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.

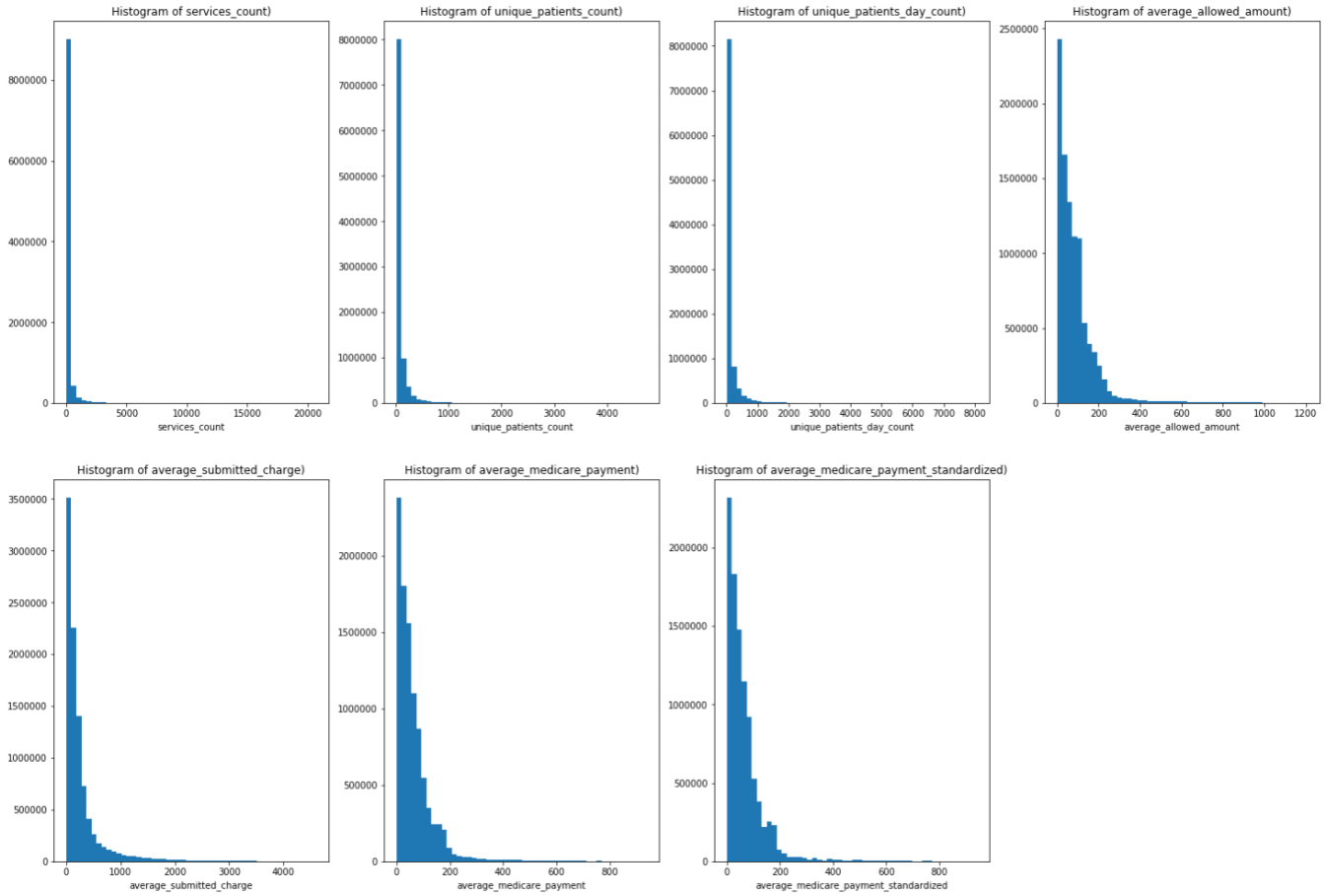

Figure 1: Histograms of Features

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

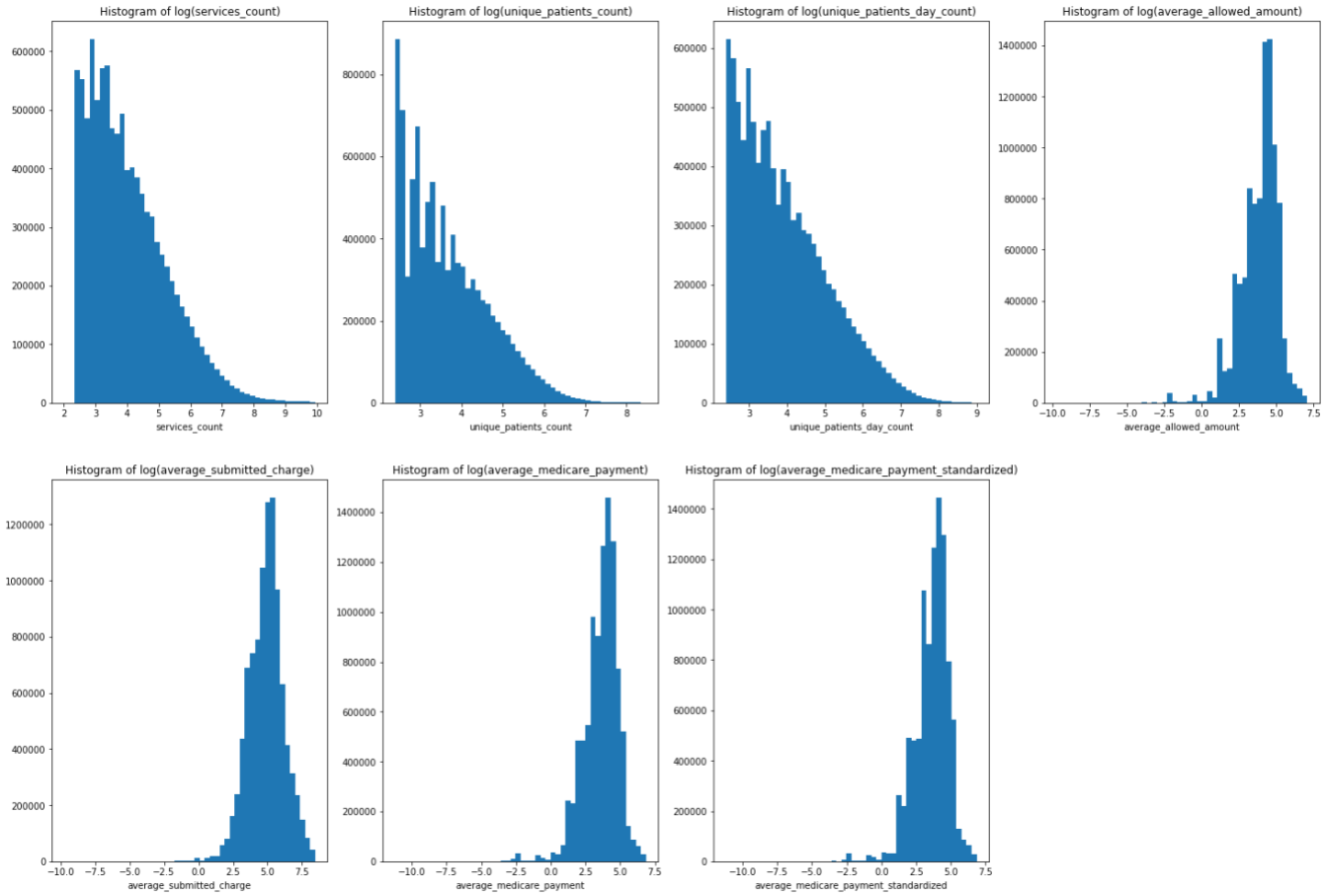

Figure 2: Logarithmic Histograms

