JABBS
Journal of Applied Business & Behavioral Sciences

Credamo PRESS

Leveraging Machine Learning for Customer Segmentation and Personalized Recommendation in Online Retail

Jiajun Li[a,*]

[a] Computing, Engineering and Digital Technologies , Teesside University, Tees Valley, Middlesbrough, TS1 3BX, UK

How to cite: Li, J. J. (2026). Leveraging Machine Learning for Customer Segmentation and Personalized Recommendation in Online Retail. *Journal of Applied Business & Behavioral Sciences, 2(1)*, 256-284. https://doi.org/10.63522/jabbs.201013

Abstract

This study explores how artificial intelligence (AI) can support the full online sales process, not just a single feature. Using a public transactional dataset from a UK retailer, we build a practical framework that moves from simple customer profiling to clear, actionable recommendations. We first create customer features and review their relationships, finding a strong "purchase intensity" pattern: frequent buyers tend to spend more, buy a wider range of products, and show larger month-to-month variability, while long gaps since the last purchase align with lower activity. We then group customers into segments and, within each segment, list the best-selling items. Each shopper receives suggestions for top products they have not yet bought. The pipeline includes basic data cleaning and outlier handling to keep the signals stable. Although the method is simple, it is transparent, easy to deploy, and works well when individual histories are thin. We outline practical checks for success — time-based splits for offline tests and A/B experiments online — and note limits, such as popularity bias and modest personalization. We also describe straightforward upgrades: giving more weight to recent activity, considering revenue or margin, adding variety to the list, and re-ranking with lightweight personal signals. Overall, the work offers a clear, reproducible path that links customer understanding to day-to-day recommendation decisions, supporting both business action and future research.

Keywords: Machine learning; Online retail; Purchase intensity

1. Introduction

1.1 Research Background of the Era

In the context of the rapid iteration of the digital economy and information technology, Artificial Intelligence (AI) has gradually moved from the technical exploration stage to the real economy fields such as e-commerce, marketing, and retail operations, becoming a core engine driving industrial structure upgrading and reshaping the business ecosystem. With the increasing maturity of technologies such as big data and deep learning, AI is no longer a mere auxiliary tool, but penetrates the entire process of business activities, comprehensively reshaping the operational logic of traditional business

* Correspondence: jiajundh@163.com

Received 19 January 2026; Revised 2 March 2026; Accepted 2 March 2026

from demand mining, customer connection, service supply to operational optimization, and injecting new vitality into industrial development.

Under the tide of digital transformation, the e-commerce and retail industries have taken the lead in launching AI-enabled practices. The consumption scenarios integrating online and offline have become increasingly abundant, and consumers' demand for personalized and efficient services has continued to increase, forcing enterprises to break through traditional operational bottlenecks with the help of AI technology. At the same time, the breakthrough development of Generative AI (GenAI) has further broken the boundaries of content production and marketing communication, bringing new innovation space and development opportunities to the business field, and also giving rise to a series of urgent theoretical and practical issues, laying an important era foundation for this research.

1.2 Industry Application Status and Development Trend

In the field of e-commerce, the application of AI technology has formed a large-scale implementation pattern, showing significant technical value from basic personalized recommendations to complex supply chain optimization. Bawack et al. (2022) pointed out through research that the application of AI in e-commerce scenarios has covered core links such as user profiling, precision marketing, and inventory scheduling, gradually building an intelligent business operation system and promoting the transformation of the e-commerce industry from traffic competition to efficiency competition. The application of deep learning technology has further strengthened this trend. Zhang et al. (2019) found that deep learning-based recommendation systems can accurately capture users' potential needs, significantly improve the user conversion rate and retention rate of e-commerce platforms, and have become an important part of the core competitiveness of e-commerce platforms.

The marketing field has also benefited from the innovation of AI technology, and the shortcomings of large-scale and homogenized traditional marketing models have been gradually addressed. Davenport et al. (2020) prospectively pointed out that AI is fundamentally reshaping marketing processes, customer relationship management, and value creation models, transforming marketing activities from "blanket" promotion to data-driven precision operations. In particular, the rise of generative AI has brought disruptive changes to the marketing field. Kshetri and Dwivedi (2024) proposed that the application of GenAI in scenarios such as content creation and personalized interaction can quickly respond to changes in market demand, but GenAI is also accompanied by new challenges such as data security and ethical compliance, making the practical exploration of AI-enabled marketing more complex.

As an important application carrier of AI technology, the retail industry is closely related to the depth of integration with AI technology. Shankar (2018) clearly stated that AI is reconstructing the retail industry from multiple dimensions such as supply chain management, store operations, and customer experience, promoting the industry's transformation from traditional offline models to online-offline integration, and from large-scale supply to personalized services. Grewal et al. (2017) also emphasized in their prediction of the future development of the retail industry that AI has become the core driving factor for the transformation and upgrading of the retail industry, and the depth of its application directly determines the position of enterprises in market competition. This trend has been fully verified in industry practices in recent years.

Dynamic pricing and customer relationship management, as core operational links in the e-commerce and marketing fields, have also achieved optimization and upgrading under the empowerment of AI technology. The early research by Elmaghraby and Keskinocak (2003) laid the

foundation for AI-enabled dynamic pricing, while the empirical research by Feng et al. (2019) further verified the application value of multi-dimensional data such as online reviews in AI dynamic pricing models, making pricing strategies more in line with changes in market demand. In terms of customer relationship management, Rahman et al. (2023) showed that the technological readiness of B2B enterprises directly affects the application effect of AI in customer relationship management, and enterprises with high readiness can effectively improve customer satisfaction and loyalty through AI technology, building long-term stable customer relationships.

1.3 Necessity and Significance of the Research

Despite the significant progress in the application of AI in e-commerce and marketing, existing practices still face many urgent problems to be solved. From the perspective of technical application, AI dynamic pricing mostly focuses on a single influencing factor, insufficiently considering the synergistic effect of multiple factors, and the adaptability of deep learning technology in cross-scenario applications still needs to be optimized. From the perspective of industry practice, the application ethics, data security, and regulatory mechanisms of generative AI are not yet sound, and enterprises face the dilemma of balancing ethical compliance and innovative development in the process of technology implementation. Verma et al. (2021) also pointed out through systematic research that there is still a research gap in the in-depth integration of AI and marketing models, and the guiding role of existing theories in practice needs to be further strengthened.

At the same time, the rapid iteration of AI technology and the continuous upgrading of industry demand have led to the continuous emergence of innovative practices in the e-commerce and marketing fields. There is an urgent need for relevant research to sort out existing phenomena, analyze core issues, and provide theoretical support for industry practice. The AI marketing strategic framework constructed by Huang and Rust (2021) shows that clarifying the application path and boundary of AI technology is the key to promoting technology implementation and realizing commercial value. In this context, carrying out this research, based on the current application status of AI in e-commerce and marketing, focusing on core issues and development bottlenecks, has important theoretical and practical value. It can not only fill the gaps in existing research but also provide scientific guidance for enterprises' AI-enabled practices.

1.4 Research Directions and Innovations

This study takes the transaction dataset of a UK-based retailer (available from the UCI Machine Learning Repository) as the research object to explore the rapidly expanding online retail sector. The core objective of the research is to enhance marketing performance and drive sales growth by leveraging customer segmentation techniques. To achieve this goal, we convert the raw transaction data into a customer-centric format through feature engineering, thereby enabling the application of the K-means clustering algorithm to divide customers into multiple differentiated segments. The clustering process reveals the unique characteristics and purchasing patterns of distinct customer groups. Building on these research findings, we further develop a recommendation system that accurately identifies and pushes popular products to customers within each segment who have not yet purchased them, ultimately boosting marketing impact and supporting sales growth. This algorithmic analysis method is straightforward and intuitive; even individual distributors on online retail platforms can effectively conduct customer data analysis using it.

2. Research Framework

2.1 Feature Engineering

The RFM framework—short for Recency, Frequency, and Monetary—serves as a widely adopted approach for evaluating customer value and segmenting a customer base into meaningful groups. More importantly, the RFM method is characterized by its simplicity and intuitiveness, rendering it more efficient and convenient when dealing with static data such as that used in this paper (Feng, Li, Sun, & Zhang, 2019).

Recency (R) measures the elapsed time since a customer' s most recent purchase. Customers with lower recency values have engaged with the business more recently, signaling stronger brand interaction. To quantify this dimension, we construct the feature Days Since Last Purchase, representing the number of days since the last transaction. Smaller values imply recent activity and higher engagement, whereas larger values may indicate customer dormancy. Understanding recency enables businesses to identify disengaged customers and implement targeted campaigns aimed at reactivation, thereby improving retention and loyalty.

Frequency (F) captures how often a customer interacts with the retailer within a defined period. Higher frequency values typically correspond to greater brand loyalty or satisfaction. To represent this dimension, two features are introduced: Total Transactions, reflecting the number of purchases made by a customer, and Total Products Purchased, denoting the total quantity of items acquired across all orders. These metrics collectively provide insight into the intensity of customer interactions and form the basis for distinguishing high-engagement segments from low-engagement ones.

