Medicare Clustering Analysis Report

Rosanne Harrison

Executive Summary

This analysis utilized data sourced from The Centers for Medicare and Medicaid Services (CMS). It is a public data set called the 'Provider Utilization and Payment Data Physician and Other Supplier Public Use File', abbreviated as 'Physician and Other Supplier PUF'.

The analytical tool applied in the following analysis is clustering, a method for parsing large data sets into groups (clusters) of data points with similar characteristics. These clusters can then be further analyzed to identify trends or anomalies within clusters, or unexpected differences between clusters. In the case of the Medicare data set, clustering is a useful method for understanding which data points contribute most to overall Medicare spend in the US. By creating clusters based on variables that indicate overall Medicare costs, we are able to analyze smaller groups of data points and identify their shared characteristics (specifically their provider types) and thus draw overarching conclusions about the data in each cluster.

After exploring, cleaning and concatenating the data to prepare it for analysis, I went through a variable selection process to choose 'bene_unique_cnt', the count of Medicare beneficiaries, and average_Medicare_standard_amt as the best variables to perform clustering with. The k-means clustering algorithm was then applied to all remaining data points, and after generating and analyzing a scree plot, 5 final clusters were created.

In analyzing the final clusters, I was able to draw several conclusions about which clusters (and thus which provider types) contribute most to Medicare spending in the US. Each of the five clusters had different characteristics that allowed them to be classified (i.e. highest dollar spend on Medicare but fewest beneficiaries, medium dollar spend on Medicare but very high number of beneficiaries, etc.) As I had hoped, many provider types were only present in one or two clusters. As such, I was able to apply these overarching conclusions about the clusters to the contained provider types, leading to valuable information about which provider types contribute most to overall Medicare costs.

Problem Statement

I initially kept my analysis broad and sought to identify a categorical splitter of Medicare costs for the Social Security Administration. That is, I wanted to see if clustering could split the data into groups that cost the Medicare program significantly different amounts of money, and then identify categorical variables that differed between clusters. This analysis would allow me to identify which category contributes most to differences in cost to the Medicare program.

Once this category (provider type) was identified, my clustering analysis focused on identifying the provider types that most contribute to Medicare spending in the US.

Assumptions

- No errors or inaccuracies in data set, except for rows with obvious errors from loading data set into R program (i.e. rows where data was not tab delimited)
- The removal of rows with 1) obvious data loading errors or 2) missing values in any of the numeric columns (last seven columns of data set) could be removed without impacting the validity of results.

- Data can be concatenated among providers that are identical based on broad categorical variables (i.e. gender, state, provider type, etc.) Patients examined and services offered by these providers should be nearly identical, and thus combining these data points to shrink the data set will not significantly impact overall results.

Methodology

My first step was to pull the 'Physician and Other Supplier PUF' data set into an R data frame. Although the data set contains over 4.7 million rows, I was able to process the entire data set into R, using the 26 column names listed in the original data set.

The next step involved exploring and cleaning the data set. When parsing through the first several thousand rows, I began noticing abnormalities and errors within the data. Upon further inspection I realized that these errors were due to rows that had been incorrectly entered into the original data set – namely, rows that hadn't been correctly tab delimited.

In order to clean and prepare the data set for analysis, I first pulled data from only the 50 US states. I decided that based on my problem statement, I only wanted to analyze Medicare transactions made within the 50 states. I then transformed any non-numeric values in the seven numeric columns into NAs, and then eliminated all rows with NAs in any of those columns. This served the dual purpose of eliminating all rows that been loaded incorrectly, and also any rows that had missing data in the columns required for my analysis. This shrunk my data set from about 4.72 million to 4.69 million data points.

At this stage I also examined any remaining 'outliers'. However, after inspection I decided not to remove any more data points. All the rows I examined that contained extreme values were loaded consistently and did not appear to be erroneous; instead they were just legitimate large data points. Since my business problem relates to identifying categories that contribute the greatest amounts to Medicare spend, I decided these large data points were important to my analysis.