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.

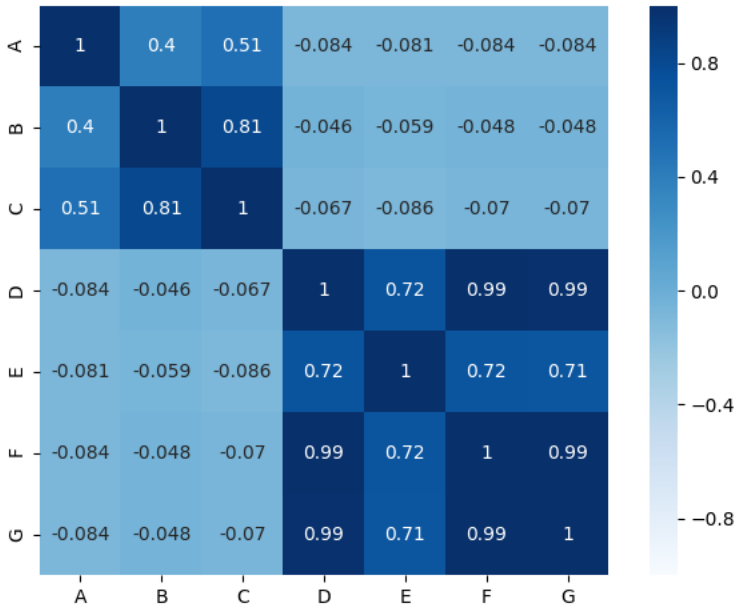

Figure 3: Correlation Heatmap

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 can we categorize different providers 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.

