Aneesh Kudaravalli

IEMS 308 HW 1

Clustering

**Business Question:**

A large healthcare system is looking to expand its practices to the Chicago area by either building a clinic or hospital in the area. A common problem the health system faces is when their submitted charge for a service is much higher than the allowed Medicare amount. When it comes to patients that don’t have other insurance, the health system often resorts to asking the patient to cover the rest of the charge. If that’s the case and patients aren’t able to cover that difference, the health system has to write off the extra costs as sunk.

In order to avoid these sunk costs, this healthcare system wants to understand – **what types of practices in Chicago have the highest Medicare allowable amounts for services relative to the costs of those services**? Since the rates that Medicare allows for services differ by practice depending on a variety of quality measures and factors including geography, physicians’ practice patterns, and patients’ willingness to get care, it is important to understand type of practice as it relates to location within Chicago as well as types of services offered.

**Methodology:**

The data that was used to conduct this analysis comes from the Medicare records on practices and their services: 2017 Medicare Physician and Other Supplier Public Use File. The samples (Medicare allowance info for medical services aggregated by both provider and unique service) were clustered based on location, types of services offered, average submitted charge for the service, and average Medicare allowable amount for the service. We then compiled the results and insights from clustering the data to understand if/why certain types of practices in Chicago have higher Medicare allowable amounts.

The first step was subsetting the national dataset using a SQL query to select only providers who were based in Chicago (or had the word Chicago in their listed city). Then, the procedural codes for services were categorized into 6 different fields (Evaluation and Management, Anesthesiology, Surgery, Radiology, Pathology and Laboratory, and Medicine) according to Category 1 CPT code specifications. Alphanumeric HSPCS codes that were coded on a different system were discarded (a small portion of the data). From there, the data was further categorized by the first 3-digits of the zip code of the provider. At this point, it became apparent that some of the samples were not located in Chicago, but cities elsewhere with Chicago in the name. **(Appendix A)** These samples were removed from the data set. After this, 77,263 samples were left in the dataset.

The distributions for both the average submitted charge for a service and for the Medicare allowable amounts for a service were heavily skewed, so they were transformed using the log function to turn them normally distributed in order to cluster properly **(Appendix B)**. To finish preparing all the features for clustering the data was normalized. **All of the above data preparation and analysis was conducted in R, and further cluster analysis was conducted in Python.**

It should be noted that at this point many clustering attempts were performed using different subsets of the features, but these solutions had issues. For instance, clustering that included a 5 digit zip code for location had results with far too many clusters to be helpful. And clustering on types of services and costs/Medicare amounts only led to clusters weighted too heavily on the types of service despite the categories being extremely broad making any insights ambiguous.

Thus, we employed the k-means algorithm to cluster the data on all 4 features discussed (type, location, submitted charge, Medicare payment). The optimal amount of 15 clusters was determined by comparing the results with different numbers of clusters to find the point where the reduction in error was diminished. This number was then confirmed by looking at the average silhouette scores for differing numbers of clusters and recognizing that at 15 clusters, the average silhouette score is the highest. Furthermore, we examined the silhouette scores of each of the 15 clusters in the ideal case and saw that only 2 of the 15 have values below .5 which further confirmed the accuracy of these clusters **(Appendix C)**.

**Clustering Results:**

Outliers detected by clusters with small amounts of samples in them were thoroughly investigated. We found that each cluster is composed entirely of samples from one 3 digit zip code **(Appendix D)**. The small clusters make up the entirety of providers in the outskirts of Chicago while the other clusters are composed entirely of samples from the 606 3 digit zip code which makes up the vast majority of the dataset. This allowed us to generate insights on what types of services are popular with providers in the outskirt of Chicago. It also allows us to understand the differences in rates for types of services in the center of Chicago. For that reason, we kept the outliers in the data set.

**Table 1. Shows the size of each cluster, the zip code the samples are located in, the types of operations that make up the majority of the cluster, the mean ratio of Medicare allowed amount to submitted charge, and the mean total non-Medicare allowed costs (the difference between the average submitted charges and average allowable amount times average amount of times that service is performed in a day).**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cluster Size** | **Zip Code** | **Type** | **Average Medicare Allowed amount/submitted charge** | **Average Total non-Medicare allowed costs** |
| 31205 | 606 | Evaluation | .388 | 23049 |
| 13433 | 606 | Medicine | .309 | 25280 |
| 11932 | 606 | Radiology | .207 | 16110 |
| 7016 | 606 | Surgery | .174 | 91995 |
| 5123 | 606 | Surgery | .394 | 25963 |
| 4205 | 606 | Pathology and Laboratory | .242 | 33921 |
| 2066 | 606 | Anesthesiology | .136 | 37963 |
| 1559 | 604 | Mixed (Evaluation focused) | .265 | 46453 |
| 337 | 600 | Mixed (Evaluation focused) | .394 | 10957 |
| 207 | 607 | Mixed (Evaluation focused) | .388 | 23347 |
| 94 | 601 | Mixed (Medicine focused) | .361 | 20064 |
| 37 | 608 | Mixed | .175 | 149932 |
| 37 | 605 | Radiology | .231 | 8067 |
| 9 | 602 | Evaluation and Surgery | .254 | 19849 |
| 3 | 603 | Evaluation | .396 | 3096 |