My final step before beginning clustering analysis was concatenating the data to shrink the overall size of the data set. I combined the data on the following seven variables: "nppes_credentials", "nppes_provider_gender", "nppes_entity_code", "nppes_provider_state", "provider_type", "medicare_participation_indicator" and "place_of_service". For the seven numeric columns, I took the mean of all data points within each new row. I chose to take the mean (as opposed to the sum) of each column because I believed it would contain less bias and create fewer extreme outliers (ie. control for population of states, types of medical service that are more commonly used, etc.) This shrunk my data from 4.69 million rows to 63,384 rows. This concatenation made the data set a much more manageable size while still containing a very large number of points and maintaining integrity of data.

Analysis

I began my analysis by exploring the combined data to look for indicators of potential categorical splitters. To do this, I looked a variety of plots and did some preliminary clustering.

I first investigated splitting by state. My hypothesis had been that some states would overall spend more on Medicare, based on the overall health of their residents. However, I found that the only observable difference was population-based (states with larger populations had highest Medicare spend), and once I controlled for population the means across states for most numeric variables were generally similar.

I next attempted splitting by provider type; this led to much more promising results. Boxplots of many numeric variables, including the two I eventually selected, showed significantly different means across provider types. This also made logical sense, and therefore I decided to move forward with clustering.

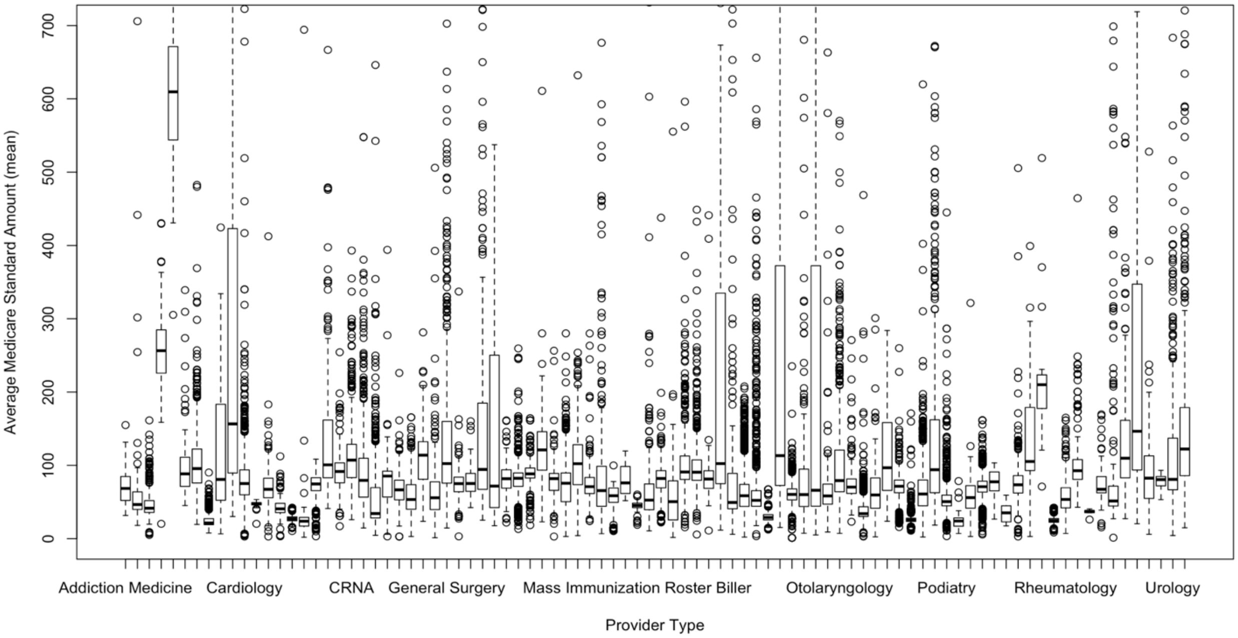

Figure 1: Boxplot of Average Medicare Standard Amount (mean) across Provider Types

I selected my variables for clustering based on the business problem I had posed. I wanted to only include variables that would indicate which data points contribute most (and least) to how much the Social Security Administration spends on Medicare. I tried a variety of variables in different combinations, including all relevant options in the data set and some that I created myself. One that I created but eventually removed was ‘percent coverage’, which indicated what percent of a submitted charge was covered by Medicare. In the end I decided to use the count of Medicare beneficiaries (bene_unique_cnt) and the standardized average Medicare dollars spent on the service (average_Medicare_standard_amt) – I chose the standardized amount in order to minimize any geographic bias in the amounts. Those two factors combined seem to cover the full explanation of why a particular service would be expensive for the SSA.

