**Clustering Project1**

Hao Zhu

**Executive summary:**

The data for the Physician and Other Supplier PUF provides the information of provider demographics, provider charges, Medicare allowed amounts and payments. In our report, we focus on the situation of Medicare payment coverage in different states. So, we choose and aggregate the same state of multiple records features: the number of providers, the number of services, average of the Medicare allowed amount for service and one term we defined as Medicare fee cover rate (*“average\_Medicare\_standardized\_amt”* / “*average\_submitted\_chrg\_amt*”). The first three features are used in the clustering to classification, the last one is a feature to measure the extent of Medicare service cover payment.

Using the result of clustering, we make a conclusion that for some states of small population, like AR and MT, they do a really good work on the Medicare service cover payment, probably due to the small population and the lower chance of expensive services required, however there also has some states because of the small population, devoting to the Medicare system less far than any other states, which has the lowest Medicare fee cover rate such as AK, WI. As for the states of large population, they really pay a lot and have lots of providers, but they also have a large average of the charges that the provider submitted per service. There is a restriction that limits the Medicare fee cover rate increasing among those states.

**Problem Statement:**

We want to know how well the different states commit to cover the Medicare service fee. So, we try to analysis the correlation between Medicare fee cover rate and provider, service, amount allowed situation by different states. Here Medicare fee cover rate is simplified as *“average\_Medicare\_standardized\_amt”* / “*average\_submitted\_chrg\_amt*”.

**Assumptions:**

* States are all in America. We don not take the foreign countries into consideration. Also, except the fifty U.S. states and the District of Columbia, we also do not consider other states in America.
* The data we get is correct. There is no mistake on number recording or other things different from the fact.

**Methodology**

Our first step is performing data exploration. First, we aggregate the number of providers(this feature can be replaced of the number of “*NPI*”) ,the number of services (“*line\_srvc\_cnt*”), average of the Medicare allowed amount for the service (“*average\_Medicare\_alllowed\_amt*”), average of the charges that the provider submitted for the service (“*average\_submitted\_chrg\_amt*”), average amount that Medicare paid after deductible and coinsurance amounts have been deducted for the line item service and after standardization of the Medicare payment has been applied (“*average\_Medicare\_standardized\_amt*”), these five features by states. The last three features divided by the number of services is the average amount per service. Then draw the histogram of these features, analysis the correlation and outliers among these.

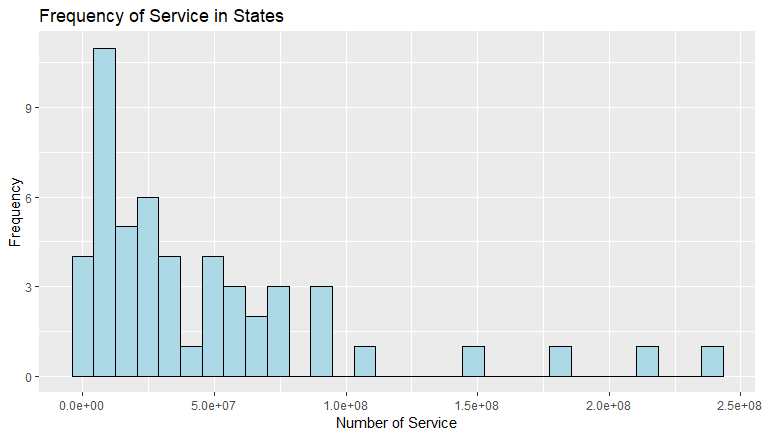
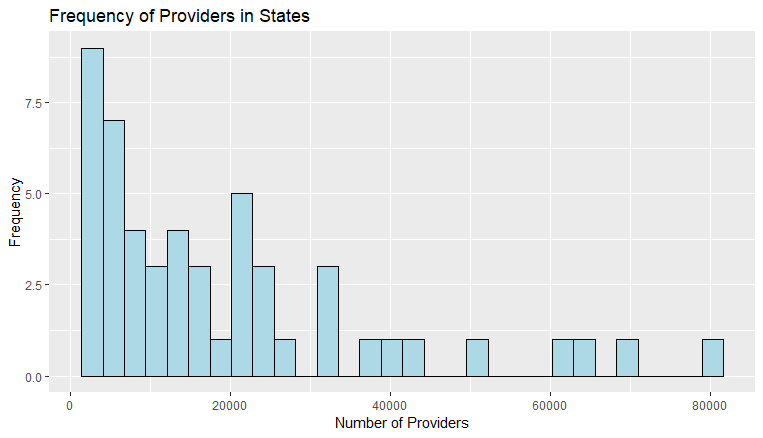
Second step we use a new feature we create called Medicare fee cover rate to measure the extent of Medicare cover, which equals to *“average\_Medicare\_standardized\_amt”* / “*average\_submitted\_chrg\_amt*”. When select features, we choose the left three features except two used one to do the clustering.

