Capstone 1: Final Report

Problem statement:

Can you predict how much Medicare will pay on average for diagnoses in the hospital? Additionally, can you predict the Medicare payment for a top ten diagnosis or diagnosis of interest?

This would be useful to a hospital budgeting department when anticipating upcoming costs or to an insurance company when planning to adjust reimbursements (as most insurance companies follow what Medicare does).

May also be useful for any hospitals not currently treating a given diagnosis who are looking into adding treatments/programs to their facilities.

Obtaining, cleaning, and wrangling the data

I downloaded the center for Medicare services (CMS) data sets from their website detailing the average Medicare reimbursement for every hospital in the country that treated and discharged more than 10 people with a given diagnosis (denoted by their DRG-diagnostic resource group-code) in fiscal years 2014-2017.

I first downloaded each of the excel spreadsheets into its own dataframe, removed columns I deemed not relevant to my purpose, added a column with the appropriate fiscal year to allow for separation and comparison later, and separated the DRG definition code to allow for organization by either the numeric code or string definition.

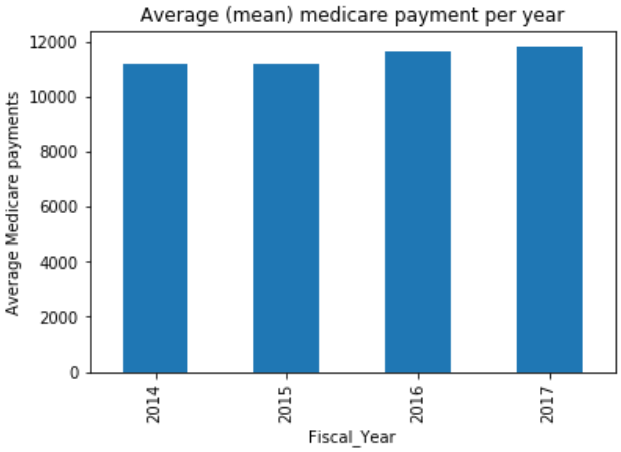
This was followed by merging all the data into one, complete dataframe: full_df.

I called .info() on the full dataframe to make sure I didn't have any missing values or less useful data types. I had exactly 798140 entries on each of 11 columns and felt confident that I did not have any empty cells in my data. I do note that I do not have data from all 50 states nor does each state that is included have data on each diagnosis. To overcome this, I decided to look at the data from a more regional perspective using zip codes. This way a region average can encompass a given state. There were data points that could potentially be outliers when I

plotted the data, but when I looked deeper, these data points corresponded to highly costly procedures such as heart transplant.

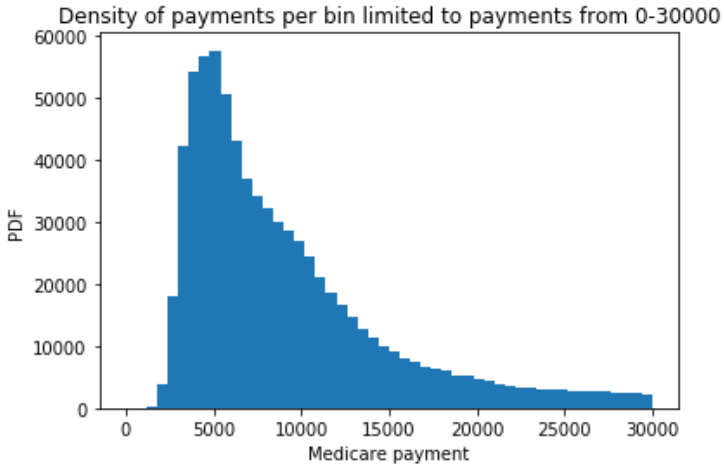
Relevant Exploratory Data Analysis, Statistical Inference, and Linear Regressions

I looked at the mean and median Medicare payment per year. The values themselves were different but both graphs looked very similar as below:



2014 mean: $11218.47
2015 mean: $11213.91
2016 mean: $11656.63
2017 mean: $11816.24

I checked to see if the data is in a normal distribution. It is rather noticeably right tailed, which means that my calculated mean and median will read high. I will need to take this into account for my analysis.



