IEMS308
HW1 - Clustering
Yujia Zhai

Executive Summary:

Based on the Medicare data description, various information features of the Medicare provider and coverage amount are available in this data source. To fully utilize this information, different types of problems can be further investigated. For example, in this case study, Medicare cover rate, which is defined as the percentage of total payment covered by Medicare, can be studied by applying clustering analysis.

Clustering analysis is grouping of a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. In this case, by using the available features of Medicare allowed and submitted amount, we set: $Medicare_{cover\ rate} = Medicare_{allowed\ amt} \div Medicare_{submitted\ amt}$. The Medicare cover rate and number of services, as the two features adopted in this case study, are grouped by city and included in the clustering analysis. Namely, a 2-dimensional clustering model is developed.

According to the results, three clusters are obtained from the analysis. The three clusters show the trends of low number of services with high cover rate variance, medium number of services with medium cover rate variance, and high number of services with low cover rate variance, respectively. The result suggests that cities with larger number of services show much more stable Medicare cover rate, which might be due to different levels of market competition with different city sizes. Thus, problems of market regulation and policy should be emphasized especially in small cities.

Problem Statement

Various Medicare information, including allowed and submitted amount, number of services, and city are available. In this case study, we aim to analyze the number of services and Medicare cover rate grouped by city. By interpreting the result obtained from clustering analysis, insights about the distribution of number of services and Medicare cover rate of different types of cities can be summarized, and relevant policy-related insights can also be inspired, which are important from the market regulation's point of view.

Assumption

In this case study, we don't require essential assumption of the data, since we focus on the nature of the Medicare data of each city itself. Therefore, theoretical assumption is not needed.

Data Exploration:

Fig 1 shows different numbers of occurrences, i.e., numbers of records in the dataset, in different states. California, Florida, and Texas are the top three states according to Fig 1.

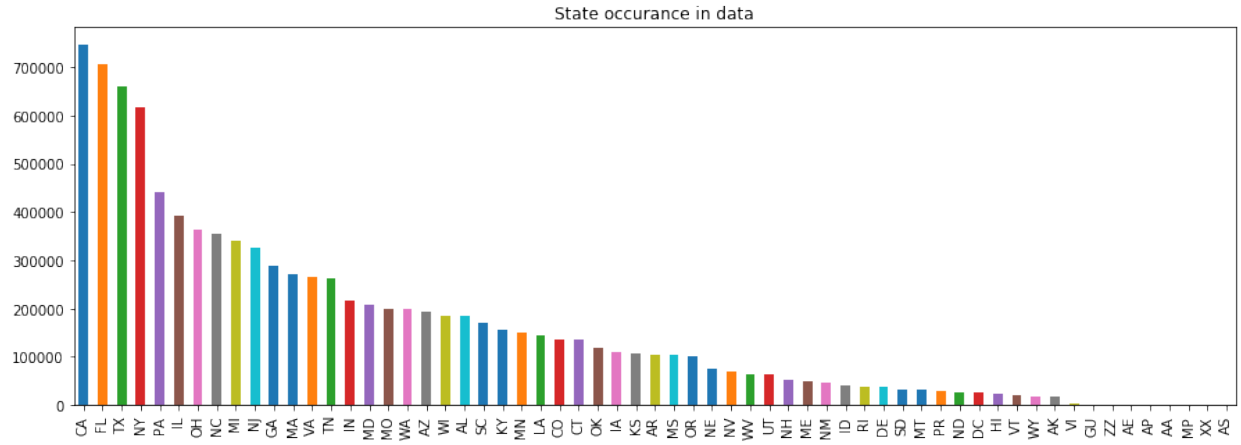

Figure 1. State occurrence in data

Fig 2 to Fig 5 present the histogram of four features either directly obtained or derived from the data – Medicare submitted amount, Medicare allowed amount, number of services, and Medicare cover rate – all averaged to city level. As can be seen from Fig 2 to Fig 5, distributions of all four features are right skewed.

