**Customer Segmentation Analysis: Clustering Customers for Targeted Marketing**

Dustin Casey

Department of Data Science, Bellevue University

DSC 680: Applied Data Science

Professor Catie Williams

September 18th, 2022

**Abstract**

Using customer data to identify and group together customers with similar characteristics. By doing so, we can make better decisions on targeted marketing and increase sales. To do this, we will explore using K-Means (centroid-based) Clustering to identify and classify our different customer groups.

**Business Problem & Background/History**

As the owner of a retail store, I have customer data, but I don’t know why some customers purchase items and others do not. To determine which customers should be targeted with marketing efforts, I want to be able to identify and group customers with similar characteristics to determine which customer group is the most likely to purchase.

We record data and assign a Spending Score to every customer. We want to cluster our customers together to target those that have the highest spending score and try to identify those with a low score to know what characteristics drive those customers. This practice is known as Customer Segmentation or Market Basket Analysis. By identifying customers with similar characteristics, you make an assumption that they make similar buying behavior as the rest of the group. If this is true, you can target that group with more specific advertising, marketing, coupons, deals, etc. to drive additional purchases. As said by Sammi Wong on her blog post entitled “Importance of Customer Segmentation”:

“Segmentation helps businesses customize their marketing strategies to provide shoppers with the brand experiences they’re looking for. This can range from sending targeted email campaigns to designing in-store displays to serving personalized digital ads. As a result, customers feel more connected and engaged. Personalization can lead to customer loyalty and brand advocacy, especially when companies nurture the business-consumer relationship in the right ways.” (Wong, 2022).

Making the customer feel special and getting that personalized touch could be the difference between making a sale or not. As a store owner, I want to utilize all of the data I have available to me to modernize and innovate my segmentation practices to drive my business to more success.

**The Data**

For our analysis, we have a few different metrics recorded for each customer including their Gender, Age, and Annual Income. We also assign them a Customer ID and a Spending Score on a scale of 1 to 100 (higher scores are better). Based on these metrics, we will find which customers we would like to have return for business in the future. The dataset is the metrics of 200 customers from the store. Age can range from 18 to 70, with a mean of 38.85. Annual Income ranges from $15K to $137K, with a mean of $60.56K. Spending Score ranges from 1 to 99, with an average of 50.2. For Gender, the split is 112 women to 88 men.

**Methods/Analysis**

The first step in our analysis was pulling the data into R and doing a bit of data clean-up such as dropping the Customer ID for clustering training. I also set the Gender field to be 1 for Male and 2 for Female so it was easier to work with when putting it through the machine learning algorithm.

I started out by doing some general Exploratory Data Analysis (EDA) in the form of histogram plots to see both, what the data looks like, and how variable it is in terms of outliers. When analyzing the results for the Age metric, the largest bucket of our customer base fall between 30 and 35 years old, with the overall the customer base heavily favored to be under 50 years old. In looking at Annual Income, the largest bucket here shows customers making between $70K and $80K per year, with a small amount who make more than $80K per year.

Chart, histogram

Description automatically generated

*Figure 1.1*

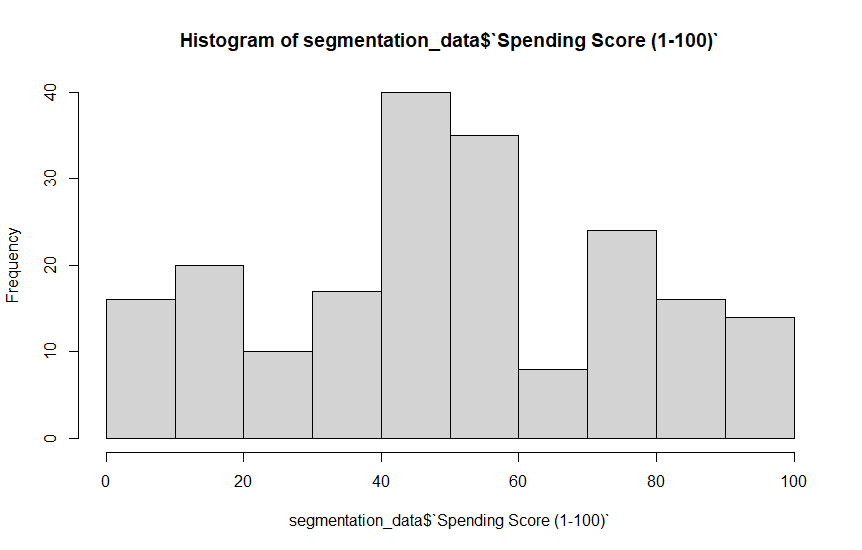
*Histogram of Age from the Segmentation Dataset*

Chart, histogram

Description automatically generated

*Figure 1.2*

*Histogram of Annual Income from the Segmentation Dataset*



*Figure 1.3*

*Histogram of Annual Income from the Segmentation Dataset*

Next, I started running K-Means clustering on different datasets. First, I wanted to look at clustering by Age. I took my dataset and only kept Age and Spending Score columns to train my model. Next, I needed to see what the ideal number of clusters would be. I determined using the elbow method that 4 clusters would be right for Age.