Monetary (M) assesses the total financial contribution of a customer over time, with higher values indicating greater revenue impact and potential lifetime value. Two features are proposed for this dimension: Total Spend, calculated as the sum of the products of unit price and quantity across all transactions, and Average Transaction Value, defined as the total spend divided by the number of transactions for each customer. While the former captures overall revenue contribution, the latter reveals spending patterns on a per-transaction basis, allowing for more precise personalization of promotions and offers.

By integrating these three dimensions, the RFM methodology provides a comprehensive understanding of customer purchasing behavior and preferences, laying the groundwork for both tailored marketing strategies and the development of an effective recommendation system.

2.2 K-Means Clustering

In this study, the K-Means clustering algorithm is employed to perform customer segmentation, with the aim of uncovering distinct characteristics and preferences among different consumer groups. K-Means is a widely recognized unsupervised learning method, whose central principle lies in partitioning the dataset into a predefined number of K clusters through iterative optimization. The algorithm seeks to minimize the Within-Cluster Sum of Squares (WCSS), thereby ensuring high similarity among data points within the same cluster while maintaining significant differences across clusters.

The implementation process consists of several steps. First, the number of clusters K is initialized based on the study objectives, and K centroids are randomly assigned. Next, each data point is allocated to the cluster with the nearest centroid. Following this assignment, the centroids are recalculated as the mean of all points within each cluster. This process is repeated iteratively until the centroids stabilize or a maximum number of iterations is reached, at which point convergence is

assumed. To mitigate the risk of local optima caused by random initialization, multiple runs with different starting centroids are conducted, and the most optimal clustering result is selected.

Determining the appropriate number of clusters is a critical step in achieving meaningful segmentation. An insufficient number of clusters may obscure genuine differences between customers by grouping distinct profiles together, whereas an excessive number of clusters could lead to overfitting and reduced interpretability. To address this challenge, the Elbow Method is applied, where WCSS values are computed across different K values and analyzed to identify the point at which marginal improvements diminish, thus indicating the optimal number of clusters.

By applying K-Means clustering, we are able to detect underlying distinctions among customer groups, providing a robust foundation for the development of recommendation systems and the refinement of targeted marketing strategies. This method not only enhances the interpretability of customer behavior but also equips businesses with a practical tool for achieving differentiation in competitive markets.

2.3 Elbow Method

The Elbow Method provides a systematic approach to determining the optimal number of clusters within a dataset. The process entails repeatedly partitioning the data using the K-Means algorithm across a range of k values. For each candidate k, the algorithm computes the Within-Cluster Sum of Squares (WCSS) — the total squared distance between observations and their respective cluster centroids. Plotting WCSS against k typically produces a curve with a distinct bend or "elbow." This inflection point marks the value of k beyond which additional clusters yield only marginal reductions in WCSS, thereby indicating a suitable balance between model complexity and clustering quality.

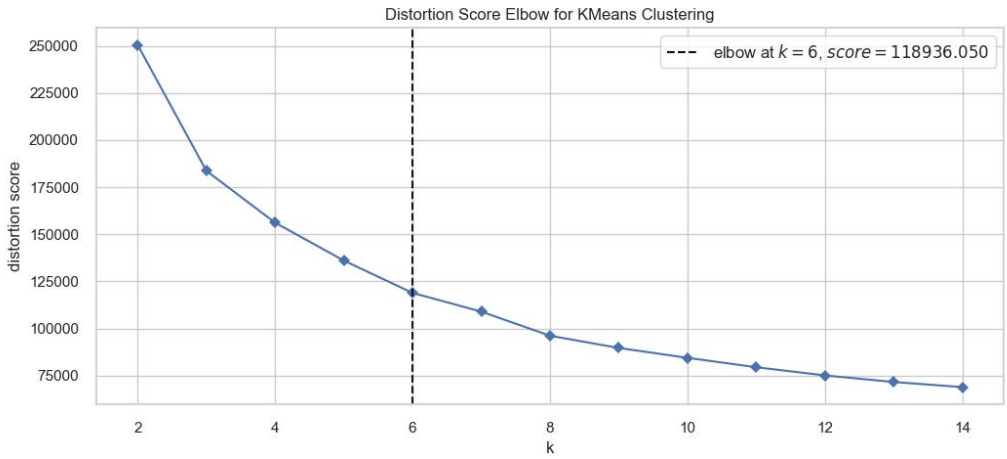

Figure 1. Distortion Score Elbow for K-Means Clustering

This plot is a distortion score elbow graph for K-Means clustering, which is employed to determine the optimal number of clusters k. The horizontal axis denotes the number of clusters, ranging from 2 to 14, while the vertical axis represents the distortion score, which quantifies the sum of squared distances from samples within each cluster to their respective cluster centroids. As observed, the distortion score generally decreases as k increases. A distinct "elbow" pattern emerges at k = 6, with a corresponding distortion score of 118936.050—prior to k = 6, the score decreases rapidly, whereas after k = 6, the rate of decrease slows down significantly. This indicates that selecting k = 6 as the number of clusters achieves a pragmatic balance between intra-cluster compactness and the parsimony of cluster count. It

neither over-simplifies customer segmentation by using too few clusters nor introduces unnecessary complexity with an excessive number of clusters, while ensuring sufficient discriminative power for subsequent customer segmentation analysis.

2.4 Silhouette Method

The Silhouette Method offers a robust approach for selecting the optimal number of clusters by simultaneously evaluating both intra-cluster cohesion and inter-cluster separation. For each observation i, the method computes a silhouette coefficient that quantifies how well the point fits within its assigned cluster relative to other clusters. The process involves three steps:

$$s(i)=\frac{b(i)-a(i)}{max(b(i),a(i))}$$

a(i): Calculate the mean distance between point i and all other points within the same cluster, representing cohesion.

b(i): Determine the mean distance between point i and all points in the nearest neighboring cluster, representing separation.

s(i): Derive the silhouette coefficient for point i using these two measures, where higher values indicate better clustering assignments.

Unlike the Elbow Method, which focuses solely on the within-cluster sum of squares (WCSS) and often relies on subjective interpretation of the “elbow” point, the Silhouette Method provides a direct, quantitative measure of cluster quality. The resulting silhouette score enables objective comparison across different k values while the accompanying silhouette plots offer a visual diagnostic of cluster consistency, highlighting potential irregularities or outliers.

In the subsequent analysis, a candidate range of k = 2 – 6—identified via the Elbow Method—will be assessed by computing average silhouette scores for each value of k. To refine the selection further, silhouette diagrams will be generated to visualize coefficient distributions across clusters, with the YellowBrick library employed to streamline both score computation and graphical evaluation.