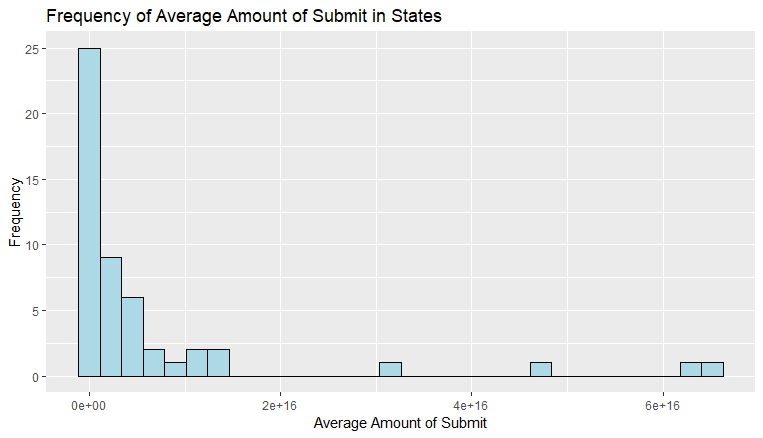
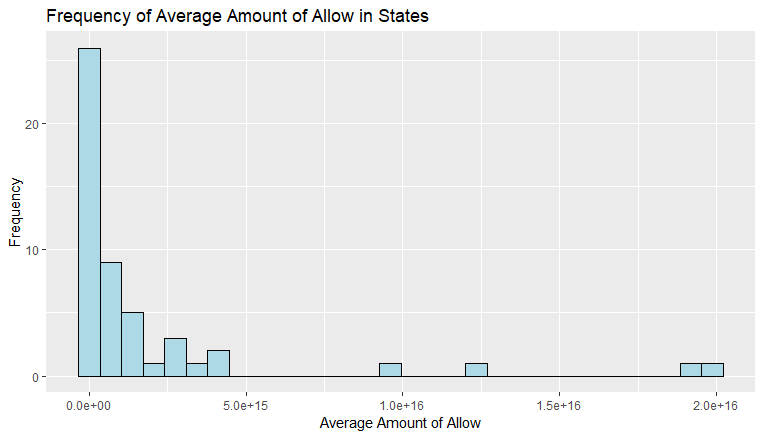
Third step is data preprocessing. The data is not normal distributed and standardized. So first do normalization to the data. Then we use method of SSE to estimate the value of K, and use the method of Gap Statistic to assure best K.

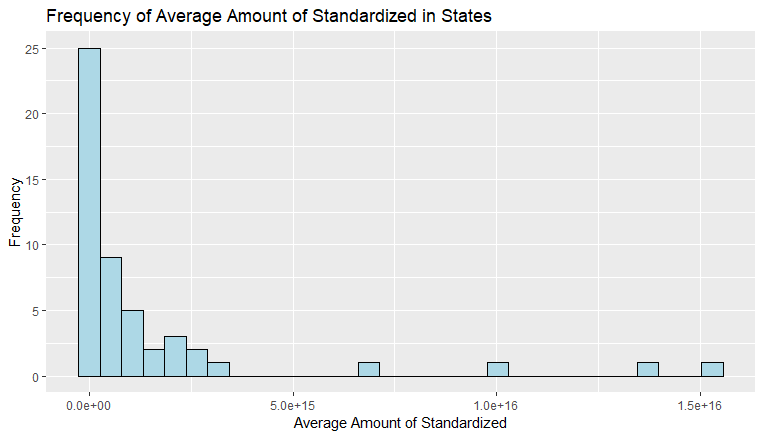
Forth step we chose the K-means clustering. Draw picture to make a conclusion. Then, we choose Silhouette Coefficient to assess the quality of our clustering.

**Analysis:**

We draw histograms of five features mentioned in Methodology before separately. Pictures are shown below.



