Clustering/Location Recommendations

Executive Summary:

A large pharmacuetical company is looking to expand in Colorado as it is growing rapidly in population. The state also has a lot of open land even in fairly populated areas. The company has a large amount of capital. Rather than just doing general research on different areas, it should utilize available data and use the numbers to determine which areas have the highest potential.

The Centers for Medicare & Medicaid Services (CMS) a public data set prepared. It provides information on services and procedures to Medicare beneficiaries by physicians and other health care profressionals all over the world. The idea is that by placing a pharmacuetical store near an area that has a lot of Medicare providers where the benificiaries are allowed large amounts of Medicare payments, the benificaries will come to the store more frequently in between visits to their physician out of convenience.

A technique commonly used to target markets is clustering. In this case, clustering can be utilized to determine which locations can bring in the most revenue. There are a large number of both numeric and non-numeric features in the dataset which can create dimensionality problems. Thus, a careful analysis of only the necessary features should be considered when running the clustering algorithm.

Problem Statement:

Our client is a large parmaceutical company that wants to expand the number of stores they have. Our job is to determine the best location for the store that will bring in the most revenue.

Methodology:

We filtered out features that we knew wouldn't give us any useful information for our problem at first glance. In this case, we removed the following features immediately: national provider identifier, first, middle initial, and last name of the provider, provider credentials (dirty values), provider gender, provider street address (too specific), provider state and country (we know we are looking at CO), and HCPS codes and descriptions (too many values for one hot encoding so no information gain).

Before doing any actual cluster analysis, we also need to examine the data itself and gather some initial insights. We will first take a general look at a histograms of the numeric data as seen below:

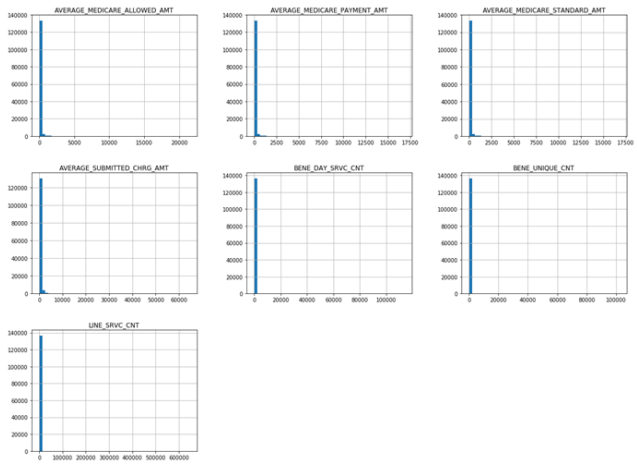

Figure 1: Histogram of Raw Data

The histograms above tells us that the data is heavily skewed and that outliers exist. However, given the size of the dataset it's likely that the few outliers that may exist shouldn't have any major effects to the clustering method.

Taking the log of the data later will reduce the skew and make the data easier to visualize and work with. However, before that, we also wanted to look at the multicolinearity of the data so that we can reduce the number of features in our dataset if certain features are high predictors of other features. We found the top 10 pairs of features with the highest absolute correlation:

Attribute Pairs		Absolute Correlation
AVERAGE_MEDICARE_ALLOWED_AMT	AVERAGE_MEDICARE_PAYMENT_AMT	0.99885
AVERAGE_MEDICARE_PAYMENT_AMT	AVERAGE_MEDICARE_STANDARD_AMT	0.998649
AVERAGE_MEDICARE_ALLOWED_AMT	AVERAGE_MEDICARE_STANDARD_AMT	0.998232
BENE_UNIQUE_CNT	BENE_DAY_SRVC_CNT	0.958517
AVERAGE_SUBMITTED_CHRG_AMT	AVERAGE_MEDICARE_PAYMENT_AMT	0.741196
AVERAGE_MEDICARE_ALLOWED_AMT	AVERAGE_SUBMITTED_CHRG_AMT	0.740471
AVERAGE_SUBMITTED_CHRG_AMT	AVERAGE_MEDICARE_STANDARD_AMT	0.735757
LINE_SRVC_CNT	BENE_DAY_SRVC_CNT	0.331825
	BENE_UNIQUE_CNT	0.270502
BENE_DAY_SRVC_CNT	AVERAGE_SUBMITTED_CHRG_AMT	0.018722

Figure 2: Top Pairwise Absolute Correlations

There are a few pairs of attributes that are almost directly correlated with each other. We decided that any one of the attributes of any pairs that have an absolute correlation of greater than 0.96 can be removed since the information we gain is essentially the same. Thus, we remove the features average medicare payment amount and average medicare standardized amount, but we kept the feature average medicare allowed amount.