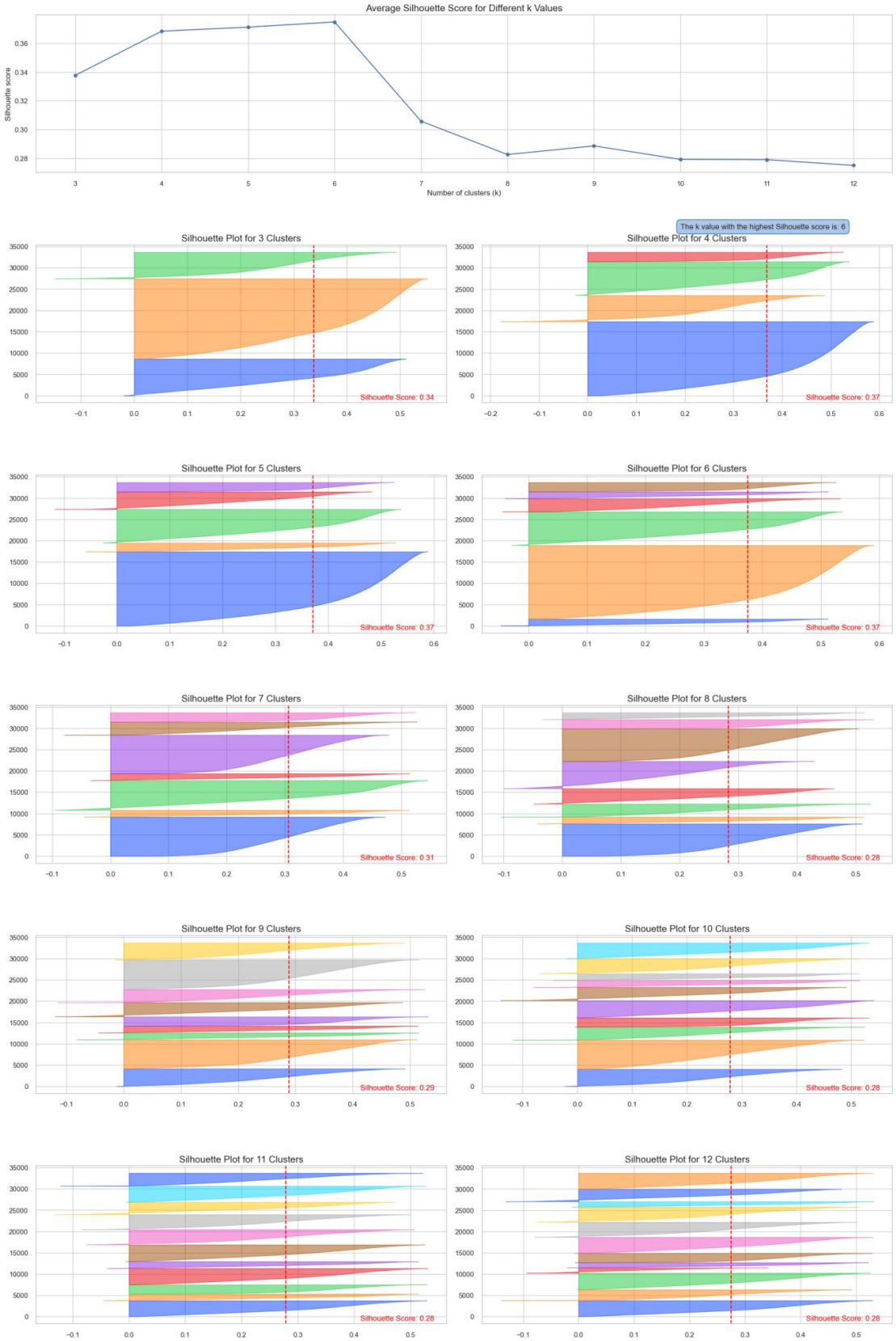

Figure 2. Average Silhouette Score for Different k Values and Silhouette Plot

When interpreting silhouette plots to identify the optimal number of clusters (k), the first aspect to consider is the overall silhouette width. Clusters in which most data points have widths approaching 1

indicate high within-cluster cohesion and strong separation from neighboring clusters, reflecting well-defined group boundaries. In contrast, clusters with narrow or even negative widths suggest ambiguous assignments where many observations lie close to the decision boundary, reducing confidence in the cluster structure. The average silhouette score across all clusters also serves as a global measure of clustering quality, with higher averages generally implying more distinct partitions.

Another important dimension concerns the uniformity of cluster sizes. Ideally, silhouette plots should display relatively consistent cluster thickness across groups, signaling a balanced segmentation where no cluster disproportionately dominates the dataset. Significant variation in width or thickness often indicates structural imbalances, with certain clusters being overly dense and others containing only a small fraction of the data. Such irregularities may compromise the interpretability and stability of the clustering solution, especially when applied in downstream analyses such as customer segmentation or recommendation systems.

The average silhouette score curve further supports the identification of the optimal k. A distinct peak in this curve suggests the number of clusters that best balances within-cluster compactness and between-cluster separation. However, this peak should be interpreted cautiously, as fluctuations in silhouette widths across clusters may also reveal inconsistencies in cluster quality. Stable width patterns across all clusters typically signal homogeneous and well-formed groups, whereas large oscillations suggest that some clusters lack compactness or exhibit considerable overlap with neighboring clusters, undermining the reliability of the segmentation.

Ultimately, selecting the optimal number of clusters requires maximizing the global average silhouette score while simultaneously ensuring that most individual clusters achieve above-average values. This prevents the inclusion of weakly defined or unstable clusters that could otherwise degrade the interpretability and practical utility of the model. Visual inspection of the silhouette plots complements these quantitative criteria, allowing researchers to verify that clusters exhibit consistent boundaries and compact structures, with most data points attaining values close to 1. Clustering configurations meeting these criteria not only achieve statistical rigor but also provide meaningful, actionable groupings for real-world applications such as targeted marketing, personalized recommendations, or behavioral analytics.

3. Data Reading and Processing

3.1 Dataset Overview

The dataset employed in this study consists of anonymized online sales transaction data, focusing on multiple dimensions including product purchases, customer details, and order characteristics in e-commerce or retail scenarios. It is designed to support research directions such as the analysis of sales trends, customer purchase behavior, and order management. Meanwhile, it can be used to explore the impacts of discounts, payment methods, and shipment providers on sales performance and customer satisfaction. Additionally, it facilitates the evaluation of the effects of discounts and payment methods on sales, the optimization of inventory by studying product demand, and the improvement of customer satisfaction through better shipping and return handling.This dataset has a multi-dimensional feature system, covering attributes related to product purchases (e.g., product category, specifications.), customer detail features (e.g., customer demographics, etc.), order features (e.g., order time, order amount, etc.), and business operation features such as discount strategies, payment methods, and shipment provider choices. These features provide rich and effective data support for academic

research on sales performance analysis, customer purchasing pattern research, and operational efficiency evaluation of order management.

First, we will perform a preliminary analysis to understand the structure and types of data columns (Table 1):

Table 1. Data Type of Dataset

Data Columns (total 17 columns)			
#	Column	Non-Null Count	Dtype
0	InvoiceNo	49782 non-null	int64
1	StockCode	49782 non-null	object
2	Description	49782 non-null	object
3	Quantity	49782 non-null	int64
4	InvoiceDate	49782 non-null	object
5	UnitPrice	49782 non-null	float64
6	CustomerID	44804 non-null	float64
7	Country	49782 non-null	object
8	Discount	49782 non-null	float64
9	PaymentMethod	49782 non-null	object
10	ShippingCost	47293 non-null	float64
11	Category	49782 non-null	object
12	SalesChannel	49782 non-null	object
13	ReturnStatus	49782 non-null	object
14	ShipmentProvider	49782 non-null	object
15	WarehouseLocation	46297 non-null	object
16	OrderPriority	49782 non-null	object
dtypes: float64(4), int64(2), object(11)			

The dataset contains 49,782 entries across 17 attributes, each serving a distinct purpose. The InvoiceNo column stores invoice identifiers, with each number potentially representing multiple products purchased in a single order, while StockCode provides the product codes for individual items. Description holds product descriptions with 49,782 non-null valid values. The Quantity column records the number of units bought, and InvoiceDate captures the transaction date and time in object format. UnitPrice indicates the price per unit of the product, whereas CustomerID, despite holding customer identifiers, has 44,804 non-null entries. Country specifies where each transaction took place, with 49,782 non-null values. Additionally, the dataset includes attributes such as Discount (49,782 non-null, float64), PaymentMethod (49,782 non-null, object), ShippingCost (47,293 non-null, float64), Category (49,782 non-null, object), SalesChannel (49,782 non-null, object), ReturnStatus (49,782 non-null, object), ShipmentProvider (49,782 non-null, object), WarehouseLocation (46,297 non-null, object), and OrderPriority (49,782 non-null, object). An initial review highlights missing values in ShippingCost and WarehouseLocation, as well as partial missingness in CustomerID, signaling the need for further data cleaning and imputation. The repeated CustomerID values suggest that customers often make multiple purchases over time. Going forward, additional preprocessing steps will be required to handle missing or erroneous data and to engineer new features that align with the project' s analytical goals.

3.2 Limitations

The dataset employed in this paper dates back quite a long time, and certain types of data essential for modern business operations may not have been recorded. Consequently, the conclusions drawn in this paper might not be fully consistent with the requirements and outcomes of modern e-commerce.

3.3 Summary Statistics

An examination of the dataset uncovers several noteworthy patterns across multiple variables (Table 2). The Quantity attribute records an average of roughly 22.37 items per transaction, with values spanning from -50.00 to 49.00. Negative entries likely correspond to returns or order cancellations, warranting appropriate treatment. The considerable standard deviation (17.917774) and the pronounced gap between the maximum value and the upper quartile further confirm the presence of variation. A comparable situation arises in UnitPrice, where the mean unit cost is about 47.537862, with values ranging from -99.98 to 100.00, implying possible data entry errors or noise, as negative prices are inherently implausible. For CustomerID, the dataset retains 44804 valid entries, leaving a portion missing; the observed IDs range between 10001.00 and 99998.000000, providing a means to differentiate individual customers. Turning to InvoiceNo, it has 49782 entries, with values ranging from 100005 to 999997. For categorical variables, StockCode encompasses 1000 unique product codes, the most frequent being SKU_1761 with 76 occurrences. Similarly, Description includes 11 unique product labels, led by "Wall Clock," which appears 4617 times. InvoiceDate has 49782 unique entries, with each transaction time being distinct as shown by "2020-01-01 00:00" occurring only once. The Country attribute documents transactions from 12 nations, with France being the most frequent with 4230 occurrences. PaymentMethod has 3 unique types, with Bank Transfer being the top with 16747 instances. Category includes 5 unique types, led by "Furniture" with 10084 occurrences. SalesChannel has 2 unique types, with Online being the most frequent at 25051. ReturnStatus has 2 types, with "Not Returned" being dominant at 44888. ShipmentProvider has 4 unique providers, with FedEx being the top at 12501. WarehouseLocation has 5 unique locations (with 46297 valid entries), led by Amsterdam with 9458 occurrences. OrderPriority has 3 types, with Medium being the most frequent at 16678. (Table 3)

Table 2. Data Count of Dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
InvoiceNo	49782	550681.2399	260703.0099	100005	324543	552244	776364	999997
Quantity	49782	22.372343	17.917774	-50	11	23	37	49
UnitPrice	49782	47.537862	33.47951	-99.98	23.5925	48.92	74.61	100
CustomerID	44804	55032.87178	25913.66016	10001	32750.75	55165	77306.25	99998
Discount	49782	0.275748	0.230077	0	0.13	0.26	0.38	1.999764
ShippingCost	47293	17.494529	7.220557	5	11.22	17.5	23.72	30

