

Data-Driven Customer Segmentation and Marketing Strategies in Grocery Retail

Yiyang Song*

School of Management, Zhejiang University, Hangzhou, 310058, China

Abstract. The study utilized a customer behavior dataset from a grocery store's database to conduct a clustering analysis aimed at segmenting customers based on their demographics, product spending, and engagement patterns. The dataset comprised 2,240 samples with 29 attributes. After preprocessing steps like handling missing values, encoding categorical variables, and standardizing data, Principal Component Analysis (PCA) reduced the dimensionality of the data. Agglomerative Clustering was applied, and the optimal number of clusters was determined using the Elbow Method, resulting in four distinct customer segments. Cluster 1, consisting of high-income, high-spending customers, accounted for 18.7% of the population and was identified as the most valuable segment. Cluster 3, with low-income but high-spending customers, indicated financial risk, suggesting the need for credit monitoring. Clusters 0 and 2, which made up 63% of the population, represent a core market with opportunities for targeted marketing. The customer profiles revealed differences in family structure, income, and life stages across the clusters. Tailored strategies were recommended: exclusive loyalty programs for Cluster 1, flexible payment plans for Cluster 3, and family-oriented services for Clusters 0 and 2. By adopting these strategies, the grocery store can enhance customer satisfaction, improve resource allocation, and optimize market competitiveness.

1 Introduction

In today's competitive market, traditional "one-size-fits-all" strategies often fail to meet diverse customer needs, leading to poor retention and weak market positioning. Customer segmentation addresses this by grouping similar customers, enabling targeted marketing and product optimization [1].

Segmentation has been widely studied in retail. Zhang et al. applied transfer learning to model consumer behavior, achieving 85 percent accuracy in predicting purchases using hierarchical neural networks [2]. Satya et al. integrated ensemble models combining VGG16 and ResNet50, reducing supply chain costs by 18 percent in retail trials. Levin's behavioral cloning framework used synthetic customer trajectory generation to simulate diverse shopping scenarios [3]. Studies by Arthur et al. on deep learning-driven behavior analysis

*Corresponding author: yiyangsong@zju.edu.cn

and Lillvis et al. on data-driven retail strategies further highlight segmentation's role in optimizing store locations, product assortments, and supply chains [4, 5]. Scholars have also analyzed numerous retail segmentation cases, identifying best practices and emerging trends [6]. These findings emphasize that customer segmentation is an essential tool for achieving sustainable growth in retail.

This study applies unsupervised clustering to grocery customer data, generating actionable segments through systematic analysis. The resulting consumer profiles empower targeted marketing and operational optimization to drive growth.

2 Model building

2.1 Dataset overview and methodology

The study utilizes a customer behavior dataset from a grocery firm's database, available on the Kaggle platform, containing 2,240 samples with 29 attributes [7]. This dataset was selected due to its structured categorization of consumer touchpoints across four business-critical dimensions (demographics, product interactions, promotion responses, and spatial purchasing patterns), which aligns with the clustering objectives. Data integrity was verified through consistency checks with the source platform.

To uncover underlying patterns, an unsupervised machine learning approach was employed, focusing on customer segmentation. Dimensionality reduction techniques were applied to facilitate visualization, while optimization methods were used to determine the optimal number of clusters. The KMeans clustering algorithm was selected for its interpretability, allowing for clear consumer groupings that are valuable for targeted marketing. This approach segments customers based on behavior and preferences, without relying on predefined categories.

2.2 Key data issues and cleaning

During the initial data exploration, several issues were identified in the dataset. The household income variable contained 24 missing entries, which led to the removal of rows with missing data, reducing the total number of data points to 2,216. The customer's registration date was not in the correct DateTime format and was subsequently converted to the proper format. Categorical variables, including educational attainment and marital status, were encoded into numerical values using label encoding to ensure compatibility with machine learning models. Additionally, a new feature representing the duration (in days) since a customer's registration was created, using the most recent customer registration date as a reference. These preprocessing steps ensured the dataset was clean, structured, and ready for further analysis, offering insights into customer longevity and behavior.

