IEMS308

Yiheng Pan

Clustering

**Executive Summary:**

Healthcare is one of the most concerned topics in this century. US government tries to make the healthcare system more clear and helpful. Therefore, Physicians and other healthcare professionals provide the services and procedures to Medicare beneficiaries and record the information. The Physician and Other Supplier Public Use File includes information on utilization, payment, submitted charges and place of service. With these valuable data access, we can take advantage from analyzing and make effective conclusion and strategy for future better healthcare system.

Cluster analysis is an iterative process of discovering the similarities between each item.

There is 9497892 observations of 26 variables in the file. After exploring the data set and choosing the meaningful numerical variables, four features(Average Medicare Allowed amount, Average Submitted charge amount, Average Medicare Payment amount, Average Medicare Standard amount) are selected to cluster the provider.

**Problem Statement:**

As we known, the government already put large effort into improving the technology of Medicare. However, we wonder if the high-quality Medicare is affordable for most people and how the Medicare allowed charge amount is. We should find out the relationship between the location often chose and the cover rate of this specific place.

**Assumptions:**

1. All the information is accurate and final. There is no processing charge or missing fee. CMS update the information as soon as possible. All the healthcare providers and patients respect the rule to submit the actual information about their healthcare services.
2. The sample data was selected in random. The selected data with the perspective of all the data can represent well.
3. We assume that the amount of services offered symbolizes the ability of facility and equipment. The larger amount, the higher quality is.

**Methodology:**

First, we explore the data set and look over the features by reading the CMS documentation and using software ‘R’. After loading 9497892 data into system, we randomly sample 10000 of them. Since our daily laptop cannot work on such large data set. By using summary function, we know the basic information of each feature, such as the mean, maximum, minimum and so on.

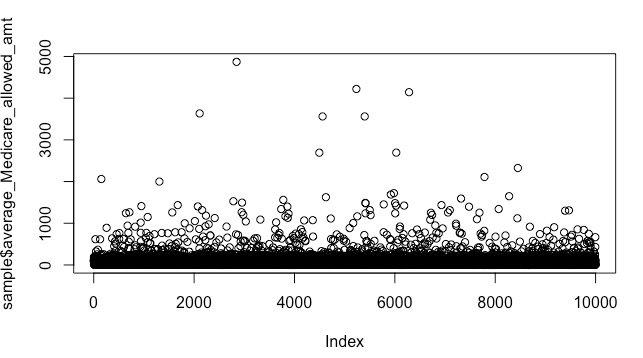
Next, in order to make sure all the data sampled are normal with only few outliers that the amount cannot make big influence on the result. We plot all the numerical variables and found the data is satisfying.

Then, the last four numerical values was chose to decide how many clusters we use. After picking them up and normalization by scale function, we use K-means clustering and draw a graph about the within sum of square. The line after 6 becomes smooth. However, I decide to choose 4 since too large number of clusters is not useful to analyze problems. And the sharp drop of first four can be observed. The sum after 4 drops slowly.

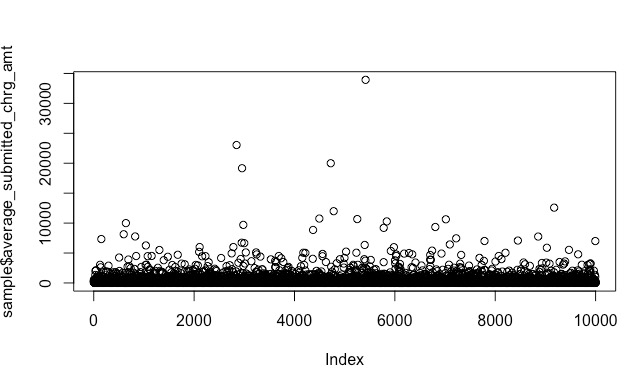
Last, we choose the information we care about and plot the relationship between each features. In this task, we focus on the location and the cover rate.