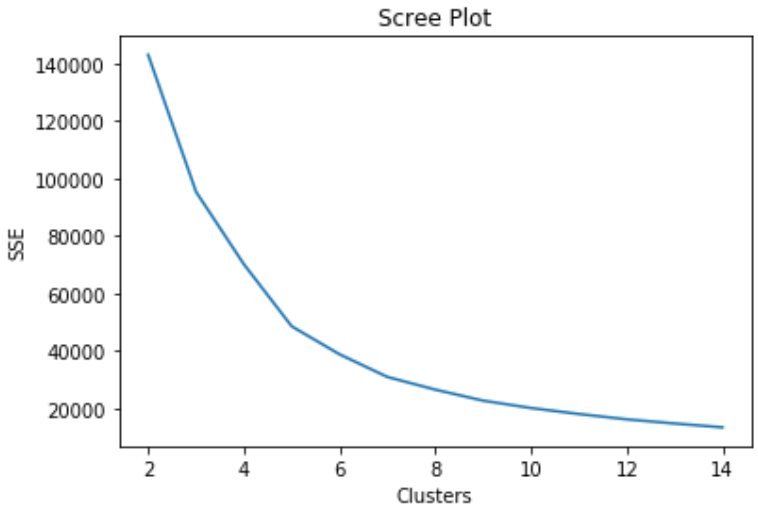

Figure 4: Scree Plot

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.

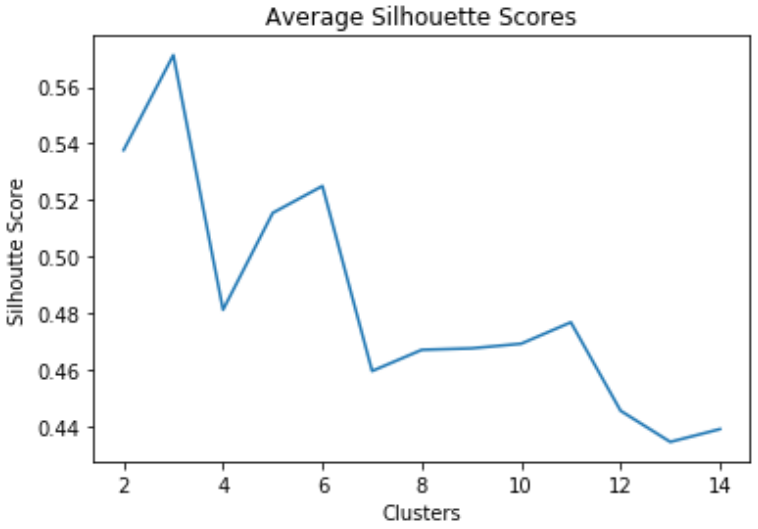

Figure 5: Average Silhouette Score

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.