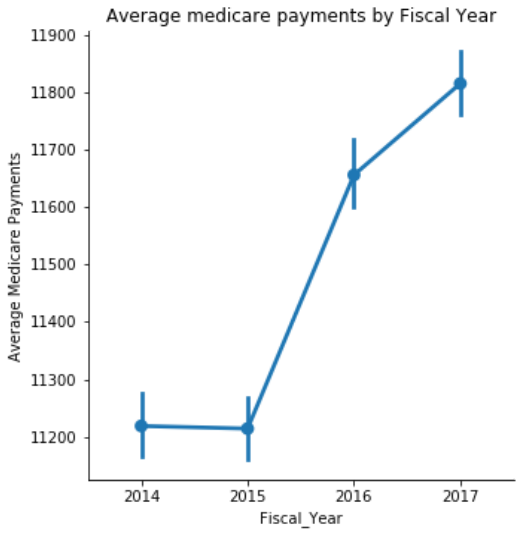
To determine if there was a statistically significant change, I completed bootstrap permutation hypotheses for 2014-2015, 2015-2016, and 2016-2017 with results as follows:

- 2014-2015: not statistically significant, cannot reject the null hypothesis (that they're the same)
- 2015-2016: statistically significant, can reject the null
    - This was also the biggest difference with a change of ~$440
    - Likely the only one that could have practical significance
- 2016-2017: statically significant, can reject the null
    - May not be practically significant, only a difference of ~$160

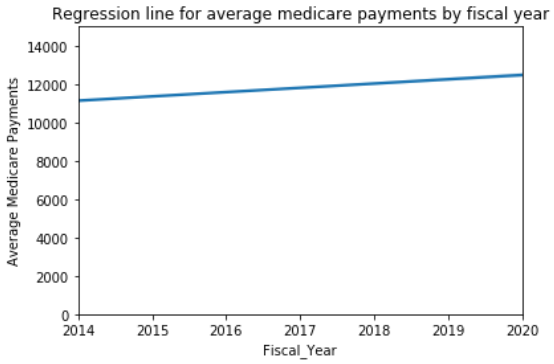
I also completed an ANOVA to provide a double check of my data. The p-values calculated with this were consistent with my conclusions from the bootstrapping testing.

I completed a linear regression to examine the rate of change and to allow for prediction of future payments. I chose linear regression because I am attempting to predict a value in the future (rather than a category that would have preferenced logistic regression). I figured that a linear regression would be a fairly simple, straightforward way to get useful data like the rate of change for payments as a whole or for an individual diagnosis. I do appreciate that typically, you would want more data points to create a more accurate regression, and that 4 data points is not ideal.

The average Medicare payment catplot looked like this:



With a regression line like this:

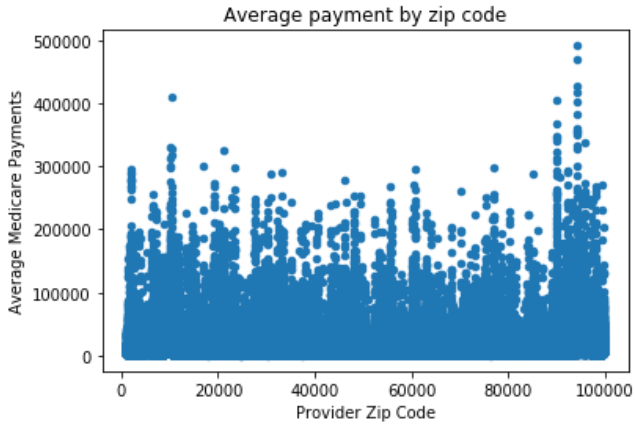


With slope and intercept as such:

```
In [27]: slope, intercept, r_value, p_value, std_err = stats.linregress(
             x=full_df['Year_cat'], y=full_df['Average Medicare Payments'])
         print('slope =', slope)
         print('intercept =', intercept)

         slope = 223.31240211269716
         intercept = 10917.675010876103
```

I turned my attention to visualizing the data in other ways.