**Analysis:**

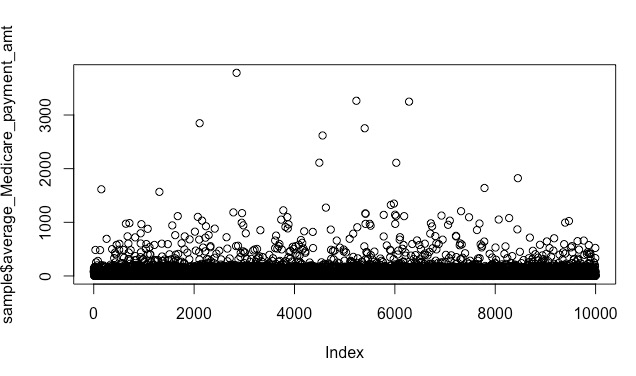
The first step of our analysis is to look over each data’s distribution and outlier. The graphs are below:



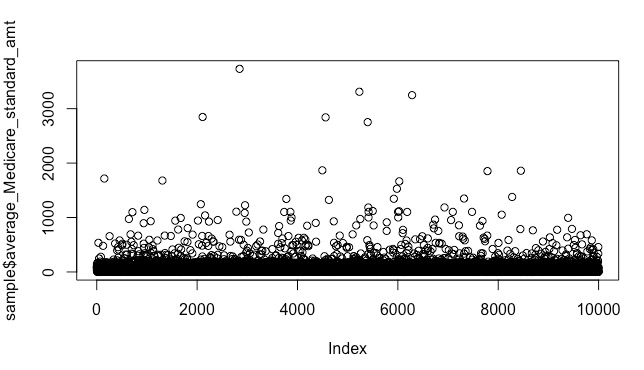
plot average medicare allowed amount



plot average submitted charge amount



plot average medicare payment amount



plot average Medicare standard amount

Then, we use summary function to explore the details of each column.

line\_srvc\_cnt bene\_unique\_cnt

Min. : 4 Min. : 11.0

1st Qu.: 21 1st Qu.: 18.0

Median : 44 Median : 33.0

Mean : 248 Mean : 90.8

3rd Qu.: 121 3rd Qu.: 78.0

Max. :5313837 Max. :741837.0

NA's :1 NA's :1

bene\_day\_srvc\_cnt average\_Medicare\_allowed\_amt average\_submitted\_chrg\_amt

Min. : 11.0 Min. : 0.00 Min. : 0.00

1st Qu.: 20.0 1st Qu.: 23.15 1st Qu.: 53.85

Median : 42.0 Median : 62.89 Median : 136.25

Mean : 147.4 Mean : 99.47 Mean : 327.39

3rd Qu.: 110.0 3rd Qu.: 113.88 3rd Qu.: 278.00

Max. :2167568.0 Max. :52601.47 Max. : 99999.99

NA's :1 NA's :1 NA's :1

average\_Medicare\_payment\_amt average\_Medicare\_standard\_amt

Min. : 0.00 Min. : 0.00

1st Qu.: 18.42 1st Qu.: 19.30

Median : 45.66 Median : 47.89

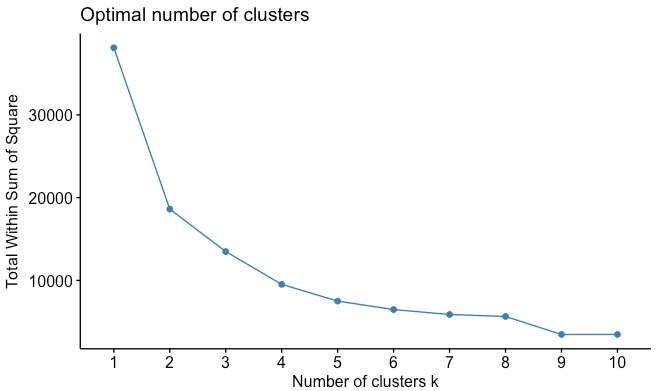
Mean : 76.31 Mean : 77.58

3rd Qu.: 85.26 3rd Qu.: 85.37

Max. :39691.58 Max. :39691.58

NA's :1 NA's :1

Next, we choose 10,000 data sample from the original table. And by using the last four numerical variables and k-means analysis, the clustering plot was drawn below to deciding the numbers of clusters.

