Customer Segmentation with Python

Maddie Bauer

Bellevue University

DSC 680: Applied Data Science

Professor Catherine Williams

July 2, 2021

Introduction & Problem Statement

Retail businesses are often faced with the unpredictability of their customer base and need to learn as much as possible about their customers in order to determine the best marketing and promotional strategies possible. Specifically, they need to determine what strategies will work for different types of customers as no two customers are completely identical. The purpose of this report is to create customer clusters that will allow retail companies to target different user bases where similarities among customers are determined. This is known as customer segmentation. A more formal definition of customer segmentation is as follows:

> "Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits" (Dataflair.training, n.d.).

Some advantages of customer segmentation include being able to determine appropriate product pricing, customize marketing campaigns, choose specific product features for deployment, prioritize new product development efforts, and design an optimal distribution strategy (Sagar, n.d.). Also, customer segmentation can help a company set more specific and measurable goals as well as help retain customers in the long run (Lintern, 2013).

This research paper seeks to understand how factors such as Age, Gender, Spending Score and Annual Salary may or may not help determine different customer segments for marketing and promotional strategy gains within the retail industry.

Data

The data set being used for this research paper was retrieved from Kaggle.com. It is a relatively small data set consisting of only 200 customers. It was created for a Kaggle competition several years ago to help learn about customer segmentation. The idea is that basic data was collected for mall shoppers through the use of their mall membership cards. The goal of this research is to analyze the following variables - Gender, Age, Annual Income and Spending Score - to help determine appropriate marketing strategies for the mall moving forward. Below is a little bit of information for each variable within the data set.

Gender – Only male or female are possible values in this dataset

Age – Customer's age (numeric value)

Annual Income – Numeric value in thousands (ex. 40 = $40,000)

Spending Score – Range of 0-100 that is assigned by the mall based on the customer's behavior and previous spending nature

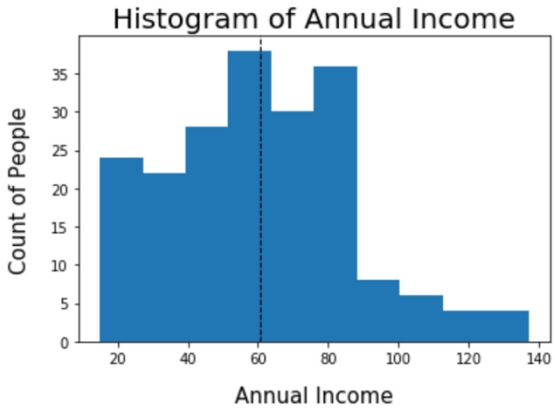
If conducting this research again from scratch, I would make sure to include more options for Gender to be more inclusive. I would also add additional variables such as time and/or date of purchase as well as the type of store (clothing, jewelry, shoes, etc.) where each purchase was conducted since Spending Score is a very broad term.

Methods

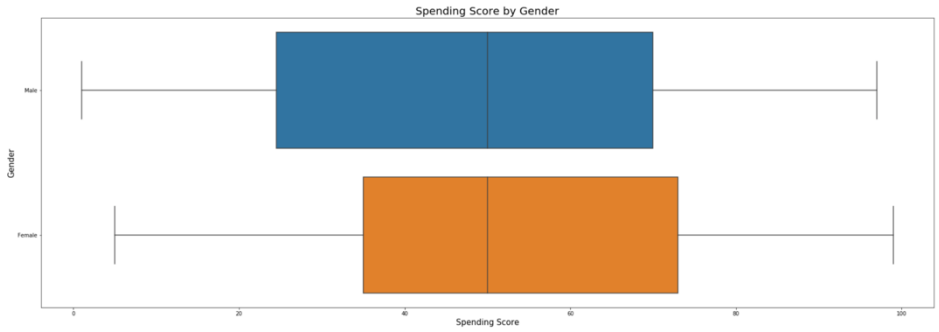
The overall plan with this data set is to explore relationships among the variables and analyze the customer base provided. I plan to start out with a simple data exploration in which I will view the summary statistics, distribution and spread of each variable. I will also compare different pairs of variables to see if there are any trends or patterns to be aware of. Next, I plan to prepare the data and make sure it is in usable format for the remaining portion of the project. Once the initial data exploration and data preparation is complete, I will then thoroughly research and implement K-Means clustering. The final step will be to summarize and interpret the results for each customer segment found.

