KMeans Clustering for Customer Segmentation Analysis

Ariana Arellano

Rogelio (Roy) Avalos

Norberto Limon

Issac Chen

Cal Poly Pomona - MSBA

GBA 6430

Dr. Koohikamali

KMeans Clustering for Customer Segmentation Analysis

Authors: Ariana Arellano, Rogelio (Roy) Avalos, Norberto Limon, Issac Chen

Introduction

Over the past two decades online retail has taken the world by storm and continues to become an increasingly prominent force in the retail space. Although raw sales on average remain favorable for traditional businesses according to US census data, there is an increasing need to augment brick and mortar retail with a digital presence, loyalty incentives or even adopt more digital first business models for old, new and emerging use cases. (U.S. Census Bureau, 2023) According to the 2018 Personalization Pulse Check from Accenture Interactive, 91 percent of consumers are more likely to shop with brands that recognize, remember, and provide them with relevant offers and recommendations. (Accenture, 2018)

Research Question

How can customer segmentation based on purchase history using K-Means Clustering enhance personalized marketing strategies and optimize customer retention?

Our objective is to develop a classification model for marketing segmentation through the use of K-Means cluster analysis on a sample dataset from an anonymous Online Retailer.

Gap / Problem

A gap in the current market for this topic is the lack of segmenting based on normalized purchase history and consumer behavior. Our focus is to find the methods for K-means to successfully segment and cluster customer groups to produce the highest expected revenue.

Background Research

Literature Review 1 - Our first paper, titled "Cluster Analysis in Marketing Research: Review and Suggestions for Application" by Girish Punj and David W. Stewarte. In this work the authors recommend a two-stage cluster analysis methodology, consisting of a preliminary identification of clusters via Ward's minimum variance method or simple average linkage, followed by the second phase consisting of cluster refinement by an iterative partitioning procedure. (Punj G.,1983)

Literature Review 2 - The authors look at cluster analysis for the customer segmentation use case, recognizing it as a tool with nearly limitless potential for guiding businesses towards effective marketing and product development. This paper also presents a two-step optimization (FSGA-FCEN) akin to topic modeling for customer division based on heredity calculation (GA) and cluster gathering (CE). (Kaur et al., 2021)

Literature Review 3 - From the paper on segmenting e-commerce customers from Zengyuan Wu, Lingmin Jin et al., a K-Medoids clustering algorithm is used to optimize the initial clustering centers using distance and correlation between samples (Wu et al.). The choice of K-medoids over K-means clustering is to avoid the noise and isolated points that come as a result from using K-means, which leads to poor clustering results (Wu et al.). While K-medoids is the main cluster method the authors used, RFM was also taken into consideration to identify the company's best customers. After the RFM model selected the top percentage of customers, a CH index took place to find the optimal *k* value to perform the K-medoids algorithm to select the initial clustering center.

Data Collection and Planned Methodology

Data Processing

Transforming and structuring the data from its raw transactional format was needed to prepare the dataset for K-Means clustering and determining the optimal number of clusters using the Elbow Method. The goal is to efficiently collect, clean, select features, and scale the data to minimize the effect of outliers in our data for our clustering analysis.

The data used was the Online Retail II dataset, collected from UC Irvine Machine Learning Repository. It consists of transactional records from a registered UK-based online retail company spanning from 01/12/2009 to 09/12/2011. The company specializes in selling distinctive all-occasion giftware. A significant portion of the company's customer base consists of wholesalers. The data consists of 8 Variables of Multivariate, Sequential, Time-Series, Text data types. Table 1 shows the variables Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country .

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	$2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	$3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	$2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	$3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	$3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/2010 8:26	$7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/2010 8:26	$4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	12/1/2010 8:28	$1.85	17850	United Kingdom
536366	22632	HAND WARMER RED POLKA DOT	6	12/1/2010 8:28	$1.85	17850	United Kingdom
536368	22960	JAM MAKING SET WITH JARS	6	12/1/2010 8:34	$4.25	13047	United Kingdom
536368	22913	RED COAT RACK PARIS FASHION	3	12/1/2010 8:34	$4.95	13047	United Kingdom

Table 1: Online Retail II dataset

The data set underwent a thorough data cleaning process to address issues like missing values, duplicates, and outliers. The process focused on ensuring data completeness, accuracy, and relevance to the problem at hand. Missing values were identified in the CustID variable and

removed for our analysis. Since the goal was to segment customers', this was appropriate. No duplicate values we detected.

