WHOLESALE CUSTOMER SEGMENTATION

By Tram Tran

22 November 2016

Purpose

In this project, I would like to conduct customer segmentation based on a dataset of a wholesale distributer in Portugal. Customer segmentation is very crucial as it helps business to allocate resources efficiently. The wholesale distributer can tailor its marketing effort (promotion, advertising), pricing, distributions to match the retailers' need. This personalization will help business to achieve maximum values from both high and low-profit customers.

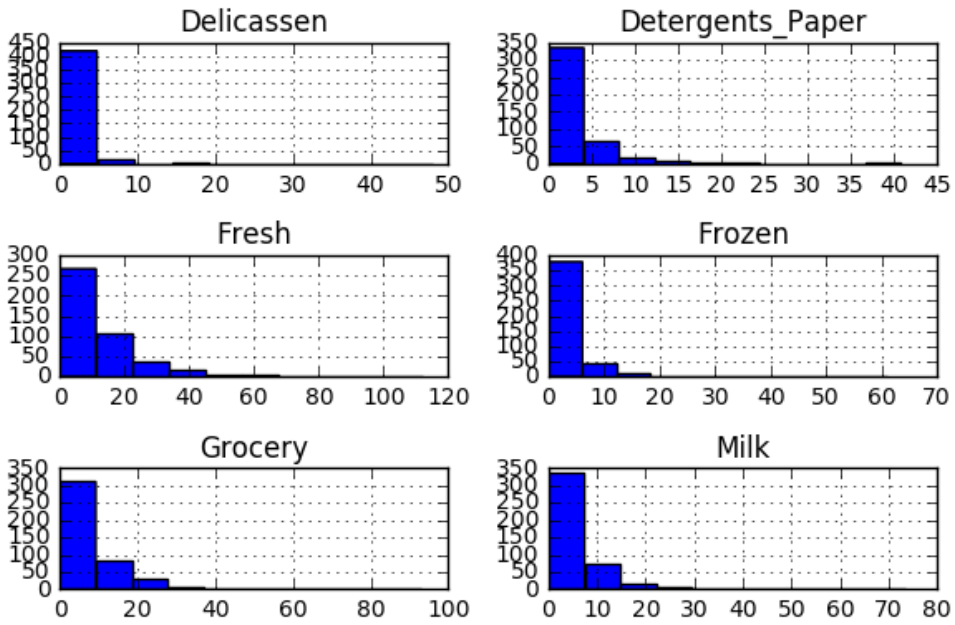
Describe dataset

The data is taken from Machine Learning Repository UCI- Link. The dataset includes 440 observations with 8 characteristics, 4 of which are continuous variables representing spending in categories, 2 of which are nominal variables (channels, regions). The data contains no missing values.

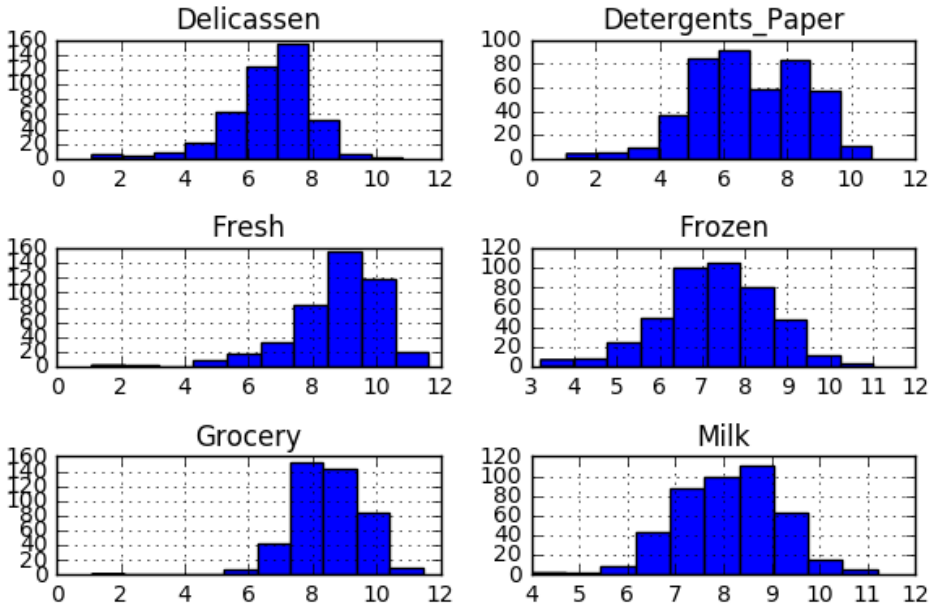
Method

Customer segmentation is the practice where we divide customers into sub-groups with similar characteristics. The more homogeneous inside each segment and the more different between segments, the better our customer segmentation. For this project, I apply K-Means clustering to find clusters (segments) for the wholesale firm.