Before beginning clustering, I also ensured that all my variables had been standardized. In order to do this, I subtracted the minimum and divided by the maximum value in each variable column.

I began my clustering analysis by calculating the WSS value for clustering analyses with 1-10 clusters to select the optimal number of clusters. The scree plot below shows the results of the WSS calculations:

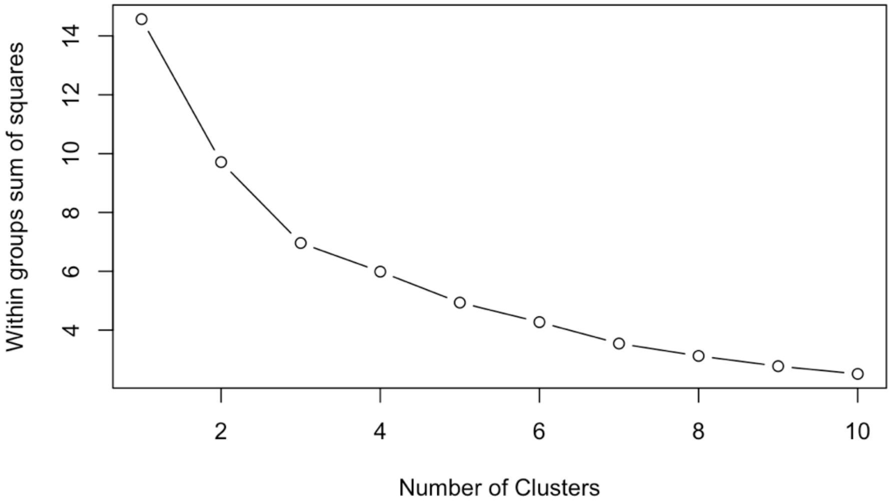

Figure 2: Scree Plot for 1 through 10 clusters

Based on the above scree plot, I decided to perform my final clustering with 5 clusters. I initially struggled to decide between 3 and 5 clusters; the first obvious kink in the graph is at three clusters, but the slope of the line between 4 and 5 clusters is also steeper than most other slopes. I initially chose to use three clusters (based on the drop between 3 and 4 clusters being less than 20% of the previous drop). However, after examining clustering results using both three and five clusters, I decided to move forward with five, as the clusters made more sense and drove more valuable results.

The clustering algorithm I applied was k-means clustering. I felt this was the most appropriate given the size of my data set, and given that I did not need my cluster centers to be actual data points (which would have required k-medoids).

An overview of the summary statistics from my clustering analysis can be seen below:

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Total**
Within Cluster SS	0.356	1.152	0.506	2.053	0.958	**5.027**
Between Cluster SS	---	---	---	---	---	**9.542**
Size	7	43563	126	1644	18044	**---**

As we can see from the above statistics, the clusters vary pretty significantly with regards to their within cluster sum of squares. We can attribute much of this to similarly wide variety of cluster sizes; however, the WCSS does not perfectly correlate with the sizes of the clusters, indicating that some clusters are stronger than others.

The extremely variable sizes of the clusters indicate that some (particularly clusters 1 and 3) could potentially contain outliers. While this may be true, clusters 1 and 3 also have, by far, the two highest center values for number of beneficiaries; this information could still be valuable with regards to my business problem, which is why I feel comfortable with their inclusion.

In order to understand the general trends for each cluster, I ranked centers (for both variables) of each the five clusters by which are highest/lowest in both clustering categories. As I had hoped, the cluster that contains most of the data (over 43,000 data points) is low in both beneficiaries and amount of Medicare dollars spent. This allows us to study the four other clusters – specifically their Provider Types – which contain either high numbers of beneficiaries or large Medicare dollar amounts.

The table below shows the ranking of the clusters' center values, for both clustering variables, from lowest (1) to highest (5).

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
Count Unique Beneficiaries	5	2	4	1	3
Average Medicare $$ Amount	1	2	3	5	4

We see that clusters 1 and 3 contain the data points that have the two highest numbers of unique beneficiaries, and clusters 4 and 5 contain the highest average amounts of spend on Medicare. It is unsurprising that the clusters that are highest for one variable are lowest for the other; it makes logical sense that the most expensive medical treatments are received by fewer people, and that the medical services provided to many people are routine and thus not very costly.

