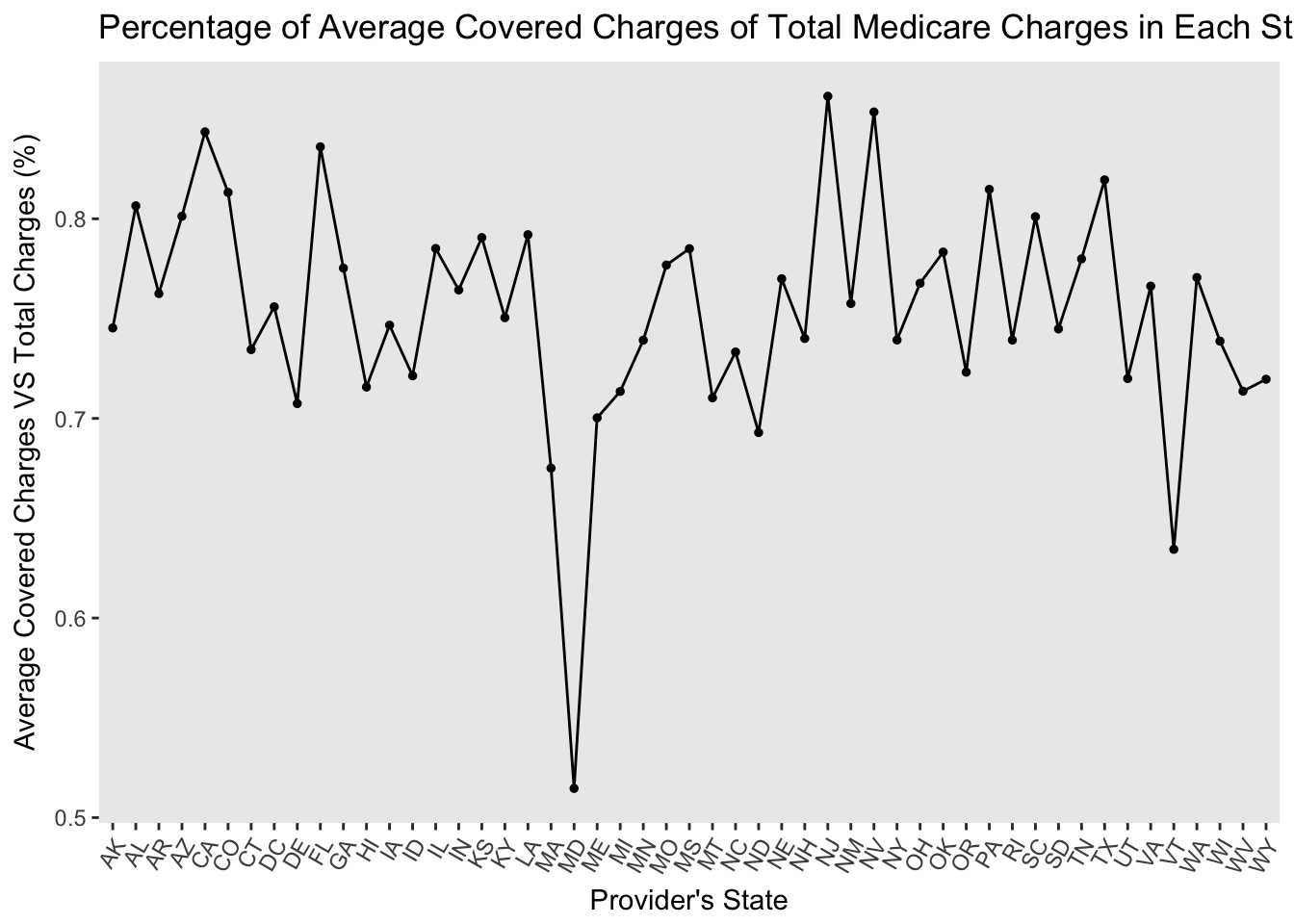
As our project was to create an application to be used for a variety of analyses, rather than to create a specific analysis, our exploratory data analysis was mainly creating several plots to see what would or wouldn’t work in our final application. We had an idea of what we wanted from the initial proposal, so we tried to turn our questions into usable plots. Missing from our initial exploration was a sample t-test for the difference between two means to be run on percentage covered or average costs of healthcare, and plots aggregated by city, which we had by the time of the exploratory analysis did not work. We did include plots for the average costs of healthcare by state and region, as well as the percentage of the cost that is covered grouped by state and region.

The following image shows the average payments for every disease treatment in the dataset, grouped by region. This is an example of a plot we came across that showed no usefulness in our application. The costs of treatments are mostly the same in each region. What the graph did show, however, were a few high outlying disease treatment average payments. This indicated to us that it was worthwhile to look at the different costs and coverage values for different diseases within the app.

Average Treatment Payment By Region



The next plot confirmed the usefulness of looking into differences in percentages covered by state. It shows a couple outlying states and some variation among the other states. We also created a similar plot grouped by region, that showed a similar variation pattern.



The plots mentioned formed the basis for the plots we used within the app. They also confirmed that we could allow the user of the application to easily perform some kind of significance test to gain statistical insight on differences on whatever aspect of healthcare costs and coverages they are looking at.

Final Analysis

Throughout this project, our group jumped from question to question, detail to detail, learning about handling real data that isn’t organized in the best way for a specific use and learning about the development of an application to be used for statistical analyses. In the final analysis and creation of the application, we gained valuable insights on the specific questions concerning the dataset itself, but we gained more valuable and general insights on how to make an application look clean, usable, and useful.

From the questions explored throughout the project, the features that made it into the final application are as follows:

1. Explore summary statistics for each state or region
2. Look at the average payment or average charged of healthcare by state or region
3. Look at the covered percentage by state or region
4. Perform a t-test for difference between two means on two states or two regions
5. Look at average payments or total charged for each disease treatment type

These features can successfully help the user of our app to answer the initial questions we found to be practically answered in the application and the questions we had along the way.

Presentation

(Here I will add labels to the screenshot describing each feature, once the app is finished)

