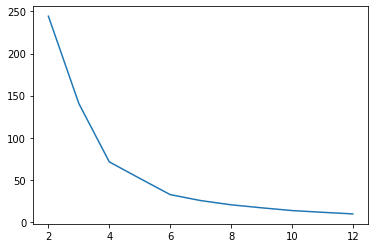
IEMS 308 Project 1 Report

When first observing the features, I realized that the dataset had quite a limited number of numerically based features that were also not extremely descriptive. As a result, I decided to take the average\_Medicare\_payment\_amt divided by the average\_Medicare\_allowed\_amt to get a ratio of coverage amount by Medicare in order to understand the amount of out of pocket expenses being incurred. I also derived a statistic that took average\_Medicare\_allowed\_amt divided by average\_submitted\_chrg\_amt to understand how much of the amount requested by the facility Medicare sees as a valid amount for the procedure. For the rest of the report I will refer to these two statistics as coverage ratio and request ratio respectively.

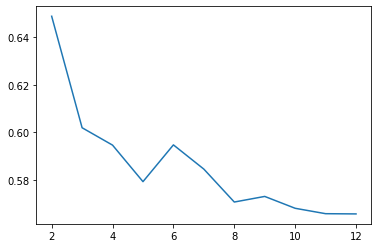
Based on these derived statistics and my knowledge of Medicare funding discrepancies between states, I thought it would be best to try and cluster around these statistics in order to look at geographic differences in Medicare operations across states. I chose to look at all 50 states, D.C. and Puerto Rico due to the states and D.C.’s importance and because Puerto Rico was the only U.S. territory with a large enough number of data points to sample properly for an experiment. I also chose to look at these statistics as average across all procedures at a facility and then observe differences on a facility by facility basis. This was because there are no geographic differences to observe among different procedures at the same facility. In order to work with a reasonable amount of data from the immense dataset I chose to take 500 samples from each state/territory to derive these statistics from and cluster. These samples were then standardized between 0 and 1 using the sklearn.preprosessing.Normalizer() module.

When beginning to try and cluster the data based on these random samples, I found that my clustering results had low silhouette scores that I didn’t feel comfortable drawing conclusions from. As a result, I employed a Kmeans outlier detection clustering technique to fix this. I looked at a high number of clusters from 15 to 30 in order to get many possible outlier groups and observed their silhouette scores and inertia plots. This ultimately led to me choosing 25 clusters for outlier detection, due to its high silhouette score and inertia plot kink. After this I cycles through removing each cluster from the sample and re-clustering the sample in a range from 2-15 cluster via k-means and recording the highest silhouette score achieved in order to determine which clusters were throwing off the clustering so much. I decided to single out all clusters that when removed resulted in a silhouette score of .6 or higher which was a clustering accuracy goal of mine. These outlier clusters tended to have rather extreme values in both statistics, but ultimately reflected geographic patterns that would be found in later clustering, so I thought that removing these deficient clusters was a good move for the sake of fit.

After identifying the outlier clusters and removing them I resampled the remaining facilities and drew out 350 facilities per territory/state. This number was chosen because after numerous trials 150 facilities were never taken out from any state during outlier detection. After resampling the data was again renormalized from the original data using the same Normalizer module as before. I then fit this data using a Kmeans clustering with cluster numbers ranging from 2 to 13. The inertial and silhouette score graphs for these clustering’s are displayed below.

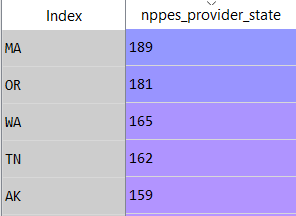
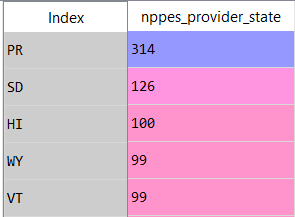


*Figure 1: Inertia Scores for Kmeans clustering on final samples*

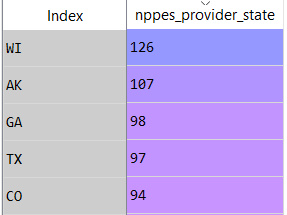
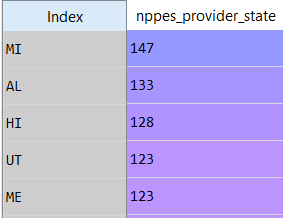


*Figure 2: Silhouette Scores for clustering on final samples*

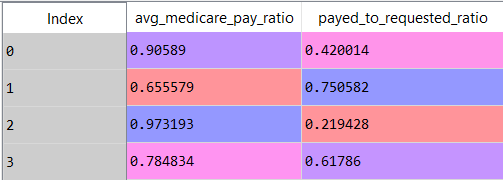
Based on these results I decided to use four clusters due to its apparent kink in the inertia plot and its relatively high silhouette score. Although it appears that three clusters might have been a slightly more accurate choice, I found the clusters to not be descriptive enough to draw results from. After obtaining the four clusters I decided to look at the frequency of each state in the clusters. Various results are displayed in the figures below.

*Figure 3: Top 5 states in cluster 0 Figure 4: Top 5 states in Cluster 1*

*Figure 5: Top 5 states in cluster 2 Figure 6: Top 5 states in cluster 3*



*Figure 7: The centers in each statistic for the Clusters*

When observing the top states in each cluster and comparing it with available data on per capita Medicare spending in states it sheds light on some interesting interactions. Three out of the top 5 states in the highest coverage cluster are in the bottom 17 states in Medicare spending per capita. The same also holds in the cluster with the second highest coverage rate. This suggests that although their relative spending per person is lower, they must have a more efficient Medicare system in order to have such efficient coverage with less cost. This may be due to the more accurate requests for funding which are very strongly correlated with better coverage. Also, very notable is the prevalence of Puerto Rico in the cluster with the lowest coverage. 90% of the sample from Puerto Rico were in Cluster 1. This is largely consistent with the differences in Medicare funding policies for the U.S. territories. U.S. territories are provided with a set amount of funding for each year instead of the consistent funding provided to the U.S. states.[[1]](#footnote-1) This was a huge factor in Puerto Rico’s recent bankruptcy because the Puerto Rico state budget has to cover the remaining lapses in Medicare payments.[[2]](#footnote-2) This is another example to add to the fact that the U.S. Medicare policy in the territories is massively flawed and requires redoing.

1. https://www.finance.senate.gov/imo/media/doc/Puerto%20Rico%20Healthcare%20Community%20(Attachment%201).pdf [↑](#footnote-ref-1)
2. https://www.nytimes.com/2019/09/27/business/puerto-rico-bankruptcy-promesa.html [↑](#footnote-ref-2)