**Analysis:**

At a high level, we can see that the 7 largest clusters in the dataset are located in the 606 region of Chicago. The other 8 clusters are separated very clearly by zip code.

The 606 clusters are all distinctly defined by the CPT classifications of the procedure, except for the case of 2 of these clusters corresponding to surgery. Surgery contains the largest amount of different codes or services, and the amounts charged can vary wildly depending on the type of surgery. We can see that these 2 clusters in 606 have Medicare amount to submitted charge ratios of completely different values one very high (.394) and one very low (.174). The one with a high ratio has an average submitted charge of around $45 while the one with a low ratio has an average submitted charge of around $1540. Furthermore, the high ratio cluster has an average unique benefactors value of 75 while the low ratio cluster has an average unique benefactors value of 44. This tells us that the surgeries performed in these clusters are very different, and that is the reason they were separated.

It also appears that across the data, *Evaluation type clusters tend to have higher ratios of Medicare allowed amount to cost of services.* Apart from the previously discussed low-cost surgery cluster, the other highest ratio clusters are made up of mostly Evaluation type procedures. The cluster with the highest ratio of .396 is the smallest cluster in Table 1 corresponding to the 603 zip code. The 3 samples in this cluster are made of 1 dermatology office that is conducting 3 different types of evaluation operations on a total of 95 unique patients. The insight here is that this Dermatology practice in the 603 area of Chicago has high Medicare allowable amounts. That said, this cluster is made up of only one provider, and there could be other factors about his clinic that are the reason he has high Medicare allowable amounts. The remaining high ratio clusters are mainly Evaluation type clusters in the 600, 607, and 606 zip codes.

After Evaluation type clusters (and the unique low-cost surgery cluster), *the next type of practice that has relatively high ratios of Medicare amount to cost is Medicine type procedures.* The Medicine type cluster in the 606 area code has a ratio of .309, and the 601 cluster (which is mainly Medicine focused) has a high ratio of .361.

It’s also important to consider the average total non-Medicare costs of services for each cluster which is calculated by the taking the difference in costs and Medicare amount and multiplying that by the amount of times the service was performed in a day. *It appears that Radiology practices tend to have low total non-Medicare costs.* The cluster with the lowest average total non-Medicare costs of $3,096 is the 603 cluster again made up of one dermatologist. Interestingly, the next lowest cluster has total non-Medicare costs of $8067 is the cluster corresponding to the 605 area code that is focused on Radiology. Further examination of this cluster shows that it is also entirely composed of one provider offering a variety of radiology services. He could have low costs due to a variety of quality measures in his practice or the demographics in the area, however, we continue to see Radiology practices in low cost clusters. The next lowest total non-Medicare cost of $10,957 corresponds to the 600 area cluster that is mostly Evaluation focused, but further investigation shows that Radiology services are the next majority in the 600 cluster after Evaluation. The cluster with the fourth lowest non-Medicare cost of $16110 occurs in the 606 area and is Radiology dominant. This cost is smaller than both the corresponding costs in the Evaluation and Medicine clusters from the 606 area.

The data also shows the insight that Anesthesiology practices aren’t covered well by Medicare in Chicago. The Anesthesiology cluster in the 606 area has the lowest ratio of coverage, .136, compared to all the other clusters, including the high-cost surgery one. Further analysis into this cluster shows that the range of Anesthesiology services vary extremely widely in submitted charges, Medicare allowed amounts, and even benefactors of the services per day to the point where the standard deviations of these sample statistics are almost or greater than the mean of these sample statistics.

**Key Insights:**

* Evaluation procedures over other types of procedures (according to Category 1 CPT specifications) in Chicago medical practices tend to have the highest ratios of Average Allowed Medicare Amounts to Average Submitted Charge amounts. This ratio is higher on average in locations that begin with 600 in their zipcode compared to the ratio for all procedures in the rest of Chicago.
* Medicine type procedures tend to have the next highest ratios of Average Allowed Medicare Amounts to Average Submitted Charge amounts. This ratio is higher on average in locations that begin with 601 in their zipcode compared to the rest of medicine type procedures offered throughout Chicago.
* Radiology services in Chicago tend to have low total average non-Medicare allowed costs
* Anesthesiology services in Chicago tend to have very low ratios of Average Allowed Medicare Amounts to Average Submitted Charge amounts.

