K-Means Clustering for Customer Segmentation in Retail Environment

Kaylynn Mosier

20 December 2024

Background

In the modern digital age, consumers are bombarded with advertising messages at all times of day. For a retailer to stand out, they must be sure to target the right advertisement to the right consumer, at the right time. This task is not possible without a thorough understanding of a customer's needs and habits. Customer segmentation is a tried-and-true method of understanding the consumer. In this practice, consumer data is used to classify consumers into different segments that are similar to each other. Depending on the data available, customers may be grouped on metrics such as age, race, family size, spending habits, frequent store locations, or any other number of metrics. Leveraging this data allows retailers to better understand their customers and ultimately capitalize on marketing spending to increase profits. This research will utilize publicly available consumer demographic data from an anonymous retailer in a customer segmentation practice with the goal of forming a better understanding of the retailer's consumer base.

Data Explanation & Preparation

In this study, I utilized 'The Complete Journey' from dunnhumby. This is a robust dataset that contains customer demographic information and purchase history of over 2,500 households in a retailer over 2 years. The data has been anonymized and any identifying information about the customers and retailers has been removed. This is a large dataset containing a broad range of information. For the purposes of this research, I focused on the demographic information. This contained a household key for each household and 7 different demographic metrics. Although this dataset has been thoroughly anonymized, metrics have been separated to retain ordinality so conclusions can still be drawn about consumers. This data was constructed for training purposes of advanced researchers, so there are no missing values. Data preparation was minimal, but OrdinalEncoder was utilized to encode all categorical variables into numerical values that retained ordinality. An examination of the data revealed most households were part of age group 4, most were homeowners, and the number of children in each household was unknown for the majority of respondents.

Methods & Analysis

Customer segmentation is, at its base, a classification problem. For this problem, I chose to employ K means clustering. This is an unsupervised method that can group data points based on similar attributes. The result of this method will be labels, or clusters, for each household. The households within each cluster will be more similar to each other than households in other clusters. Knowing which cluster each household belongs to will allow more appropriate marketing to reach each household, which should result in higher spending from the household and ultimately higher profit for the retailer.

The first step in implementing a K means clustering algorithm is applying a StandardScaler to the data. This is used to scale all values between 0 and 1 to ensure no metrics carries more weight than any other. Next, I needed to determine the correct number of clusters that data should be sorted into. This is usually done with the ‘elbow’ method in which the within cluster sum of squares is plotted against the number of clusters. The area of the plot where the within cluster sum-of-squares is reduced minimally by increasing the number of clusters is the ideal number of clusters for the dataset. The below plot is the elbow plot for this dataset. At initial glance, 4 clusters appeared to be the correct number of clusters. However, a major assumption of the K means algorithm is equal distribution of points in each cluster. Running the algorithm with 4 clusters leads to large disparity in cluster size. When the algorithm was run with 5 clusters, this disparity was minimized.

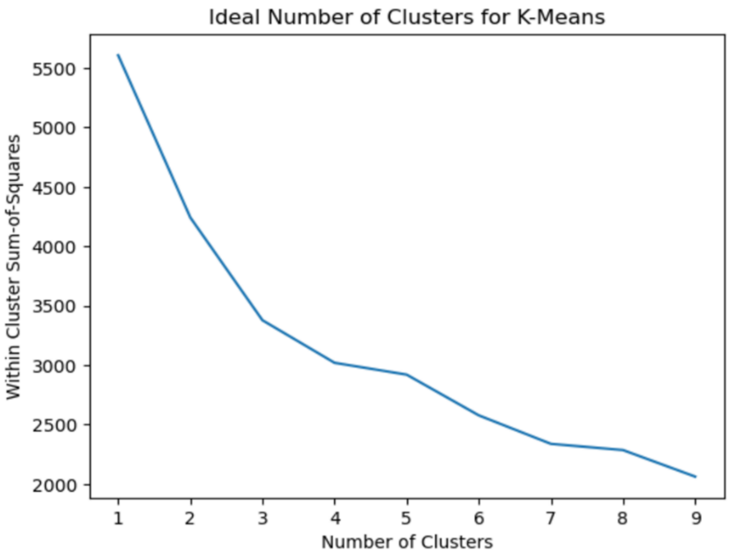

Figure 1: Elbow plot of within cluster sum of squares against the number of clusters to determine optimal number of clusters.