Table 3. Summary Statistics of Dataset

	Count	Unique	Top	Freq
StockCode	49782	1000	SKU_1761	76
Description	49782	11	Wall Clock	4617
InvoiceDate	49782	49782	2020/1/1 0:00	1
Country	49782	12	France	4230
PaymentMethod	49782	3	Bank Transfer	16747
Category	49782	5	Furniture	10084
SalesChannel	49782	2	Online	25051
ReturnStatus	49782	2	Not Returned	44888
ShipmentProvider	49782	4	FedEx	12501
WarehouseLocation	46297	5	Amsterdam	9458
OrderPriority	49782	3	Medium	16678

3.4 Handling Missing Value

We will determine the percentage of missing values present in each column, followed by selecting the most effective strategy to address them (Figure 3):

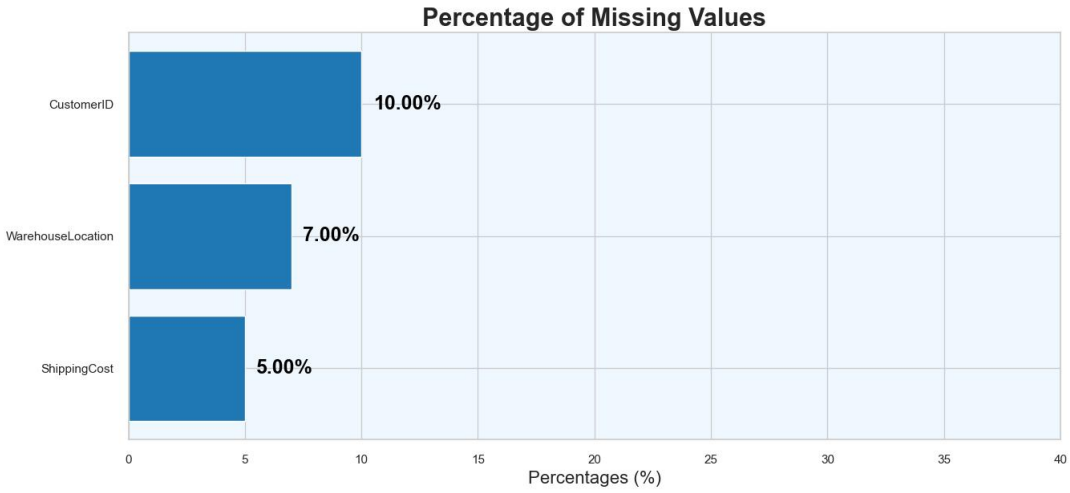

Figure 3. Percentage of Missing Values

The CustomerID attribute exhibits 10.00% missing values. Since this variable is central to both customer clustering and the development of a recommendation system, although imputing these missing identifiers requires meticulous consideration, excluding transactions with missing CustomerID entries emerges as a viable strategy to preserve the integrity of the clustering process. Meanwhile, WarehouseLocation has 7.00% missing values and ShippingCost has 5.00% missing values. For these attributes, the choice of handling methods should be guided by their specific roles in analytical tasks.

Notably, CustomerID is pivotal for understanding individual customer behavior and preferences, so removing its missing entries helps ensure the accuracy of customer-centric analyses. For WarehouseLocation and ShippingCost, if they are critical to supply chain or cost-related analyses, targeted imputation methods (such as mode imputation for WarehouseLocation or mean/median imputation for ShippingCost) could be explored, while also assessing the potential impact of such imputations on data integrity.

By addressing the missing values in CustomerID, WarehouseLocation, and ShippingCost through tailored strategies, we aim to construct a cleaner and more reliable dataset. This refined dataset will support not only accurate clustering and recommendation system development but also broader analytical endeavors, such as evaluating the efficiency of warehouse operations or the impact of shipping costs on customer behavior.

3.5 Correcting Stockcode Anomalies

In order to rectify irregularities within the StockCode attribute, the first step involves quantifying the total distinct codes and examining the ten most recurrent ones together with their relative frequencies. The dataset encompasses 1000 unique stock identifiers, indicating a diverse range of merchandise offered in this e-commerce context. This diversity creates opportunities to identify heterogeneous customer segments with distinct product preferences. A closer look at the most recurrent stock identifiers can reveal popular items or dominant product categories that resonate with consumers.

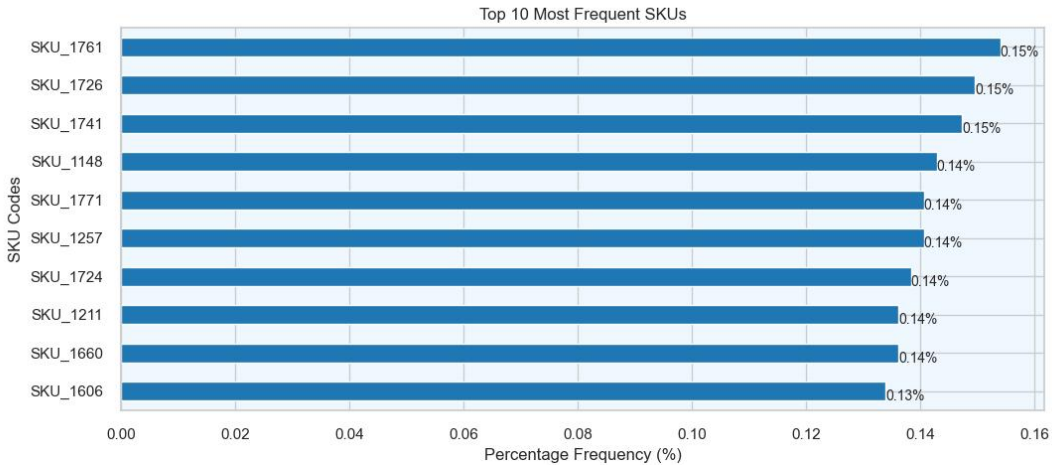

Figure 4. Top 10 Most Frequent SKUs

Among the top 10 most frequent SKUs, SKU_1761, SKU_1726, and SKU_1741 each have a relative frequency of 0.15%, while SKU_1148, SKU_1771, SKU_1257, SKU_1724, SKU_1211, and SKU_1660 each account for 0.14%, and SKU_1600 has a frequency of 0.13%. Despite this structured distribution, potential deviations may exist in the StockCode variable. For instance, some codes might not correspond to tangible products but rather to services or other non-merchandise items. Given that the project aims to segment customers based on genuine purchasing behaviors and build a product recommendation framework, retaining such anomalous entries could introduce extraneous noise into the analytical process.

Thus, filtering out these irregular records ensures that the analysis focuses solely on legitimate product transactions, leading to more accurate customer clustering and a more robust recommendation system. Moreover, understanding the frequency distribution of these top SKUs can also help in inventory management, identifying which products drive the most transactions and should be prioritized in supply chain and marketing strategies.

3.6 Product Diversity

This stage focuses on examining the heterogeneity in customers' purchasing patterns, as recognizing variations in product selection is essential for designing personalized marketing campaigns and tailored

recommendation systems. To capture this dimension, we introduce the feature Unique Products Purchased, defined as the total count of distinct items acquired by each customer. A higher value for this metric reflects broader interests or more diverse purchasing preferences, whereas a lower value implies a narrower or more specialized focus on certain products. By quantifying purchase diversity, businesses can classify customers according to the breadth of their buying behavior, thereby enabling the development of recommendation strategies that align more closely with individual shopping profiles.

3.7 Behavioral Features

This phase concentrates on uncovering the temporal dynamics of customer purchasing behavior to better understand when individuals are most inclined to engage with the retailer. Such insights can significantly enhance the personalization of marketing strategies and customer experiences. Three new behavioral metrics are proposed. Average Days Between Purchases captures the mean interval between successive transactions for each customer, providing a foundation for forecasting the timing of future purchases and enabling precisely timed promotional efforts. Preferred Shopping Day identifies the weekday on which a customer conducts the highest number of transactions, thereby informing day-specific marketing initiatives for distinct customer groups. Likewise, Preferred Shopping Hour isolates the hour of the day when customer activity peaks, allowing campaigns to be scheduled during periods of maximum engagement. By integrating these temporal attributes, the dataset gains behavioral depth, improving the clustering algorithm' s capacity to generate well-defined and practically meaningful customer segments.

3.8 Seasonality and Trends

This phase focuses on exploring seasonality and longitudinal patterns in customer purchasing behavior to uncover insights that can inform precision marketing strategies and improve customer engagement. Three features are proposed. Monthly_Spending_Mean captures the average monthly expenditure of each customer, with higher values typically associated with individuals inclined toward premium products, while lower values may signal cost-conscious buyers. Monthly_Spending_Std quantifies the month-to-month variability in spending; elevated standard deviations suggest irregular purchasing patterns with occasional large transactions, whereas smaller values indicate steady and predictable spending behavior. Finally, Spending_Trend measures the temporal direction of customer expenditure by fitting a linear regression line to monthly spending data. A positive slope implies rising expenditure—often reflecting increasing loyalty or satisfaction—whereas a negative slope points to declining engagement, and a near-zero slope signals spending stability. Integrating these temporal and behavioral metrics into the segmentation model allows for the identification of nuanced customer segments, enabling marketing strategies tailored to specific spending trajectories and seasonal behaviors.

