INFO 5100 Project 1 Report
YuanYuan Li (YL2577), Richard Quan (RQ32), Akshay Tata (AT758)

Analysis of Hospital Costs for Top Diagnoses in the United States from 2011 to 2013

We found our original dataset, "Medicare Provider Utilization and Payment Data: Inpatient" from CMS.gov (https://goo.gl/XFUtm6), a Medicaid & Medicare agency of the U.S. Department of Health and Human Services. The data provided hospital-specific charges for more than 3,000 U.S. hospitals that received Medicare Inpatient Prospective Payment System (IPPS) payments for the top 100 most frequently billed discharges. This data is essentially a breakdown of procedural costs and coverage for the most common diagnoses in the United States. We found data for three fiscal years: 2011, 2012, and 2013.

The original dataset had variables including (1) the name of the diagnosis/procedure, (2) location of the provider, (3) average covered charges for the services, and (4) average total payments paid to a provider by Medicare. We took the difference between average covered charge and average total payment to a provider as the average cost of a procedure for a provider/hospital. So we calculated the difference of (3) and (4) for specific diagnoses, and then compared the differences from 2011 to 2013. (NOTE: Since the top 100 diagnoses/procedures are different from year to year, mapping the costs from 2011 to 2013 shrank the data from the top 100 diagnoses/procedures to the top 95.) We were surprised to find that regardless of the diagnosis, the difference always increased, meaning that the costs of procedures for U.S. providers/hospitals were increasing from 2011 to 2013.

This is the trend of increased costs we attempt to show in our first graph. The different colors of the circles represent specific diagnoses/procedures. The x-position of each "bucket" of circles represents the year of the cost. The y-position of each circle, mapped from a linear scale called yScale, represents the average cost per hospital of the diagnosis/procedure. To differentiate circles in the same year, we also used a random function for the x-position per bucket to avoid excessive cluttering and overlap. Finally, even though the increase trend was common for all 95 diagnoses in our dataset, we chose to only show costs of the the top 10 diagnoses to decrease the range of data and thus improve the visibility of the trend.

We wanted to analyze this trend deeper and decided to break it down by location, specifically by state due to the influence of state tax and healthcare policies. We used (2) and the difference of (3) and (4) to find the trends of each procedure's costs per state. For the most popular diagnosis (major joint replacement or reattachment of lower extremity), we found that even though the cost of the procedure increased in most states, its cost actually decreased in Massachusetts.

This major joint replacement procedure, the most popular diagnosis in the U.S., showed an interesting outlier that we mapped out in our second graph. We used a color scale called "changeScale", which is a color scale mapping of percentage change from to red to white to blue. A red colored state signifies that the cost of the procedure increased in that state from 2011 to 2013. Meanwhile a blue colored state means that the cost of the procedure dropped. The saturation of the color shows the intensity of the increase or decrease rate, where rate is calculated as the difference of the cost from 2011 to 2013 as a percentage of the original cost of the procedure in 2011. This is again mapped to a linear scale. If a state is deep red, the increase rate of the cost is high. In contrast, if the state is mostly white and only a little red, the increase rate of the cost is low. In our second graph, only Massachusetts had a decrease in the cost of major joint replacement, so it is the only state labeled blue.

We wondered if some states were more "successful" at reducing costs of procedures than others, or if Massachusetts was just a random outlier. To measure this, we counted the number of procedures that benefitted from a cost reduction for each state, and how much the cost was reduced on average. We combined this with latitude and longitude data (http://dev.maxmind.com/geoip/legacy/codes/state_latlon/) to add an extra layer of visualization for a third graph.

Our third graph shows a per-state analysis of the number of procedures that have lower costs as well as the intensity of the cost decrease. Again, we used linear scales. The size of each circle in each state represents the number of diagnoses there out of the top 95 that have lowered in costs from 2011 to 2013. The color of the circle shows the average reduced cost per diagnosis. If the color is deeply blue, the average cost of procedures lowered a lot from 2011 to 2013 (up to around -$4,000). If the color of the circle is totally white, the cost reduction is very low or nonexistent. We created the graphic to optimize the visibility of the circles and saturation, making design decisions such as moving Alaska and Hawaii near the U.S. to zoom in closer to map while moving several state circles away from each other to avoid overlap. As seen in the graph, we found that hospitals in states with big circles like Massachusetts were benefitting from reduced costs for many procedures, while states with really dark blue circles like Alaska were benefitting from very large cost reductions.

The analysis of this procedural cost per year and per state has been very eye-opening for us as a team. Our initial plans of just graphing out the dataset, as we did in our first graph, showed us a pattern of increased costs of procedures all across the nation. The time frame of this data is within the passing of Obamacare, although none of us are well versed enough in healthcare enough to draw any direct causations (or lack thereof). Although it is worth noting that the increase of procedural costs we saw was more significant than national interest rates and inflation. Breaking down the data by state and intensity in the latter two graphs further piqued our interests. Even though there was a national trend of increased costs for procedures, there were hospitals in certain states that were benefitting from significant cost reductions.

Does Massachusetts have really good policies? Why is California so darn expensive? Those are questions that would require a lot of further research. We struggled early on in the assignment to even understand the basics of the U.S. Medicare system. But working together, with Yuanyuan focusing on dataset cleanup and analysis, Richard on visualization prototyping, Akshay on UI engineering, and group coding sessions, we were able to create visualizations of the dataset on three different levels. One of our biggest learnings from the first iteration has been to portray complex data in an as simplified way as possible. When you have a large set of data points, as stated by the professor, it is a tough decision to take on what to really display so that a layman finds it worthwhile. We have learned not just the fact that some of the d3 calls are asynchronous but also the fact that how important is simplicity. I am sure that our data could also be presented in a simple bar-graph but the learnings from this project and the design lectures have been to use simple shapes and forms to convey data meaningfully. Hopefully our visualizations can provide insights for other users and help guide them towards future research and analysis on this topic.