Northwestern University

Medicare Provider Utilization & Payment Data Clustering with K-Means

Aarij Rehman
January 27, 2020

Exploratory Data Analysis

After cleaning the data, we can inspect the histograms of various numerical features to see where there may be patterns of interest.

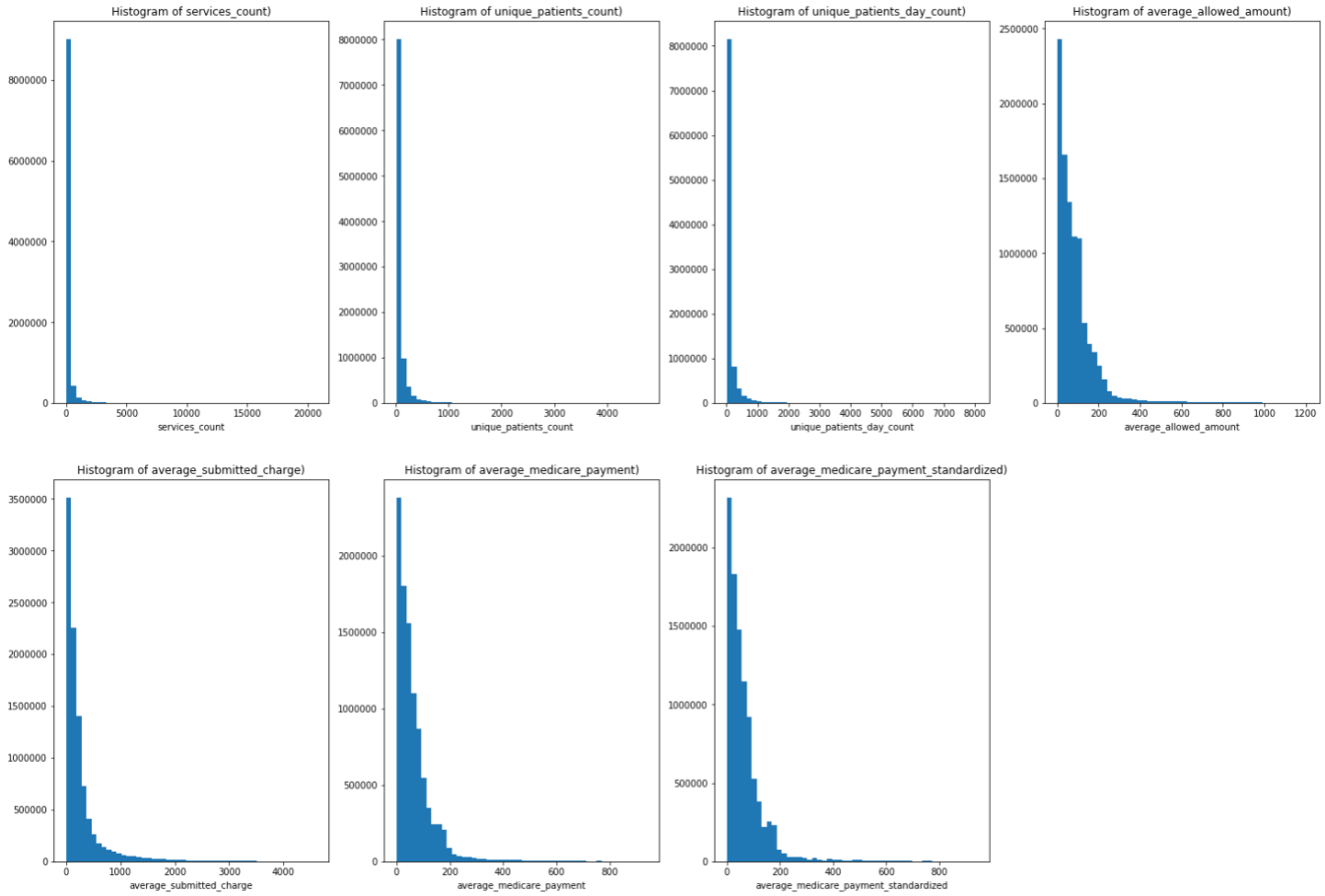

Figure 1: Histograms of Features

Looking at the plots suggests taking a logarithm would be an appropriate transformation to normality. Doing so produced the following the results:

