# Clustering Medicare Data

## Executive Summary

Medicare is a single-payer, national social insurance program administered by the US federal government since 1966 and is currently being administered using 30-50 private insurance companies across the United States under contract. As part of Obama administration’s efforts, the Centers for Medicare & Medicaid Services (CMS) prepared a public data set, the Provider Utilization and Payment Data Physician and Other Supplier Public Use File, with information on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals. The entire Medicare program can benefit immensely from thoughtful and efficient analysis of this data to find out scope of improvements to the program.

Clustering is the task of grouping a set of objects in such a way that objects in the same group (cluster), are more similar to each other than those in other groups (clusters). In case of Provider Utilization and Payment Data, clustering was used to find cluster of frequently and ubiquitously used medical services which are not being provisioned properly. These services may be candidates for future investment by the program, since they are utilized by many people.

After aggregating Provider Utilization and Payment data to find features for per service usage and costing, the dataset was clustered using k-means algorithm and visualized by converting to 2 dimensions using PCA reduction. The segmentation of services dataset revealed a cluster which contained similar services which are being frequently used but are not being provisioned fully by the Medicare program.

## Problem Statement

The Provider Utilization and Payment Data contains data about medical services provided to the beneficiaries. The goal was to segment these services into various clusters based on existing features in the data and hybrid features and study the clusters to glean meaningful information about the services. This would in turn provide insights into how well the services are being provisioned under the Medicare program.

## Assumptions

1. There is no data-entry error in the file database. The source of the data is reliable and the information in the dataset is correct.
2. Another assumption is that if we aggregate the data of medical services across all the providers, it would make the data agnostic from geographical discrepancies, e.g., the average submitted charge (average\_submitted\_chrg\_amt) could be aggregated across all providers for a medical service to find the national average submitted charge amount.
3. There are no fraudulently high charges filed by providers or the number of non-fraudulent providers far outweigh the fraudulent ones.
4. There is no duplicate hcpcs code for the same medical service.
5. Even without other information (quantitative or qualitative) about the medical services, we can find out sufficient insights from the features present in the dataset.
6. A single beneficiary goes to the same physician/facility to avail all his/her treatments.

## Methodology

1. The first step was to download and extract the data from CMS website at: https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Physician-and-Other-Supplier.html
2. The data was analyzed for missing/nan values. Only a single row was found containing missing values and was removed from the dataset.
3. The dataset was the explored using multiple methods like histograms, correlation and general skimming through to find out feature relationships and distribution.
4. The subset of data containing 'hcpcs\_code', 'hcpcs\_description', 'line\_srvc\_cnt', 'bene\_unique\_cnt', 'average\_Medicare\_allowed\_amt', 'average\_submitted\_chrg\_amt' was then selected for solving the problem of clustering and segmenting medical services.
5. The following transformations were done to calculate total amounts per row:
6. ‘total\_Medicare\_allowed\_amt' = 'average\_Medicare\_allowed\_amt' \* 'line\_srvc\_cnt'
7. ‘total\_ submitted\_chrg\_amt ' = 'average\_ submitted\_chrg\_amt' \* 'line\_srvc\_cnt'
8. The per provider rows were then aggregated find sums of 'line\_srvc\_cnt', 'bene\_unique\_cnt', ‘total\_Medicare\_allowed\_amt' and ‘total\_ submitted\_chrg\_amt ' giving per service values.
9. Another feature of interest was added to the dataset: ‘num\_physicians\_cnt’ which was number of physicians providing the specific service.
10. The unique 'hcpcs\_description' for each 'hcpcs\_code' was then joined with the above data.
11. Finally, the following features were developed to complete the data preprocessing part:
12. ‘average\_ submitted\_chrg\_amt ' = 'total\_ submitted\_chrg\_amt' / 'line\_srvc\_cnt' (Per service average submitted charge amount)
13. 'submitted\_chrg\_to\_allowed\_amt' = 'total\_submitted\_chrg\_amt' / ‘total\_Medicare\_allowed\_amt’ (This would give a sense of provisioning for a medical service. The larger the ratio, the less the provisioning.)
14. After applying the above transformations, 5 features were selected for further exploration and analysis of medical services. The clustering problem then boils down to the following:

Objective: Cluster medical services identified by ‘hcpcs\_code’

Features:

'line\_srvc\_cnt': Total instances of services provided across all providers

'bene\_unique\_cnt': Total number of unique beneficiaries getting a specific service

‘num\_physicians\_cnt’: Total number of physicians/facilities providing a specific service

‘average\_submitted\_chrg\_amt ': Average submitted charge for a service across all providers

'submitted\_chrg\_to\_allowed\_amt': Ratio of total submitted charges to total allowed amount across all providers for a specific service

## Analysis

### Exploratory Analysis

After creating the dataset described in the above section, the first step was to perform exploratory analysis on the aggregated data.

