Customer Segmentation

The overall idea is to identify key features and find explanations for clustered shapes by K-means clustering algorithm. The date of when the job is created alongside the first time the job is realized is given. The year, month and the difference between these two dates are extracted as a part of the data preparation process. The difference is represented in days and not scaled between 0 to 1 because a big difference may bring meaningful patterns to the clusters.

In the next phase, I organized the categorical features. Year and JOBSTATUS are encoded using one-hot encoding method. However, attributes such as LOCATION and SERVICENAME are neither label encoded nor one-hot encoded. I simply removed the pre-fix term and scaled the numeric value between 0 to 1 for the sake of normalization. The reason is to reduce the variance among numbers with considerable differences. Some of the features such as PRICE, REVENUE, etc. are not coded because their actual values are important to the K-means model. Such values, in this case, could make the main difference in the patterns. Throughout the data preprocessing stage, regular expressions are used as much as possible to reduce the complexity of the solution.

In the next step, I make sure there is no missing value, extra features are removed and the data is very well prepared for the K-means model. However, before running the K-means model, I use the elbow method to figure out an efficient number of clusters that must be used in K-means. The elbow is located at 4, so I chose 4. Notably, I could also choose 5, but I did no significant difference in the clusters. So, I decided to stick with 4 clusters.

In the next phase, I visualized the distribution of attributes across each cluster and compared them together. My comparisons show that if there is a long difference between the JOBCREATEDATE and JOBDATE, there is a higher chance that the price is high, or the JOBSTATUS is unsuccessful. Furthermore, the REVENUE is going to be centered around 0, which is expected. On the contrary, the jobs that were from various locations, or dominant locations, chose services with lower prices and the rate of success was higher. Therefore, the following is the interpretation for each cluster exclusively:

Cluster 3: Rate of success is 0. Location tendency is centered at 0.4. Price is ranged low between 1 to 2.5 and normal to expect that revenue is centered at zero.

Cluster 2: 29266 successful registrations. The tendency of the location follows the main pattern in the data. Registrations are with regards to low price services. Revenue and price are highly correlated yet low and do not vary considerably. Therefore, it is more likely that there are more successful jobs with low prices, and these jobs are registered in mostly two particular and dominant locations.

Cluster 1: Prices ranged 2-8 which makes it distinguishable from cluster 2. There is more variety in revenue and price compared to all other clusters, and location tendency is centered at 0.4. It should be noted that there is no success in this cluster and apart from cluster 2, different clusters, including cluster 1 are very low in terms of data samples.

Cluster 0: Locations are more variant than clusters 3 and 1. 3 successful registrations exist within this cluster. Prices are high, yet revenue is low.