Thomas Cho

IEMS 308

Clustering

**Executive Summary**

In an article called “Once again, U.S. has most expensive, least effective health care system in survey” by The Washington Post, a report ranks the United States dead last in the quality of health care systems.[[1]](#footnote-1) The United States spends more on health care than any other country and has more specialized physicians than any other country. Health insurance and premiums are increasingly, leaving many Americans without the proper health care plan they need. While prices have gone up, the general consensus is that care and performance are not increasing. In an effort to make our healthcare system more transparent, affordable, and accountable, the Centers for Medicare & Medicaid Services (CMS) has provided a public data set with information on services and procedures to Medicare beneficiaries by physicians and professionals.

**The goal of the analysis is to reduce the health care costs and to find focal areas to optimize current resources.**

Cluster analysis is performed on this data set to divide data into groups that are meaningful and share common characteristics. In the case of the CMS data, information on utilization, payment, and provider data are provided. The features that were used in the clustering analysis were the cost difference between service charge and Medicare allowed, place of service, gender, and HCPCS drug indicator. K-means clustering resulted in 4 groups. One group had a high cost difference of more than $400, and that group’s services were in- facility and were all conducted by male physicians. Those services that were HCPCS indicator had small cost differences of no more than $40, and those that weren’t had very high cost differences starting from $150. There were more male physicians than female physicians in total. Splitting the United States into regions, East Central and New England regions had a higher percentage of huge cost differences of more than $400. These findings are good starting points for further analysis on improving the costs of US healthcare system.

**Methodology**

The first step to any analysis is to understand the data. That includes looking at data size, variable types, variable descriptions, etc. Since the dataset is rather large, random sampling is used to make analysis computationally easier and eliminates any form of bias in sub-setting the data.

The second step is data cleaning and feature selections. In order for K-means clustering to work properly, rows with NA values must be removed. Looking at basic statistics and histograms, outliers are removed, and variables are transformed to fit a normal distribution, and these steps result in more accurate clustering groups. Correlation matrix is used to remove correlated variables because correlated variables give more weight to the decision-making process. Features are standardized on the same scale with mean 0 and standard deviation of 1. For nominal features, categorical data that don’t imply order, one hot encoding technique is used.

After the data is preprocessed and the features are chosen, K-means algorithm is used because it’s computationally very efficient than other clustering algorithm. Because the number of clusters is an input before the result, elbow method and silhouette plots are used to decide the optimal number of clusters.

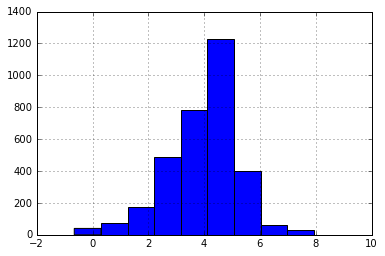
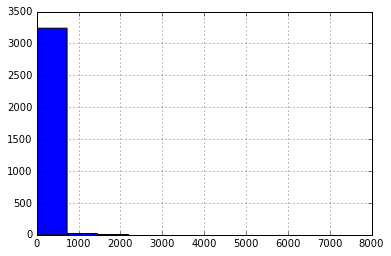
**Analysis**

*Understanding the data*

The data contains more than 9 million records and is too big (memory extensive) and time-consuming to analyze wholly. Pandas is used to load and read in a sample 10,000 random records. Numpy and matplotlib are also used to manipulate data and plot graphs. There are 26 variables, and the variable descriptions are available in the *Medicare Physician and Other Supplier PUF* Methodology[[2]](#footnote-2).

*Preprocessing*

I remove records with NAN values. That reduces the 10,000 records to 3,281 records. Then, I plot histograms for each numerical variable. For example, plot 1 shows the histogram of the average\_Medicare\_allowed\_amt. The data is heavily positively skewed. K-means work best when all the features are normally distributed. Therefore, a log transformation is done, as shown in plot 2. Because there might be zero values and log 0 is undefined, I add a small constant of 0.5 to every log transformation to ensure that any sampled data will take the log of non-zero values.