Results

Beginning with the exploratory data analysis, I determined that none of the variables had any missing values or outliers that needed to be dealt with. This allowed me to move onto further exploration quickly. I first explored Gender and Age. There are 112 females and 88 males represented in this data set with an average age of 38 years old. Next, I explored the Annual Income. The majority of annual incomes fell at or below $80,000 with the average falling just above $60,000 per year. This is illustrated below.



I wanted to then determine what the spending score looked like in terms of spread broken down by Gender. Male shoppers have a wider spread of spending scores ranging from 1 - 100 whereas the spread for females is slightly smaller at 5 – 100.



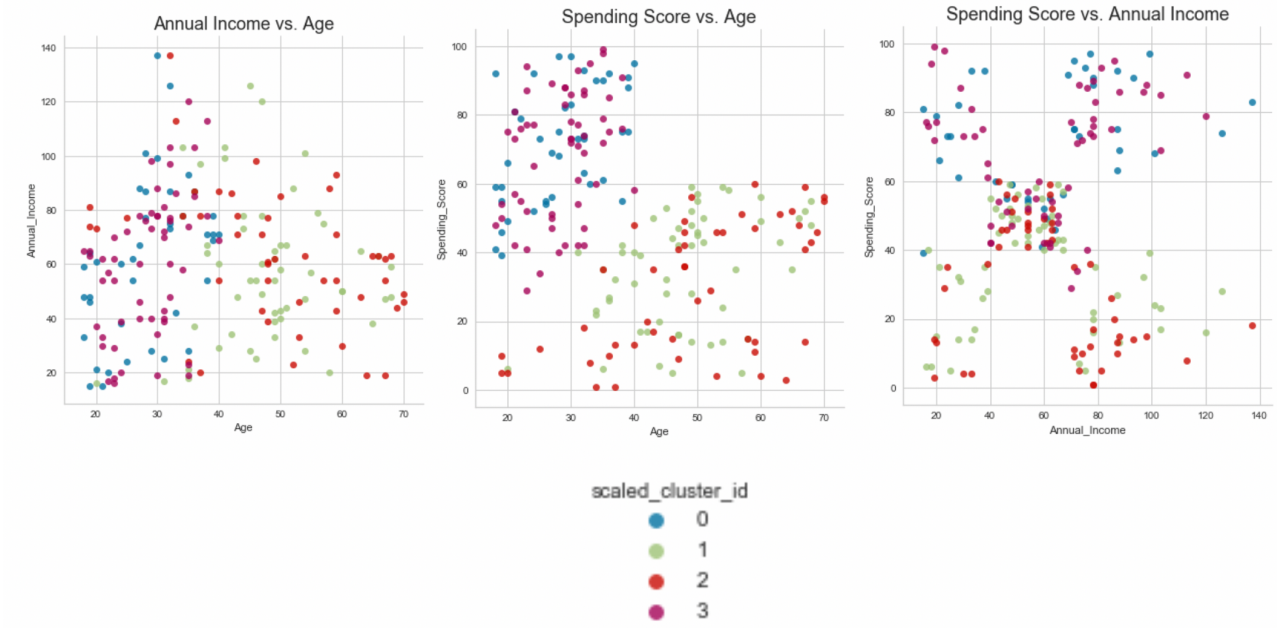
After exploring these variables, I chose to one – hot encode the categorical variable, Gender, to make it usable for my clustering approach. Females are represented as 0 and males are represented as 1. I also looked at a heatmap to determine if any of the variables are highly correlated to one another. Spending score and Age were the most prominently correlated at -0.33. The remaining pairs of variables were a mixture between positive and negative correlations.

The final step in my analysis is to actually create the different clusters of customers and determine which clusters are or are not valuable to the mall. In order to do this, I'll use K-Means Clustering. K-Means clustering is a popular unsupervised machine learning technique used to group similar data points together and to discover underlying patterns (Garbade, 2018). In order to achieve this objective, K-Means uses a fixed number of clusters, denoted by k, within a data set which can be determined graphically by using the Elbow Method (Garbade, 2018). The center of each cluster is called the centroid and can be an actual data point or an imaginary data point. Each data point is assigned to a cluster by its relative location to the nearest centroid. After

the data points are assigned to a cluster, you can start analyzing each cluster in depth to determine what patterns might evolve.

For my data set, there ended up being 4 different clusters that customers could fall into. The clusters were relatively even in size, with cluster 0 notably being the smallest. The table and scatterplots below show important information for each cluster and can help us determine some underlying patterns for each cluster group.

