**Clustering**

**Executive Summary**

Current president of the United States, Barrack Obama put significant effort to transform the healthcare system of the United States. He tried to make healthcare system more transparent and universal. As part of the Obama Administration’s efforts to make our healthcare system more transparent, affordable, and accountable, the Centers for Medicare & Medicaid Services (CMS) has prepared a public data set, the Medicare Provider Utilization and Payment Data: Physician and Other Supplier Public Use File. With this data being available, CEOs and executives of healthcare company can plan a strategy to make Medicare more beneficial and effective.

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

One million data out of the whole data set was sample in order to ease the analysis. After exploring the data and selecting two features (Average submitted charge & payment ration) that would categorize well the big data set, two features that would well represent the subset of the group would be used as the parameters for clustering. K-mean clustering was used to categorize the data set.

Taking a brief look at the result, significant number of Medicare beneficiaries who were charged of higher bill for treatments tends to be undercovererd by Medicare. At the same time, beneficiaries who were charged of lower bill for treatments tends to be very well covered by Medicare. By incorporating these insights into the social responsibility of Medicare term, Medicare companies should plan an effective strategy to be responsible for patients in serious health condition, who need expensive treatments, and make profit at the same time.

**Problem Statement**

Medicare companies should be responsible for not only making profit but also aiding patients in need. The biggest value proposition of the company is that patients can financially prepare themselves for heavy surgeries, which cost a lot; at the time they were having time they are having surgeries, they would be able to work properly to finance their treatments. By clustering randomly selected data sets, how healthcare companies are fulfilling their social responsibility was measured.

**Assumptions**

* The omitted data was ignored. Some healthcare providers omitted data for some categories. Omitting the data can bias the data. However, with the omitted data being present, clustering can’t be performed. Thus, those data were removed before analysis.
* All data submitted and stored is correct. It is a law that Medicare providers should report accurate information about their healthcare services to the CMS. If a healthcare provider did not provide accurate data, he is not abiding by the law. If there is a crash on the data, the CMS should consider upgrading their database system.
* The data was randomly sampled. There is not any pattern among selected data. Thus, one million data selected for analysis appropriately represents the whole data set.

**Methodology**

The software ‘R’ was used for data analysis. Our first step to begin clustering analysis was to select one million data out of the whole data set, which has about 9.6 million data. First, the whole data table was loaded with function ‘fread’. Then one million data were randomly sample.

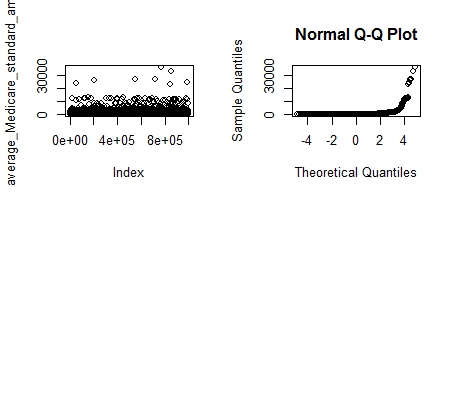
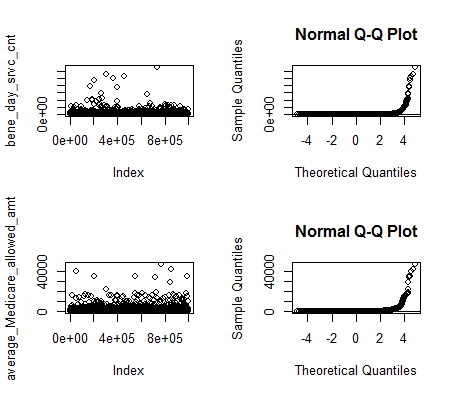
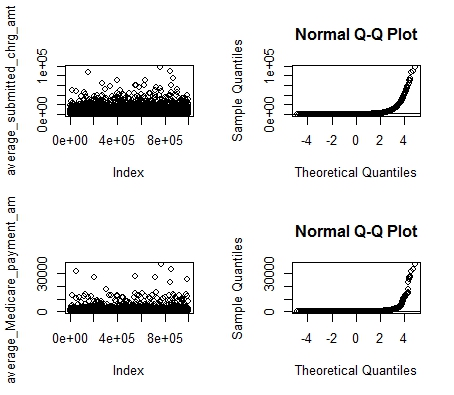
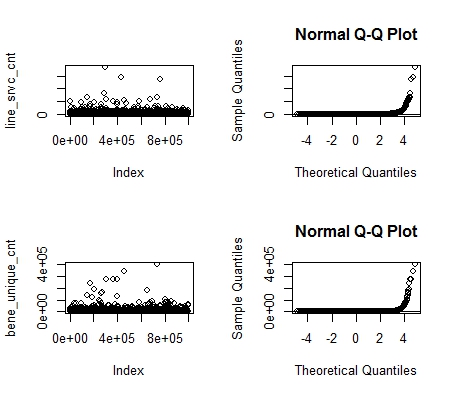
All the rows were explored to look for outliers or normality. For each column of the data set, scatter plot and Normal plot was created. All the data looked normal. Even if portion of the data were located outside the error line, they were not significantly off the line. Some outliers were observed. However, number of outliers were very miscellaneous

