Dec 17, 2021

Customer Segmentation with K-Means Clustering

University of Minnesota
STAT4893W Research Report
Sihyeon An

Introduction

Companies try various strategies and marketing methods to sell their products. In order to attempt marketing strategies, it is essential to understand the characteristics of customers. However, each consumer has their own needs and wants. So, identifying and understanding the needs of all individual customers is costly in terms of time and cost and it is practically impossible. This problem can be solved through customer segmentation. Customer segmentation is clustering consumers of the same needs through their characteristics. In short, marketing costs can be reduced through customer segmentation. Through customer segmentation, a company can use different marketing methods for each separate consumer group. In addition, a group of potential customers can be found. As a result, marketing costs can be reduced by marketing only to targeted groups, not all customers. There are various ways to sort customers. The cluster analysis method is one of the most commonly used methods for customer segmentation. With cluster analysis, the characteristics of each group can be identified by grouping similar groups through the characteristics of the given data. Among the various cluster analysis methods, we chose to use the K-means clustering method for the research. K-means clustering method is to classify observations with similar characteristics into k clusters. Then, identifying the characteristics of each cluster. One of the advantages of K-means clustering is that K-means clustering is easy to conduct and does not require detail about the information of data. Also, it is easy to apply to various data.

Research goal

The goal of the research is to cluster customers with the given data information. Thus, we visually inspected the clusters by making plots using variables in a dataset. Then, we estimated which variables can distinguish clusters clearly. With the plot results, we also evaluated which clustering plot is informative for customer segmentation.

Dataset

The dataset has customer information about wholesale distributors. The data is collected in Lisbon, Portugal. The data includes the annual spending in currency units on 6 products and 2 variables that are about customers' residence and the place they buy the products. In detail, the data include annual spending of 6 items (fresh, milk, grocery, frozen, delicatessen, detergent with paper products) and 2 categorical data (channel and region) for 440 customers. Channel means the places customers spend their money. It is divided into two sections. The first channel combines hotel, restaurant and café. The other channel is a retail channel. For the region variable, it has three parts. Lisbon, Oporto, and others are the parts of the regions where customers reside (Abreu, 2014). We will determine how many number of clusters are the clearest and appropriate with K-means clustering method with the dataset.

Methods

For the research, K-means clustering method was used. K-means clustering is one of the unsupervised learning clustering methods. Unsupervised learning is a type of machine learning. The purpose of unsupervised learning is to determine a characteristic (pattern or structure) of data. It means that there are no response variables and there are only explanatory variables for unsupervised learning (University of Cincinnati, 2021). K-means clustering method divides data values into several clusters. Data values belonging to each cluster have similar characteristics. In this way, customers with similar characteristics can be divided into several clusters to compare and grasp the characteristics of each cluster. K-means clustering equation is as follows.

$$J = \sum_{k=1}^{K} \sum_{i \in C_k} d(x_i, \mu_k)$$

J : Objective function

K : Number of clustering

C_k : Set of data belonging to the cluster of k

μ_k : Centroid of cluster of k

d : Distance between two data points

For the research, Euclidean distance was used. Thus,

$$d(x_i, \mu_k) = \parallel x_i - \mu_k \parallel^2$$

For the steps of K-means clustering algorithm, select k for the cluster center points. Next, after allocating each observation to the nearest cluster center then a cluster center is newly calculated. The algorithm repeats the process until the center is the same as the center of the newly calculated cluster. So, categorical data cannot be used because of the characteristics of the k-means clustering method that the center value must be defined. So, the mean value is used to define the center value. Categorical variables are discrete, so it is not meaningful when calculating Euclidean distance for categorical variables (University of Cincinnati, 2021).

Euclidean distance :$d_{euc}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

Thus, to conduct the research, categorical variables that are channel and region in the dataset were excluded from the analysis because categorical data cannot be used for K-means clustering. The k-means clustering method has various advantages. First of all, the method is easy to apply because prior information for the data is not required. In other words, only the distance among observed data is the value required for analysis. Also, with many variables, it is a computationally faster way than hierarchical clustering when the number of clusters is small (Priya Pedamkar, 2021).

