Clustering Analysis for Medicare Data

Executive Summary

Since Medicare was enacted in 1965, Medicare program has been helped elder people above 65 years old and disabled people to afford health costs including hospital and medical expenses. The Centers for Medicare & Medicaid Services (CMS) served as a federal agency that manages Medicare program and has been able to collect data about information related to what beneficiaries received from Medicare. To optimize the utility of Medicare program, we should take advantage of the dataset, such as analyzing characteristics that private insurance companies owned.

Clustering Analysis is an analytical strategy to group objects so that objects in the same group are similar within the same group. In Medicare program scenario, Medicare providers can have ubiquitous characteristics that can be studied, even though they might be in different locations. By doing this, we might be able to see which can be improved to increase the efficacy of the Medicare Program.

After cleaning Provider Utilization and Payment Data and running descriptive statistics on the data, we selected two features - cover ratio and average amount that Medicare providers submitted, to run the k-means algorithms on the randomly selected 1 million data. In brief, for extremely expensive services, Medicare providers are less likely to provisioned their plans accurately. However, low-cost services doesn't imply that Medicare providers provisioned their plans well enough as there is plenty of room for improvement.

Problem Statement

The data from CMS contains information about the types of services Medicare providers provided as well as how well they provided to the patients. Our job is to cluster different providers so that features and characteristics within the same group can be further studied, implying what can be improved for the Medicare Program from economic view.

Assumptions

1. There is no error in the dataset since Medicare providers should report real & accurate information to the federal agency. Meantime, CMS should also have the most updated data so that its information is the most updated final version.

2. Loading 9714898 lines into laptop is not feasible after trial and error. One of the alternatives is to randomly select a subset of the data set so that the subset is representative of the overall data. Therefore, we assumed that built-in random generator in Python is random enough so that the selected 1 million rows of data represent the total image well enough.

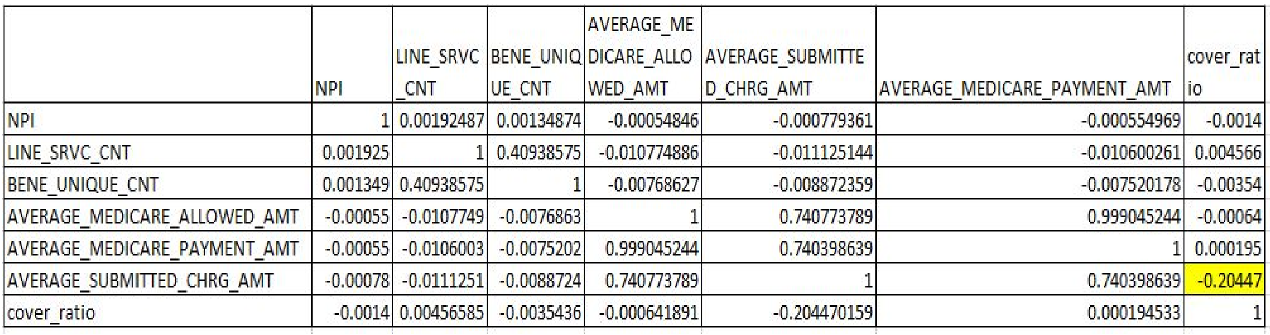
3. Standing on policy perspective, we care more about maximizing the utility of the system, which usually refers to how economically efficient the system is.

Methodology

1. After downloading data from the website, the dataset was loaded into the dataframe, using Pandas module from Python. As mentioned, loading full dataset is not feasible so 1 million data was randomly selected.
2. We first checked if there was any missing value. It turns out to be that all information is complete. We did see there was a nan value so we drop that row.
3. Following calculation is done to calculate the cover ratio:
   ‘cover_ratio’=’average_Medicare_allowed_amt’/’average_submitted_charge_amt’
   By definition, average_Medicare_allowed_amt is the amount that Medicare providers set so that this amount of money is allowed in the health care plan or the government for beneficiaries. Average_submitted_charge_amt is the real money that Medicare providers submitted for the service. Therefore, the cover ratio implies how well the Medicare providers provisioned their plans. The larger the ratio, the better predicted ability that providers have.
4. Basic descriptive statistics were examined. We ran correlation matrix and histograms to see if there is any relationship within different features.
5. Standing from ***economic*** perspective, two features really align with our interests: cover ratio as well as average_submitted_charge_amt. Other two numerical features, ‘line_srvc_cnt’ and ‘bene_unique_cnt’ won’t provide any information about how Medicare providers should improve the system in terms of economic efficiency. So we further pre-process the dataset based on this criteria. More specifically, Log transformation and standardization were conducted.
6. In order to choose how many clusters should the dataset have in terms of the above two features, scree plot and average silhouette score was calculated for each potential k. The k(=5) that gives the kink in the scree plot was chosen.
7. Additional coloring plot with 5 clusters was drawn to understand the characteristics of clusters.