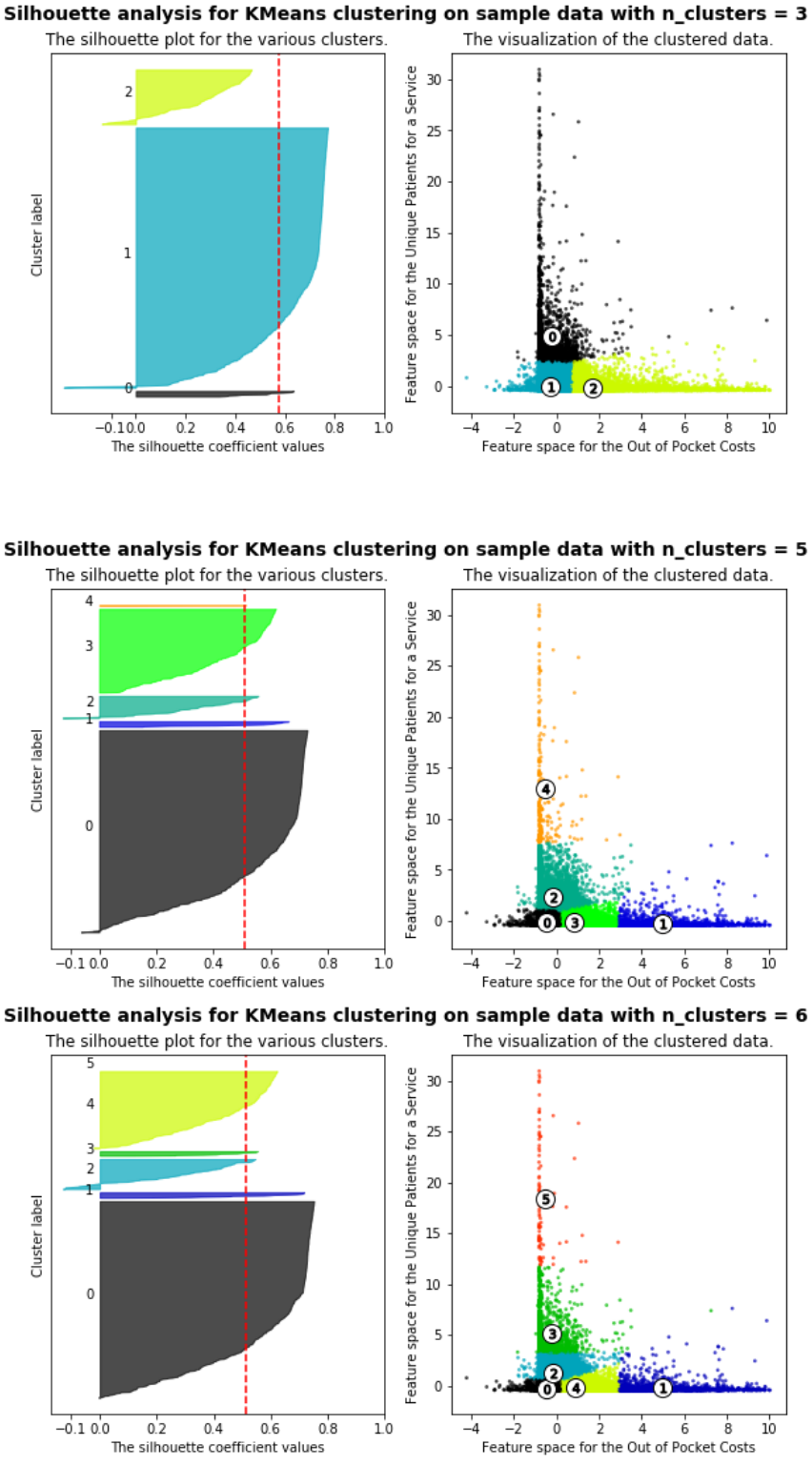

Figure 6: Silhouette Values with Feature Space

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.

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

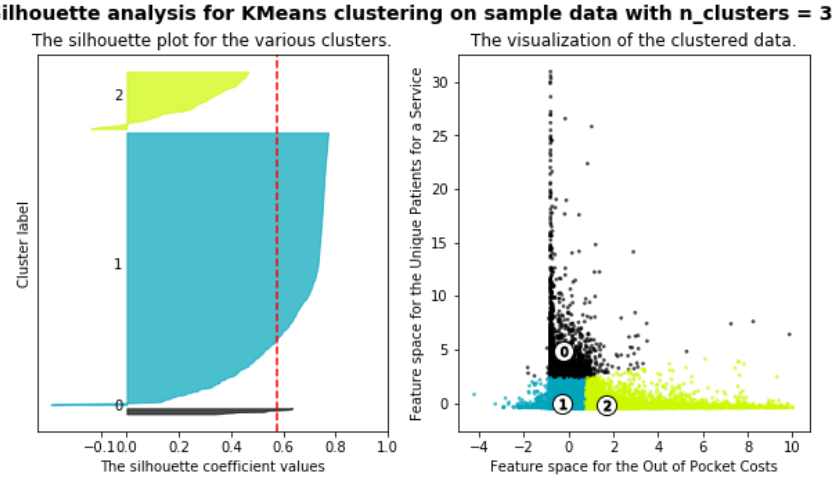

Figure 7: Three Cluster Mapping

Table 1: Cluster Demographics

Cluster	unique_patients_count	pocket	services_count	average_submitted_charge
0.0	70.294264	21.288636	157.404489	280.353135
1.0	75.455158	20.988174	170.904861	285.327295
2.0	75.686030	21.384065	170.333675	295.204041

As providers tend to fall into one of three clusters, we can also see patterns with the number of services they've done and the average amount they charge Medicare. Most providers fall into Cluster 1, but Clusters 0 and 2 are the most interesting to consider from a business perspective.

Cluster 0 has providers with low patient counts, low service counts, and low submitted charges. These providers are doing less services and charging the least for them. For that reason, this demographic has the *least* to lose by not accepting Medicare. In particular, they would be most susceptible to being approached by an insurance company that would prefer them to accept their insurance over Medicare's. This is meaningful because an insurance company should be targeting potential clients in Cluster 0 more than anywhere else.

Cluster 2 on the other hand has providers with high patient counts, high service counts, and high average submitted charges. This cluster is interesting from the perspective of insurance companies because these high-volume providers offer the most potential gain if they were to accept other insurances. Insurance agencies should offer discount deals to these providers because they are seeing many patients and completing many services. It would be meaningless to offer these deals to other Clusters because they wouldn't fit the criterium for 'high-volume'.

In doing this sort of analysis, an insurance agency has the opportunity to select different provider types and offer them catered services. Ultimately, this form of customer segmentation offers a company flexibility and precision.