The following table shows the results of mutual correlation between different features:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **line\_srvc\_cnt** | **bene\_unique\_cnt** | **num\_physicians\_cnt** | **submitted\_chrg\_to\_allowed\_amt** | **average\_submitted\_chrg\_amt** |
| **line\_srvc\_cnt** | 1.000000 | 0.652451 | 0.562266 | -0.003775 | -0.050029 |
| **bene\_unique\_cnt** | 0.652451 | 1.000000 | 0.890454 | -0.015823 | -0.048547 |
| **num\_physicians\_cnt** | 0.562266 | 0.890454 | 1.000000 | -0.020269 | -0.056629 |
| **submitted\_chrg\_to\_allowed\_amt** | -0.003775 | -0.015823 | -0.020269 | 1.000000 | 0.049920 |
| **average\_submitted\_chrg\_amt** | -0.050029 | -0.048547 | -0.056629 | 0.049920 | 1.000000 |

We can see that 'line\_srvc\_cnt', ‘bene\_unique\_cnt’ and ‘num\_physicians\_cnt’ have high positive correlation with each other. This is expected, as if a service is relative common, it will be provided ubiquitously, will have more unique beneficiaries and will have high service counts. Thus, one of the 3 features should really be input into the k-means algorithm, otherwise we will get biased results. All 3 features were further analyzed to find the best feature to use in k-means algorithm.

Another observation is that ‘average\_submitted\_chrg\_amt ' and 'submitted\_chrg\_to\_allowed\_amt' have very low correlation. This means that it is not necessary that more expensive treatments are necessarily being under-provisioned (which is normal expectation). Thus, we can theoretically find common, low-cost medical services which are under-provisioned by the Medicare program.

Next step was to find distribution of each of the 5 features and transform them if necessary. Each of the features seemed to be lognormal and thus for all of the features log was taken. The following results were found:

|  |  |  |
| --- | --- | --- |
| Feature | Original Distribution | Transformed Distribution |
| 'line\_srvc\_cnt' |  |  |
| ‘bene\_unique\_cnt’ |  |  |
| ‘num\_physicians \_cnt’ |  |  |
| ‘average\_submit ted\_chrg\_amt ' |  |  |
| 'submitted\_chrg\_ to\_allowed\_amt' |  |  |

The new features were added in the original aggregated dataset for easy reference to original data values. They were named by prepending the feature names by ‘l\_’. Thus, the following new features were created in the dataframe:

l\_line\_srvc\_cnt = log(line\_srvc\_cnt)

l\_bene\_unique\_cnt = log(bene\_unique\_cnt)

l\_num\_physicians\_cnt = log(num\_physicians\_cnt)

l\_submitted\_chrg\_to\_allowed\_amt = log(submitted\_chrg\_to\_allowed\_amt)

l\_average\_submitted\_chrg\_amt= log(average\_submitted\_chrg\_amt)

The correlation between the log features was again calculated. The correlation between log of 'line\_srvc\_cnt', ‘bene\_unique\_cnt’ and ‘num\_physicians\_cnt’ increased which is intuitive.

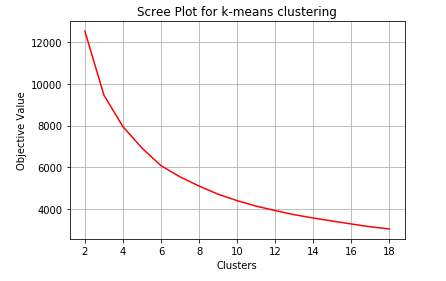
(The below steps were run 3 times, each time with one out of 'l\_line\_srvc\_cnt', ‘l\_bene\_unique\_cnt’ and ‘l\_num\_physicians\_cnt’, giving average Silhouette metric of 0.25, 0.26 and 0.277 respectively. Since ‘l\_num\_physicians\_cnt’ maximized the average Silhouette Metric, ‘l\_num\_physicians\_cnt’ was chosen for reporting.)