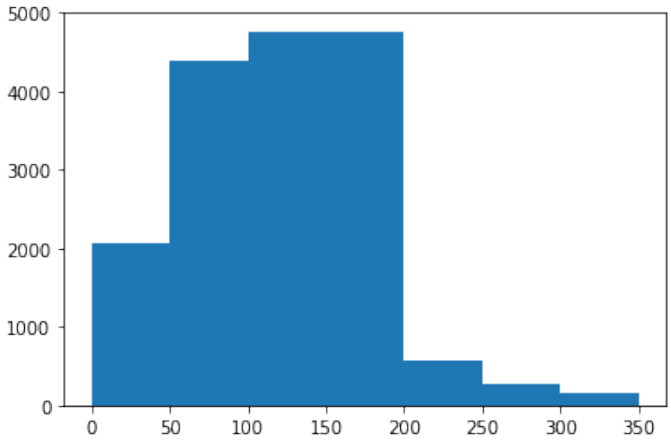

Figure 2. Histogram of average Medicare submitted amount of city

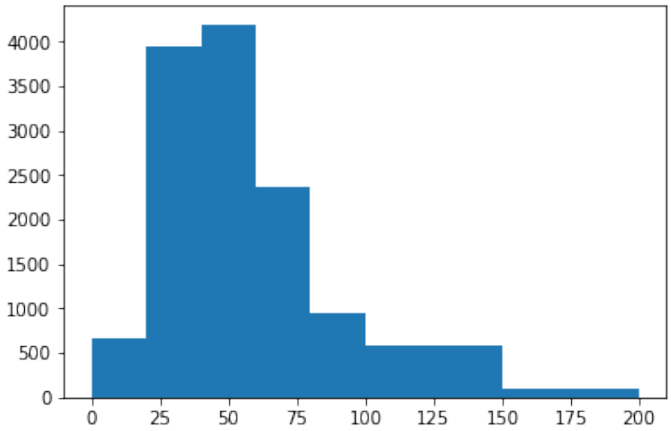

Figure 3. Histogram of average Medicare allowed amount of city

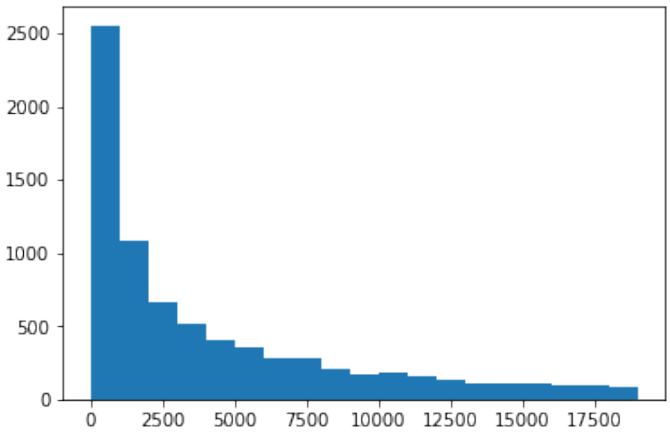

Figure 4. Histogram of number of services of city

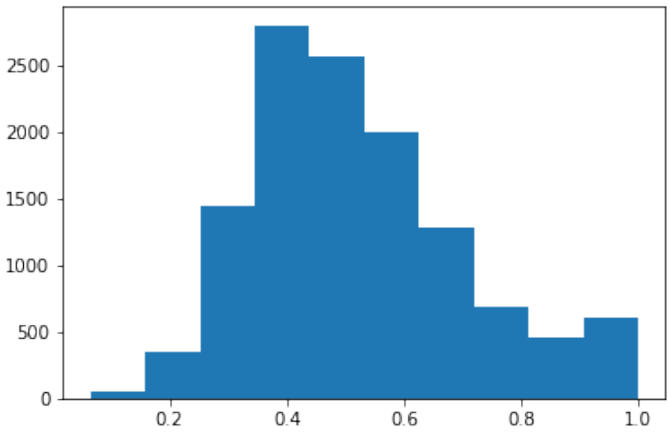

Figure 5. Histogram of average Medicare cover rate of city

Then a correlation test is conducted. Based on result of the correlation test, number of services and average Medicare cover rate are selected as the clustering features. Namely, a 2-dimensional clustering is conducted. Fig 6 shows the 2-dimensional distribution of number of services and average Medicare cover rate.

