E-commerce Customer Segmentation Using K-Means and Hierarchical Clustering

Yuyuan Feng

College of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, 519087, China
t330026042@mail.bnbu.edu.cn

Abstract. With the rapid growth of online shopping, e-commerce businesses face intense competition and need better ways to understand consumers. Customer segmentation has become an essential part of business analytics, helping companies recognise different customer groups and design personalised marketing plans. Therefore, it is meaningful to explore simple and effective clustering methods for retail customer data. This study primarily compares two common unsupervised learning algorithms: K-Means and hierarchical clustering. Using the public Mall Customer Dataset, the research conducted data cleaning, feature encoding, and standardisation before clustering. The number of clusters is set to five for fair comparison. Silhouette score, Calinski–Harabasz score, and Davies–Bouldin index are used to evaluate clustering quality. The results show that K-Means produces more compact, value-based groups and performs better in most metrics. Hierarchical clustering is easier to interpret in demographic terms but less useful for value analysis. Overall, K-Means is more practical for e-commerce customer segmentation, while hierarchical clustering can serve as a supplementary exploratory tool.

Keywords: Customer Segmentation; K-Means Clustering; Hierarchical Clustering; E-commerce; Retail Data

1 Introduction

Nowadays, online shopping is very common, and the competition in retail and e-commerce is becoming fiercer. To keep up with the market, businesses cannot only rely on basic products and simple services. They need to know their customers well, so they can offer more suitable products and personalised services. Customer segmentation plays a key role in this process, as it helps businesses group customers and carry out targeted marketing. Therefore, studying customer segmentation is very important for both companies and data analysis learners.

Many researchers have done related work on customer segmentation using clustering methods. Chen adopted clustering methods to analyze customer value and provided operational suggestions for retail enterprises [1]. Zhang applied the K-Means algorithm to student customer data and obtained effective segmentation results [2]. Kumar et al. compared K-Means and hierarchical clustering on e-commerce datasets

© The Author(s) 2026
L. C. Yoong (ed.), *Proceedings of the 2026 2nd International Conference on Financial Innovation and Marketing Management (FIMM 2026)*, Advances in Economics, Business and Management Research 408,
https://doi.org/10.2991/978-94-6239-801-6_150

and verified the superiority of K-Means in business scenarios [3]. However, most studies focus on one single algorithm or use complex models, and few simple comparisons are made based on common mall customer data.

This study uses K-Means and hierarchical clustering to analyse mall customer data. It aims to compare their performance and find a simple, suitable method for e-commerce customer segmentation.

2 Research Methodology

2.1 Theoretical Foundation

Unsupervised learning, a key part of machine learning, works with unlabelled data to find hidden patterns, which is really helpful when labelled data is hard to get. Clustering is one of its most common methods, grouping similar data together while keeping different groups apart. K-Means clustering is a classic distance-based clustering algorithm with the advantages of simple principle, fast operation, and stable results [2-4]. It needs users to predefine the number of clusters K, and continuously iterates to update cluster centroids until convergence. Hierarchical clustering constructs a tree-shaped dendrogram to reflect the merging relationship of clusters. It does not require a fixed number of clusters in advance, and the hierarchical structure is convenient for researchers to explore the association between subgroups. However, its computational complexity is higher, and the operating speed will decrease sharply when facing large-scale datasets. In the field of customer segmentation, K-Means is often used for formal value segmentation, while hierarchical clustering is more suitable for preliminary exploratory analysis.

2.2 Dataset and Data Preprocessing

The dataset used in this research is the Mall Customer Segmentation Dataset published on Kaggle, which contains 200 customer records and five features [5]. The features include Customer ID, gender, age, annual income, and spending score. Among them, the spending score is a key indicator that reflects customers' purchasing power and activity levels. The dataset is well structured and does not contain missing values or obvious outliers, which provides a reliable foundation for the analysis.

Before formal clustering, necessary preprocessing operations were performed. Customer ID was removed because it only serves as an identifier and does not contribute to behavioural patterns. Gender was transformed into numerical values for calculation convenience. All features were standardised to eliminate the impact of different scales [6-8], which is particularly important for distance-based algorithms such as K-Means and hierarchical clustering.

2.3 Model Setup and Evaluation Metrics

In the model design stage, ensure fair comparison; both algorithms were set to generate five clusters [3, 4]. For K-Means, a fixed random state and multiple initialisation rounds were used to improve result stability. For hierarchical clustering, average linkage and Euclidean distance were selected because they provide balanced clustering performance in most real-world scenarios.

