**Gene Chen - Assignment 1 Clustering**

**Executive Summary**

Medicare is a federal funded health care program that provides health insurance for people who are 65 or older and others with certain disabilities. It is one of the largest federal programs and helps make up over 20% of the federal governments annual spending. There are 3 different types of Medicare. Medicare A which covers inpatient hospital stays and other hospital insurance, Medicare B, which covers doctor’s services and outpatient care and Medicare D, which covers prescription drug. Because of these various types of Medicare coverage, more than 44 million beneficiaries, totaling over 15% of the U.S. population, are currently enrolled in the Medicare program. With such a large amount of Medicare beneficiaries comes quite a bit of data, which is compiled by the Centers for Medicare and Medicaid Services or CMS.

Using this data, we can apply a grouping method called clustering which allows us to see data points that are in theory “more similar”. We can also use this method to determine relationships between the different features of the datapoints. By doing this we can more accurately visualize the different types of Medicare providers and target these subsets with business driven reasonings.

With this methodology, I was able to cluster together the data points using the K-means algorithm to look at highest average Medicare spending by State. The results show that the cluster which was characterized by a high average Medicare spending also happened to consist with most data points from the state of New York, New Jersey and Florida. It happens to be the case that these states are the 3 largest Medicare spender per enrollee as stated by the Kaiser Family Foundation, which further supports my results. Using these insights, we can now look deeper into the average health of Medicare beneficiaries and determine whether large spending for Medicare is a driver of how healthy these individuals are.

**Problem Statement**

The Centers for Medicare and Medicaid Services have a large dataset with information on services and procedures provided to Medicare beneficiaries. Our job was to group together the data points in a method called clustering to determine the different ‘types’ of beneficiaries and which features characterize them. With this information, we will be able to pursue certain goals with operational and/or financial intent.

**Assumptions**

• The data provided by the CMS is valid and correct

• We dropped ~2% of datapoints that had corrupt and/or missing data. This assumes this will not skew the true distribution of features.

• The 10% random sampling of the entire dataset has a comparable feature distribution of the entire population

• We replaced the K-means algorithm with the Mini Batch K-means algorithm. This sacrifices accuracy for processing power and memory. This assumes the new algorithm doesn’t skew our results

**Methodology**

We first began by pulling the public data off the CMS website and initiating data cleaning. This comprised of properly formatting the tab delimited data as well as dropping rows with improper data in the dataset matrix. We then took a random subset of 10% random samples from the dataset due to the memory limitations of our computer. These ~900,000 samples will still allow us to properly cluster the data while providing much needed performance to run our analysis. We finished our data cleaning by extracting the columns or “features” of the dataset that would be most beneficial to solving our business problem. These features were [Provider State, Provider Type, Provider Gender, Average Payment Amount, Unique Beneficiary Amount].

The next step in our methodology was data exploration. Plotting our non-categorical features using histograms, we can see the distribution of points within each feature. Since these were heavily skewed, we decided to log transform the values. Now that the distribution of these values is more bell-shaped, they can be more easily interpreted.

Finally, we began the clustering step. We first completed preprocessing steps, consisting of changing the categorical features into numerical values using one hot encoding as well as normalizing each feature. This allows all features to be compared properly as they are now all on the same scale. We then initialized K-means by running the algorithm for k=1 to k=9. Compiling the objective values for each k, we were able to produce a scree plot as shown below. The plot shows that before k=7, the sum square error of the clusters drops steadily. However, as we reach k=7 and beyond, the benefits of each additional cluster begin to diminish as we reach the convergence point.

After determining the optimal no. of k clusters, we run the K-means algorithm one more time to store our results.

**Analysis**

We began our analysis by first determining the cluster with the highest average Medicare payment. We did this by taking the cluster centers of the corresponding feature and comparing them on a bar graph. As you can see, cluster no. 5 can now be easily characterized as having datapoints with high average Medicare payments. Now, looking into cluster 5 individually, we can visualize the distribution of states within this cluster. As shown below, the largest presence within this ‘high average payment’ cluster is New York, with Florida and New Jersey not far behind. This infers that these states are those that can be described as having high average payments.

**Conclusion**

We were able to successfully group states with high average Medicare Payments. Within this group, of the most prevalent states were New York, New Jersey and Florida. This infers the relationship that **New York, New Jersey and Florida have some of the highest Medicare spending per enrollee.** Some next steps with our analysis include corroborating our findings with other teams who are working on similar studies. If found to be true, we can use our findings to answer more useful financial and/or operational use cases. For instance, if we can be sure of the aforementioned relationship, we can then measure average health of Medicare beneficiaries of each state and attempt to determine the relationship between Medicare spend and health of Medicare beneficiaries.

sdfsdfsdf