K-means clustering

*Zi Mei*

Executive summary:

Administrated by U.S government, Medicare provide health insurance over 55 million people aged 65 and younger people with disabilities. People who paid when they are young and want to have a good cover when they get old and retired. Therefore, it is important Medicare does job properly to help the elderly keep healthy and enjoy their benefits.

This analysis focus on how well Medicare covers their clients, and check if there is any undercover or over payment. Understanding this question could help patients better prepare their budget when they got sick. Plus, it can also help Medicare customers improve their financing situation, diversify their health insurance plans to prevent potential risk. Furthermore, Medicare could improve its service by reviewing cover policies to make a better fit to client’s request.

RStudio is used, and the main methodology is k-means clustering. After the analysis, I found applicants who have small amount of submitted requests are being taken good care of, while who require higher amount are hardly been covered enough.

Assumptions:

* The data source and data records are correctly loaded into dataset, and data is randomly collected, not biased.
* All the blank or NA records are eliminated and not studied for the analysis. The bad data can affect analysis greatly and it would be better do explore without them.

Problem statement:

After roughly checked the dataset, I began exploring the data with this specific question: how well the Medicare covers all the applications? Is there any under or over coverage?

Methodology steps:

K-means clustering in RStudio was used to analyze this question, K-means will generate better results when applied to numeric data, and based on my coverage question, the last four variables were chosen, including: average\_Medicare\_allowed\_amt, average\_submitted\_chrg\_amt average\_Medicare\_payment\_amt, average\_Medicare\_standard\_amt

And the steps are following:

Prepare the data including exclude bad data records, normalize data and choose the sample data size to better analyze.

Explore the raw data to check if there any outliers, heavily skewed data records.

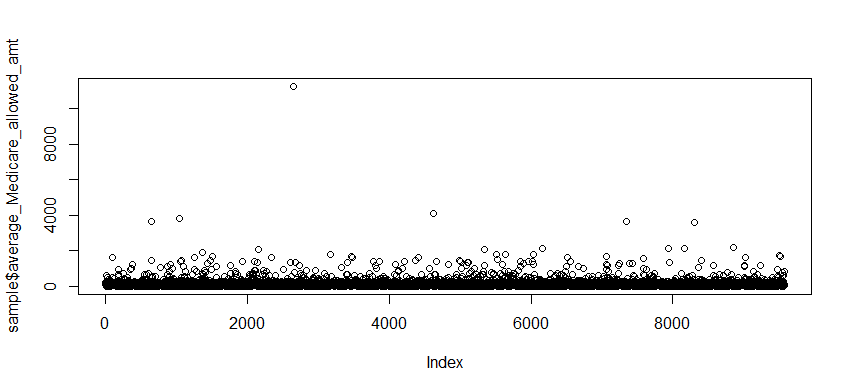
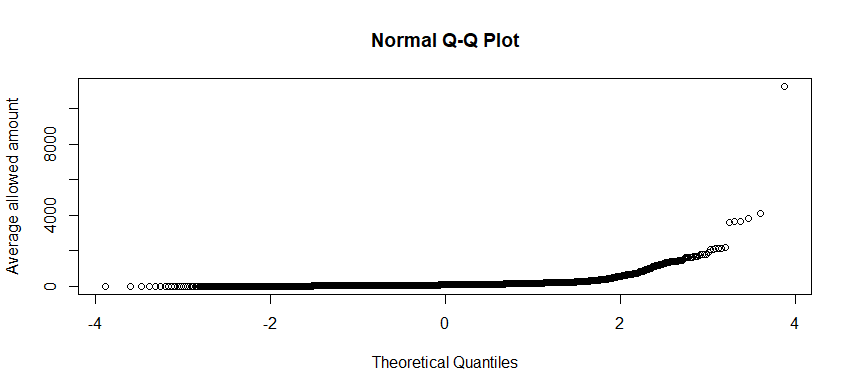
Do the clustering and plot