Analysis

After basic data cleaning about null value, we did correlation matrix for the numerical values to observe relationships between potential features. Surprisingly, average_submitted_charge_amt is negatively correlated with the coverage ratio, which might imply Medicare provider covers those expensive bills well enough. On the other hand, it might also mean we may find out that lower-cost bills are not fully provisioned by Medicare providers.



	NPI	LINE_SRVC_CNT	BENE_UNIQUE_CNT	AVERAGE_MEDICARE_ALLOWED_AMT	AVERAGE_SUBMITTED_CHRG_AMT	AVERAGE_MEDICARE_PAYMENT_AMT	cover_ratio
NPI	1	0.00192487	0.00134874	-0.00054846	-0.000779361	-0.000554969	-0.0014
LINE_SRVC_CNT	0.001925	1	0.40938575	-0.010774886	-0.011125144	-0.010600261	0.004566
BENE_UNIQUE_CNT	0.001349	0.40938575	1	-0.00768627	-0.008872359	-0.007520178	-0.00354
AVERAGE_MEDICARE_ALLOWED_AMT	-0.00055	-0.0107749	-0.0076863	1	0.740773789	0.999045244	-0.00064
AVERAGE_MEDICARE_PAYMENT_AMT	-0.00055	-0.0106003	-0.0075202	0.999045244	0.740398639	1	0.000195
AVERAGE_SUBMITTED_CHRG_AMT	-0.00078	-0.0111251	-0.0088724	0.740773789	1	0.740398639	-0.20447
cover_ratio	-0.0014	0.00456585	-0.0035436	-0.000641891	-0.204470159	0.000194533	1

Fig 1: Numerical features' correlation matrix

Another thing to notice is that 'Line_srvc_cnt' and 'Bene_unique_cnt' have positive, relatively high correlation. As these two features can both describe the number of people who received the services, it is expected that "the more individuals come to the service, the number of services provided will increase." Even though these two features are not included in the final clustering result as mentioned above, it's nice to observe the trend as they can be included in a more general report later on.

After selecting features that align with our interests, we then pre-processed the dataset so that k-means result can be the most accurate. Specifically, we plotted distributions of two features to see the scale of the dataset. For cover_ratio, since it's already in the range of [0,1], only normality was being checked. When we plotted average_submitted_amt, it is skewed to the left so that log transformation was conducted before standardization.