In order to observe how efficient are Medicare for paying off charges, the row “average\_Medicare\_allowed\_amt” was divided by another row “average\_submitted\_chrg\_amt”; the newly created row was named “Payment\_Rate”. The rows “average\_submitted\_chrg\_amt” and “Payment\_Rate” were chosen to feature clusters. Two rows were scaled; the submitted charge goes up to thousand dollars and more. Payment Rate is in 100 percentile.

To decide the number of clusters, the data were clustered by k-means clusterings, and within sum of square plot was created. The sharp drop of “within groups sum of squares” was observed. That point is the number of cluster that both minimize “within groups sum of squares” and made the cluster look less complicated. With the number of clusters selected, the data were clustered again by k-means clustering. Cluster plot was drawn to aid understanding of the analysis.

**Analysis**

First step of the analysis was to explore data to look for normal distribution of the data and outliers. For all columns comprised of numerical data, normal plots and scatter plots were drawn.



For all the columns, not much significant outliers were observed. For some normal plots such as “average\_submitted\_chrg\_amt”, most of the data points were not that off from error line. Even for “average\_submitted\_chrg\_amt”, most of the data stayed within the error line. Thus, the data was assumed to be normal without significant numbers of outliers. The correlation between each column were observed. Especially, the correlation between “average\_submitted\_chrg\_amt” and “Payment\_Rate” was observed.

average\_submitted\_chrg\_amt

line\_srvc\_cnt -0.008665657

bene\_unique\_cnt -0.009140268

bene\_day\_srvc\_cnt -0.007963009

average\_Medicare\_allowed\_amt 0.746584978

average\_submitted\_chrg\_amt 1.000000000

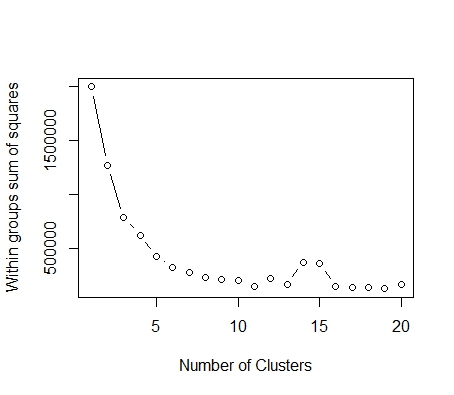
average\_Medicare\_payment\_amt 0.746525494