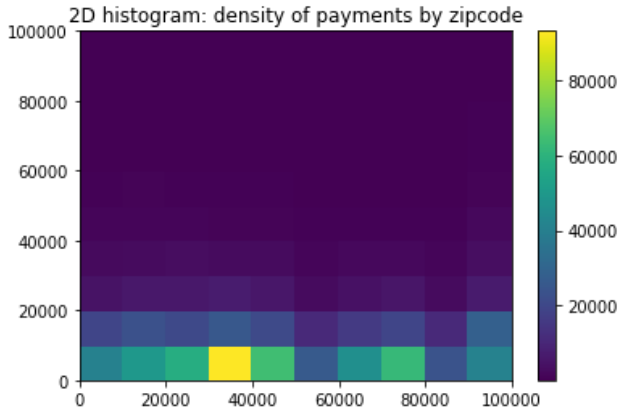
To look at the payments by zip code, I first made a scatter plot:



The above is not particularly readable, though it does appear that most of the biggest payouts are somewhere in the 9xxxx zipcode.

Also, this is an aggregate of all the data from all four years, rather than breaking it up into each fiscal year.

I decided to make a 2D histogram to better visualize the density of payments per zipcode:

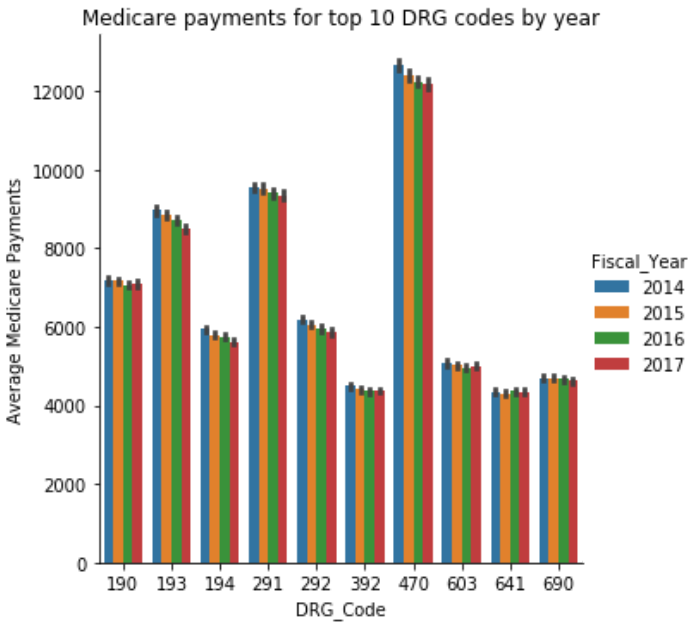


There appears to be the highest density in the 3xxxx zip code range (TN, MS, AL, GA, and FL) I do notice that the 9xxxx zip code range (AK, WA, OR, CA, and HI) appears to have one of the lower frequency of Medicare payments, but we know from the above scatter plot that this zip code range has some of the highest, single payment instances of the entire country over all 4 years.

Plotting all the DRG codes by their Medicare payments as well as looking at each individual DRG code is neither time efficient or particularly insightful, so I elected to choose the 10 most commonly billed DRG codes as well as some hand picked codes of interest to focus on.

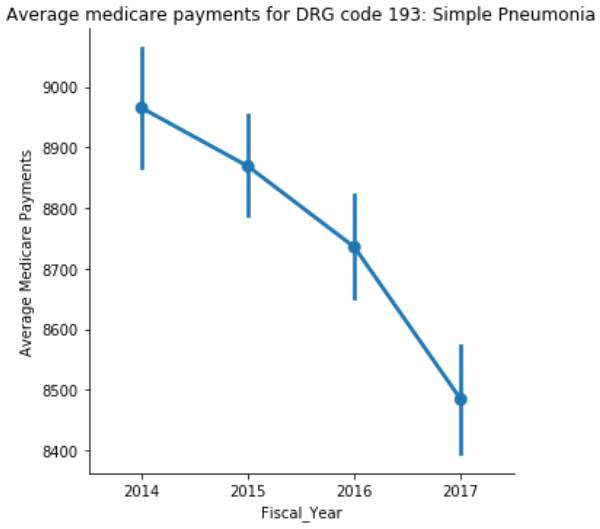
The top 10 codes are related to pneumonia with complication/comorbidity and with major complication/comorbidity, major joint replacement of the lower extremity, heart failure with complication/comorbidity and with major complication/comorbidity, esophagitis, kidney and urinary tract infections, COPD, cellulitis, and miscellaneous disorders of nutrition, metabolism, and electrolytes.

Looking at these top 10 codes broken into each year, I noticed that most are decreasing, however those for pneumonia with major complication/comorbidity (193) and major joint replacement (470) seem to be decreasing at a higher rate.