Conclusions

There are several key business insights that we can take away from this clustering analysis:

The services offered by some specific provider types are very clearly large contributors to Medicare spend. 100% of "Ambulatory Surgical Center" spend fell into the cluster with the highest average Medicare payment amount, and 100% of "Slide Preparation Facility" data points fell into the second highest cluster regarding number of beneficiaries.

Many other provider types are spread between several clusters, but clear trends are still obvious. The two clusters that contain the highest average Medicare amounts encompass the large majority of data points from providers of surgeries, including "Cardiac Surgery", "Neurosurgery", "Plastic and Reconstructive Surgery", "Thoracic Surgery" and "Vascular Surgery", among others.

We can also draw conclusions about which provider types do NOT contribute large amounts to Medicare spend as compared to other providers. The cluster that contains (on average) low numbers of beneficiaries and low Medicare spend includes 100% of data points for "Centralized Flu", "Registered Dietician/Nutrition Professional", "Chiropractic" and "Occupational therapy, among others.

To see the distributions of each Provider Type among all five clusters, please see the table in the Appendix.

Next Steps

Undoubtedly there remains a lot of variability within and between the clusters; very few provider types are contained in only one cluster, and many are spread out between 3+ clusters. It is likely that further analysis could lead to more conclusive results. There are several potential next steps that would allow this analysis to be even more valuable.

Firstly, it would be helpful to objectively measure the goodness-of-fit of the clustering model I developed; the best way to do this would likely be with a silhouette plot and calculations. Although my computer and lab computers were unable to calculate this value, likely due to the size of the data set and limited computing power, a more powerful computer could easily make this calculation.

Similarly, more computing power would allow us to carry out a similar analysis, but instead cluster all 4.6 million data points individually. This would allow us to validate our results with data that hasn't been manipulated, and would likely provide granular insights for provider types that show up many times in the data set.

Finally, it could be interesting to search for demographic categorical splitters beyond provider type. One potential area to investigate may be HPPCS codes, which identify the specific procedures and services that Medicare beneficiaries receive. By diving deeper into these individual services we would be able to separate between the many services offered by one provider, and thus create more accurate and informative clusters.

Appendix

Provider Type	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Addiction Medicine	0%	73%	0%	0%	27%
All Other Suppliers	0%	76%	0%	11%	13%
Allergy/Immunology	0%	95%	0%	0%	5%
Ambulance Service Supplier	0%	2%	30%	15%	53%
Ambulatory Surgical Center	0%	0%	0%	100%	0%
Anesthesiologist Assistants	0%	46%	0%	1%	53%
Anesthesiology	0%	36%	0%	1%	63%
Audiologist (billing independently)	0%	100%	0%	0%	0%
Cardiac Electrophysiology	0%	50%	0%	1%	49%
Cardiac Surgery	0%	22%	0%	39%	39%
Cardiology	0%	53%	0%	1%	46%
Centralized Flu	0%	100%	0%	0%	0%
Certified Clinical Nurse Specialist	0%	77%	0%	0%	23%
Certified Nurse Midwife	0%	98%	0%	0%	2%
Chiropractic	0%	100%	0%	0%	0%
Clinical Laboratory	2%	47%	48%	1%	1%
Clinical Psychologist	0%	78%	0%	0%	22%
Colorectal Surgery (formerly proctology)	0%	29%	0%	4%	67%
Critical Care (Intensivists)	0%	37%	0%	0%	63%
CRNA	0%	27%	0%	0%	72%
Dermatology	0%	48%	0%	1%	50%
Diagnostic Radiology	0%	81%	0%	1%	18%
Emergency Medicine	0%	39%	0%	0%	61%
Endocrinology	0%	78%	0%	0%	22%
Family Practice	0%	87%	0%	0%	13%
Gastroenterology	0%	25%	0%	0%	75%
General Practice	0%	80%	0%	0%	20%
General Surgery	0%	37%	0%	5%	58%
Geriatric Medicine	0%	69%	0%	0%	31%
Geriatric Psychiatry	0%	69%	0%	0%	31%
Gynecological/Oncology	0%	42%	0%	13%	44%
Hand Surgery	0%	53%	0%	15%	32%
Hematology	0%	56%	0%	0%	44%
Hematology/Oncology	0%	58%	0%	0%	42%
Hospice and Palliative Care	0%	33%	0%	0%	67%
Independent Diagnostic Testing Facility	0%	23%	0%	1%	75%