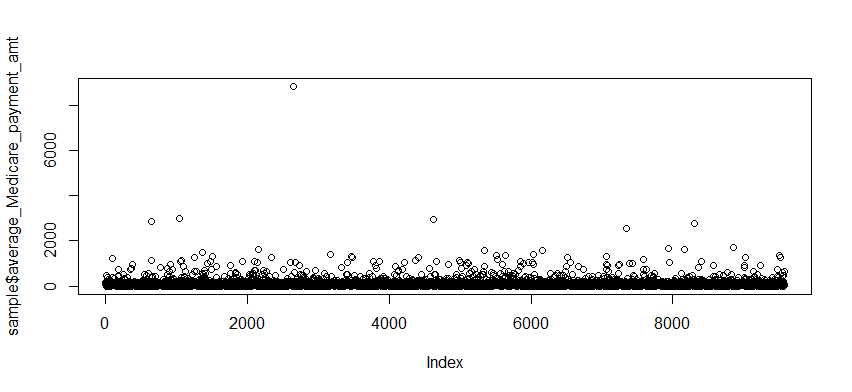
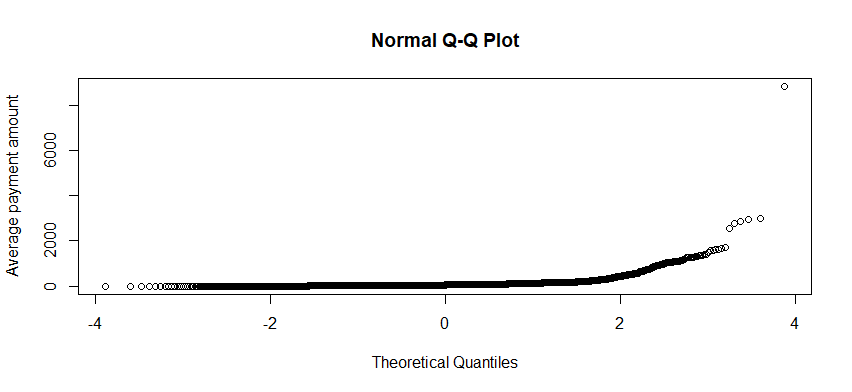
Interpret the output and extract useful insights.

