Hyung C. Park
IEMS 308

Case Study: Men and Women in Radiology

Executive Summary

Discrimination between genders in workplace and medical care have been growing concerns in the modern American society. There has been widespread discrimination in the workplace where there is a large gap between males and females in terms of pay and positions. Analyzing the Medicare provider data set using clustering, the pay gap seems to exist within doctors who perform Diagnostic Radiology. Hospitals should focus on fostering more female residents in Diagnostic Radiology field, and the US government should monitor on how much female medical providers get paid compared to their male counterparts.

Clustering is an analytics tool that observes the data set by grouping the large data set into groups based on similarity. In this instance, this strategy was used on a Medicare data set from 2012 to 2015 that provides a wide plethora of observations on the healthcare provider as well as information on standardized amount of money paid and the types of patients treated. The clustering categorized each Radiology medical provider based on his/her revenue and location.

A brief overview of the clusters demonstrates that the medical providers who get paid more are located near major metropolitan cities. This is somewhat to be expected, however taking a more detailed composition of the Medicare providers suggests some interesting trends. The US government can see a trend that clusters with higher pay tend to be filled with men. the US government can ensure that there is no discrimination of pay within Radiology Medicare.

Problem Statement

The US government posted a large data set with different feature and characteristics on Medicare and Medicaid Services. Based on some basic exploration of the dataset and description of what each feature entails, I clustered a subset of the data set on location and revenue to see if there is any significant difference in money paid to Medicare Providers in Radiology based on their gender.

Assumptions

1) Individual medical providers are also representative of those in large hospitals– the data set encompassed information based on all Medicare providers including large organizations like major hospitals. However, only data set relevant to male and female were extracted, since it is not possible to determine gender of those in organizations.
2) Revenue should be relatively same throughout - Although there could be some differences in revenue due to how many patients each medical provider gets and the type of radiology treatment that is given, the revenue should be relatively similar throughout a large data set of radiologists.

3) All the radiologists operations and average payment after standardizations are reported – We assumed that report contained all the radiologists treatment and payment information for each individual. There could have been some instances where radiologist operations performed by some individuals were not reported. If there is a significant number of unreported Medicare payment history, this analysis could be heavily affected.

Methodology

First step in clustering analysis was data mining and cleaning. Fortunately, the data was provided to us by Centers for Medicare & Medicaid Services. Additionally, in order to enumerate one of the features of the clustering, I used an external data set that matched zip codes to latitudes and longitudes.

I grouped each Medical provider by the type of services that they provide and extracted only those who perform radiology. This choice was made because there are some Medical services that are dominated by certain gender types, and medical services are high influential of the payments that the Medicare providers get. In order to avoid this problem, radiology was selected because it encompassed a large number of datasets. The outliers in terms of average Medicare standard amount were removed, and outliers corresponded to over 2 standard deviations from the mean.

Then the columns of the number of services provided and the average amount of money paid for the services provided were used to group individuals listed multiple times into one row by calculating the weighted average of average amount of money paid. The zip code data was then matched to latitudes and longitudes using an external data set mentioned above. Since this analysis was targeted at Medicare providers in the US, those in different countries or zip codes that were invalid were removed from the data set.

The total revenue for each individual was calculated by multiplying weight average amount of money paid by number of services provided. After cleaning up the data set, the data on total revenue, latitude, and longitude were normalized. I performed k-means clustering on this data set.

Analysis

The first part of the analysis was focused on looking at the centroids and observing any abnormalities or pattern within them. Determining the number of clusters has been the biggest challenge of the clustering. The number of clusters needs to be very large in order to clearly distinguish the location and revenue of each centroid. For example, if the number of centroids is too small, then the cluster will encompass Medicare providers that have very similar revenue but farther apart, making the centroid's physical location of latitude and longitude to be misrepresentative.

In order to ensure that the number of clusters is effective at capturing the similarity relationships, I performed the clustering with different numbers of clusters and when the square distances to centroids seem to decrease at a much smaller rate, I decided to stop increasing the number of clusters. I chose to run the analysis with 14 clusters.

The clusters and the centroids demonstrated somewhat of an interesting result that the Medicare providers that make the most revenue tend to be focused around rural cities and areas. The centroids with lower revenue Medicare providers tend to be around major cities and expensive areas. Even though the data set suggests that it has been standardized to geographical locations, the fact that every high revenue clusters are around rural, while other lower revenue clusters are centered around expensive areas around the US are very intriguing.

The second trend within clustering were the ratio of women to men in each of the clusters. The ratio of females to males in the Radiology has been 6577/22275 or 0.295. This is what we expect the ratio of females to males to be for the different clusters, which will suggest that there is no difference in pay between males and females. However, there was a statistically significant difference between the ratio of females to males that were paid the most. For example, the five clusters with more than hundred thousand dollars' revenue all had significantly smaller ratio of women to men. This seems to suggest that even in Medicare there seems to be a gap between money paid to men and women.

(Important relevant figures and tables are in the Appendix, other more detailed tables and information are in the jupyter notebook)

Conclusion

Clustering analysis demonstrated that on the higher end of pay spectrum, number of men seems to dominate the number of women. In addition, the Medicare providers with lower revenue seem to be concentrated near major Metropolitan cities in the US. If other lines of service demonstrate similar results, it seems necessary for the US government to intervene and implement a system where women are not under represented and under paid in Medicare business.

Next Step

The next steps of the clustering analysis would be to assess the possible loopholes with the current analysis and expanding it to other areas. First, there were some assumptions that were made for this analysis, and it would be necessary to see if the assumptions are held. If Radiology type of service that included the largest portion of the data set has such a big difference in revenue, it might be necessary to perform a similar analysis for other types of Medicare service.