3.9 Customer Dataset Description

Through the data processing described in the above chapters, the overall description of the customer dataset adopted in this study is presented in the following table(Table 4).

Table 4. Customer Dataset Description

#	Column	Non-Null Count	Dtype
0	CustomerID	35389 non-null	object
1	Days_Since_Last_Purchase	35389 non-null	int64
2	Total_Transactions	35389 non-null	int64
3	Total_Products_Purchased	35389 non-null	int64
4	Total_Spend	35389 non-null	float64
5	Average_Transaction_Value	35389 non-null	float64
6	Unique_Products_Purchased	35389 non-null	int64
7	Average_Days_Between_Purchases	35389 non-null	float64
8	Day_Of_Week	35389 non-null	date
9	Hour	35389 non-null	int32
10	Is_UK	35389 non-null	int64
11	Cancellation_Frequency	35389 non-null	float64
12	Cancellation_Rate	35389 non-null	float64
13	Monthly_Spending_Mean	35389 non-null	float64
14	Monthly_Spending_Std	35389 non-null	float64
15	Spending_Trend	35389 non-null	float64
16	Prefers_Online	35389 non-null	int64
17	PaymentMethod	35389 non-null	object

4. Machine Learning Framework

4.1 Outlier Detection and Treatment

There we concentrate on the detection and mitigation of outliers within the dataset—observations that diverge markedly from typical data patterns and risk introducing bias into subsequent analyses. Such anomalies are especially problematic for analytical techniques, as extreme values can disproportionately distort results and degrade the quality of insights. To address this, an algorithm suited for anomaly detection is employed. Here, a method like Isolation Forest can be adopted due to its computational efficiency and ability to isolate irregular instances by recursively partitioning the feature space. Each observation receives an assessment, enabling systematic identification of aberrant records. The results are encapsulated in a dedicated variable (e.g., Is_Outlier), facilitating both quantitative assessment and graphical exploration. Subsequent visualizations of inlier – outlier proportions provide a clearer understanding of the anomalies' prevalence and potential impact on the dataset.

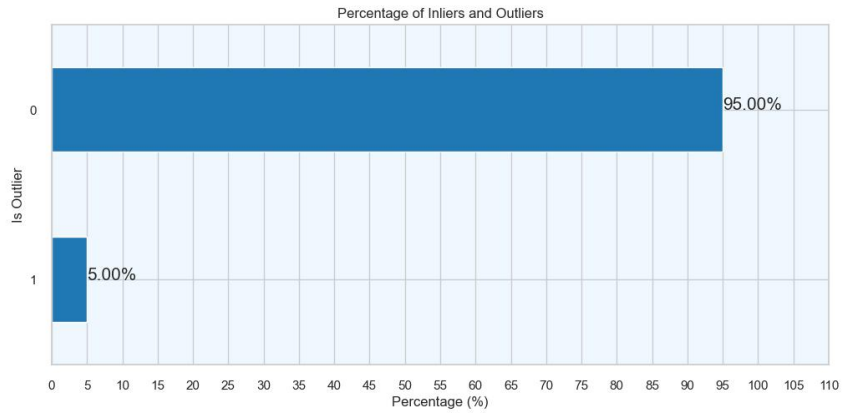

Figure 5. Percentage of Inliers and Outliers

Analysis of the visualization indicates that 5.00% of the observations have been flagged as outliers, while 95.00% are inliers. This proportion shows that outliers constitute a moderate share of the dataset—large enough to reflect genuinely anomalous data points yet sufficiently limited to avoid excessive data loss. Such detection is essential for enhancing the robustness of subsequent analyses, whether it is clustering, predictive modeling, or descriptive analytics. Given the need for accurate and unbiased results in analytical endeavors, mitigating the influence of these anomalies is imperative. Consequently, the identified outliers can be isolated for potential auxiliary analysis and stored separately if required, while being excluded from the primary dataset to preserve analytical integrity. Finally, any auxiliary variables generated solely for anomaly detection purposes can be removed to streamline the dataset prior to subsequent analytical steps.

4.2 Correlation Analysis

Prior to implementing K-Means clustering, it is necessary to examine inter-feature correlations within the dataset. High correlations—commonly referred to as multicollinearity—can obscure the true structure of the data by introducing redundant information, thereby hindering the model' s ability to identify distinct and meaningful clusters. Excessive multicollinearity may ultimately lead to poorly separated segments and reduced interpretability.

Should such correlations be detected, dimensionality reduction methods such as Principal Component Analysis (PCA) can be applied. By transforming correlated variables into a smaller set of orthogonal components while retaining most of the variance, these techniques mitigate the impact of multicollinearity, enhance computational efficiency, and improve the overall quality and stability of the resulting clusters.

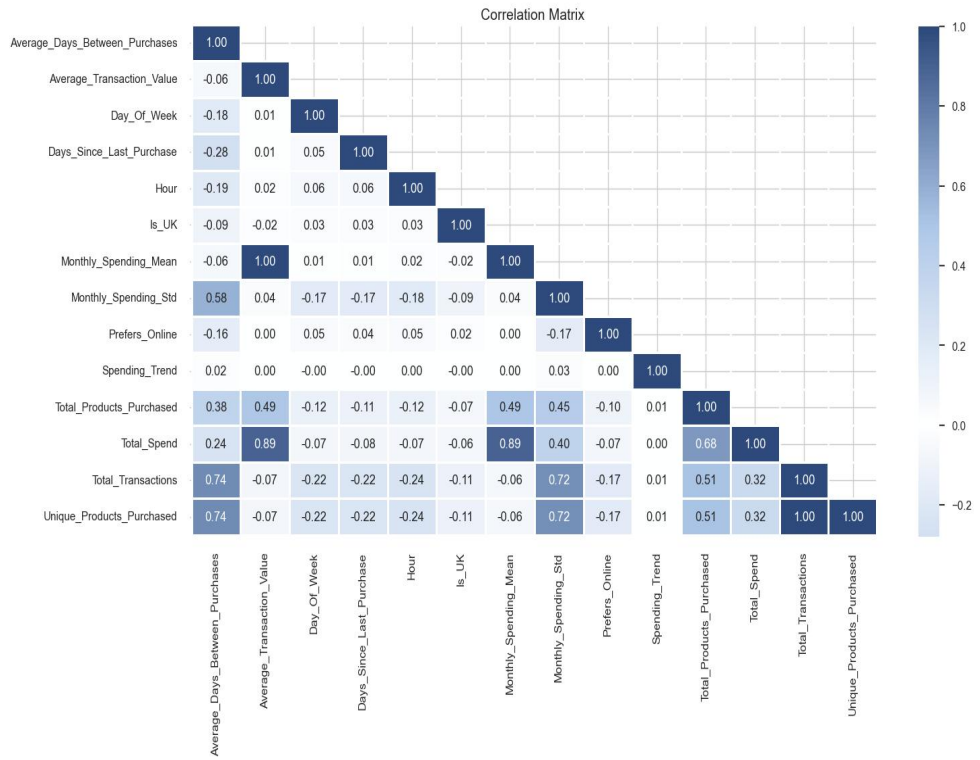

Figure 6. Correlation Matrix

This figure reports the pairwise correlations among the engineered features. A clear "purchase intensity" block emerges: Total_Transactions, Unique_Products_Purchased, Total_Products_Purchased, Total_Spend, Monthly_Spending_Mean, and Monthly_Spending_Std are all positively associated (roughly 0.45 – 0.89). For instance, Total_Products_Purchased correlates strongly with Total_Spend (~0.69) and with Monthly_Spending_Mean (~0.49), indicating that customers with higher product purchase volumes also generate higher total spend and have correspondingly higher monthly spending averages, with notable variability in month-to-month spending as well. In contrast, Days_Since_Last_Purchase is negatively related to some activity measures (e.g., ~-0.11 with Total_Products_Purchased, ~-0.07 with Total_Spend), and shows a negative link with Monthly_Spending_Std (~-0.17), while Average_Days_Between_Purchases exhibits strong positive correlations with Total_Transactions (0.73) and Unique_Products_Purchased (0.73), suggesting that customers with more time between purchases might actually have higher transaction counts and purchase more unique products, a nuance that warrants further exploration.

Temporal and contextual attributes—including Day_Of_Week, Hour, and Is_UK—display near-zero correlations with the main spending and frequency metrics (e.g., Day_Of_Week with Total_Spend is ~-0.08, Hour with Total_Spend is ~-0.07, Is_UK with Total_Spend is ~-0.06), indicating limited linear association in this dataset. This does not rule out nonlinear or interaction effects; rather, it suggests such variables are unlikely to drive segmentation on their own without additional modeling structure. Meanwhile, Monthly_Spending_Std has a moderate positive correlation with Average_Days_Between_Purchases (0.58), implying that customers with more days between purchases may have higher variability in their monthly spending. Prefers_Online shows a negative correlation with Monthly_Spending_Std (~-0.17), suggesting that customers who prefer online channels might have less variable monthly spending.

