Financially Favourable Healthcare Locations in Illinois, USA

IEMS 308 - Data Science and Analytics, Homework 1: Clustering

Saif Bhatti

January 27, 2020

Executive Summary

In the US, hospitals are privately funded institutions, and as a result must consider financial viability just as much as any other business in order to stay in operation. A key decision in opening a new medical centre is location; determining a financially feasible location may be the difference between flourishing and under performing. In order to make this decision, medical clinics may use public healthcare data to analyse locations and determine those locales that are show greater promise in demand for medical services. In this case, the analysis will be applied to Medicare Provider data, which is a public dataset required of all medical institutions in the US.

This paper details the initial analysis and pre-processing methods applied to this dataset, which was first subset to only have information from providers in Illinois. The analysis technique that this paper focuses on is k-means clustering - typically applied to find natural groupings within data under analysis. After finding financially relevant information, k-means clustering was applied to find patterns by zip code bundles.

The results were that there was no major disparity between the clusters in terms of fiscal indication, with the notable exception of the 604 zip area code. The cluster ('cluster5') that this grouping arises from is constructed entirely from data in this area. Geographically, the 604 zip code represents the north-west suburbs of Chicago, which is a particularly wealthy area of Illinois - and as a result, it bears taking notice. Further more, the primary factor that determines fiscal possibility is neither the number of patients examined nor the raw amount that medical procedures cost - but rather the location of the medical provider. In this sense, location analysis is a helpful aid in seeking optimal opportunity, but should be taken in context.

Data Cleansing

There were 4 main steps applied to the data:

0. **Data Pruning**: It was immediately apparent that the dataset was too large to deal with (2.22Gb), and as a result, pruning the data was required. First, this larger dataset was pruned to only have Illinois data by filtering on 'State of the Provider' in Excel. This dataset was still too large to work with in pandas, so next the csv was parsed with perl from the command line, the following two lines were used to randomly select 0.5% of the dataset.

```
        head -1 medicare.csv > subset\_medicare.csv
        perl -ne 'print if (rand() <.05)' medicare.csv > subset\_medicare.csv
```

Listing 1: Data Pruning with perl

The first line takes the header line from the original file, and creates a new file called subset_medicare.csv. The second line randomly samples the data and appends it to the subset file. This will create a relatively representative sample of the data that can be dealt with in Python.

1. **Exploratory Data Analysis**:

   In order to understand the data, it is useful to first perform an exploratory data analysis step. This will help determine which variables are suitable for k-means clustering analysis. The data is immediately converted to a pandas DataFrame, and then subset to only return Illinois medical provider information. From this, there are over 20k samples - which can further be subset to only return relevant financial information, as well as zip code of the medical centre.

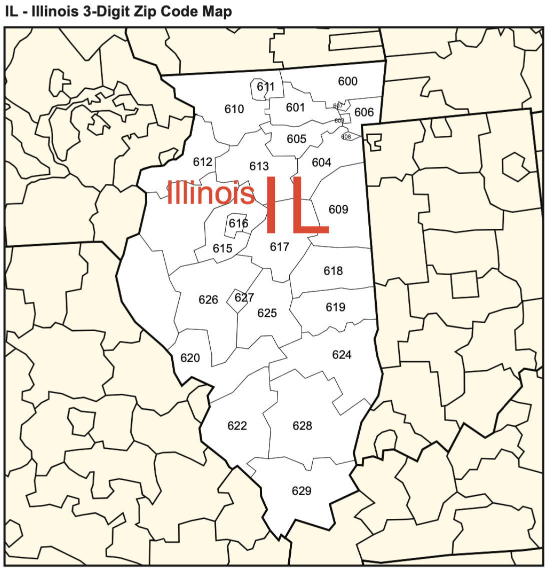
   The chosen features were the following:

a. 'Average Medicare Payment Amount': Average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service.

b. 'Number of Services': Number of services provided; note that the metrics used to count the number provided can vary from service to service.

c. 'Zip Code of the Provider': This is the specific zip code of the location of the medical provider.

   Zip codes can be converted into 3-digit versions which encompass wider geographical areas. In particular, there are 26 3-digit zip codes in Illinois, as mapped below.