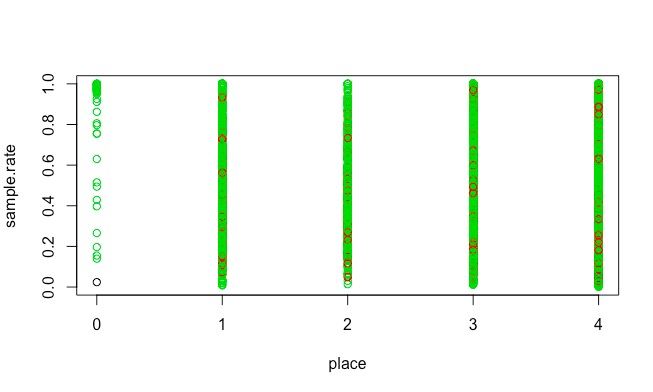


We can easily observe the drop of total within sum of square. It’s obvious that after 4 the line drops slowly. For easy analysis and minimizing the sum, we decide to choose 4 clusters.

Cover rate is a meaningful value to evaluate the ability of Medicare benefits, which equals allowed amount divided by submitted charge amount.

At first, in order to find out the relationship between the location and the pay rate, I mark all the places into zero to four by state column. Zero symbolized outside America. And one to four are west, mid-west, mid and eastern Amercia.

Set x as places and y as cover rate. The graph is shown as followed:



In this graph, we can draw conclusion that in outside America, most health service is covered in high rate. But in other areas, it’s not clear and cannot make cluster analysis. So, this way of analysis is dropped.

I think the reason about this wrong and confusing output is that I mistakenly choose the category variables to analyze. We need to think back about what problem we are going to analyze and what we expect in the plot before we find the relationship.

So I change the way to count the number of a specific place appears instead of selecting the position as a specific numerical.

After sampling the data set, by using group by function, we create a new table, in which the information are stored as specific place, number of service offering, allowed amount, submitted amount and the cover rate. And let us look over the summary of this table.

count allowed submitted rate Min. : 1.000 Min. : 0.067 Min. : 0.07 Min. :0.01024

1st Qu.: 1.000 1st Qu.: 43.868 1st Qu.: 95.00 1st Qu.:0.29005

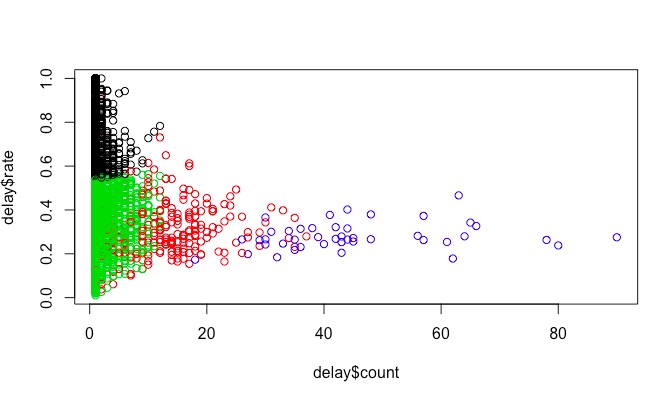
Median : 2.000 Median : 125.219 Median : 265.87 Median :0.42302

Mean : 4.016 Mean : 397.429 Mean : 1292.31 Mean :0.46294

3rd Qu.: 4.000 3rd Qu.: 358.973 3rd Qu.: 1029.46 3rd Qu.:0.60130

Max. :90.000 Max. :9173.248 Max. :37377.04 Max. :1.00000

Then plot the graph by setting count as x, setting rate as y.



From this graph, for these 10000 records, most cities offer normal amount of service like black cluster and green cluster. Most cities with mean or under mean number of times to offer service. And larger amount (green cluster) can only offer the cover rate under 0.46(mean value). And the rest (black cluster) can cover large part of the Medicare fee.

And for the cities which can offer medium amount of services (red clster), the cover rate cannot be too low and too high, always between 0.2 and 0.6.

As the blue cluster shown, the cities which shows the strong ability to offer large amount of service, cannot offer very high cover rate. The value are always below 0.5.

**Conclusion:**

We can draw conclusion that the ability of offer service does not appear with high cover rate. So even though some cities has better facilities and equipment, can offer different kinds of services, people have to pay large part of Medicare fee for themselves. There is not a balance between the medical technology and health beneficiary.

**Next steps:**

For the further Medicare development, we should focus on which type of service need to be improved the cover rate. One important next step is finding out the relationship between the amount of different provider types and each type’s cover rate and to see there is a balance between them.

We can also add more information about the patient, like age, income to analyze which group and which income level needs the service best. Then we can modify the policy of health insurance to make sure more people get fair benefits.