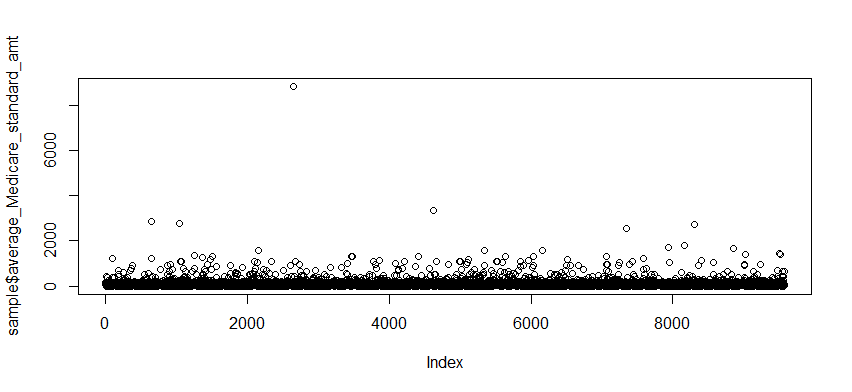
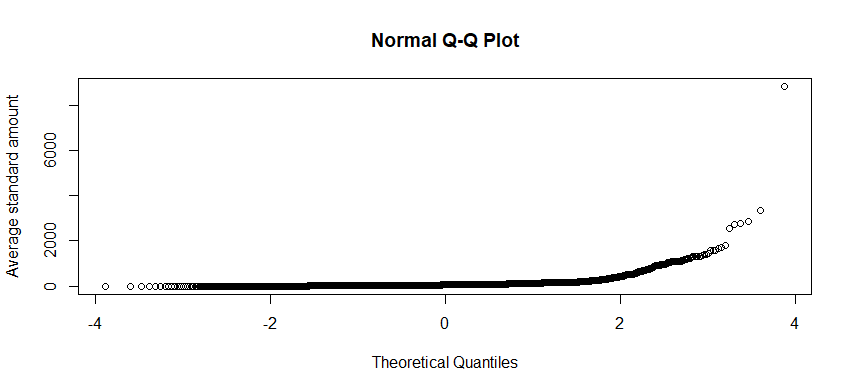
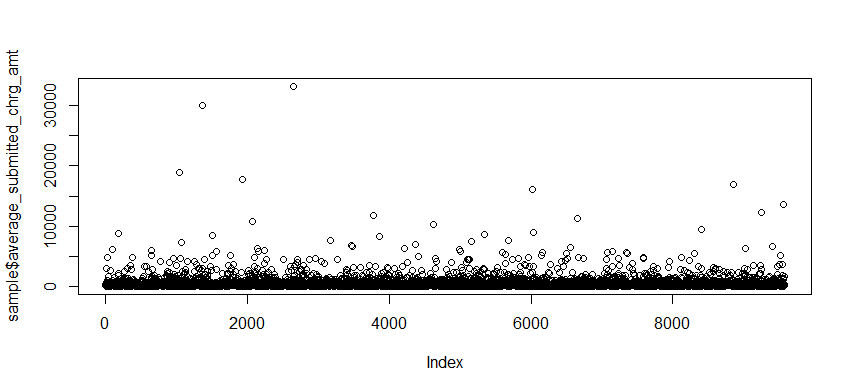
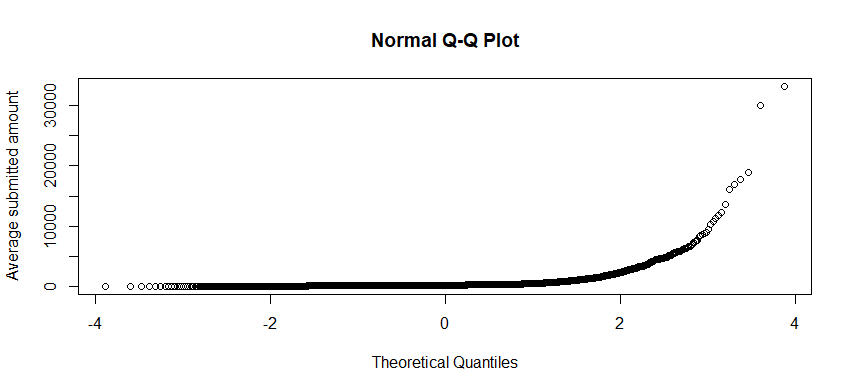
Analysis:

After done with the data cleaning, and the original dataset is huge, so I set the sample size following 1% margin of error and 95% confidence level.

The four variables’ qq plots and scatter plots are draw to check the data and I didn’t observe any heavily skewed trend or chunks of outliers. Although several points did pop up but it looks fine to me.