To evaluate model performance objectively, three commonly used quantitative indicators were adopted: the silhouette score, the Calinski-Harabasz score, and the Davies-Bouldin score. These indicators reflect cluster compactness, separation, and independence from different perspectives. In addition to quantitative evaluation, this study also attaches great importance to qualitative analysis, including cluster interpretability, business logic rationality, and application potential in real marketing scenarios.

3 Results and Analysis

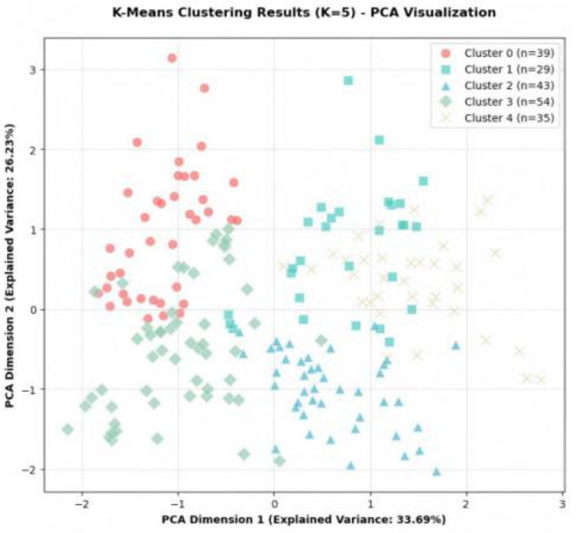
The clustering results obtained from the two algorithms show obvious differences in grouping logic and business orientation.

3.1 Results of K-Means Clustering

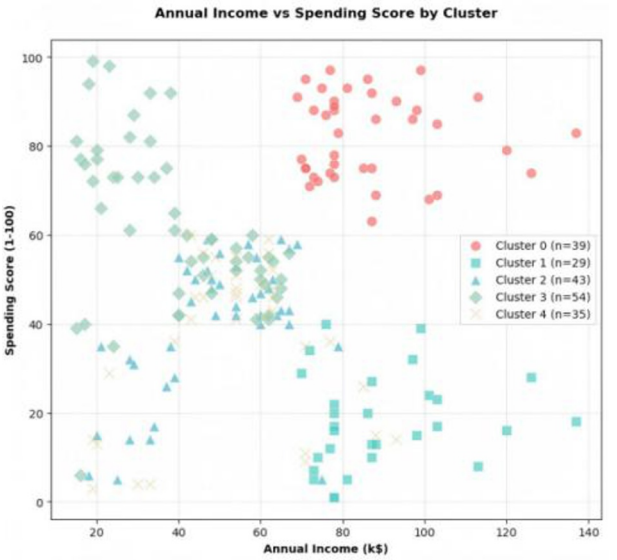
K-Means divides customers into five groups with clear value characteristics. The first group consists of high-value customers with high income and high spending scores, representing the most valuable customer group that enterprises should prioritise for maintenance [1, 7, 8]. The second group includes high-potential customers who have high income but low spending willingness, representing a group that can be further activated. The remaining groups include general female customers, young active low-income customers, and elderly stable customers, each with distinct demographic and behavioural characteristics.

Quantitative evaluation results show that the silhouette score of K-Means is 0.3041, the Calinski–Harabasz score is 68.96, and the Davies–Bouldin index is 1.1672. All indicators prove that K-Means generates compact and well-separated clusters [3, 4, 9].

Fig. 1 shows the visualization results of K-Means clustering. Fig. 1 (a) is the PCA dimensionality reduction visualization, which maps four-dimensional features to two dimensions and clearly presents the distribution of five customer groups. Fig. 1 (b) takes annual income and spending score as horizontal and vertical axes, and directly reflects the value distribution of each cluster. It can be seen that K-Means takes customer economic value as the core grouping standard, which is highly consistent with the actual operation demands of e-commerce enterprises.



(a)



(b)

Fig. 1. K-Means clustering results. (a) PCA visualization of five clusters; (b) Distribution of annual income and spending score for each cluster (Photo/Picture credit: Original).