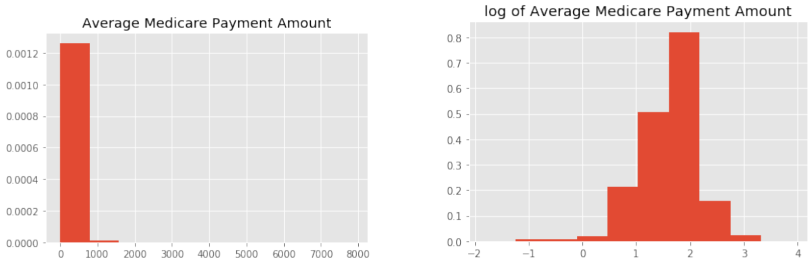
1. **One-hot encoding of categorical predictors**: For K-means clustering, it is important for the data to be well formatted. Categorical variables must be converted to a binary variable, which in this case requires the use of the pd.get_dummies() command to one-hot encode the 3 digit zip codes.

2. **Systematic outlier removal**: In order to avoid having large values for Average Medicare Standardised amount and Number of services from biasing the cluster, the outliers can be removed. Outliers are defined as any values that fall more than 3 standard deviations from the mean; all such rows were dropped from the data.

3. **Data Standardisation** It is important to determine the distributions of the data being used in clustering analysis. Standardisation works to ensure that large values do not exert excessive leverage over any analysis.

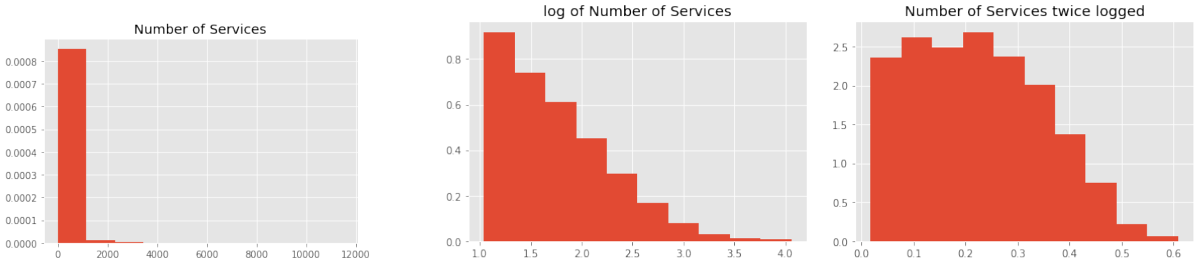
   By examining the data, it can be seen that the data is certainly not normalised.

   The data can be normalised by applying a base-10 log to the entire column. It is possible, as in the case with Number of Services, that this process may have to be repeated in order to result in a roughly normal distribution.



(a) (b)

Figure 1: Data Normalisation of 'Average Medicare Payment Amount'

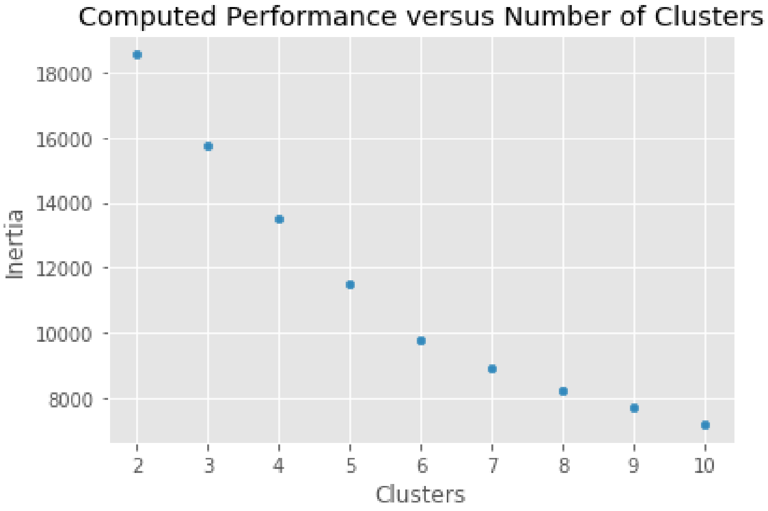


(a) (b) (c)

Figure 2: Data Normalisation of 'Number of Services'

Analysis

After the data has been cleansed, clustering analysis may commence. The first aspect is to determine an appropriate amount of clusters to split the data into. In order to accomplish this, a scree plot is used to indicate total summed distance from data points to the cluster centroids. In the computed performance measure of inertia, there is an elbow point at 6 clusters, and this will be the chosen number of clusters. This aspect in decision-making is more art than data science.