One of the most important for the K-means clustering is determining the appropriate number of k. K means the number of clusters. There are several ways to find the optimal number of clusters. Elbow, Silhouette, Gap Statistic methods were selected and used to find the optimal cluster's number and the results of each method are compared.

For the elbow method, we have to check TWSS (total within the sum of squares) to find the optimal number of clusters. As the number of clusters increases, the total within the sum of squares decreases. Then, we need to find a point where the total within the sum of squares decreases rapidly. The point means the optimal number of clusters. (University of Cincinnati, 2021).

The silhouette method is a way of checking how efficiently and well the divided clusters are separated through cluster analysis. The efficient separation of clusters means that the distance from the other clusters is separated and the data from the same clusters are closely grouped together. Silhouette analysis is based on the silhouette coefficient, and each individual data has a silhouette coefficient. At this time, the silhouette coefficient of individual data is an indicator of how the observations are clustered closely in the same cluster and how far it is separated from data in different clusters. The silhouette coefficient value s(i) is defined as follows.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

- S(i) : Silhouette coefficient.
- i: Index for each data point.
- a(i) : Value obtained by averaging the distance from other observations in the same cluster as i data point.

- b(i) : Average distance from the nearest cluster among the clusters to which the corresponding data point does not belong.

So, $b(i) - a(i)$ calculates the value of the distance between the two clusters. To normalize, divide it by max $\{a(i), b(i)\}$. The closer the silhouette coefficient is to 1, the farther it is to the nearby cluster. The closer the silhouette coefficient is to 0, the closer it is to the nearby cluster (Ankita Banerji, 2021).

For the gap statistic method, it makes a comparison the total intra-cluster variation for different values of k with their expected values under null the reference distribution of data. The reference dataset is produced using Monte Carlo simulations of the sampling process. It means that we calculate its range [min(x_i), max(x_i)] for each variable (x_i) in the dataset and generate values for the n points from the interval min to max. In short, total intra-cluster variation is calculated using different value of k. (University of Cincinnati, 2021).

$$Gap_n(k) = E_n^*\{logW_k\} - logW_k$$

E_n^* means the expectation under a sample size n from the reference distribution. It is also determined through bootstrapping (B) by generating B copies of the reference datasets and, by calculating the average log (W_k^*). Gap statistic method calculates the deviation of the observed W_k value from its expected value under the null hypothesis. Thus, the optimal number of clusters will be the value that maximizes $Gap_n(k)$ (University of Cincinnati, 2021).