Plot 1 Plot 2

Correlation matrix is computed to ensure that during feature selection, two highly correlated variables won’t be picked. As seen in image 1, this is a part of the correlation matrix and it shows that average\_submitted\_amt is strongly correlated with the other 3 average\_amt variables with values of about 0.78. The rest of the correlation matrix can be found in the Appendix.

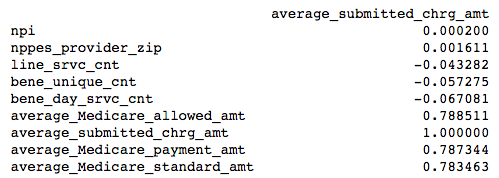


Image 1

Lastly, features are standardized to mean 0 and standard deviation of 1 to ensure that each feature has equal weight in deciding the clusters. For nominal variables, I use Pandas to make dummy variables.

*Variable selection*

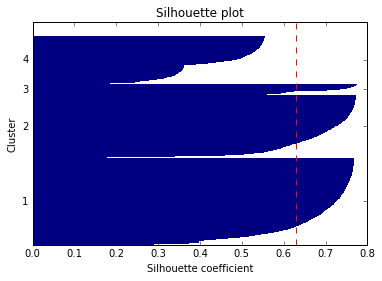
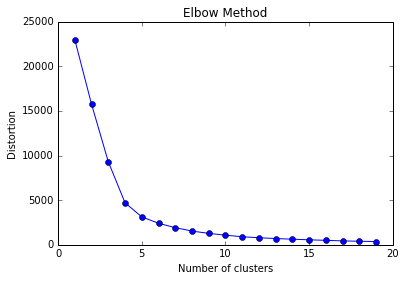
The main issue with US health care is the extravagant costs. Therefore, I focus on the 4 variables – average\_Medicare\_allowed\_amt, average\_submitted\_chrg\_amt, average\_Medicare\_payment\_amt, and average\_Medicare\_standardized\_amt. Reading the variable descriptions found online, the average\_Medicare\_allowed\_amt, average\_submitted\_chrg\_amt are most interesting because in a situation in which the submitted charge is more expensive than the allowed Medicare amount, which is the sum of the amount Medicare pays, deductible and coinsurance amounts that the beneficiary is paying, and any amounts that a third party is paying, then questions such as who pays for the difference and why is there a difference occurs. Therefore, I create a new variable that takes the difference between the submitted charges minus the allowed charges, and try to understand which groups have the biggest difference.

Second, we choose variables that might affect the differences. In this case, provider gender, place of service, and HCPCS drug indicator are variables that might affect the price of service. Since all 3 are binary variables, I create two dummy variables for each nominal, and the total number of features is 7.

*Clustering*

The fallback of K-means clustering is that we have to specify the number of cluster prior to our analysis. We use elbow method and silhouette plots to evaluate the quality of the clusters and determine the optimal number of clusters. The idea is to find the value of k where the distortion begins to increase most rapidly, and that point seems to be at k=4, shown in plot 3.

We can confirm this by looking at the silhouette plot in plot 4. The closer the silhouette coefficient is to 1, the clusters are better separated from each other. The average silhouette coefficient is around 6.2, which is a good number. Also, there are no negative coefficients, which means that there are no incorrectly placed clusters. These two plots give a strong indication that 4 clusters is a good number.



Plot 3 Plot 4

**Findings**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Count | Mean of submitted minus allowed | Place of Service F | Place of Service O | Female | Male | HCPCS drug ‘NO’ | HCPCS drug ‘Yes’ |
| 0 | 1367 | $148.58 | 0 | 1367 | 0 | 1367 | 1367 | 0 |
| 1 | 983 | $414.09 | 983 | 0 | 0 | 983 | 983 | 0 |
| 2 | 173 | $35.94 | 0 | 173 | 49 | 124 | 0 | 173 |
| 3 | 758 | $149.61 | 274 | 484 | 758 | 0 | 758 | 0 |

Table 1

The optimal number of clusters is 4 groups that best represent the data. Several tabulations were done grouped by the clusters, and they are compiled in table 1. The key insights and possible hypotheses from the clustering are as follows:

**Cluster 2, which are HCPCS** **approved drugs services, have the lowest difference between service charge and Medicare allowed amount.** This is likely due to the fact that HCPCS approved services cost the least, and therefore the difference in cost, which may stem from inefficiencies in the system are the least. Also this is the only group that both female and males share.