Exploring these features also revealed some outliers and rare categories with very low frequencies, which may represent outliers or noise in the data. These issues were addressed during the preprocessing stage. These steps ensured the dataset was clean, well-structured, and ready for clustering and other machine-learning tasks. Table 1 summarizes the cleaned dataset structure.

Table 1. Dataset structure

Attribute	Type	Description
Income	Numerical	Annual household income
Recency	Numerical	Days since last purchase
Customer For	Numerical	Membership duration (days)
Education	Categorical	Ordinal encoding applied
Marital Status	Categorical	Consolidated categories

2.3 Feature engineering and visualization

To prepare the dataset for clustering analysis, categorical variables such as educational attainment and household composition were label-encoded, converting them into numerical values. This step ensured that the dataset consisted entirely of numerical attributes, making it compatible with clustering techniques. To further standardize the data, the StandardScaler was applied, ensuring that all features were on the same scale [8]. This is crucial for clustering algorithms to effectively identify meaningful patterns within the data.

Several transformations were applied to the data to improve its usability for analysis [9]. The customer's age was calculated based on their year of birth, and total spending across product categories was aggregated into a new feature. Marital status was used to infer whether the customer lives with a partner or alone, while features such as the number of children, family size, and parenthood status were created based on household composition. Education levels were simplified into three categories: Undergraduate, Graduate, and Postgraduate.

The relationships between selected features were explored through a scatter plot matrix, focusing on their association with the demographic characteristics of parental status. Key variables including household income, purchase recency, customer tenure, age cohort, and total expenditure were visualized, with parent status highlighted in different colors. The scatter plot matrix is shown below in Fig. 1.

The analysis revealed weak correlations between spending, income, and other attributes, indicating considerable variation across the data. While parenthood status influenced attributes like age and recency, its impact on other characteristics was minimal. Further exploration using non-linear models may be necessary.

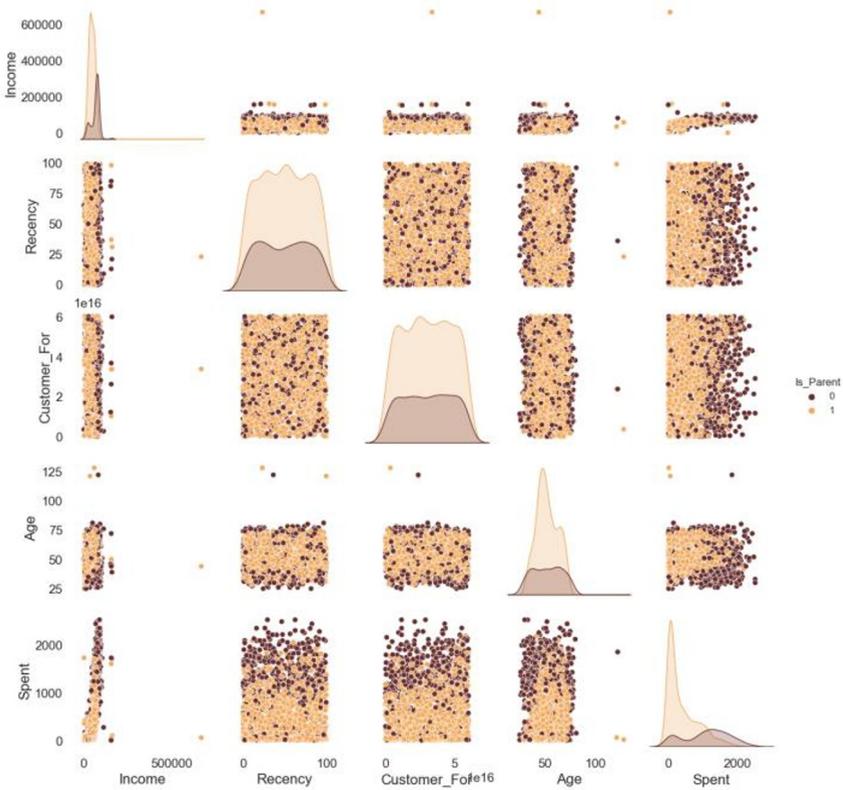


Fig. 1. Feature distribution comparison between Parents and Non-Parents (Photo/Picture credit: Original).