Results

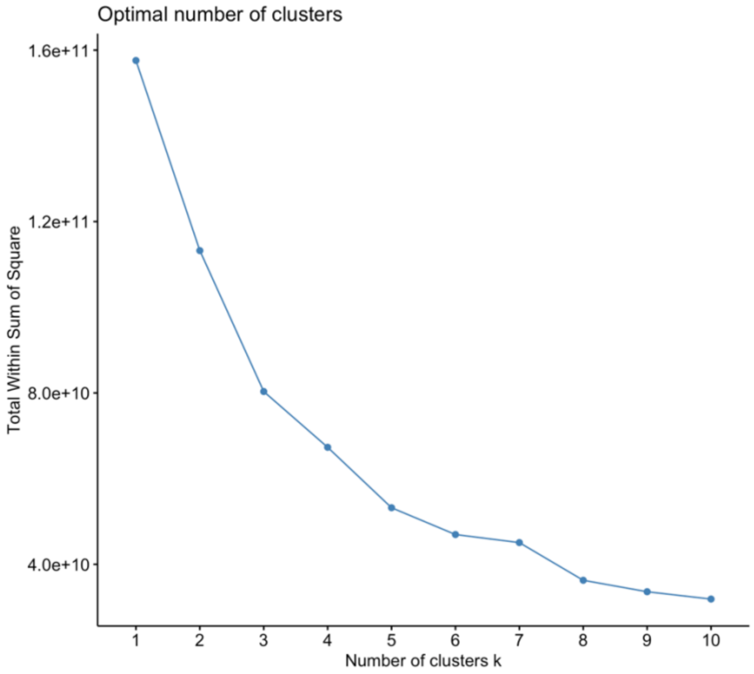

Figure 1. Elbow method

Based on the elbow method graph, when k=1, the total within sum of square is the highest. As the number of clusters increases, TWSS decreases. Also, when k=3, the decrease rate of TWSS rapidly decreases. So, the optimal number of k is 3 with the graph result.

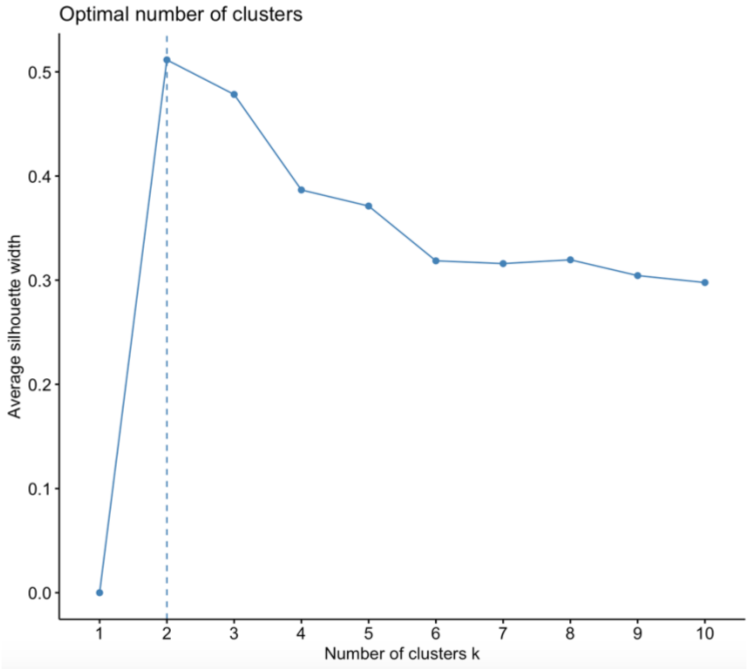

Figure 2. Silhouette method

However, using the silhouette method, the average silhouette width is the highest when the number of clusters is 2. Also, when the number of the cluster is 3, the average silhouette is the second highest. In other words, 2 clusters are the optimal number of clusters and 3 clusters are the next optimal number of clusters.