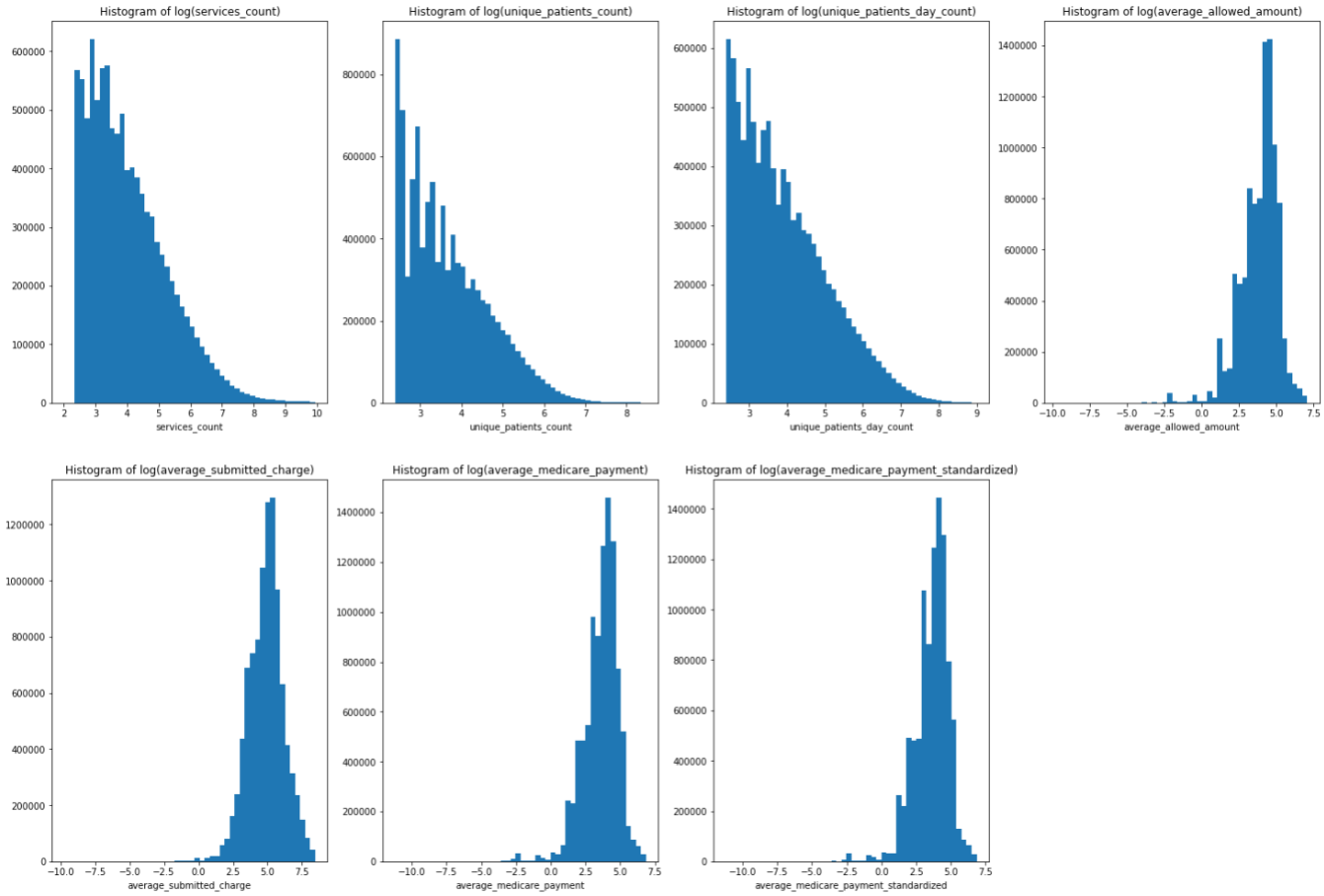

Figure 2: Logarithmic Histograms

Of our seven features, four display a gaussian distribution. These features are

1) Average Allowed Amount – *Average of the Medicare allowed amount for the service; this figure is the sum of the amount Medicare pays, the deductible and coinsurance amounts that the beneficiary is responsible for paying, and any amounts that a third party is responsible for paying*.