**Recommendation:**

Based on the results of our analysis, you will have the best success in avoiding low Medicare Allowed amounts on services (relative to the submitted charge for the service) in Chicago if you focus on offering Evaluation, Medicine, and Radiology type services. We recommend further location analysis on where there’s a market need for these types services in Chicago to determine the specific location that’ll be best for expansion.

**Appendix A – 3 digit Zip Codes in data set**

x freq

1 463 8

2 600 337

3 601 94

4 602 9

5 603 3

6 604 1559

7 605 37

8 606 74980

9 607 207

10 608 37

11 617 13

12 660 2

> summary(all\_data$nppes\_provider\_city)

CHICAGO CHICAGO HEIGHTS CHICAGO HTS CHICAGO RIDGE

75303 662 124 770

EAST CHICAGO NORTH CHICAGO SOUTH CHICAGO HEIGHTS WESTCHICAGO

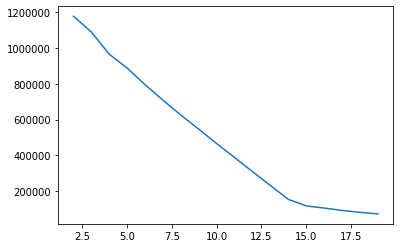
8 335 3 81

**Appendix B Distributions for Medicare allowed amount and submitted charges**

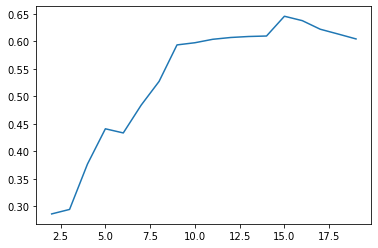
****

****

**Appendix C – Data used to decide number of clusters**

****

* Reduction in SSE for each number of clusters used to cluster the data

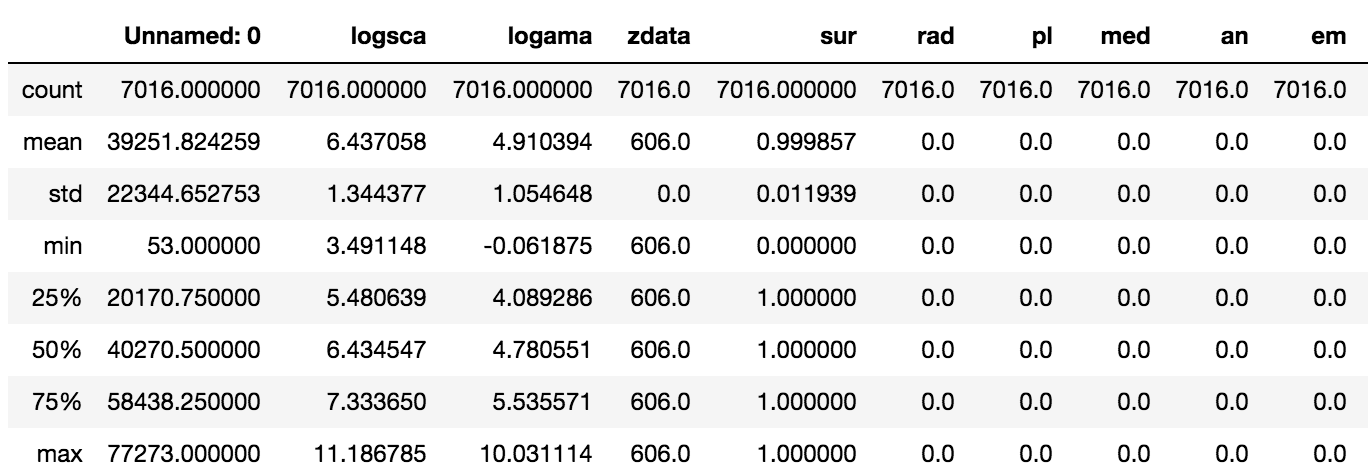
****

* Average silhouette plot used to determine optimal number of clusters at 15

[0.6923585258241877, 0.6292872018658935, 0.4323954102820523, 0.5428790915238955, 0.6146773453575007, 0.6718261255864675, 0.8748649824128635, 0.8389552962720637, 0.9210153911193065, 0.8239642038443253, 0.8799661484944578, 0.9619100077315619, 0.9766123275987769, 0.4732470219867189, 0.9944976430758533]

**-** Silhouette values for each cluster in 15 cluster solution.

**Appendix D – Cluster Zip Codes**

****

****

Each of the clusters have the same min and max 3 digit zip code (only two are shown here), letting us know each cluster corresponds to a specific location.