At this point, we realizedt that the only numeric data that we are interested in is the average medicare allowed amount and the average submitted charge amount since these will likely be the most telling indicators which providers are utilizing the most amount of money. Thus, we remove the rest of the numeric features and keep only these two.

At this point, we are comfortable with the features we have now. As mentioned before, we want to scale down the data and reduce skew, which we do so by taking the log of the numeric data. The histogram of the log transformed data shown below:

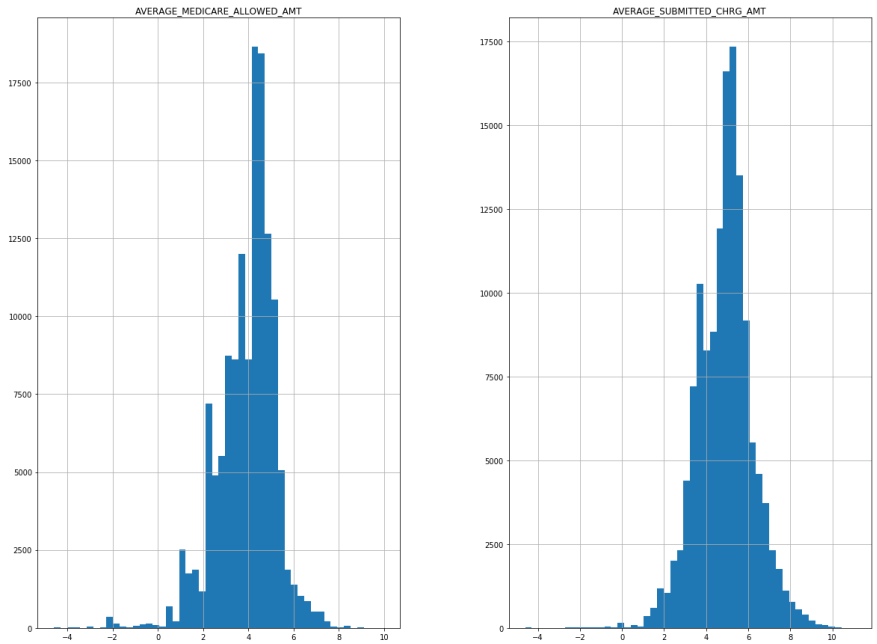

Figure 3: Histogram of Log Transformed Data

The data now has a more distinct distribution without any extreme skew. In fact, both features appear to have a somewhat normal distribution. This will aid in making better clusters. The next step, is the convert our non-numeric data into numeric data which we do using one hot encoding to maintain the orthogonality of the data as much as possible which we lose with ordinal encoding. This process can be problematic as it can drastically increase the number of features. It will not only processing time but also increase the sparcity of the data making it harder to find good clusters without sufficient data points (curse of dimensionality). As such, we need to remove many of the categorical variables that we are not interested in. The only categorical features that give us useful geo-information are the city and zipcode features so these are the only two categorical features that we will use.

Zipcodes are more or less random so we treated them as strings as opposed to numbers to analyze. We only use the first 3 digits of the zipcode as those are the most important digits in providing geographic information. This further reduces dimensionality as well.

After log transforming the numeric data and one hot encoding the non-numeric data, we finally normalize the data. This makes it so that all the features have values of similar range resulting in more meaningful clusters. The data is now ready to be input into our model.

Modeling and Results:

We want to target the location that will bring in the most revenue, but doing so will likely require generalizing the potential demand as much as possible. As such, we don't want too many clusters because it may find locations that are too specific. We felt that 3-5 clusters would be enough to effectively generalize without "overfitting." To determine which number of clusters will provide the most significant and reliable results, we use both the silhouette scores and plots as our performance metrics. Below are the results of the different models.