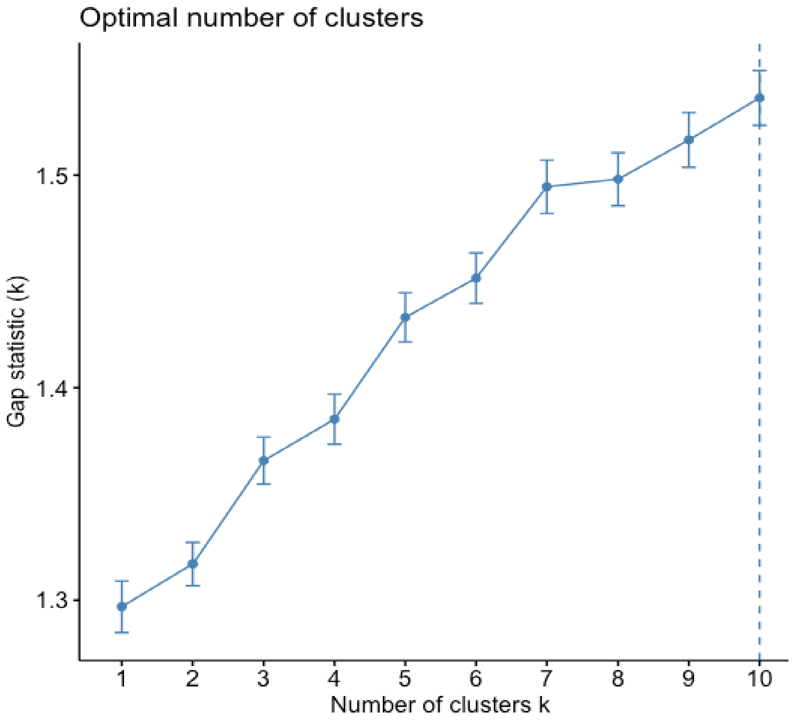

Figure 3. Gap statistic method

Also, the gap statistic value is maximized when the number of clusters is 10 when we used the gap statistic method. The gap statistic value can be increased and decreased as number of the clusters increases. With the dataset, gap statistic increased until the number of clusters is 10 and it recommends 10 is the optimal number of clusters. As a result, the optimal number of clusters is 10.

Based on the cluster plot results, clusters were made with the numbers of 2, 3, and 10 using 6 variables (fresh, milk, grocery, frozen, detergents with paper, and delicatessen) to find well-divided results between clusters.

Then, it can be found well-divided clusters result with 2 clusters using milk and fresh variables. In the cluster plot below, observations in both clusters are rarely overlapped and separated well. Also, most observations in the same cluster are gathered closely. However, some observations like data point 87 and 182 in cluster 2 are too far away from the main observations.

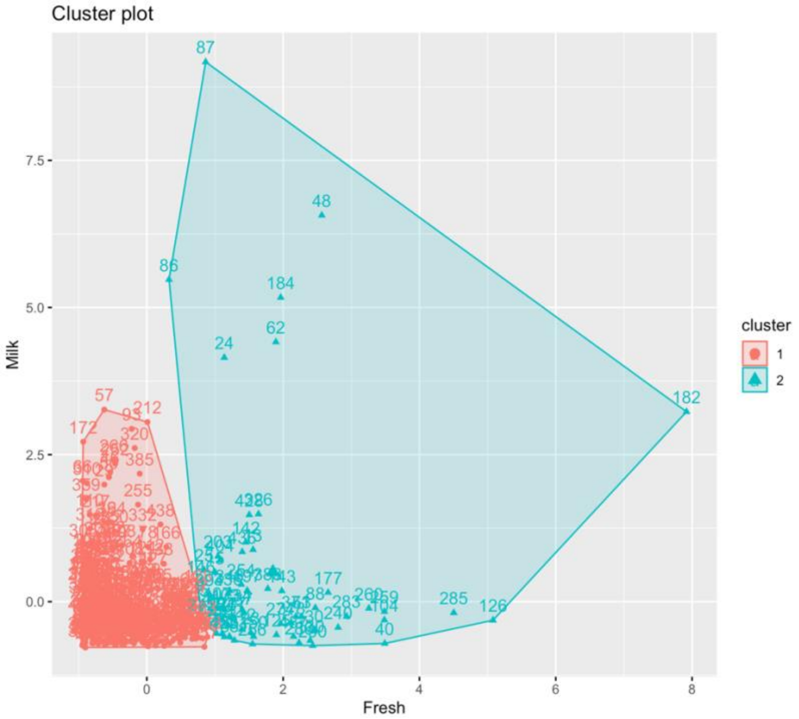

Figure 4. Cluster plot with 2 clusters

Also, clusters using fresh and grocery variables are relatively divided well than other cluster results and observations in the same cluster are closely well gathered. However, some observations in clusters 1 and 3 are overlapped each other.