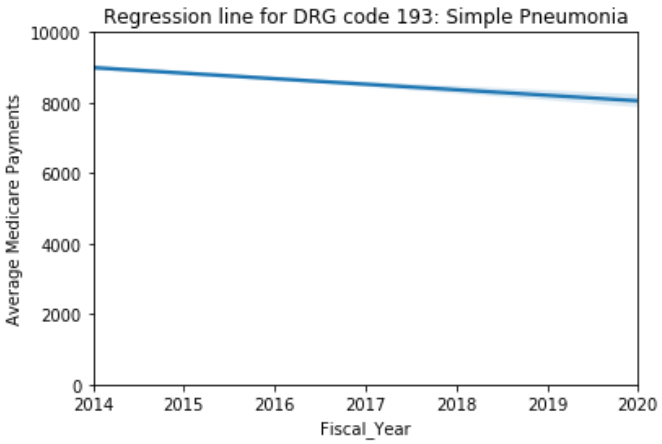


I completed a linear regression for those two that appeared to be decreasing at a greater rate to better estimate the rate of change.

I started by looking at a catplot of the data generated using seaborn to better visualize where I thought a regression line might go. Here is the simple pneumonia plot:



I then plotted a regression line for the data using seaborn to see if they would visually appear consistent, which it did.



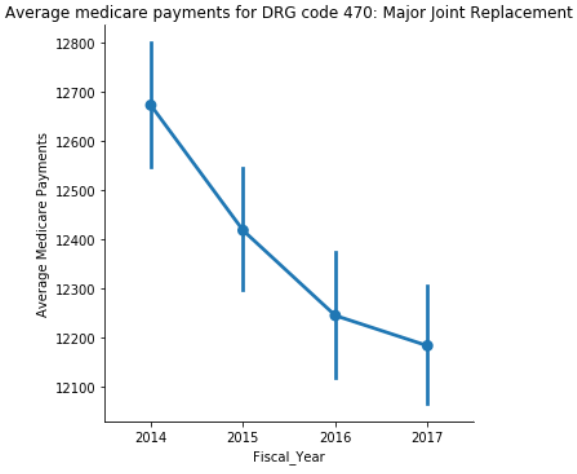
A problem I ran into was realizing that seaborn does not allow you to extract some of the statistical data, such as a slope or intercept. So I calculated those points using scipy.stats.

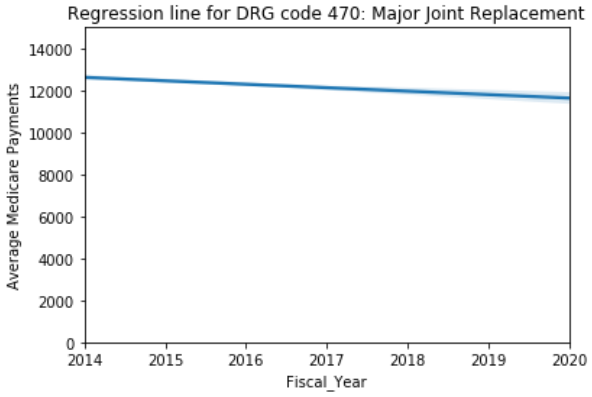
```
In [35]: slope, intercept, r_value, p_value, std_err = stats.linregress(
             x=pna['Fiscal_Year'], y=pna['Average Medicare Payments'])
         print('slope =', slope)
         print('intercept =', intercept)

         slope = -157.03622093794195
         intercept = 325270.4755498979
```

I determined that using fiscal year was messing up my calculations, so I created a column to treat the year as a categorical value of 1-4. After I made that change, the results were consistent with the graphical analysis I completed earlier:

```
In [37]: slope, intercept, r_value, p_value, std_err = stats.linregress(
             x=pna['Year_cat'], y=pna['Average Medicare Payments'])
         print('slope =', slope)
         print('intercept =', intercept)

         slope = -157.0362209379425
         intercept = 9156.562801820737
```