To understand how K-Means work, read this article which illustrates the process in the algorithm. Intuitively, K-Means is a distance-based method clustering. Imagine we put all attributes as our criteria to segment on different dimensions and observations are points that can be measured on these dimensions. In K-Means algorithm, we need to choose number of clusters that are supposed to find in our dataset as centroids. We then measure the similarity of observations to group into a cluster by the distance from points to different centroids and allocate points to the closest centroids. From this illustration, we can easily see that the result clusters will depend on the number of k cluster that we specify in the first place and on the distance method. As K-Means is distance-based method, it works well with numerical data and the data is normally distributed. So first, let's take a look at the distribution of spending on different categories.

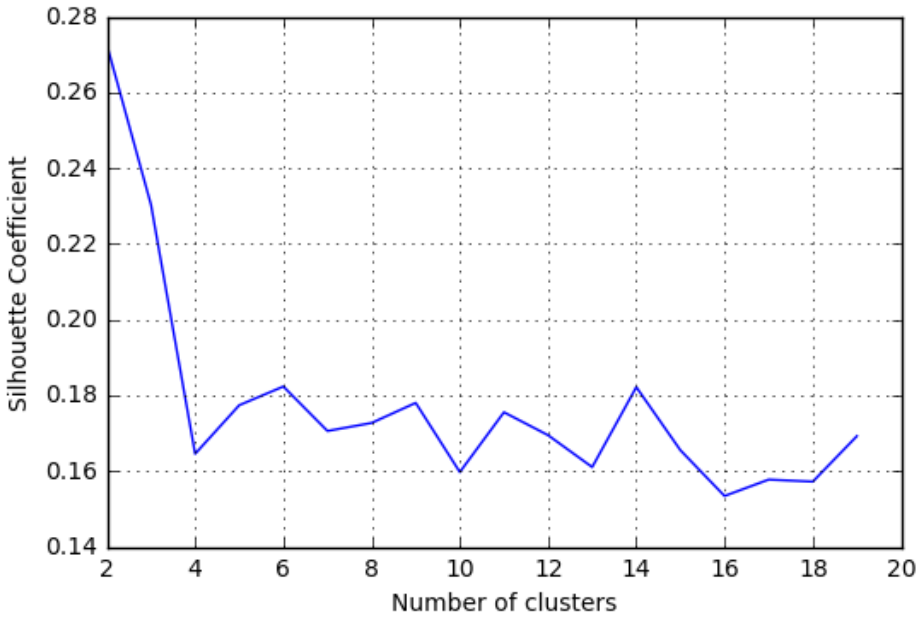


The above histograms show that the spending for all categories are highly right-tailed skewed. Hence, I will transform this data by taking log. We find that with this transformation, the spending data are now rather normally distributed.



The next step is to rescale the these transformed data as clustering will only work well when data have similar scales. As we have done with the pre-processing stage, now we can apply K-Means clustering method using scikit-learn. As mentioned above, the number of k will alter the result clusters, so we need a measurement to verify our choice. Here, I use silhouette score, for technical understanding read this article. The idea is to check whether we classify each point to the closest cluster correctly and it is done by comparing distance of intra-cluster and distance from an observation to the nearest cluster that it is not assigned

to. Silhouette score is from -1 (worst) to 1 (best). We find that k=2 achieves that highest score so we choose to have 2 segments.



Now, we can do profiling for these two clusters.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
cluster						
0	0.218190	-0.606251	-0.642083	0.259537	-0.662475	-0.147432
1	-0.289766	0.805127	0.852714	-0.344676	0.879795	0.195795

Notice that the numbers above are for scale data and we keep them to easily differentiate clusters.

- Cluster 0: Prefer fresh and frozen products and buy through Horeca channel.
- Cluster 1: Buy mainly detergent papers, groceries, and milk. Buy delicatessen more than group 0 but not much. Retail channel.

Though the finding is not so surprising, we have verified our assumption that customers who buy in Horeca mainly buy fresh and frozen products, and generally much more than in the retail channel. While customers buy mainly milk, groceries, detergent papers and delicatessen in retail stores.

The code can be found here.