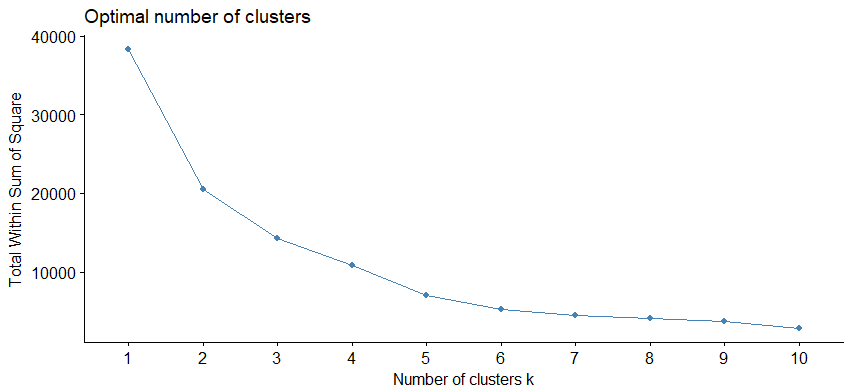


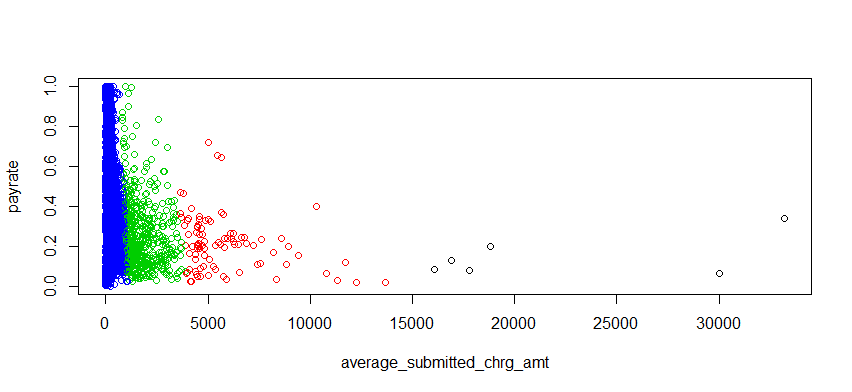
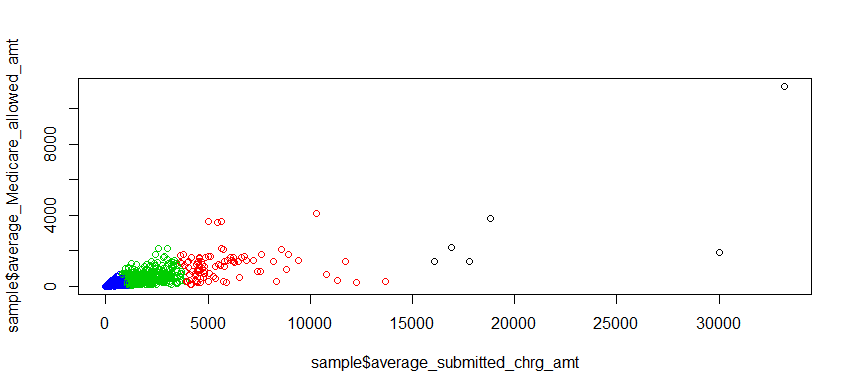




The optimal number of clusters suggested by graph is 4, and I plot the clusters by two graphs. The left one I go with submitted amount and allowed amount, and do the clustering with the four selected features that stated before. For the right one, I created a new variable payrate with allowed amount divided by submitted amount. And I redo the clustering with the payrate and the selected four numerical variables stated before. Then I plot the cluster with payrate and submitted amount.

Based on the two clusters plot, I observed that most of the submitted amount is not very large, and when the submitted amount increase, the allowed amount decreases significantly, indicating Medicare could not cover very ‘expensive’ requests. For the ‘middle’ request, Medicare did provide a good fit for them and take care most of them.





Conclusion:

From the analysis, I have the following findings:

1. Most of applicants don’t count on Medicare to cover serious diseases, since the submitted amount more concentrate under 5000
2. Medicare did well for those ‘inexpensive’ requests, but when submitted amount goes high, the payrate decreases and it is rare that very ‘expensive’ requests got a good payment.
3. For consumers, I suggest only use Medicare to prevent small illness, still need to have another insurance to prevent potential risk.

Next step:

The next step could involve more data such as geographical features, for example, different countries may pose impacts on their payrate due to the different living expenses and currencies. And according to analysis output, it may be better to divide the data to different countries and study them respectively to dig more insights. After all, identify more other features affecting the payrate can help Medicare to improve its payment and benefit applicants more practically.