There exists a check measure to determine whether the cluster choice was representative and differential across groupings within the data: the silhouette score as a measure of error. The average silhouette score computed was 0.57, indicating that the clusters are relatively differentiated from each other. While this may leave improvement to be desired, this is sufficient in order to continue the analysis.

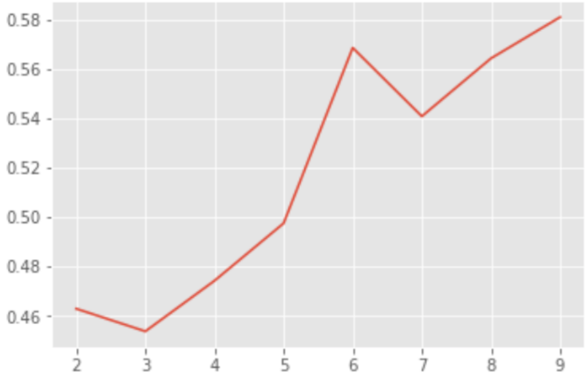


Figure 3: Average Silhouette Score versus Number of Clusters

The objective of the analysis is to determine patterns between clusters to assess fiscally beneficial opportunities - and as such, it is important to construct a metric to assess standardised revenue. The number of services multiplied by the average medicare payment amount will result in a financial metric roughly representative of revenue possibility in this region.

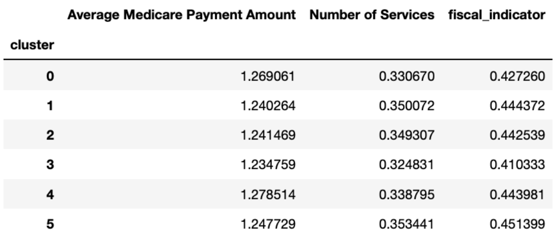
$$\text{fiscal\_indicator}_j = \frac{1}{n_j} \sum_{i \in C_j}^{n_j} \text{`Average Medicare Payment Amount'}_i * \text{`Number of Services'}_i$$

$$j \in \{\text{number of clusters}\}, i \in \{1...n_j\}$$

where n_j is the records assigned to cluster j.

Conclusions

The main insights that can be garnered from clustering analysis applied to the Medicare dataset with a view on finding financially beneficial opportunities are that location is not a key determinant in viability for healthcare providers alone. As shown by the below table, the resultant clusters from the analysis did not show very disparate fiscal_indicator scores, which hints that there is no stand-out area more prone to revenue surges.



	Average Medicare Payment Amount	Number of Services	fiscal_indicator
cluster			
0	1.269061	0.330670	0.427260
1	1.240264	0.350072	0.444372
2	1.241469	0.349307	0.442539
3	1.234759	0.324831	0.410333
4	1.278514	0.338795	0.443981
5	1.247729	0.353441	0.451399

The clustering analysis also shows that doctors in different zip code areas are differentiated more strongly on their location than on the cost of procedures or the number carried out. It is possible that this arises out of the fact that there is not a massive wealth discrepancy across Illinois zip code areas, and each zip code tends to be fairly representative across urban and rural zones. This is off from expectation, since Chicago is located in the zip code range 606-608, and there was an anticipated difference in fiscal opportunity for this area. However, the cluster with the highest performance based on the fiscal_indicator display was cluster5. Upon further examination, it pointed at zip code area 604 as being the location of higher revenue. The 604 area represents the northwest suburbs of Chicago, which is a notably affluent area, so this makes sense.

Next Steps

- While the public healthcare dataset may be a useful starting point, this is reflective of not only outdated information in regions that do not necessarily reflect resource allocation methods actually used. While zip codes form a suitable interim understanding for area comparison, there are larger areas that may be more representative of the context of a region (for instance the issue with 606-608 versus 604 zip codes, are all in the Chicagoland area). Instead, perhaps a better method is to use County areas, which will also result in a more even urban/rural split and allow for more comprehensive differentiation.
- While the data provided information about past supply, there is no indication on the estimates for demand in various areas. Economics points to basing a key analysis on indicators for demand to ensure not only a snapshot but indicators for growth over time. By establishing strong predictors for expansion in the healthcare industry of a region, there will be greater confidence surrounding the decision based on location and fiscal indicators.

Appendix

1. Data can be found at this hyperlink: Medicare Data 2017
2. Source code can be found at this hyperlink: saif1457-github