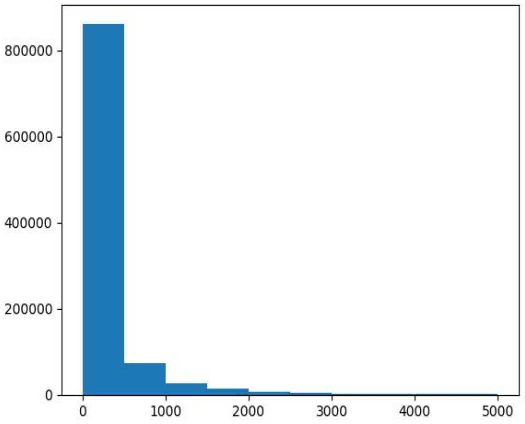


Fig 2: Average_submitted_amt before log

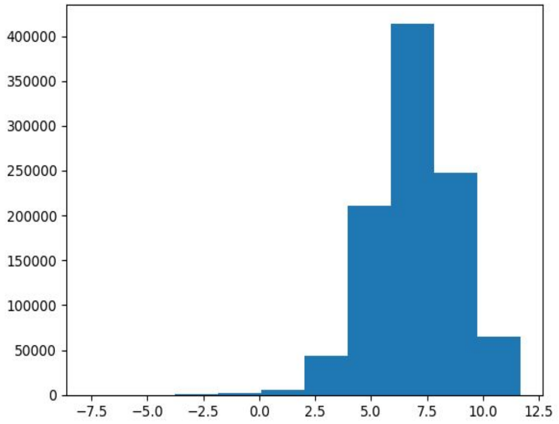


Fig 3:Average_submitted_amt after log

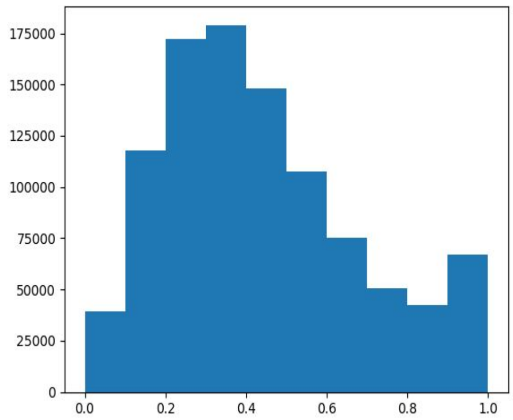


Fig 4: Cover Ratio

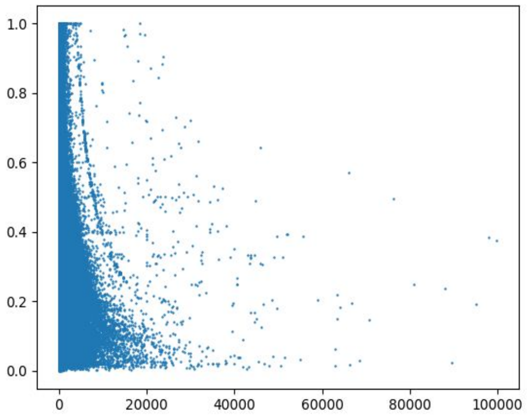


Fig 5: Scatter Plot

After preprocessing, we then used the k-means algorithm to cluster group, using Scikit Module. K was selected between the range of 2 to 9 because more clusters than 10 would be hard to interpret. Clearly the "within-group sum of squares" is always decreasing, but we need to find the sharp decrease (kink) in the graph to interpret the result.

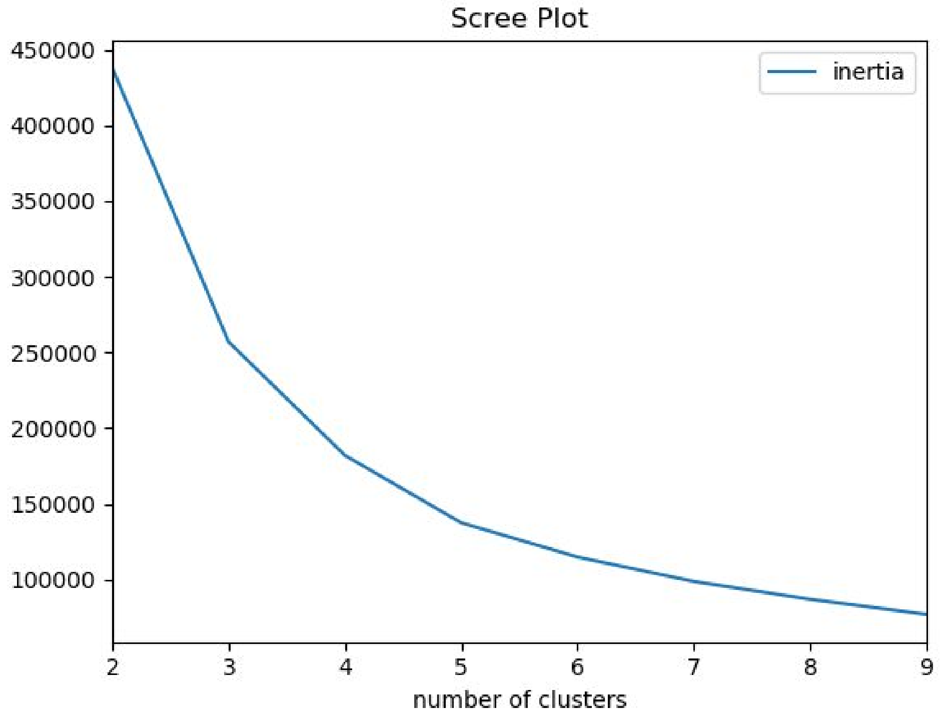


Fig 6: Scree Plot to choose best k

From the Scree plot, we chose to use k = 5 to proceed.