2) Average Submitted Charge - *Average of the charges that the provider submitted for the service*

3) Average Medicare Payment - *Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service*

4) Average Medicare Payment Standardized - *Average amount that Medicare paid after beneficiary deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied. Standardization removes geographic differences in payment rates for individual services, such as those that account for local wages or input prices and makes Medicare payments across geographic areas comparable, so that differences reflect variation in factors such as physicians' practice patterns and beneficiaries' ability and willingness to obtain care.*

After analyzing the histograms, we can consider correlations between different features. Below, a correlation heatmap shows correlation coefficients for different features. The labels have been replaced with letters for spacing purposes.

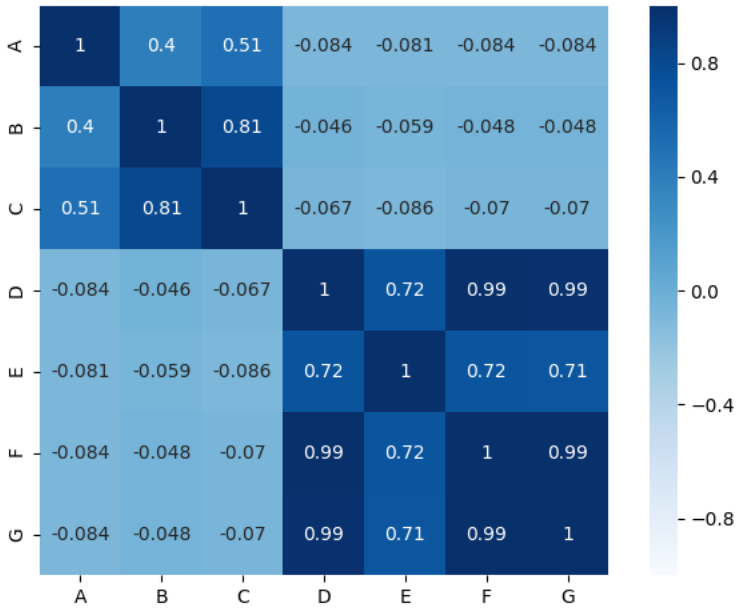

Figure 3: Correlation Heatmap

The labels correspond with features as follows:

A- Services Count
B- Unique Patients Count
C- Unique Patients per Day Count
D- Average Allowed Amount
E- Average Submitted Charge
F- Average Medicare Payment
G- Average Medicare Payment Standardized

The results confirm some intuition regarding the data like how higher submitted charges correlate strongly with higher Medicare payments. We also discover that certain things like service counts don't have a significant correlation with Medicare payments, suggesting that Medicare doesn't discriminate against doctors with less activity.

It's interesting that there is *no* significant correlation between the payments and charges with the number of patients or services a provider has. This confirms Medicare doesn't consider patient volume in pricing, something private insurance agencies likely do differently. From a funding perspective, this may offer further rational for supporting Medicare; however, by not offering any form of volume discount, Medicare may be influencing certain providers to not accept their insurance.

Clustering

In the context of our business question:

How do different provider types differ in terms of their unique patients and out-of-pocket costs for those patients?

We consider three features:

1) Average Allowed Amount
2) Average Medicare Payment
3) Unique Patients for a Service

We can take the difference between the Average Allowed Amount and the Average Medicare Payment to see how much the average out-of-pocket cost for a service is, leaving us with two features.

1) Average Out-of-pocket Cost
2) Unique Patients for a Service

With the K-Means algorithm, the most important consideration is the number of clusters to use. This scree plot charts the decrease of the sum of squared errors with respect to the number of clusters.