Methodologically, these patterns have multiple implications. First, the strong positive associations within the purchase intensity block (e.g., Total_Transactions and Unique_Products_Purchased both at 1.00, Total_Spend and Monthly_Spending_Mean at 0.89, Total_Products_Purchased and Monthly_Spending_Mean at 0.49) signal potential multicollinearity. For supervised modeling, one should either (i) apply regularization (e.g., L1/L2), (ii) reduce dimensionality (e.g., PCA), or (iii) retain a minimal, nonredundant subset (e.g., use Total_Spend or Monthly_Spending_Mean, but not both, or choose between Total_Transactions and Unique_Products_Purchased). Second, for clustering methods such as K-means, it is advisable to standardize features and consider down-weighting or removing near-duplicates (like Total_Transactions and Unique_Products_Purchased) to prevent the distance metric from being dominated by redundant constructs related to purchase frequency and product variety. Third, the nuanced correlations of temporal variables and spending variability suggest that while linear relationships are weak, there might be nonlinear patterns or interactions that could be explored in more advanced modeling.

From a business perspective, the matrix validates segmentation axes centered on purchase intensity and spending patterns. Customers with high Total_Transactions, Unique_Products_Purchased, and Total_Spend form a coherent high-value profile, whereas those with longer Days_Since_Last_Purchase show mixed engagement signals (negative with some volume metrics but positive with transaction counts). Because temporal markers (day of week, hour) are weakly correlated with spend, campaign timing alone is unlikely to drive significant outcomes without targeting the underlying intensity cohorts. The correlation between Average_Days_Between_Purchases and high transaction counts is a counterintuitive insight, suggesting that "less frequent" buyers might actually be heavy transactors

over time, which could reshape how customer loyalty and engagement are defined. Additionally, the link between Prefers_Online and lower spending variability hints that online-preferring customers might be more predictable in their spending, opening avenues for tailored marketing. Overall, the correlation structure supports a segmentation strategy anchored in multi-dimensional engagement (frequency, product variety, spend) and spending variability, provides a principled basis for feature selection in downstream modeling, and uncovers unexpected relationships that can refine customer understanding.

4.3 Feature Scaling and Dimensionality Reduction

Before proceeding with clustering and dimensionality reduction, it is essential to standardize the features to ensure comparability across variables. This step is particularly important for distance-based algorithms such as K-means and dimensionality reduction methods like Principal Component Analysis (PCA), both of which rely on the relative scale of input features.

For K-means, the algorithm partitions observations based on distances between data points. Without scaling, attributes with larger numeric ranges can dominate the clustering process, skewing the resulting groupings and obscuring true behavioral patterns. Similarly, PCA seeks to identify principal directions of maximum variance in the data. When features vary in magnitude, variables with inherently larger scales can overwhelm the principal components, masking underlying relationships.

To mitigate these risks, all continuous features are standardized to zero mean and unit variance, while three variables are excluded for specific reasons: CustomerID serves purely as an identifier with no analytical value, Is_UK is binary and unaffected by scaling, and Day_Of_Week is categorical, rendering standardization unnecessary.

Given the previously observed multicollinearity, dimensionality reduction is introduced for several reasons:

(1)Alleviating Redundancy: Transforming correlated variables into a reduced set of orthogonal components minimizes information overlap.

(2)Improving Cluster Quality: Lower-dimensional representations help K-means uncover more compact and interpretable clusters.

(3)Noise Reduction: Retaining only dominant components filters out irrelevant variability, stabilizing the segmentation process.

(4)Enhanced Visualization: Mapping data into two or three principal dimensions supports intuitive interpretation of customer groups.

(5)Computational Efficiency: Reducing dimensionality decreases processing time, especially beneficial for iterative algorithms.

Among various techniques—KernelPCA, ICA, ISOMAP, t-SNE, and UMAP—PCA serves as the initial choice due to its efficiency and strong performance in capturing linear structures, particularly valuable in datasets affected by multicollinearity. PCA reduces dimensionality while preserving most of the data' s variance, producing more interpretable and computationally tractable clustering outcomes. Nevertheless, if the leading principal components fail to explain sufficient variance, alternative nonlinear methods may be explored to capture more intricate data relationships, albeit at the expense of higher computational complexity.

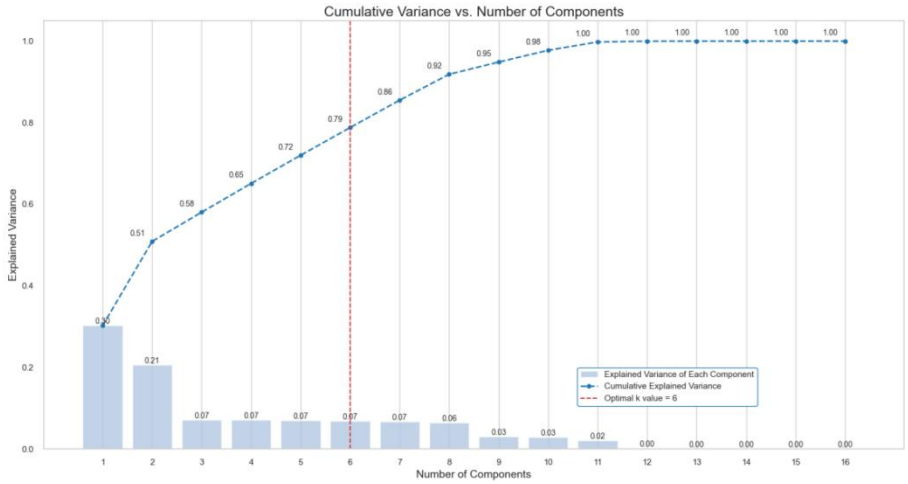

Figure 7. Cumulative Variance vs. Number of Components

The scree plot and cumulative variance curve illustrate the proportion of total variance accounted for by each principal component as well as the cumulative variance explained when multiple components are considered jointly. The results reveal that the first principal component captures roughly 30% of the variance, the first two components together account for about 51% (30% + 21%), the first three explain approximately 58% (51% + 7%), the first four reach 65% (58% + 7%), the first five hit 72% (65% + 7%), and the first six components collectively explain around 79% (72% + 7%). Beyond the sixth component, additional components continue to contribute to variance explanation but at a drastically diminishing rate, which is the characteristic "elbow point" in the curve.（Table 5）

Table 5. Customer Characteristics

CustomerID	PC1	PC2	PC3	PC4	PC5	PC6
10001	-0.166675	0.921743	0.017017	-0.562854	1.218084	0.047253
10003	-2.235664	-0.741391	0.13135	-0.575846	-0.298576	-0.875755
10005	3.662566	-0.866698	1.78398	-0.320558	1.133507	-1.110316
10008	-2.084912	-1.458041	0.144024	-0.865431	1.86907	1.013964
10009	-1.257699	0.052756	-0.036009	-0.431694	0.222663	-0.323656

In this case, the curve shows that the first six components together explain 79% of the total variance, and by the eleventh component, the cumulative variance approaches 1.00. For customer segmentation, retaining sufficient information to differentiate customer groups is critical while avoiding unnecessary dimensionality. Thus, preserving the first six components (as indicated by the optimal k value = 6 in the plot) offers a pragmatic balance, maintaining most of the dataset' s informational richness (nearly 80% of total variance) while facilitating more efficient and interpretable clustering. This selection also aligns with the need to reduce multicollinearity among features, as PCA transforms the original correlated variables into uncorrelated principal components, ensuring that each subsequent component captures unique variance. From the PCA-transformed data sample, we can observe that each customer (identified by CustomerID) has distinct values across PC1 to PC6, indicating that these components effectively capture heterogeneous customer characteristics, laying a solid foundation for subsequent clustering analysis.

5. Model Result and Discussion

After identifying the optimal number of clusters — three in this case — through both elbow and silhouette analyses, the next phase focuses on evaluating the quality and interpretability of the resulting clusters. This step is essential to confirm that the segmentation process has produced coherent, well-separated groups that reflect meaningful patterns in the data rather than arbitrary partitions.

To accomplish this, several complementary approaches are employed. First, a three-dimensional visualization based on the top principal components provides an intuitive representation of the cluster structure, allowing for direct observation of how distinct and compact the groups appear in reduced-dimensional space. Second, a cluster distribution plot is used to illustrate the relative size and density of each cluster, enabling the identification of any imbalances or anomalies across the groups.

Finally, multiple quantitative evaluation metrics are calculated to rigorously assess cluster quality. The Silhouette Score measures both cohesion within clusters and separation between clusters, with higher values indicating more robust partitions. The Calinski – Harabasz Index evaluates the ratio of between-cluster dispersion to within-cluster dispersion, rewarding configurations with dense, well-separated clusters. Conversely, the Davies – Bouldin Index penalizes clusters with high intra-cluster similarity relative to inter-cluster separation, where lower scores signal better-defined structures. Together, these metrics provide a comprehensive assessment of clustering performance, ensuring that the chosen segmentation aligns with both statistical validity and practical interpretability.