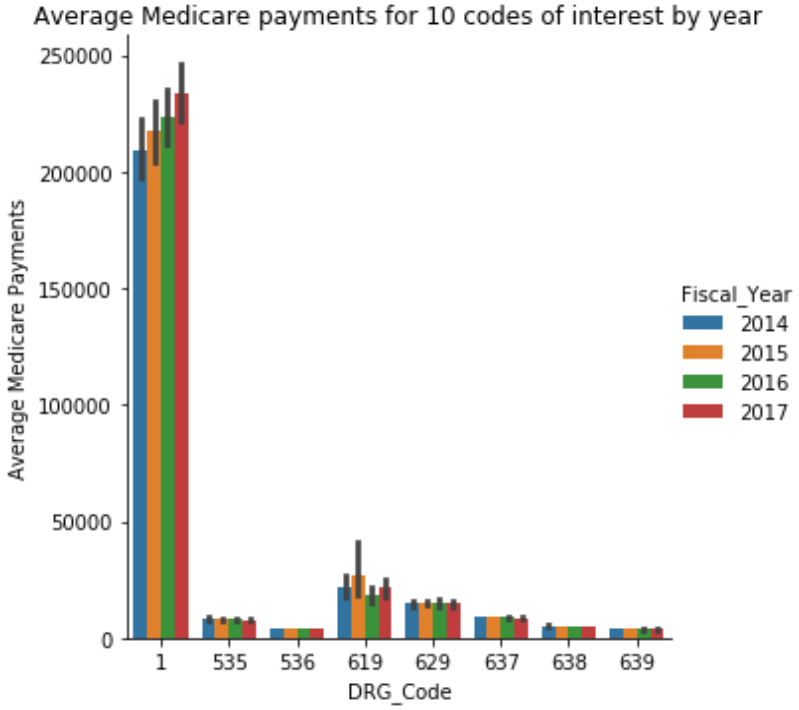
I utilized that same technique for major joint replacement with regression line and stats below:



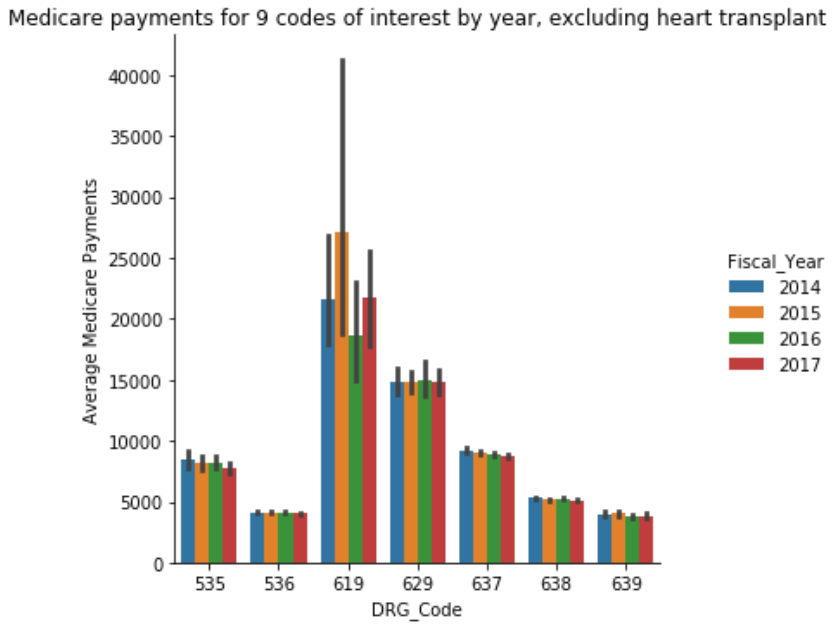


```
In [45]: slope, intercept, r_value, p_value, std_err = stats.linregress(
             x=joint['Year_cat'], y=joint['Average Medicare Payments'])
         print('slope =', slope)
         print('intercept =', intercept)

         slope = -164.44879052998758
         intercept = 12790.607618178752
```

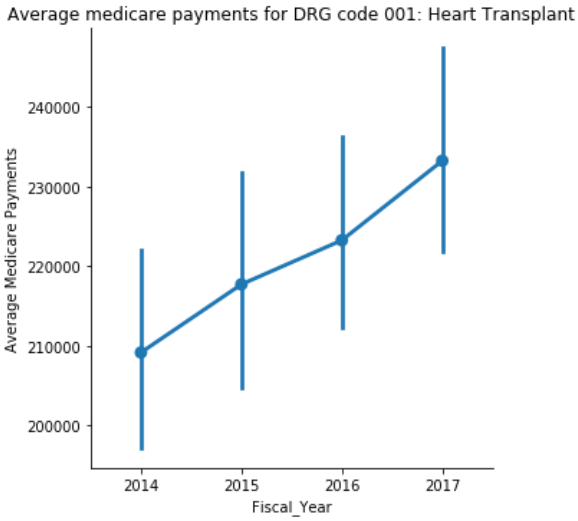
I also looked at 10 hand selected codes that I thought would be interesting: heart transplant, fractures of the hip and pelvis with and without major complication/comorbidity, OR procedures for obesity with complication/comorbidity and major complication/comorbidity and diabetes with and without complication or major complication.