Once the ideal number of clusters was determined, and StandardScaler was applied, the K means algorithm was applied to the data to fit each household into a cluster of other similar households. *Figure 2* below shows the difference in spread of classification 1 values across cluster 1 and 2. Cluster 1 has more households in higher age groups than cluster 2; cluster 1 contains older consumers than cluster 2. *Figure 3* shows the difference in spread of housing status across cluster 1 and cluster two; cluster 1 is composed mostly

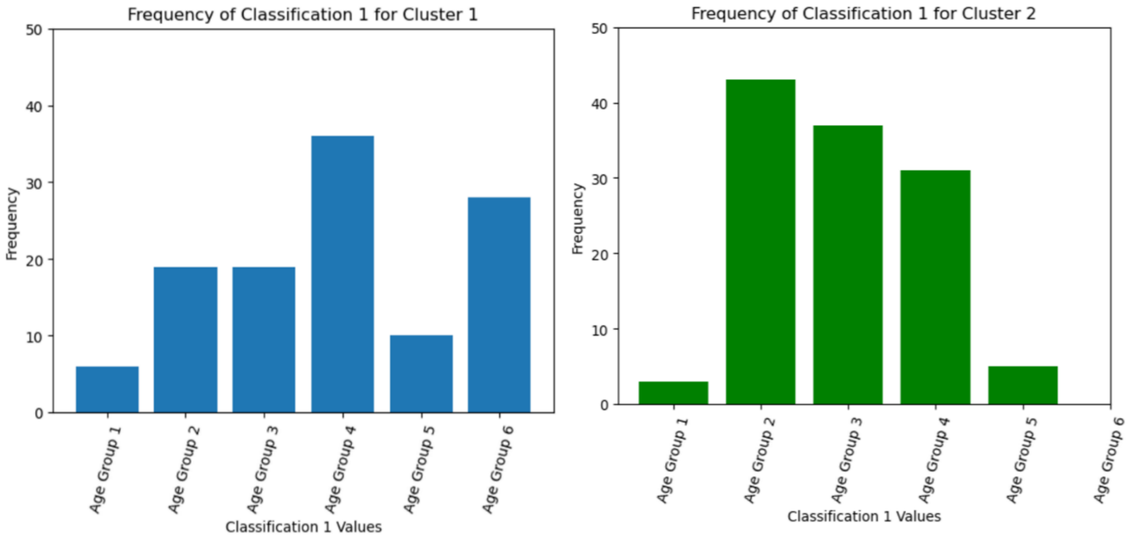

Figure 2

of homeowners while cluster 2 is composed of mostly renters.

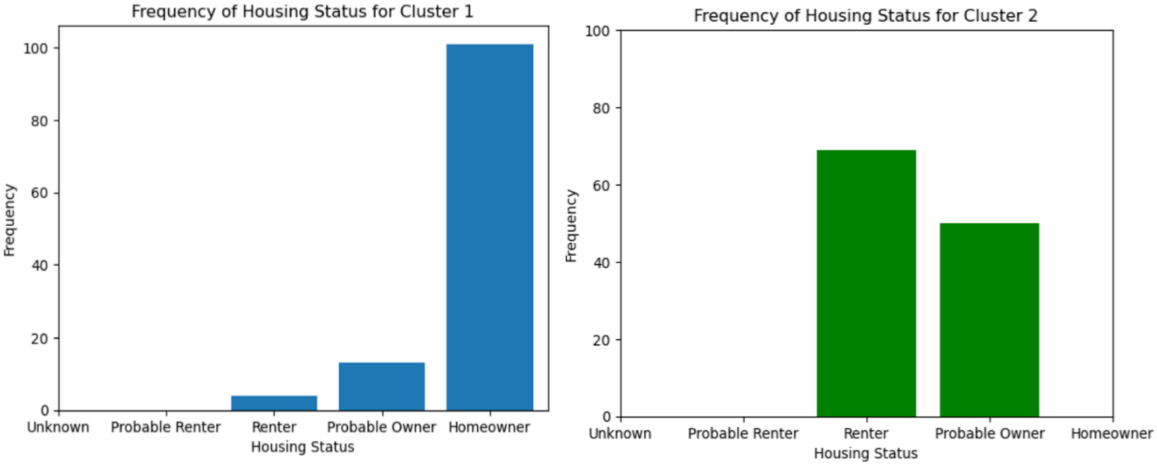

Figure 3

Conclusion

This dataset is best separated into 5 segments. A brief dive into the segments reveals stark differences between cluster 1 and cluster 2, as discussed above. Further analysis will likely reveal other differences. A true evaluation of each cluster should be conducted before any decisions are made.

Assumptions

The K means algorithm assumes clusters form spherical shapes due to each datapoint having similar variance from each cluster's centroid. The algorithm also assumes each cluster is roughly the same size. Observing the elbow plot alone, led me to believe 4 clusters was appropriate for this model. However, adjusting the number of clusters to 5 allowed for a more equal distribution of households through clusters so 5 clusters was more ideal.