Next, we had to aggregate the transaction by CustId to see the customers' ordering patterns. We also enhanced the data by creating seven new variables. Unique_Invoice_Orders was created to measure how many orders were placed per customer. Unique_Stockcodes was created to see how many unique stock codes per customer. Total_Stockcode_Orders was used to see how many items ordered per CustID were. The Average_Price_Per_Stockcode was used to see the average price per item ordered. Average_Item_per_order was used to determine the average items order per customers' order. Total_Amount_Ordered was used to see the total amount spent by each customer in the analysis time frame. And lastly we created Average_order_Amount to see the average amount per customer order. Table 2 shows the new aggregated data with new variables.

CustID	Unique_Invoice_Orders	Unique_Stockcodes	Total_Stockcode_Orders	Average_Price_Per_Stockcode	Average_Item_per_order	Total_Amount_Ordered	Average_order_Amount
12347	8	126	435	$ 2.59	54	$ 9,943.32	$ 1,242.91
12348	5	25	82	$ 4.53	16	$ 3,816.64	$ 763.33
12349	5	139	253	$ 8.34	51	$ 6,162.09	$ 1,232.42
12350	1	17	34	$ 3.84	34	$ 668.80	$ 668.80
12351	1	21	21	$ 2.36	21	$ 300.93	$ 300.93
12352	13	70	208	$ 21.52	16	$ 3,434.62	$ 264.20
12353	2	23	28	$ 3.12	14	$ 495.76	$ 247.88
12354	1	58	116	$ 4.50	116	$ 2,158.80	$ 2,158.80

Table 2: Aggregated Customer Data with New Variables

K-Means clustering is sensitive to outliers and relies on the distance between data points; data scaling was applied to ensure that all features have equal importance. The StandardScaler method was chosen to scale the data. Each feature will have a mean of 0 and a standard deviation of 1 to ensure that no feature dominates the clustering process due to differences in scale. Table 3 displays the customer id and the scaled features.

CustID	scaled_features
12347	{"vectorType": "dense", "length": 7, "values": [0.5015376802761374, 1.077594420565729, 0.6985943862487726, 0.37314319600704454, 1.782444582056247, 0.4906992452183899, 2.222933080536895]}
12348	{"vectorType": "dense", "length": 7, "values": [0.3134610501725859, 0.21380841677891452, 0.14082337430311226, 0.5545932207320388, 0.5570139318925772, 0.17590302979309122, 1.27497367477632]}
12349	{"vectorType": "dense", "length": 7, "values": [0.3134610501725859, 1.1887747972907647, 0.4970236740109844, 1.2233243602427029, 2.005250154813278, 0.38366442054316235, 2.7808682278082797]}
12350	{"vectorType": "dense", "length": 7, "values": [0.06269221003451718, 0.14538972340966186, 0.04694112476770408, 0.5619097539870789, 0.9469236842173812, 0.029128440706551302, 1.0556383005972105]}
12351	{"vectorType": "dense", "length": 7, "values": [0.06269221003451718, 0.1795990700942882, 0.057986095301281515, 0.34534036963789216, 1.169729256974412, 0.026212983438464368, 0.9499797661444934]}
12352	{"vectorType": "dense", "length": 7, "values": [0.8149987304487233, 0.5986635669809607, 0.31202041757356247, 2.933929835271076, 0.5013125387033195, 0.16456262400485583, 0.45874807967340503]}

Table 3: Scaled Features

K-Means and Validation

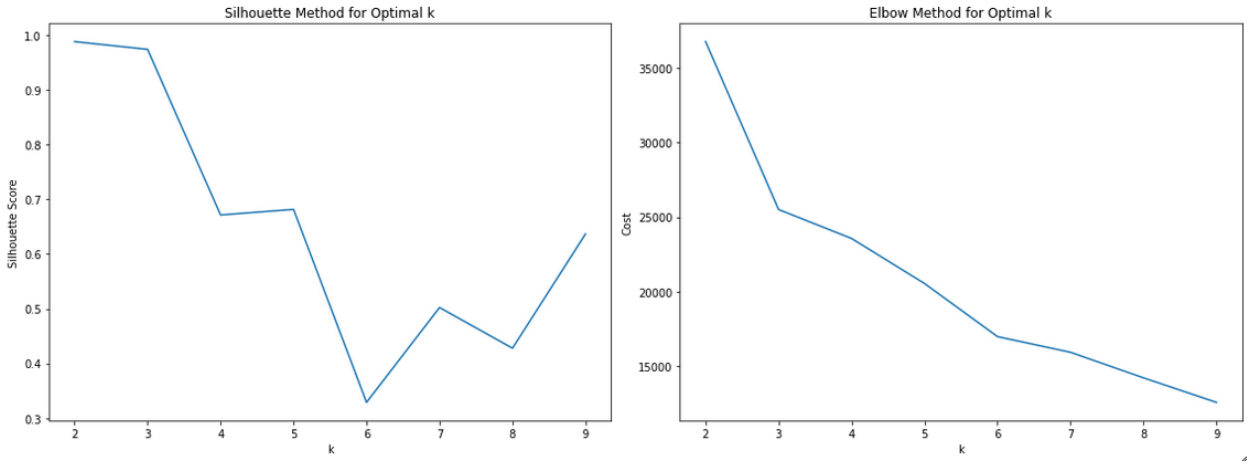
The proposed model that will address our research question is K-Means for cluster analysis. K-means is an unsupervised learning algorithm that can be applied to a wide range of datasets such as customer segmentations. Since our focus is customer segmentation, K-Means will create customer clusters to group our customers into marketing categories that best suit their history with the company.