**Cluster 1 has the highest difference in the submitted charge and allowed Medicare amount.** This leads to several hypotheses on why the difference is so high. Why is the physician charging so much? Or in another perspective, why is Medicare allowing so little for a service that costs more? What is certain is that these services happen in facilities, most likely big hospitals that have the necessary resources to perform the services that happen within this cluster. Possibly, in-facilities are more inefficient between the physician and Medicare in cost and resource management. What inefficiencies or miscommunications lead to such a difference is one area that should be addressed to reduce this number.

**There are more male physicians than female physicians.** Cluster 0 and cluster 3 provide the same HCPCS services and similar difference in charges and Medicare allowed amounts, but differ in facilities and gender. While cluster 0 has males only in office settings, cluster 3 has women in both big facilities and office settings, but the number of women physicians is much smaller. That leads to questions on why there are fewer female physicians and why males solely work in offices.

For further analysis, I was very curious about cluster 1 because of the big difference of $400. I divided the US into regions to see if any regions had interesting numbers.

**The Midwest region had the biggest percentage of data in cluster 1.** As shown in table 2, the Midwest had a 33% of data in cluster 1 (I did not include pacific region because of too small numbers). This could mean that more people found in the Midwest performed certain services that resulted in high cost differences. On the other hand, the West had the lowest percentage in cluster 1, and that may mean that methods in the West lead to the least cost difference.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Regions/ Cluster | Midwest | Northeast | Pacific | South | West |
| 0 | 221 (34%) | 261 (42%) | 5 | 590 (43%) | 283 (45%) |
| 1 | **215 (33%)** | 188 (30%) | 6 | 406 (29%) | 167 (27%) |
| 2 | 44 (6%) | 26 (4%) | 2 | 68 (5%) | 33 (5%) |
| 3 | 169 (26%) | 145 (23%) | 4 | 301 (22%) | 137 (22%) |

Table 2

I wanted to delve deeper. Therefore I split the regions even more.

**East Central and New England have more percentage share in cluster 1, both regions not in the Midwest.** As shown in table 3, East Central and New England regions had a 35%, while all the other percentages were around 27% to 30%. By delving into the 2 regions, this is a starting point to discover why there is such a big difference between the submitted charge and the allowed Medicare amount.

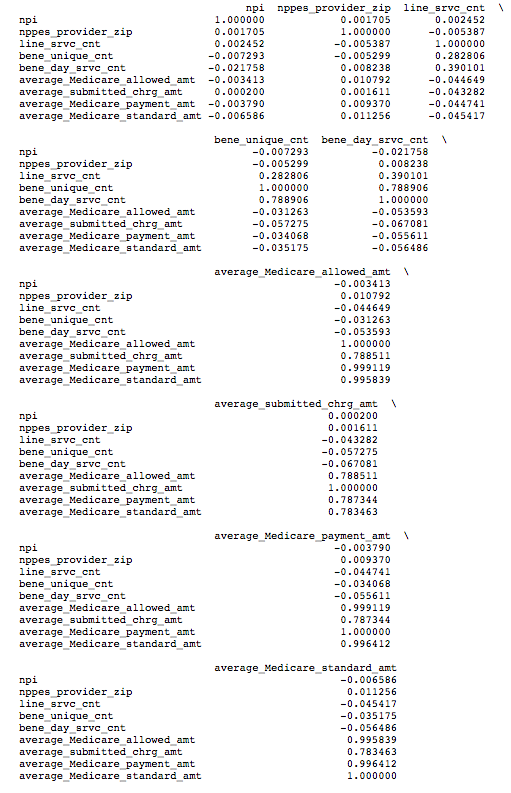
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Specific regions/ cluster | East central | East south | Middle Atlantic | Mountain | New England | Pacific | South Atlantic | West Central | West South |
| 0 | 153 (34%) | 111  (42%) | 186  (45%) | 103  (44%) | 75  (35%) | 185  (46%) | 332  (44%) | 68  (33%) | 147  (41%) |
| 1 | **158**  **(35%)** | 73  (27%) | 114  (27%) | 62  (26%) | **74**  **(35%)** | 111  (27%) | 224  (30%) | 57  (28%) | 109  (30%) |
| 2 | 26  (6%) | 17  (6%) | 18  (4%) | 15  (6%) | 8  (4%) | 20  (5%) | 37  (5%) | 18  (8%) | 14  (4%) |
| 3 | 106  (24%) | 62  (23%) | 90  (22%) | 54  (23%) | 55  (26%) | 87  (21%) | 155  (21%) | 63  (30%) | 84  (24%) |