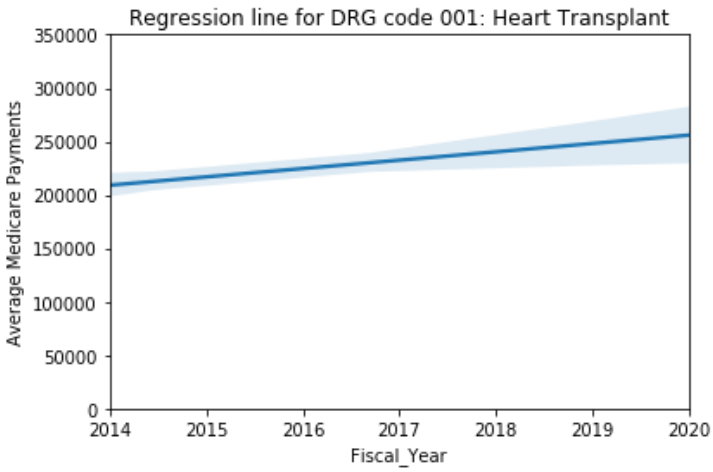


I wondered if I was getting the full picture on 9/10 codes since the reimbursement for heart transplant is so high, so I created this plot again, this time excluding the code for heart transplant.



I thought it would be worth looking at the rate of change of heart transplant, especially since it is seeming to increase so dramatically. Again, I will be looking at a linear regression to better study the rate of change. I utilized the same technique as above with the following results:





```
In [53]: slope, intercept, r_value, p_value, std_err = stats.linregress(
             x=heart['Year_cat'], y=heart['Average Medicare Payments'])
         print('slope =', slope)
         print('intercept =', intercept)

         slope = 7800.364318859858
         intercept = 201338.8012678635
```

Storytelling

Any hospital system in the country is concerned about meeting the bottom line. For many systems, Medicare reimbursement is a major driving force for the money that keeps a hospital going. As the cost of providing quality healthcare continues to increase, stagnating or downward trends in reimbursement can have a huge impact on that bottom line.

Looking at the four most recent years of data (2014-2017), we can see that the average payment is increasing very slightly at an average rate of +$223/year. In fact, the smallest difference was only around $5, while the biggest difference was around $440.

In terms of keeping the doors open to patients, even the biggest difference may not be a practically significant change, especially when considering that the 10 most commonly charged codes nationally are, in general, decreasing. Indeed, the second most common code, the one for major joint replacement of the lower extremity (think of your total hip, knee, and ankle joint replacements) is decreasing at an average rate of $164/year. For a system that might see as few as 10 patients per week, this is an annual loss of $85,000 per year.

If your hospital or system has a lot of those high dollar reimbursements, like a heart transplant, then perhaps you're seeing more of the increase. On the flip side, if your hospital is most well known for its total knee replacements, then you may be feeling the strain of the decrease in payment. For example, when we compare the 2015-2016 data, we see that Medicare payments increased by $440 on average, yet major joint replacements decreased by around $173 for that same period.

In conclusion, in terms of budgeting for upcoming fiscal years, the predicted rate of reimbursement should play a role in your figures and calculations. As more data becomes available, this predictor should become more and more accurate. For now, we can predict that the upcoming years, based on the past trends, might expect an average increase of $223/year, but that individual procedures such as the major joint replacement might expect a decrease of $164/year. This is especially important for those areas that see the highest amount of Medicare payments, those in the 3xxxx zip code range (TN, MS, AL, GA, and FL)

For future exploration, I could utilize a regularization like lasso regression to attempt to determine which one or few diagnoses are most contributing to overall cost or which state(s)/regions/zipcodes are most impacting the average cost. I could also tailor this approach to a given diagnosis or set of diagnoses of interest to a particular hospital or system.