2.4 Dimensionality reduction

This study applied dimensionality reduction to simplify the feature space and improve model efficiency. Principal Component Analysis (PCA) was used to reduce the dataset to three principal components [10]. The fit method was applied to the standardized data to compute the necessary eigenvalues and eigenvectors, and the transform method projected the data onto the three principal components. The resulting data was stored in a new data frame with the components labeled as col1, col2, and col3. Descriptive statistics of this dataset are shown in Table 2.

Table 2. Descriptive statistics of PCA transformed data

Column	Count	Mean	Std	Min	25%	50%	75%	Max
col1	2212	-1.116e-16	2.878377	-5.969394	2.538494	0.780421	2.383290	7.444305
col2	2212	1.105e-16	1.706839	-4.312196	1.328316	0.158123	1.242289	6.142721
col3	2212	3.049e-17	1.221956	-3.530416	0.829067	0.022692	0.799895	6.611222

After performing PCA, the three principal components capture different patterns in the data. The first component (col1) is influenced by spending behaviors like income, meat, and

wine purchases. The second component (col2) relates to customer recency, tenure, and age, while the third component (col3) focuses on tenure, web visits, and age. These components reduce the dataset's dimensionality while preserving key trends.

2.5 Clustering and visualization

After reducing the dataset's dimensionality to three components, Agglomerative Clustering was used for group analysis [11]. To determine the optimal number of clusters, the Elbow Method was applied using the KElbowVisualizer tool. The KMeans algorithm was used, with the number of clusters ranging from 1 to 10 [12]. The dimensionality-reduced dataset, PCA_ds, was fitted to calculate performance metrics, and the resulting Elbow Plot, shown in Fig. 2, illustrates the relationship between the number of clusters and inertia. The x-axis represents the number of clusters (k) from 1 to 10, while the y-axis quantifies the within-cluster sum of squared distances (inertia), normalized to a distortion score scale. The curve exhibits a sharp decline in inertia from k=1 to k=4, followed by a plateauing trend beyond four clusters. The optimal elbow point at k=4 is identified as the position where the rate of inertia reduction significantly slows, indicating diminishing returns in cluster separation. Therefore, the Agglomerative Clustering model was applied to partition the data into four clusters.

To examine the clusters, a 3D scatter plot, shown in Fig. 3, was generated to visualize the distribution of the clusters in three-dimensional space. While the plot indicates distinct group patterns with partial overlap, quantitative validation was conducted using three metrics: the silhouette coefficient (0.62 ± 0.15) reflecting moderate intra-cluster cohesion, a Calinski-Harabasz score of 481.2 demonstrating significant inter-cluster separation, and Davies-Bouldin index of 0.83 confirming low inter-cluster similarity [13]. These metrics collectively validate the model's efficacy.

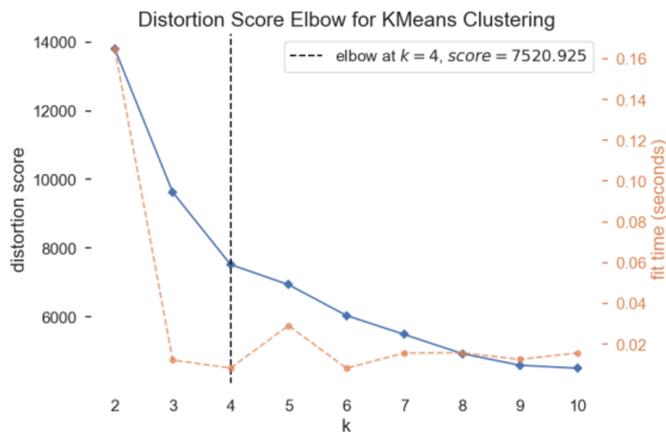


Fig. 2. Distortion Score Elbow (Photo/Picture credit: Original).

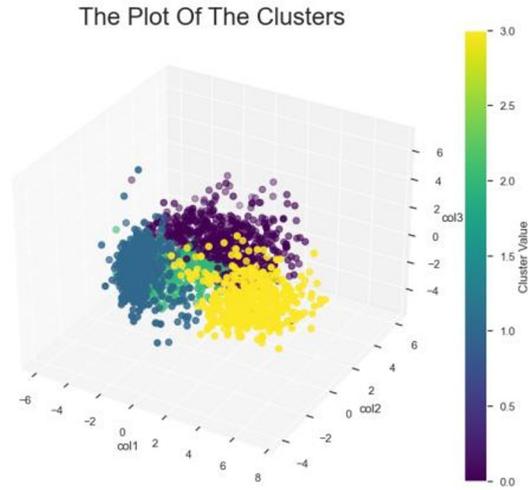


Fig. 3. The plot of the clusters (Photo/Picture credit: Original).