Cluster Number	Number in the Cluster	Age (Mean)	Annual Income (Mean)	Gender (Mean)	Spending Score (Mean)
0	40	28.25	62.000	Male	71.675
1	56	47.80	58.071	Female	34.875
2	48	49.45	62.417	Male	29.208
3	56	28.34	60.429	Female	68.179



The following insights can be concluded based on the information above:

- **Cluster 0** – Highest spending score, younger in age, varying income levels, more males

- **Cluster 1** – Moderate/Low spending score, older in age, varying income levels, more females
- **Cluster 2** – Lowest spending score, older in age, varying income levels, more males
- **Cluster 3** – High spending score, younger in age, varying income levels, more females

Gender and Annual Income do not play a key role in determining which cluster a customer is assigned to whereas Age and Spending Score are important factors in determining which cluster a customer is assigned to. This result is confirmed by the correlation heatmap discussed above as well. Clusters 0 and 3 are the most valuable for this retail mall and clusters 1 and 2 are the least valuable. This means that the mall may want to spend more time, money and energy on marketing towards younger customers as they are represented in both clusters 0 and 3. The marketing and advertisements should also be gender neutral and contain a wide range of goods of all prices. These are just the initial findings in regard to the clustering technique used. The mall could continue to research their customer's Spending Score to keep improving their marketing strategies. This cluster analysis should be conducted quarterly to determine if the behaviors of customers change drastically or if cluster centroids change as well.

Conclusion

All retail companies, large and small, can benefit by grouping their customers into smaller subgroups for the benefits of providing targeted marketing, advertisements, coupons, etc. Customer segmentation is a tool that companies can utilize to organize their audience into meaningful groups to customize the overall consumer experience. Throughout this research we were able to identify four clusters within our data set and

the key attributes of the customers within each one. We can now use these attributes and information to start customizing our marketing strategies. By tailoring a customer's experience, you make them feel welcomed, appreciated, heard, etc. which leads to brand loyalty, great consumer retention and higher profits for the company.

Acknowledgments

I would like to thank Kaggle for providing the data set used in this project. Also, I would like to thank our Professor Catherine Williams from Bellevue University as well as the students of DSC680 for their support throughout this course. Finally, I would like to express my gratitude for Medium.com, kdnuggets.com, stackoverflow.com and their countless authors for their research and posts on the subject matter. This report would not have been possible without their contributions and sharing of information online.

References

Brownlee, J. (June 30, 2020). *Why one-hot encode data in machine learning?* Retrieved June 24, 2021 from https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

Choudhary, V. (n.d.). *Mall customer segmentation data*. Retrieved June 12, 2021 from https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python#

Dabbura, I. (September 17, 2018). *K-means clustering: Algorithm, applications, evaluation, methods, and drawbacks*. Retrieved June 24, 2021 from https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

Dataflair.training. (n.d.). *Data science project – Customer segmentation using machine learning in R*. Retrieved June 23, 2021 from https://data-flair.training/blogs/r-data-science-project-customer-segmentation/

Garbade, Dr. M. (September 12, 2018). *Understanding K-means clustering in machine learning*. Retrieved June 24, 2021 from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

K-Means Clustering. (June 22, 2021). In Wikipedia. Retrieved June 24, 2021 from https://en.wikipedia.org/wiki/K-means_clustering

Lintern, M. (2013). To segment or not to segment: We weigh the pros and cons. Retrieved June 28, 2021 from https://blogs.oracle.com/marketingcloud/post/to-segment-or-not-to-segment-we-weigh-the-pros-and-cons

Sagar, A. (n.d.). *Customer segmentation using k means clustering*. Retrieved June 23, 2021 from https://www.kdnuggets.com/2019/11/customer-segmentation-using-k-means-clustering.html

Questions (Answers Provided in Presentation):

1. Based on the clusters found, what are the next steps for targeting each group?

2. Will a customer shift to a different cluster if their spending score changes in the future?

3. Are there other clustering methods that could have been used? How do they compare with K-Means?

4. If conducting this on a different data set, would you expect there to be a spending score variable?

5. What other variables could companies use to determine their clusters?

6. How often should a company run this clustering approach?

7. Would it hurt to use more or less than 4 clusters?

8. Shouldn't Gender have other options available to include everyone?

9. What are some examples of spending behaviors for someone with a high spending score? Low spending score?

10. What are some reasons that a person's annual income isn't a deciding factor in their cluster? Wouldn't annual income and spending score be correlated?