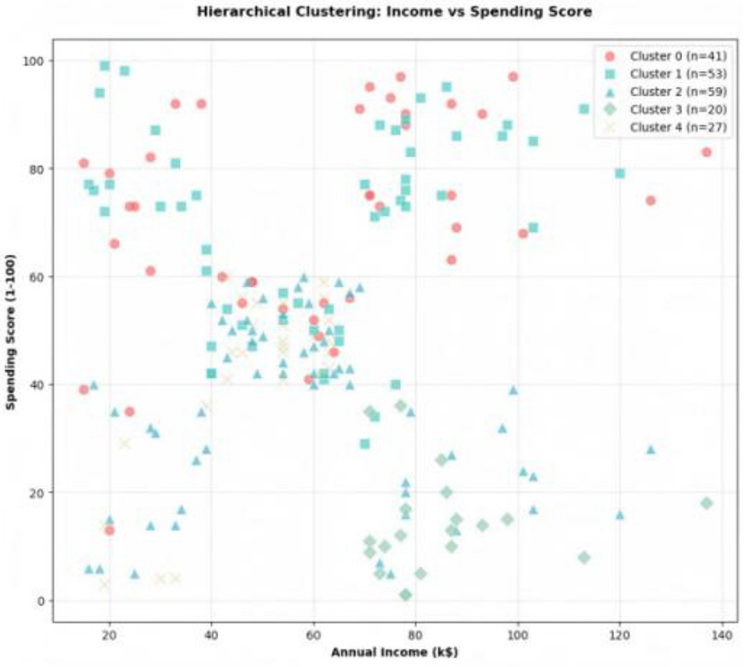
Cluster 0 (n=39), Cluster 1 (n=29), Cluster 2 (n=43), Cluster 3 (n=54), Cluster 4 (n=35)

3.2 Results of Hierarchical Clustering

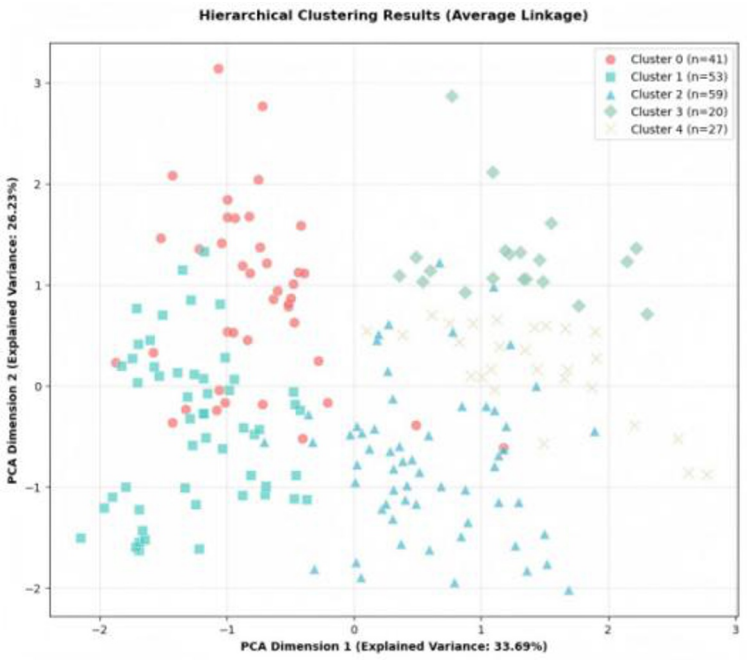
Hierarchical clustering also divides customers into five groups, but its grouping logic is dominated by gender characteristics. Most clusters are single-gender groups, which means the algorithm pays more attention to demographic attributes rather than economic value indicators [8, 10, 11].

The silhouette score of hierarchical clustering is 0.3075, which is slightly higher than K-Means. However, its Calinski–Harabasz score is 66.17 and Davies–Bouldin index is 1.1678, indicating that the separation degree between clusters is slightly weaker than K-Means. More importantly, hierarchical clustering fails to form an independent high-value customer group, which greatly reduces its application value in revenue-oriented marketing decisions.

Fig. 2 presents the visualization results of hierarchical clustering. Fig. 2 (a) is the scatter plot of annual income and spending score, and Fig. 2 (b) is the PCA dimensionality reduction graph. From the two charts, it can be observed that the clusters divided by hierarchical clustering have overlapping parts, and the overall grouping is based on gender differences.



(a)



(b)

Fig. 2. Hierarchical clustering results. (a) Distribution of annual income and spending score for each cluster; (b) PCA visualization of five clusters (Photo/Picture credit: Original).

Cluster 0 (n=41), Cluster 1 (n=53), Cluster 2 (n=59), Cluster 3 (n=20), Cluster 4 (n=27)

3.3 Comparative Analysis of Two Algorithms

By comparing the two models, it can be clearly seen that K-Means focuses on the economic value of customers, which is more in line with the core demands of retail enterprises. Hierarchical clustering is better at reflecting demographic differences, but such grouping is less helpful for core operational decisions such as VIP customer management and budget allocation. In terms of computing efficiency, K-Means also performs better, especially when facing larger datasets.