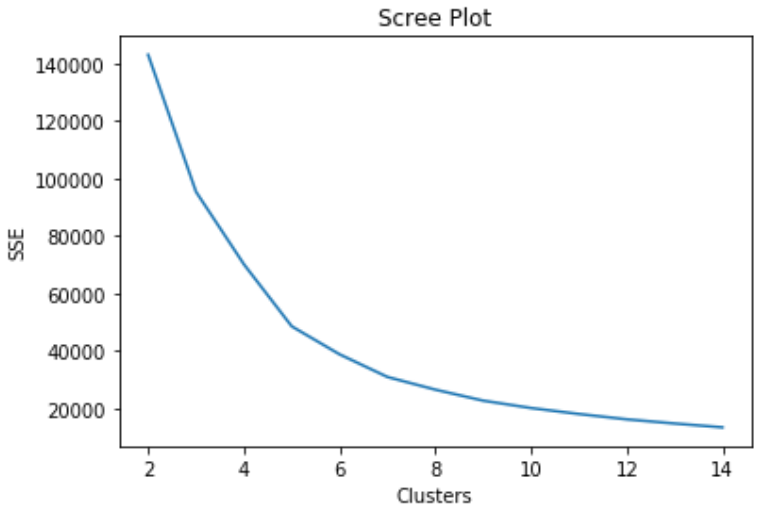

Figure 4: Scree Plot

The 'Elbow Method' suggests choosing a number of clusters where there begins only a small marginal decrease of SSE. However, due to the subjectivity of the Elbow Method, silhouette scores can offer a confirmation on the number of dimensions.

So here, we graph the average silhouette score for a number of clusters, looking for higher values.

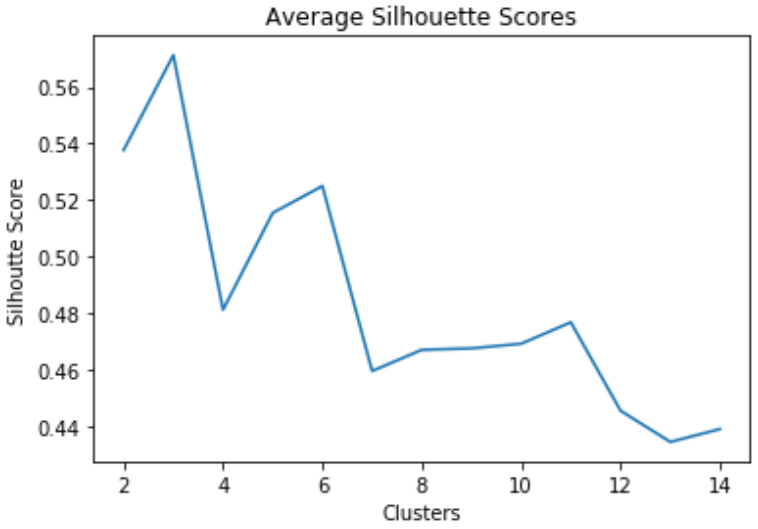

Figure 5: Average Silhouette Score

The plot above suggests using three clusters as opposed to using five or six, which is what the scree plot shows. Because of the conflicting information presented, we should observe the actual silhouette scores for each cluster size. Furthermore, since we were able to reduce our number of features to just two, we can plot the data points to get an intuitive understanding of the clusters.