To perform K-Means, we first must identify the optimal number of k clusters. To do this, we will use the elbow method. The elbow method is an empirical method that takes the previously mentioned scaled data and finds the optimal number of clusters for the dataset. To do this, steps include calculating the average separation between each point in a cluster and its centroid, then plot that information. Next, we choose the value of k at which the average distance abruptly decreases. The average distance decreases as the number of clusters (k) rises. In our case we can observe the plot to determine the value of k for which the distance falls off sharply and steeply in order to determine the ideal number of clusters (k). This will be the ideal location for an elbow to occur. A second method that will be used to find the ideal number of clusters (k)

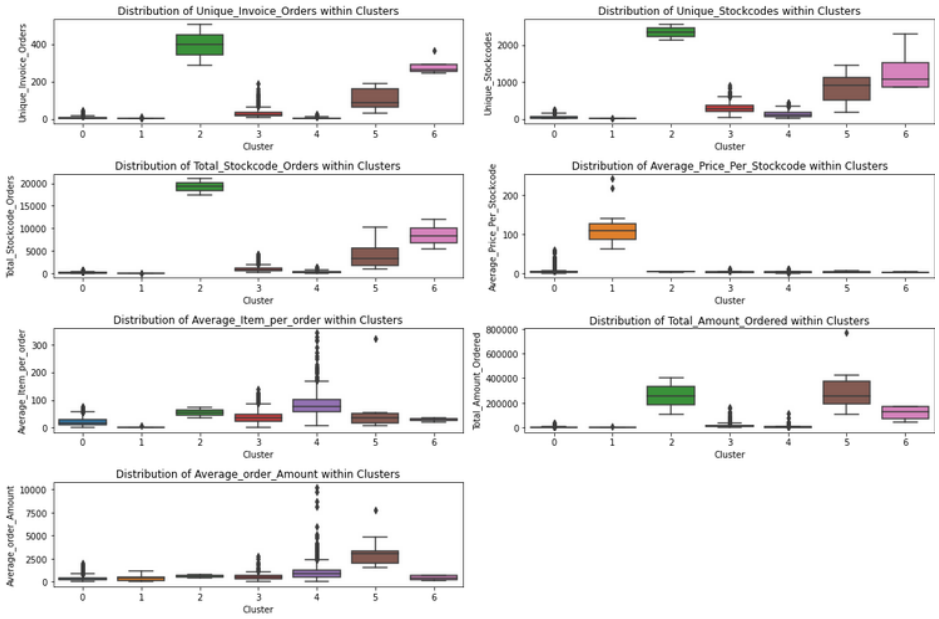
is by using the silhouette method. By comparing the distances between data points in their allocated cluster to those in the closest neighboring cluster, the silhouette score evaluates how well the data were clustered. A number near to 1 indicates that the points in a cluster are far from the points of other clusters and close to the other points in the same cluster. Its range is between 1 and -1. A higher silhouette score represents better clustering, the optimal k is the one that maximizes the silhouette score. The k we validate from both the elbow and silhouette method will be the value we use in K-Means analysis.

Analysis and Results

After obtaining our K-value of seven from the elbow method and validating it with the silhouette method, we are now ready to conduct segmentation analysis to help inform the business's marketing plans.