5.1 3D Visualization of Customer Clusters

In this stage, the three principal components explaining the highest proportion of variance are selected to construct a three-dimensional visualization. This representation provides an intuitive means of examining the spatial distribution of observations, enabling a preliminary assessment of both inter-cluster separation and intra-cluster cohesion. By projecting the data onto these principal axes, we can visually verify whether the clusters identified through the algorithm exhibit clear boundaries and structural compactness, offering additional insights beyond purely quantitative metrics.

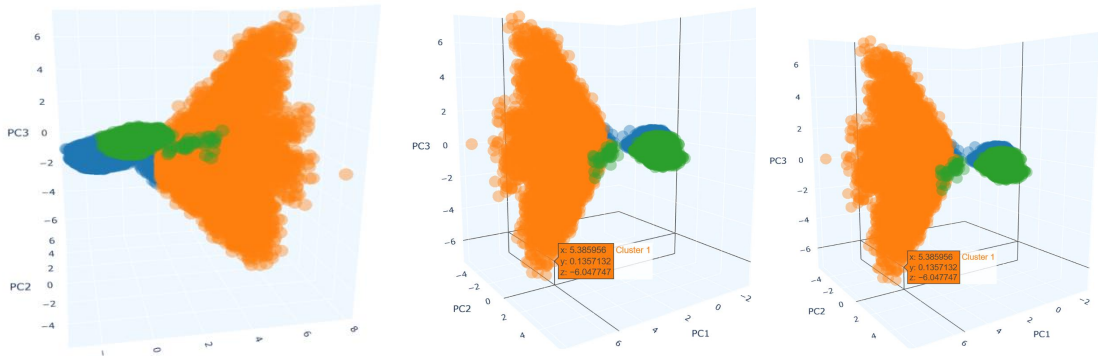

Figure 8. 3D Visualization of Customer Clusters in PCA Space

5.2 Cluster Distribution Visualization

Subsequently, a bar chart will be employed to depict the proportion of customers assigned to each cluster. This visualization facilitates an intuitive assessment of whether the clusters are balanced in size and sufficiently representative, ensuring that no single cluster dominates the dataset while others remain underpopulated. Such analysis is critical for evaluating the practical significance and interpretability of the clustering results.

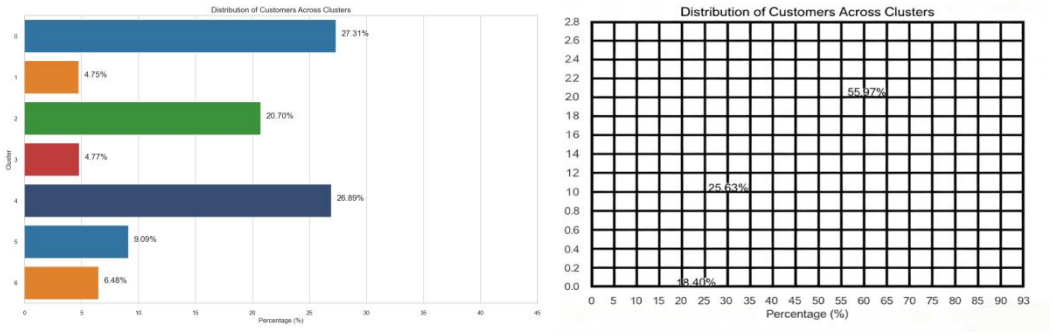

Figure 9. Distribution of Customers Across Clusters

The bar chart illustrating customer distribution across clusters reveals a varied segmentation. Cluster 0 accounts for 27.58%, cluster 3 makes up 27.34%, cluster 2 comprises 20.66%, cluster 4 constitutes 11.27%, while clusters 1 and 5 have relatively smaller proportions, at 6.59% and 6.56% respectively. This distribution indicates that the clustering algorithm has identified segments with distinct scales, where clusters 0 and 3 are the largest, followed by cluster 2, then cluster 4, and clusters 1 and 5 are the smallest. Such a structure suggests that the algorithm has captured diverse customer behavioral patterns, with each cluster representing a segment of the population that varies in size but all hold actionable insights for strategic business initiatives.

Importantly, while clusters 1 and 5 have smaller proportions, they are not negligible, and no cluster consists of only a tiny fraction of customers. This confirms that each segment reflects meaningful behavioral or characteristic differences rather than anomalous or unrepresentative data. Thus, each cluster carries interpretive value, enabling granular analyses of customer behavior within each group. This supports the design of targeted, data-driven decision-making processes across marketing (such as tailored promotions for large clusters and niche strategies for smaller ones), product development (aligning offerings with the preferences of each cluster), and customer relationship management (customizing engagement strategies based on cluster-specific traits).

5.3 Evaluation Metrics

To conduct a more rigorous assessment of clustering quality, three complementary performance metrics are employed. The Silhouette Score quantifies the degree of separation between clusters, with values approaching 1 indicating well-separated and internally cohesive groups, while values near 0 suggest overlapping boundaries, and negative values imply possible misclassifications. The Calinski‐Harabasz Index measures the ratio of between-cluster dispersion to within-cluster cohesion, rewarding configurations with compact and distinctly separated clusters through higher scores. In contrast, the Davies‐Bouldin Index captures the average similarity between each cluster and its most comparable neighbor, where lower values reflect stronger partitioning and reduced inter-cluster overlap. Together, these metrics offer a comprehensive framework for evaluating both the compactness and the distinctiveness of the clusters obtained.（Table 6）

Table 6. Evaluation Metrics

Metric	Value
Number of Observations	33619
Silhouette Score	0.305805573
Calinski Harabasz Score	14082.37955
Davies Bouldin Score	1.145511079

The Silhouette Score of 0.3058055731282483, while not reaching the ideal value of 1, still reflects a reasonable level of separation between clusters. This result suggests that although the clusters are largely distinguishable, there might be some overlaps at the boundaries. Ideally, higher values would indicate more compact and well-separated clusters, but this score is acceptable for exploratory segmentation, implying that the clusters have a moderate degree of internal cohesion and inter-cluster distinction.

The Calinski–Harabasz Index, with a value of 14082.379553160692, is relatively high, signifying that the identified clusters exhibit strong internal cohesion and substantial inter-cluster separation. As this metric generally increases with better-defined partitions, the result implies that the clustering procedure has successfully captured meaningful structural patterns within the dataset, demonstrating that the clusters are well-differentiated in terms of the variance within and between them.

Similarly, the Davies–Bouldin Index yields a score of 1.1455110790046572, indicating a moderate degree of similarity among the most comparable clusters. Since lower values correspond to clearer separation, this outcome suggests a satisfactory level of distinctiveness across groups, though there is room for improvement.

Taken together, these metrics collectively demonstrate that the clustering solution achieves a solid balance between interpretability and separation quality. The Silhouette Score indicates moderate cluster separation, the Calinski–Harabasz Index highlights strong cluster distinctiveness, and the Davies–Bouldin Index shows reasonable cluster uniqueness. Nevertheless, there remains potential for further refinement, perhaps through the application of alternative clustering strategies or dimensionality reduction techniques to enhance segmentation precision and boundary clarity, so as to make the clusters more compact and well-separated.

5.4 Cluster Analysis and Profiling

This section focuses on conducting a detailed characterization of each cluster to uncover the behavioral patterns and preference profiles that distinguish one customer segment from another. By systematically examining the defining attributes of each group, the analysis aims to construct comprehensive cluster profiles that capture the key traits, purchasing tendencies, and engagement levels unique to the customers within each segment. Such profiling not only facilitates a deeper understanding of customer heterogeneity but also provides actionable insights for targeted marketing and strategic decision-making.

5.4.1 Radar Chart Approach

The analysis begins with the construction of radar charts to visually compare the centroid values of each cluster across multiple features. These charts provide an intuitive overview of the distinctive profiles characterizing different customer segments. To generate them, the first step involves calculating the centroid for every cluster, representing the mean value of all variables within that group. Once computed, these centroids are plotted on the radar charts, enabling a direct visual assessment of

the central tendencies across features and highlighting the contrasts among clusters in a clear and interpretable manner.

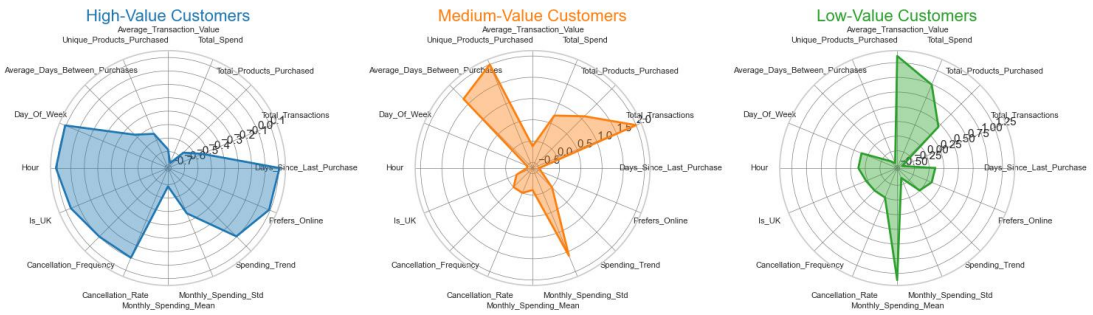

Figure 10. Distribution of Radar Chart

Customer Profiles Inferred from Radar Chart Analysis

Cluster 0 (Red Chart): Sporadic Shoppers with Weekend Preference

Customers assigned to this segment exhibit relatively low engagement levels, reflected in fewer transactions, limited product diversity, and reduced overall spending. Their shopping activity shows a slight inclination toward weekends, as suggested by elevated Day_of_Week values. Spending patterns remain stable yet modest, accompanied by minimal monthly variability (Monthly_Spending_Std). Cancellation frequency and rates are notably low, implying fewer order reversals. Furthermore, their average transaction value is modest, indicating that individual purchases tend to be small in scale.