average\_Medicare\_standard\_amt 0.742261263

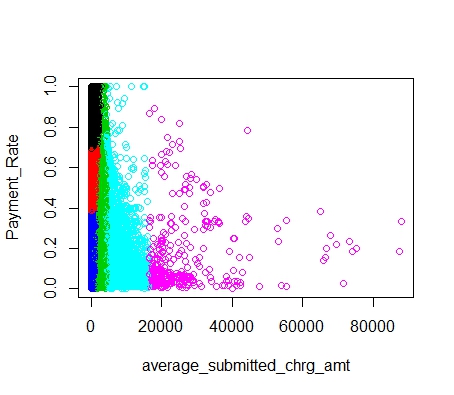
Payment\_Rate -0.208041126

There were some correlation between two features. It was negative correlation as well. That was as expected. Thus, two features were clustered together in order to demonstrate subgroups and relationship between data.

The data were clustered with different number of centers by k-means clustering, and “within groups sum of squares were calculated and plotted on the graph. Then the steepest drop on “within groups sum of squares” were pinpointed and used for further analysis.



The steepest drop on “within groups sum of squares” occurred when number of clusters were 3 or 4. However, there were still noticeable decrease in “within groups sum of squares” till 6 clusters. Thus, I decided to use 6 clusters for k-means clustering. The data was clustered and plotted.



Payment\_Rate average\_submitted\_chrg\_amt

Min. :0.0000096 Min. : 0.0

1st Qu.:0.2636470 1st Qu.: 52.0

Median :0.4094630 Median : 130.0

Mean :0.4570517 Mean : 317.1

3rd Qu.:0.6162956 3rd Qu.: 270.7

Max. :1.0000000 Max. :87978.1

Noticeable observations were made. The median and mean value of ‘average submitted charge amount’ were 130 and 317.1. The median and mean value of ‘Payment Rate’ were 40.94% and 45.70%. For data with submitted charged amount lower than median(130.3), lots of portions of data were payed higher well covered by Medicare(higher than median, 40.94%); this were represented with red and black clusters. However, as average submitted charge amount got larger, portions of patients covered higher than median Payment Rate were substantially decreasing; that downward pattern can be observed within a sky-blue cluster. In sky-blue and pink clusters, it could be observed that most of the data points were located below median Payment rate.

**Conclusions**

**Medicare in the United States generally covers inexpensive medical services very well**(below median average submitted charge amount). This trend can be observed at red, black and green clusters.

**Medicare in the United States is not doing a good job in paying for expensive and important surgeries**. This trend can be observed in sky-blue and pink clusters. This conclusion might arouse some argument, because some surgeries such as cosmetic surgeries are generally not covered under insurance. Also, 30% of total submitted charge amount, even if it is below the median value of Payment Rate, is still large amount. However, the purpose of the health care is to prepare for sudden and lethal illness, which gives people huge finance burden they are not ready for. Thus, the efficiency of Medicare should be doubted.

**Next Steps**

Possible next steps for the analysis was to divide the medical services into different subgroups by the average submitted charged amount. Then sample services that belong to each subgroups and analyze demographics: which age group needs that service and what is the most common income level of the patients using that service(to see if they can afford it)?

Healthcare companies can personalize their Medicare plans. For young and healthy customers they can decrease the benefits on heavy surgeries, which cost a lot. But they can increase the benefits on inexpensive but common medical services. Young customers are unlikely to get expensive and heavy surgeries. They are mostly unable to pay for expensive healthcare plans.

For older customers(from mid 40 to 70), they can pay for miscellaneous and cheap medical services without healthcare. However, what threatens them is sudden illness and heavy surgeries. Thus, healthcare company can focus the benefit of Medicare plans on expensive heavy surgeries.

Health care services in the United States is very capitalized. Medical treatment is hardly payed by government, unlike Canada or Norway. Healthcare company exists in order to lighten the burden of the patients, but their core goal is to generate profit. However, Medicare Company should keep in mind of their social responsibility and effectively support patients financially. This can be done with thorough statistical analysis and personalize healthcare plans.