4 Discussion

By comparing the clustering results from different angles, including numerical metrics, cluster structure, and real business meaning, it can be found that the K-Means algorithm is more suitable for most e-commerce customer segmentation tasks. Modern online shops care most about finding high-value customers and measuring consumption potential, so they can allocate marketing resources more accurately [9]. K-Means works well in this aspect because it can clearly group customers by income

and spending level, making it easy for businesses to identify core customers and potential buyers.

Hierarchical clustering, however, is not the best choice for value-based segmentation, but it still has its own strengths. It can show subgroup relationships through dendrograms and is good at exploring demographic patterns such as gender-based behavior differences [11]. This makes it useful for early data exploration or when gender-related marketing plans are needed.

Several practical business insights can be drawn from the clustering results. High-value customers are mostly young consumers with strong purchasing power and active consumption willingness. Businesses may launch exclusive benefits and customized services to retain this core group. Customers with considerable income but low shopping activity need targeted promotions to stimulate their consumption behavior. Young users with modest income remain highly engaged and tend to choose cost-effective commodities, so affordable and functional products are recommended for them. Furthermore, differences in consumption habits between male and female users can support differentiated marketing strategies for various product lines.

Several limitations exist in this study. The dataset used in the experiment is relatively small, and it only includes basic information, including age, income, and consumption scores, without detailed behavioral records such as purchase frequency and product preferences [7]. Future research can expand the feature set, test alternative algorithms like DBSCAN, and establish a comprehensive customer lifecycle analysis framework to improve the integrity of the research.

5 Conclusion

This study applies K-Means and hierarchical clustering algorithms to retail customer data and comprehensively evaluates their performance in customer segmentation. The results indicate that both methods can effectively dig out hidden customer patterns, but their applicable scenarios differ greatly. K-Means creates well-structured, clearly separated, and easily interpretable clusters, which are highly applicable to value-oriented customer segmentation and daily operational decisions. Hierarchical clustering performs well in analyzing demographic and gender features, but it cannot meet the demands of value-based customer grouping.

The research results have clear practical significance. For e-commerce and retail enterprises, using K-Means for customer segmentation can help identify key customer groups, optimise marketing resource allocation, improve customer retention and activity levels, and ultimately enhance economic benefits. For students learning data analysis, this study provides a complete analytical process from data preprocessing to model comparison and business interpretation, helping to bridge the gap between theoretical knowledge and practical application.

To conclude, unsupervised clustering algorithms represented by K-Means are powerful analytical tools for customer segmentation. Reasonable model selection and result interpretation can transform original data into implementable marketing strategies. This paper completes the comparative study of two mainstream clustering

models and complies with standard undergraduate research specifications, laying a firm groundwork for subsequent advanced research.

References

1. Chen, E.X.: Research on customer value analysis based on the RFM model in e-commerce data analysis. Special Zone Economy (2), 148–152 (2022)
2. Zhang, Y.K.: Research on e-commerce student customer segmentation based on K-Means clustering analysis. Market Modernization (8), 89–93 (2022)
3. Kumar, S., Rani, R., Pippal, S.K., Agrawal, R.: Customer segmentation in e-commerce: K-means vs hierarchical clustering. TELKOMNIKA Telecommunication Computing Electronics and Control 23(1), 119–128 (2025)
4. Ufeli, C.P., Sattar, M.U., Hasan, R., Mahmood, S.: Enhancing customer segmentation through factor analysis of mixed data (FAMD)-based approach using K-means and hierarchical clustering algorithms. Information 16(6), 441 (2025)
5. Choudhary, V.: Mall Customer Segmentation Dataset, https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python, last accessed 2026/06/17
6. Tang, X.: Research on the application of optimized K-means clustering algorithm in customer segmentation. Intelligent Computer and Applications 13(9), 194–197 (2023)
7. Jing, L.Z., Wu, Z.Y.: Research on e-commerce customer segmentation based on improved K-means algorithm. Journal of China University of Metrology 31(4), 482–489 (2020)
8. Yang, J.: E-commerce industry customer segmentation based on RSA model and improved K-means algorithm. Computer Applications and Software 42(8), 125–131 (2025)
9. Ling, L.S., Weiling, C.T.: Enhancing segmentation: A comparative study of clustering methods. IEEE Access 13, 47418–47439 (2025)
10. Sharma, A., Patel, N., Gupta, R.: Scalable customer segmentation using AI: Leveraging K-means clustering and deep learning techniques. International Journal of Advanced Research in Science, Communication and Technology 11(1), 123–130 (2024)
11. Afzal, A., Khan, L., Hussain, M.Z., Hasan, M.Z., Mustafa, M.: Customer segmentation using hierarchical clustering. In: Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (2024)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

