Arshya Srinivas

IEMS 308

January 26th 2020

**Clustering Assignment Executive Summary**

**PART I: UNDERSTANDING THE DATA**

While looking at the data initially, it was noticed that there were an incredibly large number of rows and columns, and most of the data was not clean. This was especially seen in the “nppes\_provider\_credentials” column, as similar degrees were not inputted in the same manner. Hence, it was decided that a subset should be taken from the data set. After looking more deeply into the data using SQL, I came to the conclusion to subset the data based on Midwestern states and the most popular provider types. My process of understanding the data and finally subsetting the data can be seen in the document “Output code from SQL”.

**PART II: DATA EXPLORATION**

The data exploration was conducted in three separate ways: by creating scatter plots, heat maps and bar plots.

I was initially intrigued to see if there was a relationship between the average Medicare payment amount and the average Medicare standard amount. However, I was not surprised to see that there was an almost linear relationship between these two variables.

To look deeper into the relationships between the numerical values of the data, I constructed a heat map. There are strong relationships between the average Medicare payment allowed and the average Medicare payment amount, as well as the average Medicare standard amount. Hence, I am going to explore how these correlations differ along different parameters.

Finally, I looked at bar plots of the distribution of provider types in the subset of data. It appears that diagnostic radiology is the most common provider service in the dataset, followed by family practice and internal medicine. Hence, when examining the results of the clustering analysis, I will be vigilant about these provider types.

**PART III: BUSINESS QUESTION**

The business question that I proposed is the following: Does the gender of the provider type affect the difference in costs of Medicare and the amount able to be requested? Which genders should individuals avoid in order to ensure they get the most out of Medicare?

After deciding on my business question, I conducted one hot encoding on all the variables that I was interested in using that were not numerical in nature. I also split the datasets such that there was a female only dataset and a male only dataset.

**PART IV: CLUSTERING SOLUTION**

The methodology I followed for conducting the clustering solutions is below. I conducted the same process for both male and female datasets.

1. Standardize the dataset
2. Calculate the SSE for KMeans clusters with the maximum number of clusters set to 30
3. Selected the number of clusters with a suitable SSE (i.e. the number of clusters at which there is a kink on the SSE curve)
4. Conducted KMeans clustering analysis again
5. Proceeded with the analysis of the clustering solutions.

Looking at the female dataset, the number of clusters that were chosen was 12 according to the SSE graph. One of the most interesting things to note from this analysis is that cluster four had the lowest average Medicare allowed amount (42.04) but also had the highest average Medicare submitted amount (234.92). Additionally, it has the lowest average Medicare payment amount (30.64) and the lowest average Medicare standard amount (29.36). Looking at the provider types, it can be seen that this cluster has a large amount of data for providers that provide the service of diagnostic radiology (0.256). This suggests that there may be a pattern behind the difference between the Medicare received and the Medicare requested for female providers of diagnostic radiology. However, it is important to notice that both cluster 11 and cluster 4 are anomalies in the dataset, with cluster 11 having 1 data point and cluster 4 having 52.

Looking at the male dataset, the number of clusters that were chosen here were also 12. In this dataset, there are a couple more anomalies, including cluster 9, cluster 8 and cluster 10. Upon first glance, it can be seen that the Medicare costs are generally higher than that of the female providers, but so are the average submitted costs. Looking at the highest average submitted charge amount (i.e. cluster 12, with a value of 327.42), it can also be seen that the submitted cost is significantly lower. While there was only a marginal difference between the numbers of data points in this cluster to other clusters for the provider type orthopedic surgery, it is likely that these costs are linked to this type of Medicare service due to the higher number of data points in this cluster for the provider type (0.081).

In conclusion, there appears to be an underlying pattern behind the gender and Medicare costs. As a general statement, female providers may be correlated to charging more for services linked to diagnostic radiology while male providers may be correlated to charging more for services linked to orthopedic surgery. However, a more in depth analysis is required to be completed in order to find a statistically significant correlation.