Pre-speech

Our project name is called customer segmentation by clustering algorithms.

First, let me introduce our dataset.

Our dataset is from Kaggle, the title is "credit card dataset fro clustering" .

The dataset contains 9000 active credict card holders usage behavior during the last 6 months. And it has 18 behavioral variables.

I put the data dictionary here to show some of them. For example, the balance is the left money in the bank account and the purchases is the total amount they've spent using the card. And ONEOFF_PURCHASES is like buying something big or expensive in one shot

Now, let me introduce the procedure of data processing.

- First, we check the the missing value. we can see that the minimum_payments has about 300 missing entries. But it makes up only 3.5% . So we decide to remove these values
- Then, we draw a boxplot to show the data distribution. We can see that these 5 variables 's outliers are over 1000 counts. But we do not simply remove all of them, for example, people from different class would have different credit card behavior. So it makes sense that we have some outliers in here, we just remove some very extreme values in the figure.
- And we find some strong positive skewness like minimum payments, purchases, cash advance..........
    - we use log-transformation to change the skewness of the distribution

K-means:

I use k-means for the clustering. The basic principal of k-means is that we choose initial centroids and assign each point to a neareast centroid and update the cluster centroid, then we keep repeating until the centroids stop changing.

I use the elbow method to find the best clustering number K. We calculate the sum of square error and draw the figure like this. The best k is the point that start to flatten , in the fig is 4, so I set the k to be 4.

And then I use t-sne to reduce the dimensions to 2D space for visualization. We can see that we have 3 cluster, the perple one is cluster 0,the green one is cluster 1, the yellow one is cluster2.

And I calculate the evaluation metrics.

- The Silhouette is 0.2396, which means
- The CH is high, which means the points in one cluster are similar.
- The DBI is 1.53

We see the features distribution in 3 clusters.

The cluster 0 has low balance, but high frequent purchase and minimal