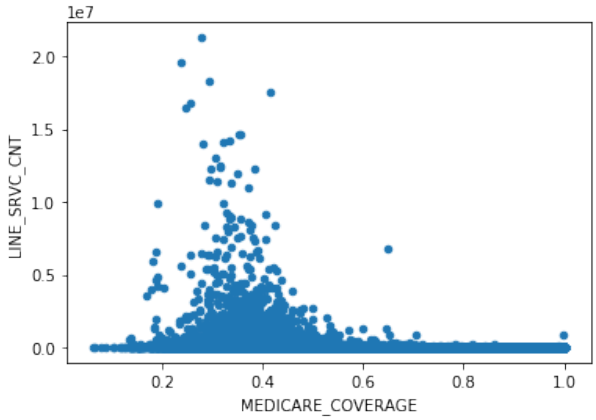

Fig 6. Distribution of number of services and average Medicare cover rate

Analysis

First we develop different clustering models with different number of clusters. Fig 7 shows different number of clusters with corresponding criterion, which is sum of squared distances of samples to their closest cluster center. When number of cluster is larger than 3, we can see that the reduction of sum of squared distances is getting much smaller. Therefore, we choose number of clusters of 3 in the analysis.

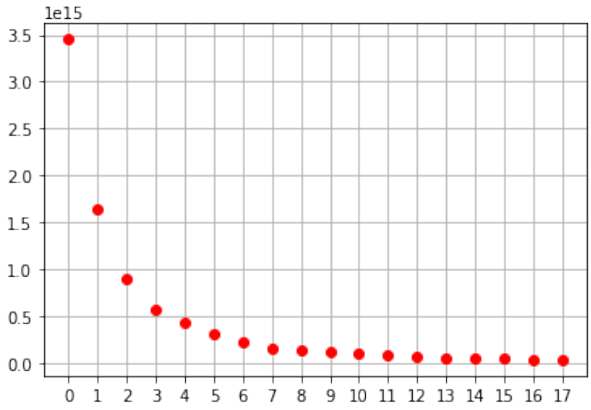

Figure 7. Sum of squared distances versus number of clusters

Result of clustering is shown in Fig 8, which the cluster centroids presented as black dots. Here we denote purple cluster as Cluster 1, yellow cluster as Cluster 2, and green cluster as Cluster 3. As can be observed, the overall distribution shows a pyramid shape. Cluster 1 shows a large range of average Medicare cover rate, while the numbers of services are all small. Both two features of Cluster 2, on the other hand, are both in the medium range. Cluster 3 represents the top of the pyramid, which has the largest number of services and more stable average Medicare cover rate.

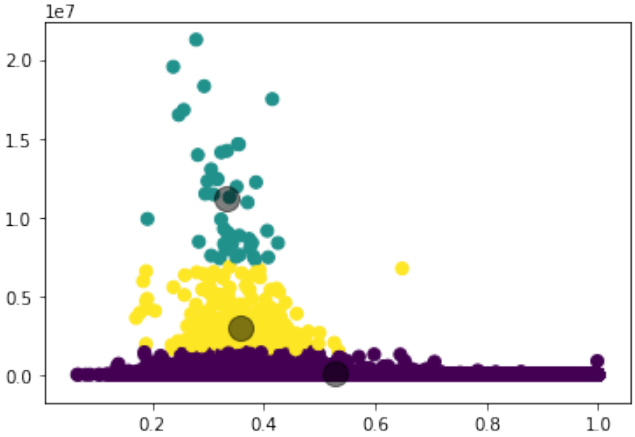

Figure 8. Result of clustering in 2-dimensional distribution

This result can be interpreted in the way that, cities with larger number of services are likely to provide a more stable and regulated Medicare cover rate than those who have smaller number of services. This trend can be explained by the fact that, smaller market that provides less service

might also be poorly regulated compared with larger market, who also provides large amount of service and has formed high level of market regulation. Thus, the level of Medicare coverage of Cluster 3 is much more fixed than Cluster 1.

The discussion above may lead to the policy-related insight that, small medical markets do require attention to make them more regulated and stable. Relevant policies are also needed to encourage market competition to improve such situation.

Conclusion

The key finding of the study is that, cities with larger number of services are likely to provide a more stable and regulated Medicare cover rate than those who have smaller number of services. This might be related to the level of local market regulation and competition. Based on the finding, relevant policy can be inspired to solve such problem that exists in cities with small number of services.

Next Steps

More detailed features and deeper investigation are needed to further analyze the factor that lead to such problem. For example, does the number of services highly depend on city size (since population and city size data are not available in the data, this needs to be further studied)? If so, then small cities with small market should be emphasized. If not, then other factors that have significant impact on the market situation should be investigated. Also, local economic factors should be brought into consideration. These factors might have even larger effect on the Medicare operation than city size or population.