Basta, Michael

Kent State University

K-means Using R

Customer Segmentation

Introduction:

Understanding customers has been the main focus of many organizations. they realized that to better serve customers you have to understand not just what they want to buy but you have to understand their backgrounds and be able to suggest things that they might be interested in and also tweak your production and marketing plans to better serve them. In order to understand customers, you have grouped them based on different parameters like demographics, age group, salary range, etc., and then target them with a suitable marketing strategy that fits this group. One of the ways that companies use to group customers is called the customer segmentation strategy. Customer segmentation is an effective tool for businesses to closely align their strategy and tactics with, and better target, their current and future customers. Every customer is different, and every customer journey is different, so a single approach often isn’t going to work for all.

Customer segmentation is the process by which you divide your customers up based on common characteristics – such as demographics or behaviors, so your marketing team or sales team can reach out to those customers more effectively.

These customer segmentation groups can also be used to begin discussions of building a marketing persona or product user persona. This is because effective customer segmentation analysis is typically used to inform a brand’s messaging and positioning and helps organizations know what new products or services they might want to invest in and uncovers ways to improve how a business sell. Because of this, marketing personas need to be closely aligned to those segments in order to be effective. The reason why customer segmentation is popular now is that it helps identify new products and services to create next, as well as how to market and sell existing offerings more effectively. This is because you can develop a better understanding of your customer segmentation and customer segmentation analysis are critical for businesses.

Objective:

The data used in this project is downloaded from Kaggle. The content of the problem is that you own a supermarket mall and through membership cards, you have some basic data about your customers like Customer Id, age, gender, annual income, and spending score. A spending score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data. The problem statement is that you own the mall and you want to understand the customers like who can be the target market so that the sensor can be given to the marketing team and plan the strategy accordingly. I think this problem is a very important one that can be considered a real-world example because it is something that every mall owner faces not just when they try to market to their mall but also when they try to pick what brands to include in their mall to attain the highest profit possible. Also, the data available is very reasonable for real-world examples such as mall membership where you don’t have much info about customers but you still have some info to make customer segmentation and market to your mall accordingly. The main objective here will be to segment customers based on the data into several segments and will give some insights on the segments for the marketing team to be able to target each segment according to the analysis.

Approach:

The approach that will be used here to accomplish the objective of segmenting customers into meaningful segments is the K-means clustering algorithm. The k-means algorithm is a partitional clustering algorithm to group n objects based on attributes into k partitions, where k < n. Each cluster is represented by a cluster centroid. The grouping is done by minimizing the sum of distances between data and the corresponding cluster centroid. Euclidean distance is commonly used with the K-means algorithm. The way the algorithms work is that it starts with a group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative calculations to optimize the positions of the centroids.

Analysis:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedAs specified earlier the data consists of 5 columns ( Customer Id, Gender, Age, Annual Income, and Spending Score). As seen from the histograms of age, the highest number of customers are between the age of 30 and 50, and that gives an initial insight into the customers that most of the people that come to the mall are in the adult segment. Also, there are more people from younger ages under 30 who come to the mall than the older group over 50. Regarding income, the histogram shows that the majority of customers are in the range between 40k and 80k yearly. This information is useful even before segmenting the customer into groups because it helps the mall know the customers and offers the right products for the groups that come to the mall regularly. Not only the annual income is important in the analysis, but also the spending score.

Chart, histogram

Description automatically generatedThere might be high-paid customers in the mall but if they are low spenders, the mall won’t make the profit needed although they have enough money to spend, so the spending here means the willingness to buy the products from the mall. The spending score is a score from 1 to 100 made to express as a percentage of how much the customer could be considered a spender. It can be seen from the histogram of spending, that the highest range of spending is between 40 and 80.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedFurthermore, from the line plots between the numerical variables, it looks like there’s no correlation between Income and spending or age and income. The only slight correlation that shows on the graph is a negative correlation between age and spending. This shows the older a customer gets the less they spend on purchasing from a mall.

**K-means**

Chart, line chart

Description automatically generatedSince there was no direct correlation between the variables that will be used in the analysis, then Euclidean distance should be suitable in this case. Before applying the k-means algorithm on R, data has to be scaled since it is scale-dependent. Now after scaling the data, we have to determine the number of clusters that will be used in the algorithm. Some methods can be used in R to specify the appropriate number of clusters; one of these methods is “fviz\_nvclust()”. The result that we get back is shown in the graph. The chart shows that the elbow point 4 provides the best value of k. While WSS will continue to drop for larger values of k, we have to make the tradeoff between overfitting, i.e., a model fitting both noise and signal, to a model having bias. Here, the elbow point provides that compromise where WSS, while still decreasing beyond k = 4, decreases at a much smaller rate. In other words, k = 4 provides the best value between bias and overfitting.

**Results:**

Text

Description automatically generatedWe use the k-means package with k = 4 and 25 restarts. The solution indicates 4 clusters with centroid values as specified in the output. Note again that the values have been normalized.

Chart

Description automatically generated with low confidenceWe can now visualize the final cluster

It’s now clear from the data and the clusters resulted that the first cluster represents mid-range people with very high annual income but low spending scores. The second cluster represents older people with low income and relatively low spending scores. The third cluster represents young people with low-income and mid-range spending scores. The fourth cluster is relatively younger people with high incomes and very high spending scores.

Conclusion:

The objective of this analysis was to do a customer segmentation for the customers of a mall that we have the data for its member customers. The purpose of customer segmentation is to help any organization have a more in-depth understanding of their customers and better serve them according to their needs and desires. Here we analyzed 200 customers based on their age, gender, annual income, and spending score. The analysis could be used by the marketing team to put a marketing plan to target every cluster of the four clusters with a different strategy according to their parameters. Finally, to summarize the results of the analysis, four clusters resulted from the k-means clustering algorithm that was applied to the data. The first cluster represents mid-range people with very high annual income but low spending scores. The second cluster represents older people with low income and relatively low spending scores. The third cluster represents young people with low-income and mid-range spending scores. The fourth cluster is relatively younger people with high incomes and very high spending scores.