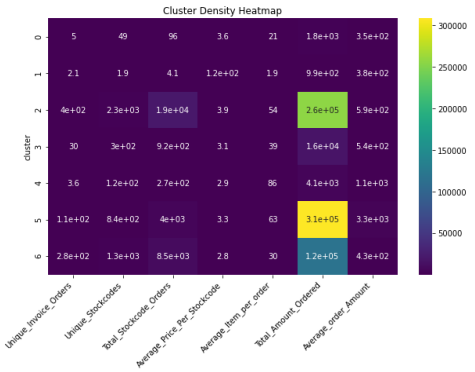


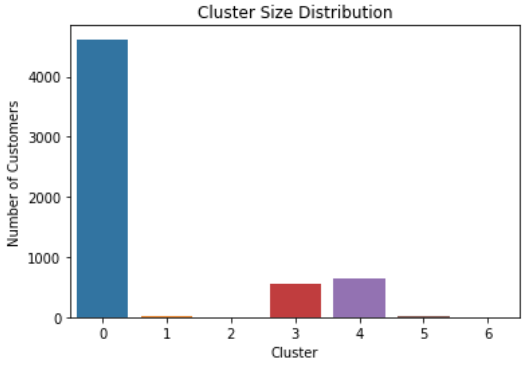
We began our analysis by taking a look at the individual normalized features and seeing how those features compare across each of our 7 clusters. We can do this with a few distinct approaches, starting with our boxplots.



By observing the box plots, it can be found that the distribution of Cluster 2 in Unique Invoice Orders is very good, but there are large outliers in Cluster 3 and the median of Cluster 5 is low. In Unique Stockcodes, it can be found that the distribution of Cluster 2 is still very good, but the median of Cluster 5 is high. Cluster 6 has a low median and a very low mean, near the bottom. If you count Total Stockcode Orders, the distribution of Cluster 2 and 6 is very even. Both the median and mean of Cluster 5 are low. After looking at the Average Price Per Stockcode, we found that although the distribution of Cluster 1 is very even, there are large outliers. Cluster 0 also has a lot of outliers above average. After analyzing the Average Item Per Order, it can be seen that both Cluster 3 and 4 have many abnormal values higher than the highest value. Cluster 5 has some very high outliers that need attention. From the box plot of Total Amount Ordered, it can be seen that Cluster 2 performs very well, while the median of Cluster 5 is low and there are very high outliers that need attention. From the box plot of Average Order Amount, we can see that Cluster 4 has many abnormal values higher than the highest value that need special attention. And the median of Cluster 5 is high.

A quick look at the Cluster Density Heatmap and Cluster Size Distribution bar graph show a significantly larger number of customers belonging to one cluster. This might help us to prioritize our marketing efforts to meet the needs of those customers first and foremost in the context of limited marketing budget or other resource constraints such as time.