Table 3

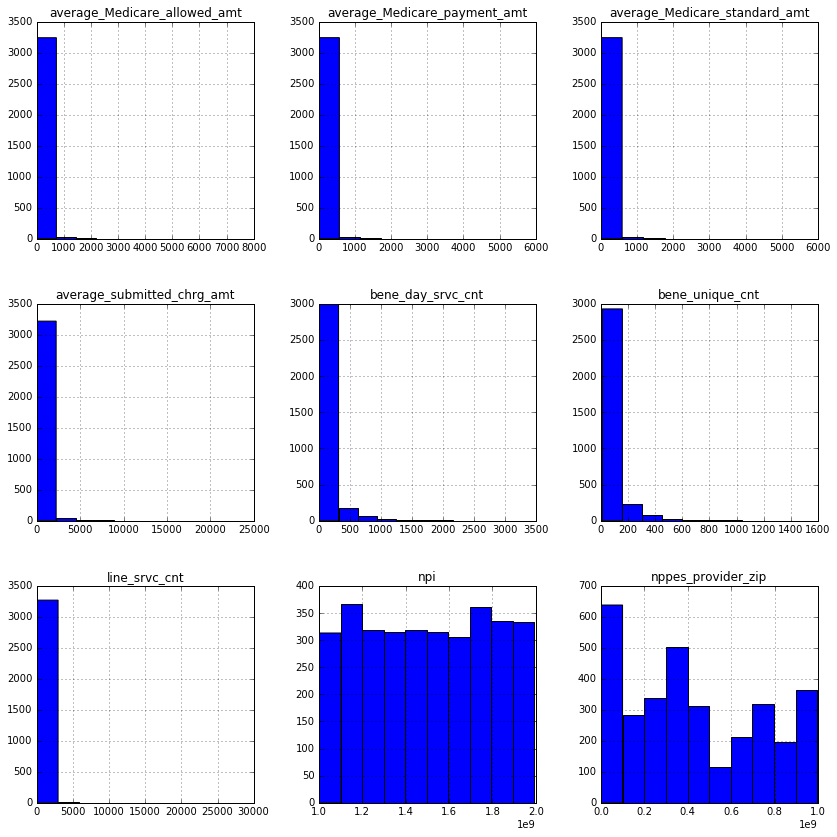
**Next steps**

Clustering is a good starting point for further data analysis. There was additional information on beneficiary data and provider demographics that can be added to the current data to gain more insight into the various groups. Furthermore given more space, more data can be used to do analysis than using samples. Given the current findings and information, implementations and efforts to reduce costs can be applied to those areas that have high costs. It would also be noteworthy to look into the gender roles of various physician roles since there seemed to be some sort of relationship between gender and price difference that certain services require.

**Appendix**



Correlation matrix



Histogram

1. https://www.washingtonpost.com/news/to-your-health/wp/2014/06/16/once-again-u-s-has-most-expensive-least-effective-health-care-system-in-survey/ [↑](#footnote-ref-1)
2. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf [↑](#footnote-ref-2)