Infectious Disease	0%	**54%**	**0%**	0%	**46%**
Internal Medicine	0%	**59%**	**0%**	0%	**41%**
Interventional Cardiology	0%	**30%**	0%	**0%**	**70%**
Interventional Pain Management	0%	**70%**	0%	0%	**30%**
Interventional Radiology	0%	**68%**	0%	**7%**	**25%**
Licensed Clinical Social Worker	0%	**100%**	0%	0%	**0%**
Mammographic Screening Center	0%	**50%**	**17%**	0%	**33%**
Mass Immunization Roster Biller	0%	**98%**	0%	0%	**2%**
Maxillofacial Surgery	0%	**81%**	0%	**2%**	**17%**
Medical Oncology	0%	**58%**	0%	**0%**	**42%**
Multispecialty Clinic/Group Practice	**4%**	**67%**	**6%**	**2%**	**22%**
Nephrology	0%	**25%**	0%	**0%**	**74%**
Neurology	0%	**35%**	0%	**0%**	**65%**
Neuropsychiatry	0%	**58%**	0%	**3%**	**39%**
Neurosurgery	0%	**35%**	0%	**29%**	**37%**
Nuclear Medicine	0%	**72%**	0%	**7%**	**21%**
Nurse Practitioner	0%	**84%**	**0%**	0%	**16%**
Obstetrics/Gynecology	0%	**84%**	0%	**2%**	**14%**
Occupational therapist	0%	**100%**	0%	0%	0%
Ophthalmology	0%	**30%**	**0%**	**32%**	**39%**
Optometry	0%	**91%**	0%	0%	**9%**
Oral Surgery (dentists only)	0%	**71%**	0%	**4%**	**25%**
Orthopedic Surgery	0%	**55%**	0%	**32%**	**13%**
Osteopathic Manipulative Medicine	0%	**79%**	0%	**2%**	**19%**
Otolaryngology	0%	**52%**	0%	**3%**	**45%**
Pain Management	0%	**76%**	0%	0%	**24%**
Pathology	0%	**96%**	**2%**	**0%**	**2%**
Pediatric Medicine	0%	**79%**	0%	**0%**	**21%**
Peripheral Vascular Disease	0%	**42%**	0%	**2%**	**56%**
Physical Medicine and Rehabilitation	0%	**74%**	**0%**	0%	**26%**
Physical Therapist	0%	**100%**	**0%**	0%	**0%**
Physician Assistant	0%	**80%**	**0%**	**0%**	**19%**
Plastic and Reconstructive Surgery	0%	**45%**	0%	**9%**	**46%**
Podiatry	0%	**91%**	0%	**0%**	**9%**
Portable X-ray	**3%**	**93%**	**4%**	0%	0%
Preventive Medicine	0%	**84%**	**1%**	**1%**	**14%**
Psychiatry	0%	**82%**	0%	**0%**	**18%**
Psychologist (billing independently)	0%	**63%**	0%	0%	**37%**
Public Health Welfare Agency	0%	**96%**	**4%**	0%	0%

Pulmonary Disease	0%	**65%**	0%	**0%**	**35%**
Radiation Oncology	0%	**16%**	0%	**1%**	**83%**
Radiation Therapy	0%	**7%**	0%	**20%**	**73%**
Registered Dietician/Nutrition Professional	0%	**100%**	**0%**	0%	0%
Rheumatology	0%	**86%**	**0%**	0%	**14%**
Sleep Medicine	0%	**29%**	0%	**1%**	**70%**
Slide Preparation Facility	0%	0%	**100%**	0%	0%
Speech Language Pathologist	0%	**78%**	0%	0%	**22%**
Sports Medicine	0%	**80%**	0%	**10%**	**10%**
Surgical Oncology	0%	**27%**	0%	**5%**	**68%**
Thoracic Surgery	0%	**18%**	0%	**31%**	**51%**
Unknown Physician Specialty Code	**1%**	**51%**	**1%**	**3%**	**44%**
Unknown Supplier/Provider	0%	**56%**	0%	0%	**44%**
Urology	0%	**47%**	0%	**3%**	**50%**
Vascular Surgery	0%	**24%**	0%	**9%**	**67%**