3 Clusters:

Average silhouette score = 0.169

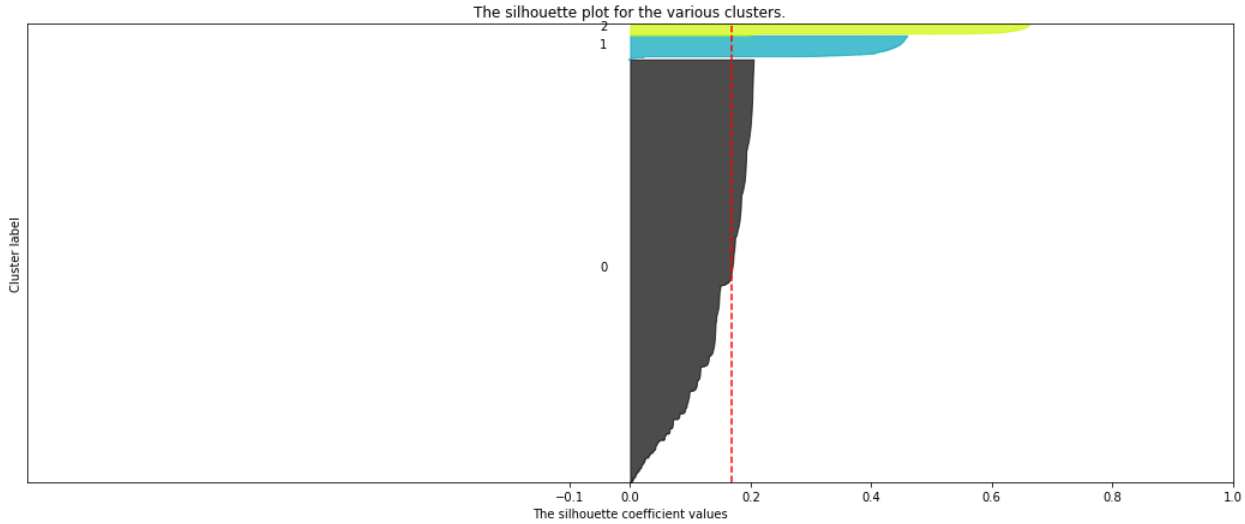

Figure 4: Silhouette Plot for 3 Clusters

4 Clusters:

Average silhouette score = -0.057

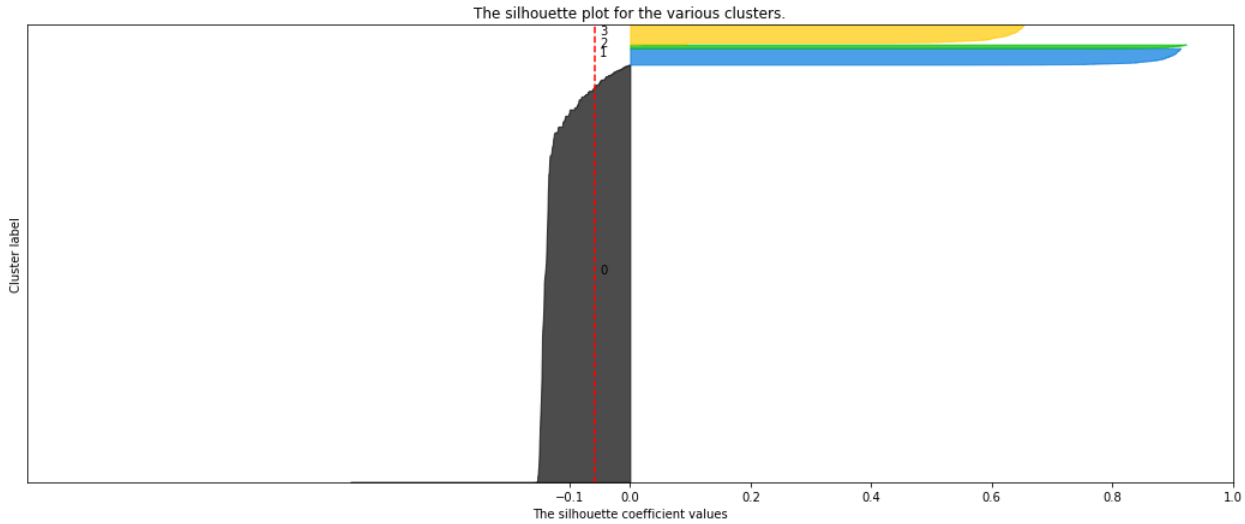

Figure 5: Silhouette Plot for 4 Clusters

5 Clusters:

Average silhouette_score = -0.108

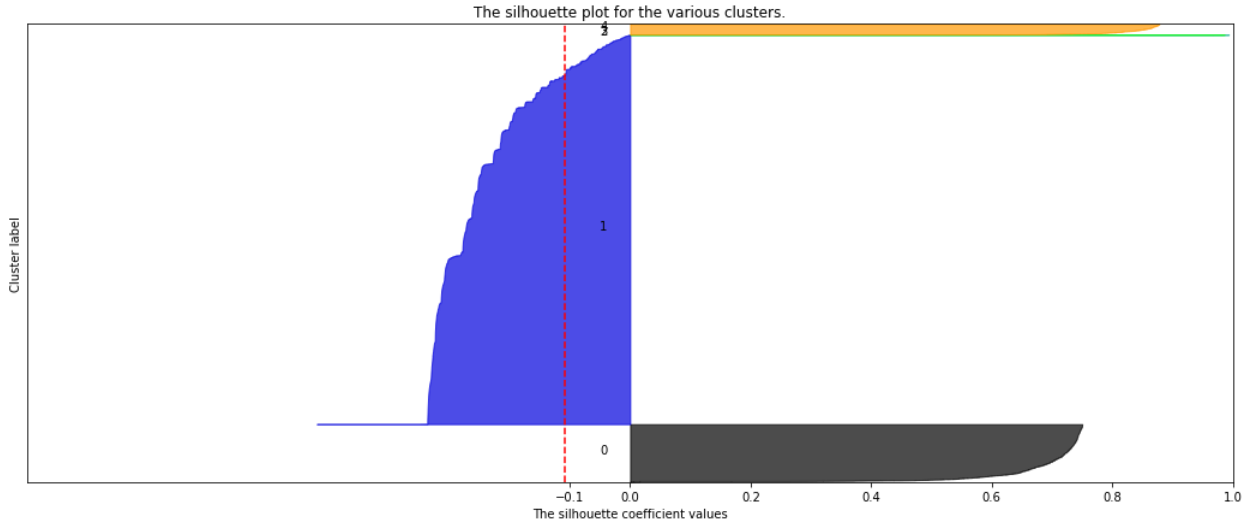

Figure 6: Silhouette Plot for 5 Clusters

Ideally, we would want a model where the average silhouette score is closer to 1 as that would indicate more distinct clusters. However, all the models that we have produced have an average score less than 0.2. While it isn't ideal, this doesn't necessarily mean that we can't use any of the models. Since the 3 cluster model was the best model from the score and plot, we will look at the results of this model. Unfortunately, due to the scarcity of the data, it is difficult to visualize the clusters. Instead, we will look at what the aggregate numbers (after reverse transforming everything) tell of us of each cluster. Cluster 0, by far, has the largest number of instances and thus also has the largest sum of average Medicare allowed amount of $13,224,556 and submitted charge amount of $43,721,567 suggesting that it will be our cluster of interest. Now from cluster 0, we need to determine what location is optimal. In this case, we will use the data point that is closest to the centroid of cluster 0 which happens to be Denver with a zip code starting with 802.

Conclusion and Limitations:

From the results, we would not have confidence that any of the models are meaningful given the poor silhouette scores. However, it's possible that finding distinct clusters for this problem is not realistic. It may just be that there isn't any strong correlation between the location and the amount the providers utilize. Still, the results from the data do seem to make sense in that **Denver with starting zip code 802** would be the best location to bring in a lot of revenue considering that Denver is the most populated area in Colorado.

For next steps, we can try doing the same analysis but with different clustering algorithms such as k-mediods or hierarchical clustering. As for the client, we recommend that they do more research on the risks on opening a store near the given location. Denver property values is fairly expensive. They can also look into a few other locations that are good and do the same risk assessment. Because the client is a pharmaceutical company, if they can get access to a data set that has similar information but more specifically information on the amount of money providers utilize for drugs, we could obtain more accurate results.