Chart, histogram

Description automatically generated

*Figure 2.1*

*“Within groups sum of squares” Graph Showing Ideal Number of Clusters for Age*

Next, I ran the data through the kmeans() function to get my clusters. I then graphed the clusters and identified target clusters. For Age, there was one cluster that had a very high average spending score (Group 4), two clusters that were okay (Groups 1 & 2), and one that was very low that we may want to avoid for marketing efforts (Group 3).

Chart, scatter chart

Description automatically generated

*Figure 2.2*

*Cluster Plot of Age*

I repeated the same process but instead of looking at Age, I now replaced it with the Annual Income data. The “Within groups sum of squares” graph showed that 5 clusters would be ideal for Annual Income.

Chart, line chart, histogram

Description automatically generated

*Figure 2.3*

*“Within groups sum of squares” Graph Showing Ideal Number of Clusters for Annual Income*

After reviewing the results of the kmeans() function, I determined that there were one cluster that had very high average spending scores (Group 1 & 2), one cluster that was okay (Group 4), and two that were very low that we may want to avoid for marketing efforts (Groups 3 & 5).

Chart, scatter chart

Description automatically generated

*Figure 2.2*

*Cluster Plot of Annual Income*

**Conclusion**

Using K-Means Clustering, I was able to identify customers who have similar characteristics and their associated Spending Scores to better target our marketing efforts. Age and Annual Income were good indicators of customers who were similar and should have similar purchasing behaviors. By using a Customer Segmentation Analysis, I will be able to be more efficient in marketing effort whilst also saving money by not pursuing customer groups that yield low Spending Scores. It’s a win-win that will only yield more accurate results as we get more customers/data added to the existing dataset and we do a look-back on how effective the new marketing efforts are in our target groups.

**Assumptions/Limitations/Challenges**

The biggest assumption made in this analysis is that Age and Annual Income are good indicators of repeat purchase/Spending Score. Also, because Spending Score is a generated value via the business, it could be that their values might not fully gauge how likely a customer is to return to purchase again. The last assumption is that customers with similar characteristics are likely to have similar buying behavior. Because of these assumptions, a follow-up/look-back study should be done to see how effective the new marketing strategy is.

A lot of the assumptions stemmed from the limitations and challenges faced with this analysis. Due to the limited amount of customer data, we had to assume that Age and Income were the driving metrics behind Spending Score. Sammi Wong cites data quality and management as one of the biggest challenges of Customer Segmentation, saying,

“Data collection processes for customer segmentation are typically complex and prone to inaccuracy, inconsistency, and incompleteness. Brands need to fully commit to the proper organization and maintenance of their data so it’s useful and meaningful. They should also focus on capturing zero-party and first-party data, rather than second-party and third-party data, as the latter two may vary in quality, relevance, and trustworthiness. Businesses should also ensure that their customer segments are clearly defined, easily understood, and simple to incorporate into their current and future strategies. They need to invest in data management and analytics tools that integrate with their existing systems and keep their information clean and standardized.” (Wong, 2022).

If possible, it would be good to get additional customer data items to analyze. Another limitation was the size of the dataset. With it only being 200 records, it’s hard to say that the dataset is fully representative of the customers and future customers. As time goes on, this should hopefully sort itself out with further records being added.

**Future Uses/Recommendations/Implementation Plan**

For how this should be implemented, I would start by sending targeted marketing to your top clusters for both Age, Income, and some combination of both. Once sent, I would do an analysis on how effective that was in driving more business to the store. I also think that the Cluster/Customer Segmentation Analysis should be done annually to see what changes happened, to work with a larger dataset, and to represent our current book of business. I would also suggest recording more customer data to determine if there are other factors at play that are affecting buying behavior/Spending Score.

**Ethical Assessment**

For this analysis, the biggest ethical issues facing the project is customer privacy. In the current analysis, we are asking for some pretty sensitive data, or, at the very least, taboo topics from the customers. Many people don’t want to share this information and will not likely want the store to have even more data on them if we were to try to collect additional metrics. But, in today’s world, customers’ data is out there in so many ways/places that it wouldn’t be out of the norm to be collecting this information and using it to make better decisions for the business.

**References**

Choudhary, V. (2018). Mall Customer Segmentation Data. *Kaggle*. <https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>

Wong, S. (2022). Importance of Customer Segmentation. *3TierLogic*. <https://www.3tl.com/blog/importance-of-customer-segmentation#:~:text=Segmentation%20helps%20businesses%20customize%20their,feel%20more%20connected%20and%20engaged>

**Appendix**

Dataset Codebook and Field Descriptions:

Graphical user interface, text, application, email

Description automatically generated