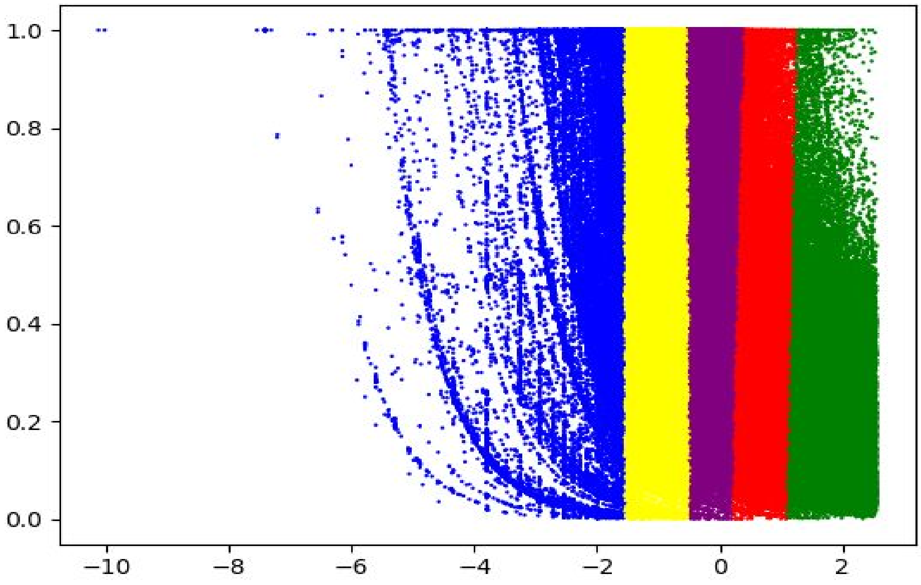


Fig 7: Original Scatter Plot that were clustered using k=5

-0.960328	0.489049
0.636937	0.392652
1.64437	0.229459
-2.15244	0.474337
-0.0705	0.526

Table 1: Corresponding Centroids positions

For interpreting results, we can see that average_submitted amt is the feature that vertically divides the cluster, as every cluster has all range of cover ratio. Furthermore, the blue cluster that has extreme low values of submitted amounted should be brought to attention since this cluster should stand for those ubiquitous services that is really common and low-fee. Another cluster to note on is the green cluster, which should stands for high-cost services group. Compare to other clusters, the cover ratio is not fully covered in the range of [0.8, 1], which does show that further improvements should be made to cover expensive services as they are extremely unaffordable.

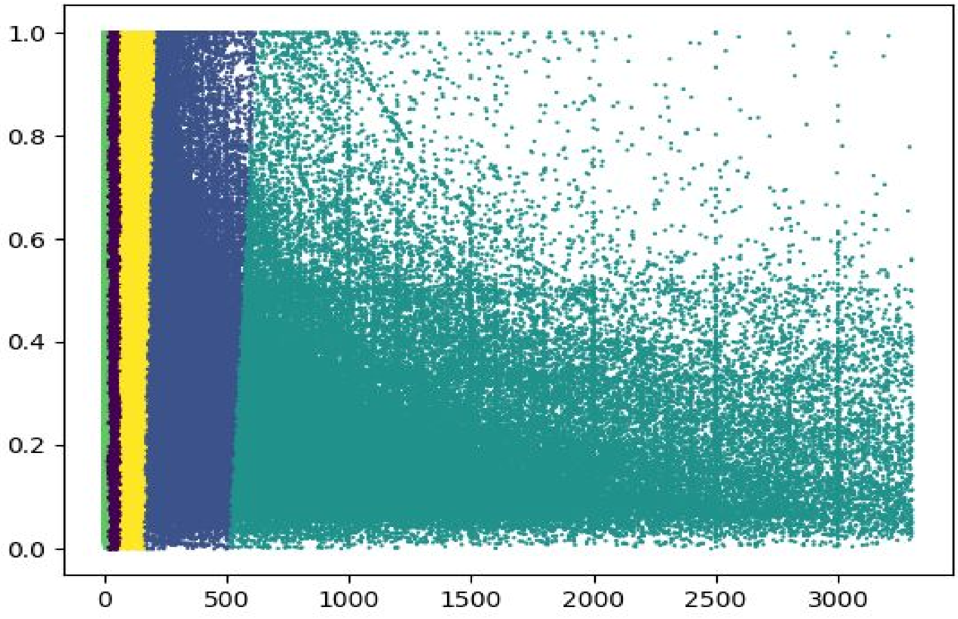


Fig 8 : Unscaled version of scatter plot with clusters

Conclusions

The key insights from using k-means to cluster the dataset are:

The capability of Medicare providers to offer services is not correlated with how costly the service is. For low-cost services, a lot of beneficiaries are still under-covered as these Medicare providers are not really good at provisioned their plans. Correspondingly, government should pay more attention to how to optimize Medicare plan coverage since current budget is not accurate for prediction.

Also, **Medicare providers are not really good at paying expensive costs for beneficiaries, such as surgeries and expensive drugs**. The lower cover ratio in the range implies that Medicare providers usually have unpredicted bills for this category. Therefore, Medicare providers in the future should consider to model this portion's variability more accurately.

Next Steps

Our next major step could be to see how clusters are being plotted on the map, implying whether there is strong correlation or not between inefficiency system and geolocations. If there is a particular region that is extremely not provisioned well, the government should dedicate to improve that region's Medicare program regulations.