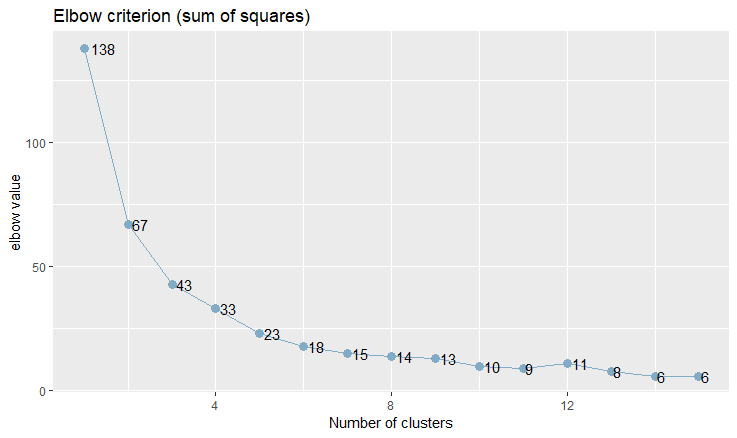




We can see four outliers clearly in the last three histograms: according to the origin data, we can conclude that the four outliers in each histogram are four same states, they are FL, CA, TX and NY. Also, the four states’ number of providers and service provided are the largest four. These four states have a far larger values of five features, so we recognize data of these four states as outliers, excluding them.

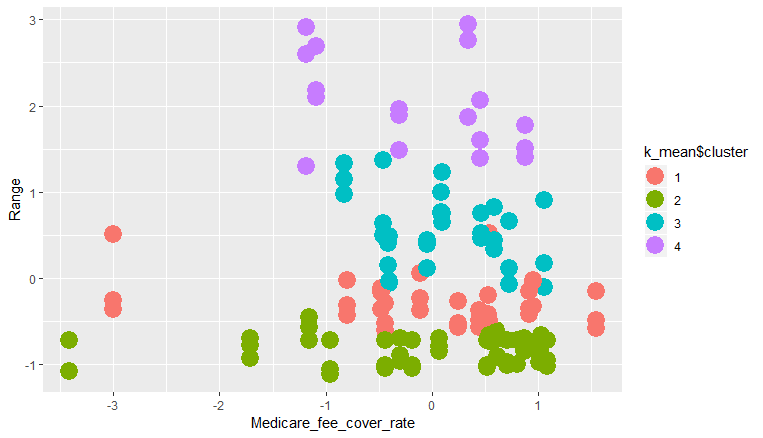
Then we analyze the histogram of five features. They have the similar shapes which means there probably is a positive correlation among these five features. To analyze the situation of the Medicare cover in different states, we try to find the correlation between Medicare fee cover rate and provider, service, average amount allowed (not per service) situation by different states. So, we use the number of providers, the number of services, average of the Medicare allowed amount for service for us clustering. After decision, we do normalization to the data of the chosen features.

Next, we use method of Elbow criterion to find the best K. We plot the SSE with different number of clustering. Picture is shown below.

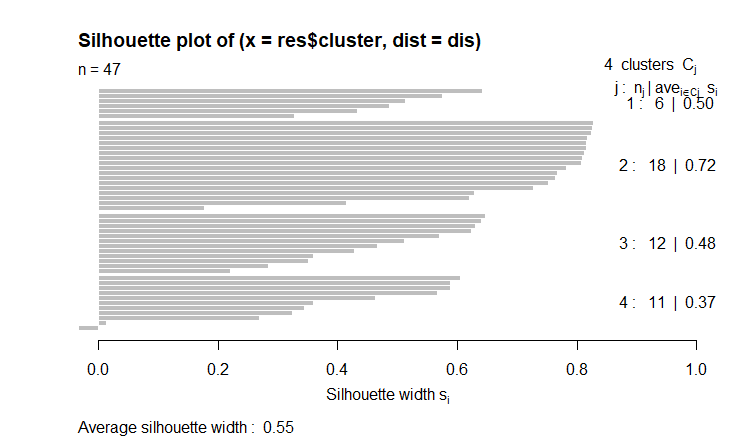


From the picture, we can estimate that K= 4 is best. Then we use the method of Gap Statistic to find best K. Because K-means have randomness, we repeat this method 30 times. At last, we get the best K = 4.

After using K-means, we get the 4 clusters. Now, take the Medicare fee cover rate (after normalization) as the x-axis, add number of providers, services and amount of allowed payment (all are after normalization) together to the picture as y axis, which means one Medicare fee cover rate has three points corresponding. Picture is shown below.



The last step is about assessing the quality of our clustering. We use the method of Silhouette Coefficient.



When k=4, there is the smallest Silhouette Coefficient 0.37. So, our clustering is quite good.

**Conclusion**

For the cluster 1 and 2, they have the relatively low providers, service and allowed payment. We can conclude that AK, WI which have devoted least to the Medicare system far than any other states due to the lowest Medicare fee cover rate while AR does the best, however, there is some states doing a really good job on Medicare fee cover. Their three features are low probably because of the smaller population, such as MT, CO, CT, etc. As for the cluster 3 and 4, these points represent those states have huge population. For those states, they devote to average a lot, however, because of the large population and an increasing chance of expensive services required, there is a restriction that limits the Medicare fee cover rate increasing.