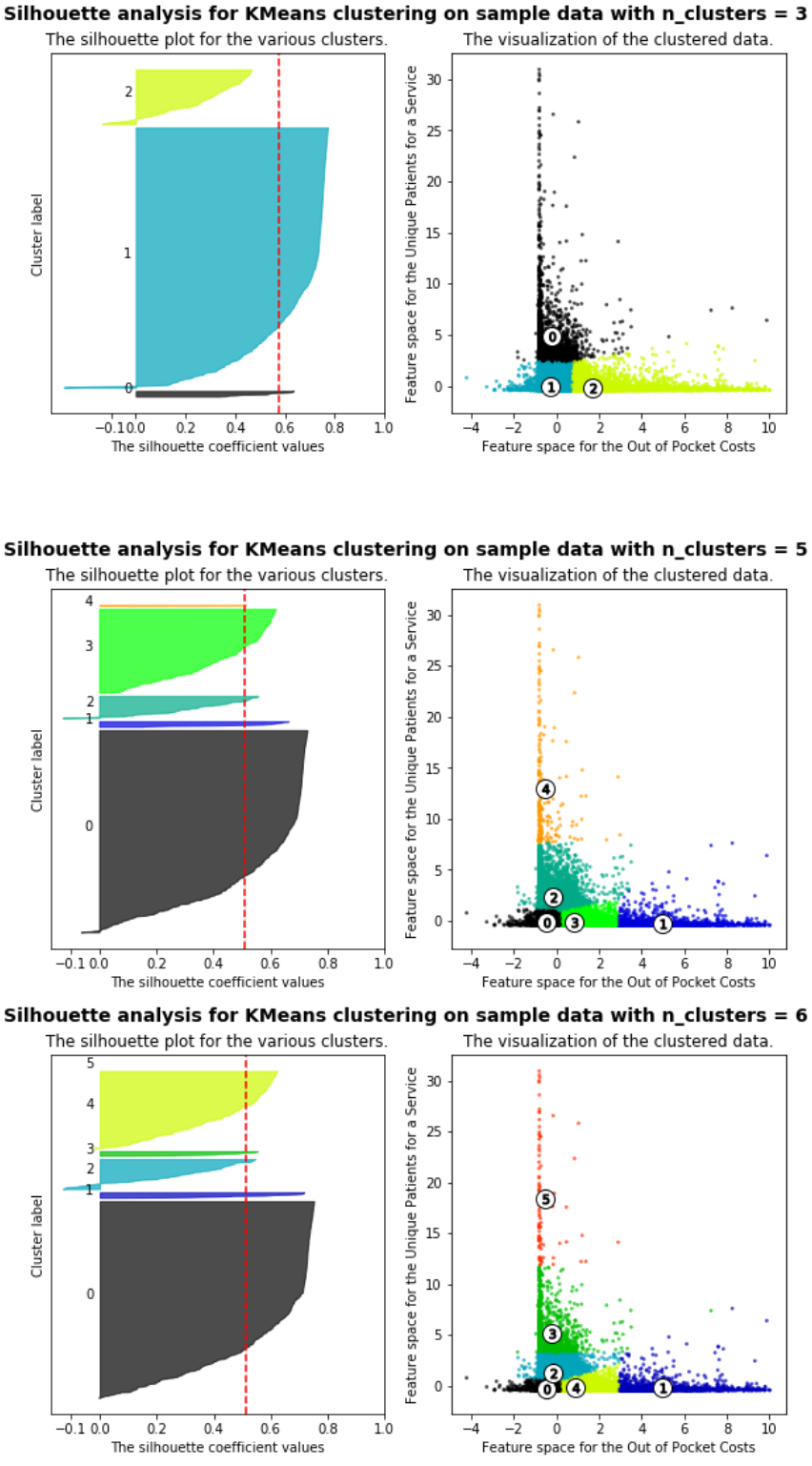

Figure 6: Silhouette Values with Feature Space

The plots here offer a confirmation that three clusters would be most appropriate. Although adding more clusters would decrease SSE, it's clear that in doing so, we reduce the overall quality of our segmentation.

Insights

When running K-Means with three clusters, we find that there are three distinct groups of provider experiences. A visual depiction and a table below show certain key features.

1) Cluster 0: High Unique Patients/Service & Low Out-of-pocket Costs
2) Cluster 1: Low Unique Patients/Service & Low Out-of-pocket Costs
3) Cluster 2: Low Unique Patients/Service & High Out-of-pocket Costs