Cluster Number	Observations
1	118
2	119
3	182
4	231
5	151

Limitations & Challenges

This research has not been without limitations. The dataset chosen, although expansive, has been fully anonymized. Although clusters can be visualized and evaluated, only general conclusions can be drawn without the true metrics for each demographic. A better exercise would be to have the true values for each metric. However, this is still a useful exercise because information can still be gleamed with the generalizations made.

A challenge with any unsupervised task is understanding the accuracy of the predictions made. True accuracy cannot be calculated because in an unsupervised model, there are no historical labels and therefore nothing to base accuracy on. Clustering tasks specifically can be difficult because the ideal number of clusters is up for interpretation and different scientists might have different viewpoints. The best approach in this situation is to seek the opinion of multiple experts and use different methods such as the elbow method and silhouette score to determine the ideal number of clusters.

Recommendations, Future Uses, & Implementation

This method can be used in any number of customer segmentation tasks. Due to the nature of the data, this research focused on demographic segmentation, however it could also be applied to geographical, behavioral and psychological segmentation tasks. The most ambitious retailer would combine all segmentation practices to have a more holistic view of their consumers. This would not be a task to undertake lightly as it would require large amounts of data collection, cleaning, analysis and modeling. However, payoffs in market share, profit, and customer satisfaction would be well worth the investment. Once clusters are obtained, similarities between groups can be observed. For example, one group might be composed of families without children, another has young children, and a third has adult-age children. Each of these families would benefit from different products. Correctly marketing to these families will result in higher sales and therefore higher profits for the retailer. A more robust understanding of customers will be obtained through recursive segmentation practices, each improving the retailers chances of a returned profit.

Ethical Considerations

Any time customer data is gathered, it must be handled very carefully. The risk of data breaches is at an all-time high, and companies must ensure data is properly protected. Additionally, customer segmentation tasks aim to group customers based on similarities. This can be a slippery slope, and if not done properly can result in highly prejudicial practices. Data scientists tasked with model building must ensure that these prejudices are not allowed to influence decisions by ensuring data quality and eliminating any problems that arise.

10 Questions from Audience

1. Why does this matter to our business?
    - Customer segmentation allows us to better understand our customers, which allows us to produce more of what they want and in turn create more profit for our company.
2. How credible is this data?
    - This data is from a well-known and reputable consumer insights company. Although the validity of the values cannot be vetted due to anonymization, it doesn’t need to be for this exercise. This is meant to be an example of the power of k means clustering in segmentation, not an actual segmentation practice on our consumer data.
3. How is this data still useful without actual values; does anonymization remove the impact of this evaluation?
    - The impact of the evaluation can still be seen by comparing the differences between each cluster. Even though no actual values are given, we can still gain understanding about the consumer.
4. How can this practice be applied to other situations?
    - This can, and should, readily be applied to our business if the correct data is obtained.
5. How can you ensure clusters are separated correctly?

- This is done by visualizing the clusters and ensuring good separation. In this instance, this is difficult to do because there is not a large enough spread of data due to the categorical nature of the data. In practice, this would be easily achievable with a scatter plot with one demographic on the x axis and another on the y axis. Clusters would then be color coded and separation between them would be easily visible.

6. What are the areas for improvement in this evaluation?
    - A big limitation here is the categorical nature of the data. Removing anonymization would allow for more real world conclusions to be drawn.
7. Can we apply this evaluation to customers that we don't have as much information on?
    - We can, but we must be careful in this approach. Applying a clustering model to incomplete data can lead to false assumptions and we may draw conclusions that are inaccurate. Ideally, we need full demographic information on any consumer we want to include in segmentation.
8. Are there other ways to view the differences between the clusters?
    - Yes, as exhibited in my presentation, there are. I chose to visualize differences between clusters using a histogram rather than the standard scatter plot.
9. What would it take to employ this model, or one like this, more often in our business?
    - We would need customer demographic data on a large number of our consumers, a trained data scientist, and ideally a contact that understands the marketing needs of the company. This would allow end-to-end communication of the modeling process and how it applies to our marketing practices.
10. Are there any other models that could do customer segmentation?

- In theory, yes. There are numerous other clustering models that could be a good fit, and I would even suggest trying them out before a final decision is made on the business approach.

References

Dunhumby (2024). *The Complete Journey*. https://www.dunnhumby.com/source-files/

Kumar, Dhiraj (2023, 19 December). *Implementing Customer Segmentation Using Machine Learning [Beginners Guide]*. Neptune.AI. Implementing Customer Segmentation Using Machine Learning [Beginners Guide]