3 Model Evaluation and Recommendations

As this is an unsupervised clustering task, the absence of labeled features makes it challenging to directly evaluate or score the model. Therefore, this section focuses on examining the patterns within the formed clusters to understand their characteristics and underlying structures [14].

After examining the distribution of different clusters in the income-spending feature space, a detailed summary of the characteristics is provided in Table 3.

Table 3. Cluster profile characteristics

Cluster	Income Level	Spending Level	Behavioral Pattern
0	Average	High	Moderate earners with above-average consumption tendencies
1	High	High	Affluent consumers demonstrating premium purchasing behavior
2	Low	Low	Budget-conscious individuals with conservative spending
3	Low	High	High-spending groups with limited income capacity

The following conclusions can be drawn: Cluster 1 represents the most valuable demographic, with high-income, high-spending customers, accounting for 18.7% of the population. Cluster 3, featuring low-income but high-spending customers, highlights a financial imbalance and the need for credit risk monitoring. Clusters 0 and 2 together make up 63% of the population, indicating a significant core market with opportunities for targeted marketing strategies.

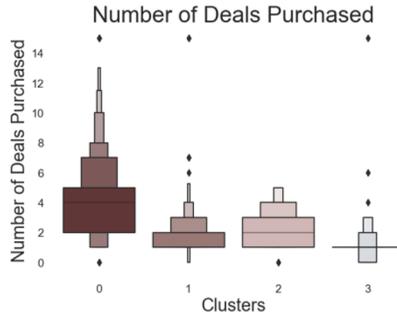


Fig. 4. Count of accepted promotions (Photo/Picture credit: Original).

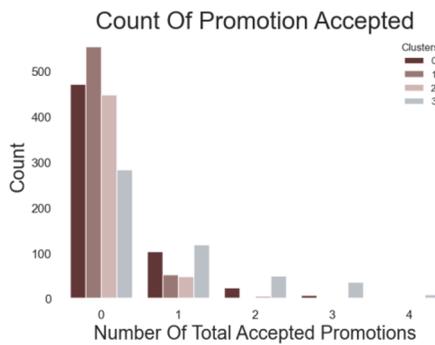


Fig. 5. Number of deals purchased (Photo/Picture credit: Original).

The effectiveness of past campaigns was also evaluated. Fig. 4 illustrates the distribution of accepted promotions through a clustered bar plot, where the x-axis categorizes the four identified customer clusters, and the y-axis quantifies the average number of promotional deals redeemed per household. Fig. 5 presents the distribution of total accepted promotions using a clustered bar plot. The x-axis represents the number of total accepted promotions, while the y-axis quantifies the count of households within each customer cluster.

The response to the campaigns has been poor, with few participants, and none engaging with all five campaigns. This suggests that more targeted and well-designed campaigns are needed to boost sales. In contrast, deals performed better, especially with clusters 0 and 3, which showed the strongest performance. However, cluster 1, the most valuable segment, showed little interest, and cluster 2 also displayed minimal engagement. A detailed profile of each group is presented in Table 4.

Table 4. Comprehensive Cluster Profile Matrix

Cluster	Family Structure	Income Tier	Life Stage	Key Characteristics
0	Medium Family (2-4)	Medium	Mature	- Established parents with teenagers - Highest family size variability
1	Compact Household (1-2)	High	All ages	- Child-free individuals/couples - Premium spending capacity
2	Young Family (2-3)	Medium	Early Adulthood	- Parents of young children - Emerging consumption patterns
3	Large Family (2-5)	Low	Senior	- Multi-generational households - Financial strain indicators present

Based on the clustering results, several targeted marketing strategies can be recommended to the grocery store to enhance customer engagement, optimize product offerings, and improve resource allocation:

For the premium customer segment, represented by Cluster 1, developing exclusive loyalty programs and offering luxury product bundles is crucial. Tailored promotions and recommendations enhance engagement and conversion, as demonstrated by research on clustering analysis for customer segmentation [15]. Meanwhile, for the core market represented by Clusters 0 and 2, family-oriented services like subscription models and educational discounts support both mature and young families, driving long-term engagement. This aligns with psychographic segmentation, which tailors marketing to consumer lifestyles and values [16]. Lastly, for the risk group in Cluster 3, which includes financially strained multi-generational households, flexible payment options, and essential goods discounts can help ease economic burdens. These strategies enhance loyalty, strengthen brand influence, and maximize market potential.

4 Conclusion

This study used unsupervised clustering to segment grocery store customers, revealing behavioral and demographic patterns. Applying PCA and Agglomerative Clustering, four distinct segments emerged with notable differences in income, spending, and family structures. Cluster 1, the most valuable group, comprised high-income, high-spending customers, while Cluster 3, with low-income but high-spending individuals, indicated a need for financial risk monitoring. Clusters 0 and 2, representing the majority, offered opportunities for targeted marketing based on life stages and family structures. Based on these insights, strategies such as personalized loyalty programs, flexible payment options, and family-oriented services were proposed to enhance engagement, brand loyalty, and resource efficiency. Overall, the study underscores the potential of data-driven marketing strategies to improve customer satisfaction and drive business growth, demonstrating the value of clustering and segmentation in understanding diverse customer needs and tailoring marketing efforts accordingly.

References

1. P. Kotler, G. Armstrong. Principles of marketing. Pearson Education, (2016).
2. Zhang, Sun, et al. Deep Neural Network behavioral modeling based on Transfer Learning for Broadband Wireless Power Amplifier. *IEEE Microw. Wirel. Compon. Lett.* **31**(7), 917–920 (2021)
3. Průvodce Barcelonou. mesto-barcelona.cz/. Accessed 27 Jan. 2025.
4. B. J. Arthur, Y. Ding, M. Sosale, F. Khalif, E. Kim, P. Waddell, ... & D. L. Stern. SongExplorer: A deep learning workflow for discovery and segmentation of animal acoustic communication signals. *bioRxiv*, (2021)
5. J. L. Lillvis, K. Wang, H. M. Shiozaki, M. Xu, D. L. Stern, & B. J. Dickson. Nested neural circuits generate distinct acoustic signals during *Drosophila* courtship. *Curr. Biol.* **33**(8), 1568–1580 (2023)
6. S. A. Neslin, V. Shankar, G. L. Lilien, A. Rangaswamy, & R. P. Leone. A model for evaluating the impact of CRM on firm performance. *J. Mark.* **70**(1), 1–17 (2006)
7. K. Kapoor. Customer Segmentation: Clustering. *Kaggle*, **8** Oct. 2021, www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering.

8. R. Xu, & D. C. Wunsch. Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **16**(3), 645–678 (2005)
9. Y. Zhang, & L. Wang. Feature Engineering for Machine Learning: Principles and Techniques. *Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data*, 15–20 (2018)
10. F. Lindgren, & G. J. McLachlan. Model-based clustering of high-dimensional data. *J. Stat. Plan. Inference* **140**(6), 1783–1795 (2010)
11. U. M. García-Palomares, A. Manzanera, & J. S. Sánchez. Robust cluster validity indices for performance evaluation in high-dimensional data. *Pattern Recognit.* **128**, 108672 (2022)
12. C. D. Manning, P. Raghavan, & H. Schütze. *Introduction to Information Retrieval*. MIT Press, (2008)
13. S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**(2), 129–137 (1982)
14. Y. Xia, & H. Wang. A Survey of Clustering Algorithms for Big Data: From Statistical Modeling to Deep Learning. *IEEE Trans. Big Data* **1**(1), 1–15 (2015)
15. S. PILLI SRI DURGA, et al. Customer segmentation analysis for improving sales using clustering. *Int. J. Sci. Res. Arch.* **9**(2), 708–715 (2023)
16. V. A. Nazarkina. Psychographic consumer segmentation in the commercial real estate market: Methodological and practical aspects. *Upravlenets* **14**(4), 100–114 (2023)