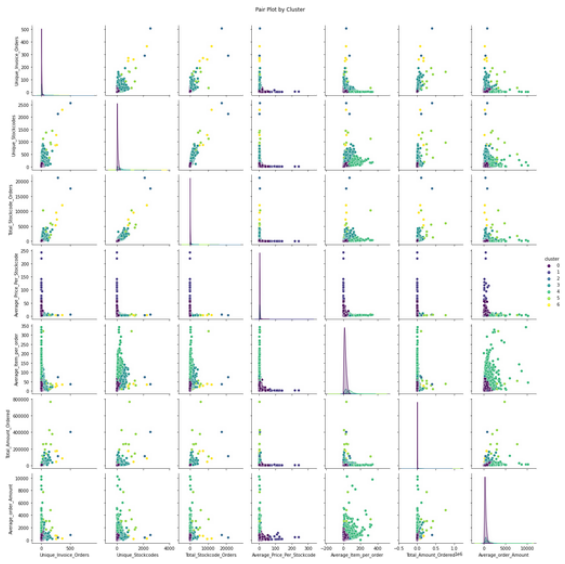


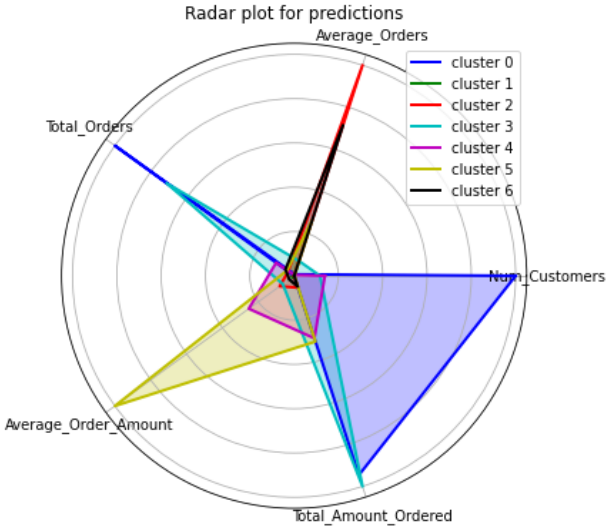
The current limitation is that some clusters have too little data for effective analysis. For example, both Cluster 1 and 5 have only 5. This makes effective analysis a challenge for those two of our seven total clusters in the absence of more data. The current limitation is that some clusters have too little data for effective analysis. For example, both Cluster 1 and 5 have only 5. This makes effective analysis impossible. But at the same time, we can know to have a more detailed understanding of these customers in order to develop a more detailed plan. This suggests we should also consider an analysis with a lower k-value candidate such at K=5. Upon careful inspection of our elbow and silhouette charts, we can see that 5 also satisfies the conditions of a K-value candidate. The broader implications and findings for our study however, remain.

Findings

Finally, in keeping with the results from our box plots and distribution charts, our pairplot and radar plot provide other interesting perspectives on our customer segments. With the pairplot visual we are able to obtain a snapshot of the relationships between each pair of features in our

normalized K-Means spending data. This not only allows us to troubleshoot our model but also to see how these relationships between features differ by cluster. The radar plot, on the other hand, showcases the feature distribution by cluster without the relationship pairs and with a greater emphasis on the cluster as a whole. This chart is particularly useful for creating marketing strategies for individual clusters, which the pairplots can help with when dealing with the particular details of a given strategy.





Conclusion

Online retail has become the new way to shop for customers across the world. Since advertising every product to every potential customer is typically not a viable option for online retailers, segmentation is an essential part of an online retail strategy. An important factor many businesses must take into consideration is, which customers are generating the most revenue. This is a good guide to see which customers to market to and for which product category, if any.

We used the elbow method to measure the sum of squared error from the center of the cluster and silhouette coefficient to measure how well each data point fits in the assigned cluster.

Impacts, Research & Business Implications

We believe cluster analysis, specifically K-Means, is an ideal option to segment customers for this purpose. Throughout the course of our work, several aspects of KMeans stood out that distinguish it as a powerful tool to develop personalized marketing strategies and optimize customer retention. First, KMeans clustering provides us with a way of grouping customers based on selected metrics around each group's shopping habits. Our approach made use of the elbow method to estimate the optimal number of clusters, which was validated by the silhouette method.

The second way our results helped to inform our research question and drive impact was by enabling us to extend our options for segmentation analysis by grouping with our original unnormalized data. This provided a means of extracting additional insights by cluster, and for certain columns, the ability to perform NLP text-based analysis by cluster. By tokenizing and lemmatizing the data for example, removing stop words and performing other preprocessing techniques, we could extract the individual product line for each cluster and expand the line of products marketed to a customer in any given cluster based on their nearest, or most similar, neighbors.

The results of our experiments could provide a foundation for future efforts around KMeans for Segmentation Analysis. The impacts however, are likely to also apply to other forms of cluster analysis like the FSGA-FCEN method or the K-Medoids method mentioned in our 3rd literature review.

References

Accenture. (2018, May 3). Widening Gap Between Consumer Expectations and Reality in Personalization Signals Warning for Brands Accentur. *Accenture Newsroom*. Retrieved July 6, 2023, from https://newsroom.accenture.com/news/widening-gap-between-consumer-expectations-and-reality-in-personalization-signals-warning-for-brands-accenture-interactive-research-finds.htm

U.S. Census Bureau. (2023, June 12). *Annual Retail Trade Survey Shows Impact of Online Shopping on Retail Sales During COVID-19 Pandemic*. Census.gov. https://www.census.gov/library/stories/2022/04/ecommerce-sales-surged-during-pandemic.html

Punj, G., & Stewart, D. J. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Joual of Marketing Research*, *20*(2), 134 https://doi.org/10.2307/3151680

Kaur, S., & Sarabjeet (2021) *Customer Segmentation Using Clustering Algorithm*. IEEE International Conference on Technological Advancements and Innovations (ICTAI) Publication | IEEE Xplore. https://ieeexplore.ieee.org/document/9673169

Wu, Z., Jin, L., Zhao, J., Jing, L., & Chen, L. (2022). Research on Segmenting E-Commerce Customer through an Improved K-Medoids Clustering Algorithm. *Computational Intelligence and Neuroscience*, *2022*, 1–10. https://doi.org/10.1155/2022/9930613