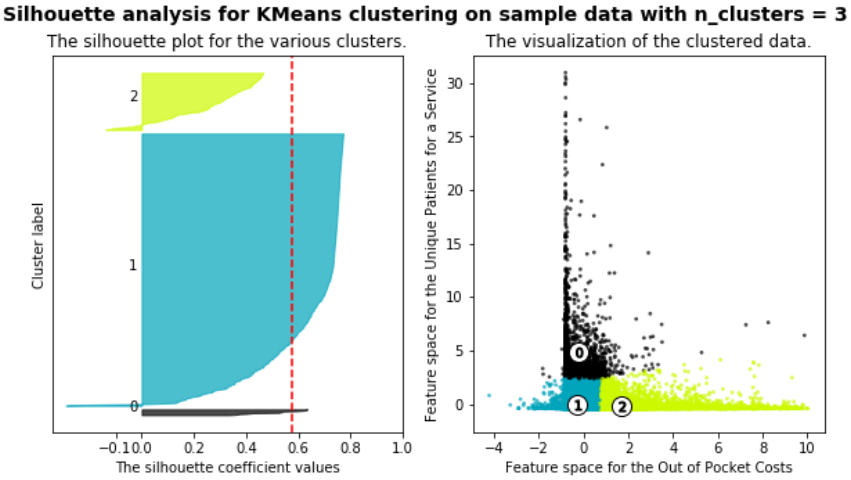

Figure 7: Three Cluster Mapping

Table 1: Cluster Demographics

Cluster	unique_patients_count	pocket	services_count	average_submitted_charge
0	788.773067	15.365776	1356.207930	175.378656
1	60.641746	13.032542	153.643996	186.932482
2	55.926073	63.827838	74.585626	812.606119

Table 2: Relative Frequency of Samples in a Cluster

1	0.826932
2	0.152467
0	0.020601

Clearly, there is a trend between what providers in how many patients they service and how much the out-of-pocket cost becomes for those patients. We can look at how different provider types are represented in each of the clusters to get a better idea of valuable business insight.

Table 3: Top 8 Provider Types for Cluster 0

Diagnostic Radiology	0.157606
Clinical Laboratory	0.116209
Cardiology	0.097756
Ophthalmology	0.084289
Internal Medicine	0.082294
Dermatology	0.059850
Urology	0.042893
Family Practice	0.041895

Table 3: Top 8 Provider Types for Cluster 1

Diagnostic Radiology	0.139435
Internal Medicine	0.116511
Family Practice	0.107974
Nurse Practitioner	0.064399
Physician Assistant	0.042096
Cardiology	0.041015
Physical Therapist in Private Practice	0.033896
Orthopedic Surgery	0.028615

Table 3: Top 8 Provider Types for Cluster 2

Internal Medicine	0.109711
Cardiology	0.062268
Family Practice	0.061460
Diagnostic Radiology	0.056405
Ophthalmology	0.050812
Emergency Medicine	0.042725
Gastroenterology	0.038075
Anesthesiology	0.037401

When looking at the 3 tables, certain patterns are immediately visible. *Cluster 0* has 11% of its samples coming from Clinical Laboratories, whereas no other cluster has Clinical Labs in

its top 8. It makes sense that these provider types have many unique patients and a low out-of-pocket cost because they likely offer simple services. From a civil perspective, Medicare likely should continue seeking out Clinical Laboratories because of how much they offer to their local communities. They keep costs low and service many different people, contributing to much of the entire purpose behind Medicare.

When looking at *Cluster 1*, we see the presence of the Nurse Practitioner and the Physician's Assistant. The less skilled medical professions have low unique patient counts and low out-of-pocket costs. Although these professions don't cater to as many beneficiaries as the providers in *Cluster 0*, we still see that they offer meaningful value.

However, from a business perspective, it makes sense to defund these groups on Medicare relative to other clusters. Since they don't serve as many patients, and the services they do are already so cheap, it may be worthwhile to move funding elsewhere to have a larger impact.

Lastly, when looking at *Cluster 2*, we see the providers that serve few patients but require the highest out-of-pocket costs. Furthermore, we see that they have significantly higher submitted charges than other clusters. Understandably, we see some of the most specialized provider types in this cluster like Gastroenterology and Anesthesiology. These highly specialized provider types require large payments from their few customers.

Because of the specialization of the different provider types in this cluster, it's important that Medicare continue funding here. Without this group, many patients would have to rely on outside insurance to meet their health requirements.

Alternatively, insurance agencies may have the most to gain by convincing the provider types in *Cluster 2* to accept their insurance. Since they are so specialized and charge such high amounts, insurance agencies would be able to make the most money from these providers.

Ultimately, the data offers valuable insights both in the form of civil and business applications. These insights can both be applied to help government efforts and increase profits for the private sector.