It would be beneficial to run this analysis with more features taken from other data sets. This data analysis would be much more accurate if we could run text analytics on the types of services that each Medicare provider provided and then adding that detail as one of the features of the clustering.

Appendix

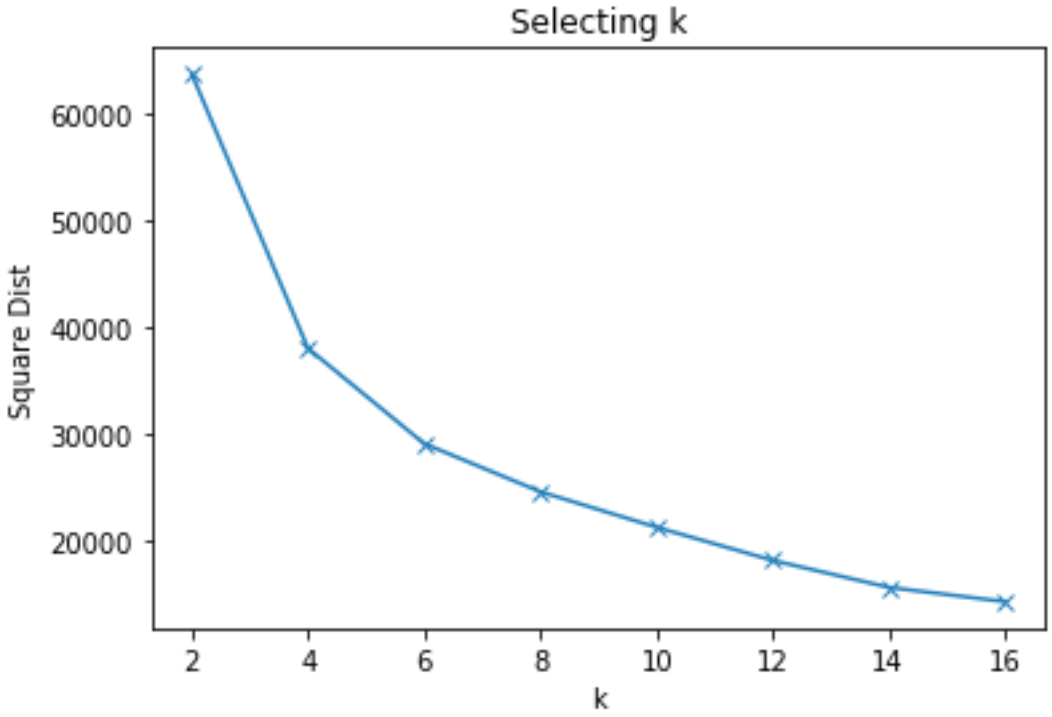

Figure 1: Selecting number of clusters or k by looking at the square distance

Figure 2: High revenue Medicare provider cluster centroid locations

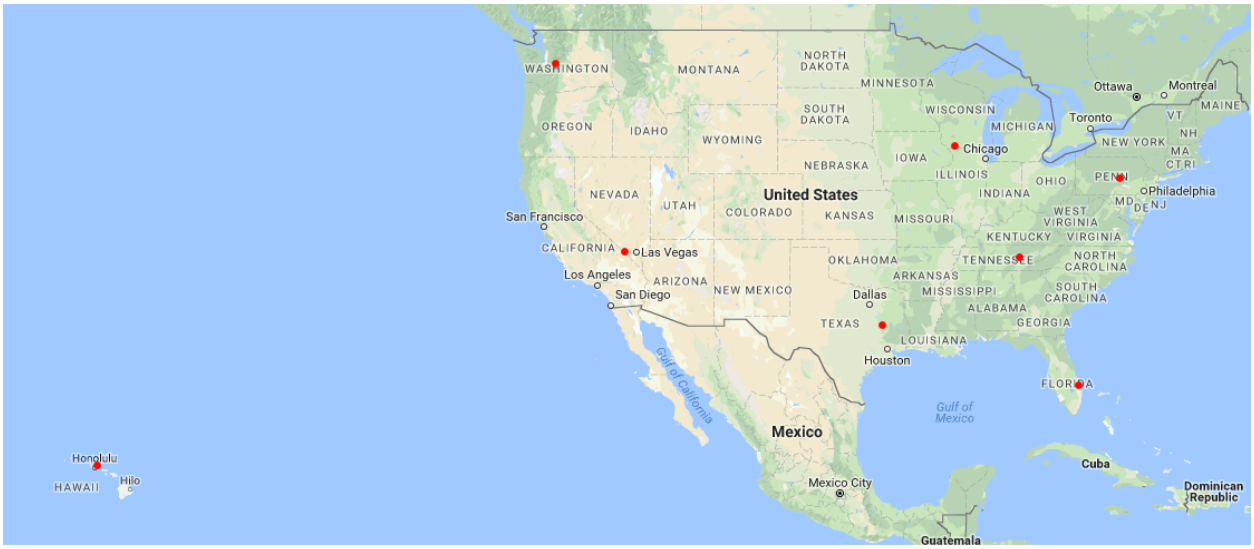

Figure 3: Low revenue Medicare provider cluster centroid locations

Table 1: The ratio of women to men for the clusters with the 5 highest average revenue (for reference, the ratio of all women to men in Radiology is 0.295)

Revenue	Ratio
$789,602	0.2179
$371,666	0.2578
$195,041	0.1236
$172,585	0.2132
$158,279	0.1283