### Data Normalization

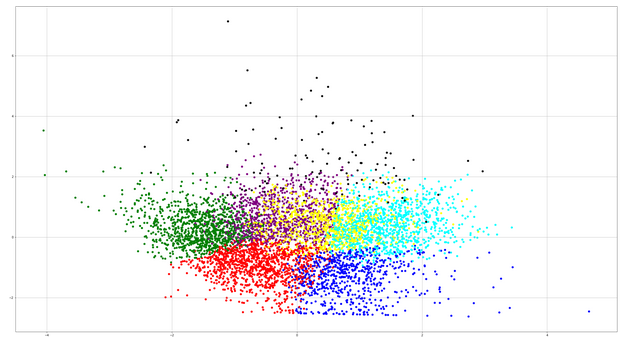
After finalizing the 3 features as ‘l\_num\_physicians\_cnt’, ‘l\_submitted\_chrg\_to\_allowed\_amt’ and ‘l\_average\_submitted\_chrg\_amt’, each of the features was normalized using Python scikit built in function – Scale. The data was extracted into an array of feature vectors to be fed into clustering algorithm.

### KMeans

Next step was to pass the feature array to Python scikit K-Means function. The chosen init method was k-means++, which runs the k-means algorithm n\_init (=10) times to find the best centroids based on objective values. K-means was run in a loop for k=2 to k=19. The resulting objective function values (inertia) was used to create a scree plot.



The kink in the scree plot seemed to occur at k=7. Thus k=7 was selected for post-processing steps. The resulting clusters were plotted by reducing the feature array into 2-dimensions using PCA analysis.



*Plot showing results of clustering reducing the feature set to 2-dimensions using PCA.*

### Post Processing

(Rest of the observations in the report are for a particular run of k-means. It produces slightly different results on each run.)

The centroids for the clusters were found to be:

['l\_average\_submitted\_chrg\_amt','l\_submitted\_chrg\_to\_allowed\_amt','l\_num\_physicians\_cnt']

**[[-0.85481101 0.51473993 0.10190296]**

[ 0.50026857 -0.60464561 -0.84080352]

**[ 1.01834327 0.97049089 -0.80040762]**

[-1.06419291 -1.1762968 -0.63088168]

**[ 0.7722547 0.31765191 0.69164175]**

[-0.64357404 -0.48672911 1.37093911]

**[-1.62504898 2.82977256 0.55764482]]**

Of these, the red highlighted cluster centroid was of particular interest as it had moderately high value for each of the 3 features ‘l\_num\_physicians\_cnt’, ‘l\_submitted\_chrg\_to\_allowed\_amt’ and ‘l\_average\_submitted\_chrg\_amt’. This translates directly to medical services which are ubiquitous, have moderately high prices and are under-provisioned.

The aggregated data was subset to find out the data associated with the selected cluster and published as the output of the analysis – services\_to\_look.csv.

The green highlighted centroids also represent medical services which are under-provisioned ('l\_num\_physicians\_cnt' is less), but these either are not being utilized by many people (which may mean that they are either specialized treatments or cosmetic treatments) or they are very inexpensive ('l\_average\_submitted\_chrg\_amt' is less), which are affordable to most people.

## Conclusions

The key insights obtained from the cluster analysis of medical service data are:

**There are ubiquitous, moderately high-priced medical services which are not properly provisioned under the Medicare program**. These services are provided quite frequently and should be under the Government’s radar for either better provisioning or optimizing the way in which they are delivered to the beneficiaries.

**There are medical services even lesser provisioned than the ones reported by the analysis** but are either not being utilized by many people (which may mean that they are either specialized treatments or cosmetic treatments) or they are very inexpensive, which are affordable to most people.

## Next Steps

The next step in the analysis will be to take the output data and validate that the data corresponds to the moderately high-priced ubiquitous medical services which are not being properly provisioned. It can further be looked at for finding the exact cause of the situation. One step further will be to actually take this data to CMS and understand their reasons for not optimizing the program for these services which incur a huge medical bill for the beneficiaries frequently.