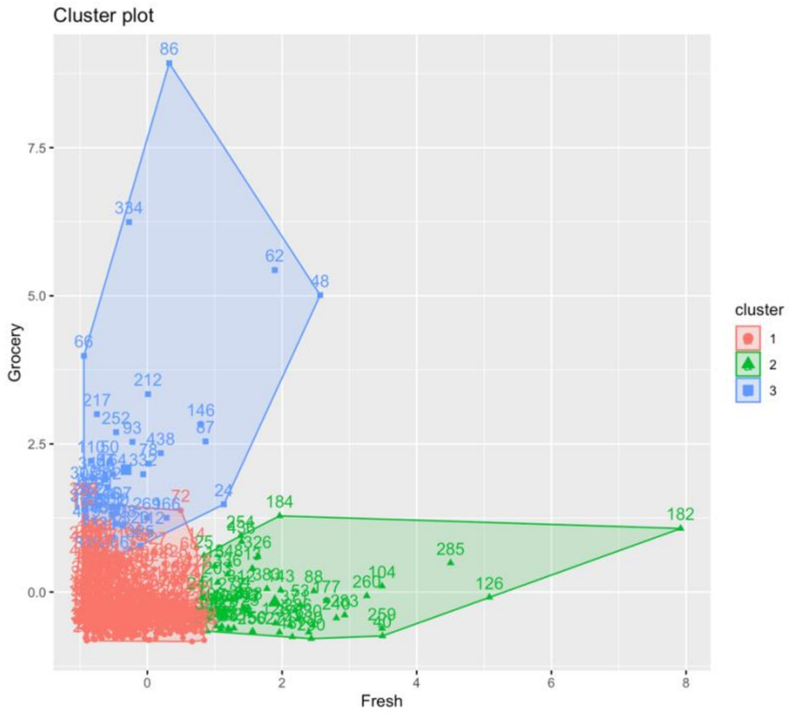

Figure 5. Cluster plot with 3 clusters

Based on the graph results, 2 clusters using fresh and milk variables and 3 clusters using fresh and grocery variables are considered as the informative results. The observations in the same cluster lie within the cluster closely. Also, clusters are well separated from each other. In addition, there are a few overlaps between the clusters. However, sales of other items could not distinguish clusters clearly. Observations in different clusters were overlapped with each other. The unclear distinguishment between different clusters cannot be used to understand the characteristics of each cluster. Other cluster results with other variables can be seen in Appendix.

Discussion

With the dataset, we could find informative cluster result using the silhouette method and elbow method. The small number of overlapping observations between clusters means that it is the easiest and clearest way to understand the characteristics of the customers about the 6 sales items.

K-means clustering method use only numerical data. It does not use categorical variables. So, I excluded the region and channel variables. However, these can influence clustering results. For example, customers in different regions can have different characteristics because each region can have different accessibility to each food product. In short, when categorical data is included, it can make different clustering result. Also, one of the disadvantages of K-means clustering is that it is difficult to determine the cluster's number. So, we had to use various methods to find the optimal number of clustering. Also, we had to compare the result with the various number of clustering.

For the research, we used two dimensions graph to cluster customers. So, we could cluster customers using two variables. However, when we use higher dimensions graphs, clustering results can be different.

Conclusion

Customer segmentation is to distinguish customers through their characteristics and needs. Through customer segmentation, the difference between clusters can be found. Through this information, companies can maximize their profits by establishing different marketing strategies for each cluster. In the research, customer segmentation was conducted through six sales items. Among these sales items, when milk and fresh variables were applied, comparatively clearer segmentation result with few overlapping observations could be found. Also, and when fresh and grocery variables were applied, we could also find informative cluster result. With these results, a company can make a marketing strategy for each cluster.

References

K- means clustering algorithm: HOW IT WORKS: Analysis & implementation. EDUCBA. (2021, October 7). Retrieved November 16, 2021, from https://www.educba.com/k-means-clustering-algorithm/.

K-mean: K means clustering: Methods to find the best value of k. Analytics Vidhya. (2021, August 27). Retrieved November 16, 2021, from https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/.

UCI Machine Learning Repository: Wholesale Customers Data Set. (n.d.). Retrieved November 17, 2021, from https://archive.ics.uci.edu/ml/datasets/Wholesale+customers.

University of Cincinnati. (n.d.). *K-means cluster analysis*. K-means Cluster Analysis · UC Business Analytics R Programming Guide. Retrieved November 16, 2021, from https://uc-r.github.io/kmeans_clustering#elbow.