Cluster 1 (Green Chart): Infrequent but Growing High-Value Shoppers

This cluster comprises customers who purchase less frequently, as evidenced by longer Days_Since_Last_Purchase intervals and higher Average_Days_Between_Purchases values. However, when they do shop, they exhibit a strong upward spending trend, signaling increasing financial engagement over time. They typically prefer late-day transactions and are predominantly UK-based. Their cancellation activity remains moderate, while the average transaction value is comparatively high, suggesting fewer but more substantial purchases per order.

Cluster 2 (Blue Chart): Frequent High Spenders with Elevated Cancellation Rates

Customers in this group demonstrate intense purchasing activity, characterized by frequent transactions, considerable total expenditure, and substantial product variety. However, they also exhibit disproportionately high cancellation frequencies and rates, pointing to less stable shopping behaviors. Purchases occur at shorter intervals, often earlier in the day, and monthly spending variability is pronounced, reflecting irregular spending habits. Despite their historically high spending, a declining Spending_Trend indicates that expenditure levels may be tapering off over time.

5.4.2 Histogram Chart Approach

To corroborate the customer profiles derived from the radar chart analysis, histograms will be generated for each feature, segmented according to cluster assignments. These visualizations facilitate a comparative examination of feature distributions across clusters, enabling the verification and potential refinement of the behavioral patterns inferred from the radar charts. By contrasting the frequency distributions, we can assess whether the cluster-specific tendencies observed earlier hold consistently across the underlying data, thereby strengthening the validity and interpretability of the identified customer segments.

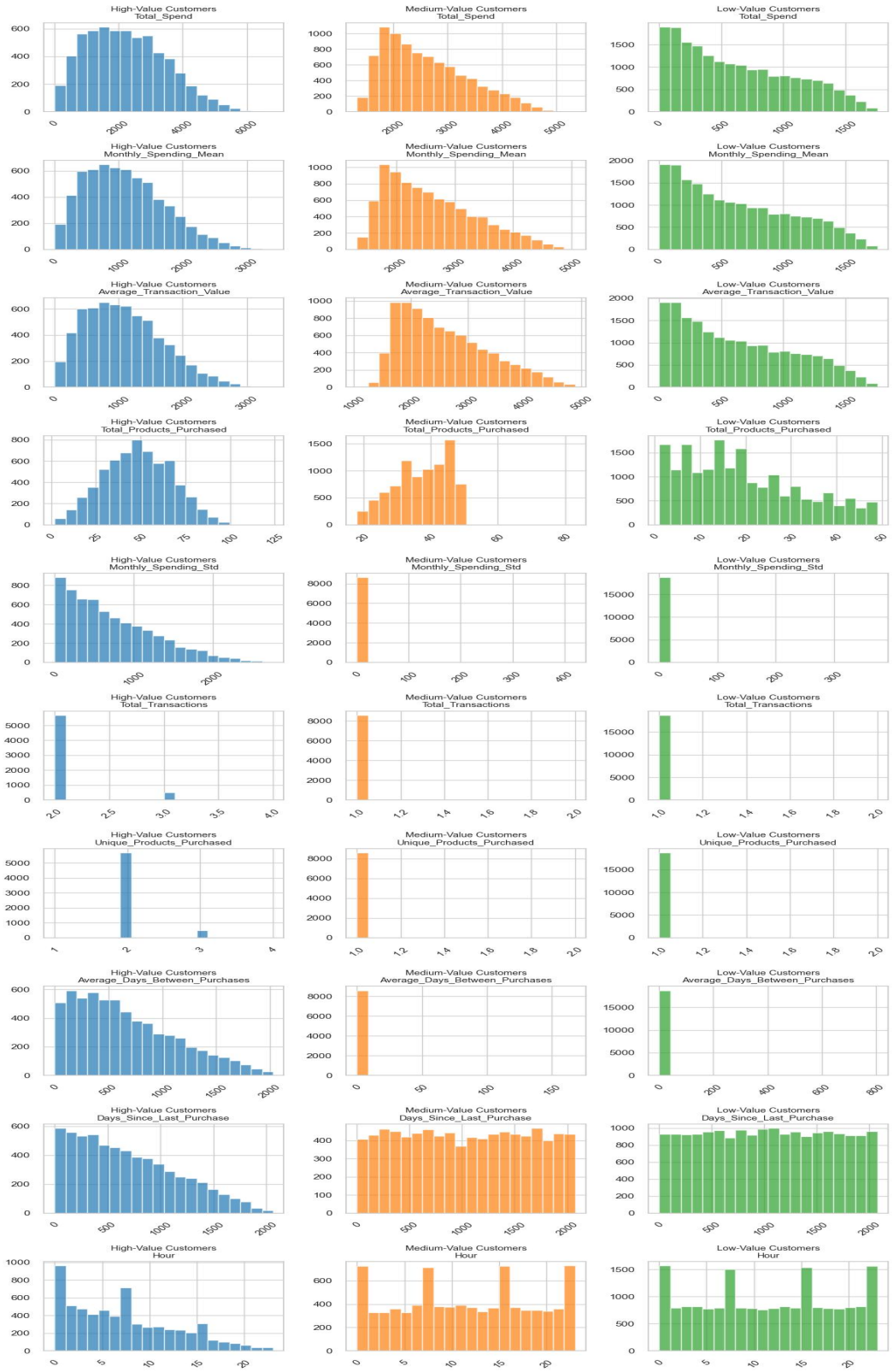

Figure 11. Feature Distribution Histograms By Cluster

The insights gleaned from the histogram analysis provide a more granular perspective on customer behavior, enabling the refinement of cluster profiles derived from both radar charts and distributional patterns. Integrating these findings yields the following revised characterizations for each cluster:

Cluster 0 – Casual Weekend Shoppers

Customers in this segment display low engagement levels, shopping infrequently and spending significantly less than other groups. They typically make fewer purchases per transaction and show a clear preference for weekend shopping, often indicative of casual or exploratory buying habits rather than deliberate, high-value transactions. Their spending behavior remains relatively stable over time, with minimal fluctuations in monthly expenditure, and they exhibit a low incidence of order cancellations. When they do make purchases, the monetary value per transaction tends to be modest, underscoring their restrained spending patterns.

Cluster 1 – Occasional Big Spenders

This cluster consists of customers who shop less frequently but allocate considerable budgets when they do engage in purchasing, often acquiring a diverse set of products. Their expenditure patterns exhibit a positive trend, suggesting increasing financial involvement over time. These customers predominantly shop during late hours—possibly post-work—and are mainly located in the UK. Their moderate cancellation rates might reflect the deliberative nature of high-value transactions, where reconsideration occurs more often. The consistently large transaction values suggest a preference for premium or higher-quality goods, differentiating them from lower-spending groups.

Cluster 2 – Eager Early-Bird Shoppers

Customers within this cluster demonstrate high overall spending levels, frequent purchasing activity, and substantial product diversity. However, they also exhibit elevated cancellation frequencies, possibly linked to impulsive or rapidly changing purchasing decisions. Their preference for early-day transactions suggests time availability before routine obligations or an inclination toward limited-time promotions. Spending variability within this segment is pronounced, with significant month-to-month fluctuations reflecting less predictable financial behaviors. Notably, the declining spending trend may indicate a potential shift in future engagement levels, warranting closer observation for emerging behavioral changes.

5.5 Recommendation System

In the concluding stage of this project, a recommendation system will be designed to elevate the online shopping experience by delivering cluster-driven product suggestions. Drawing on the customer segmentation results, the system will tailor recommendations according to the purchasing behaviors characteristic of each cluster. Earlier in the data preprocessing phase, approximately 5% of customers identified as outliers were separated into a dedicated dataset (outliers_data), ensuring that the core analysis focuses on the primary 95% of the customer base.

For this main group, the cleansed dataset will be analyzed to identify top-selling products within each cluster, capturing the most frequently purchased items associated with distinct behavioral profiles. Based on these insights, the system will recommend the top three products most popular in each customer' s respective cluster that the individual has not yet purchased. This approach supports personalized marketing strategies, strengthens customer engagement, and has the potential to enhance overall sales performance. As for the outlier segment, where purchasing patterns are less consistent, a simplified baseline strategy will initially be employed—such as recommending random products—to encourage exploratory engagement and collect additional behavioral data for future refinement.

This code is an implementation scheme of an enhanced personalized recommendation system for customer segmentation, with the core feature of stronger robustness against data missing issues. The overall execution process follows a logical closed-loop of "data verification → preprocessing →

recommendation generation → result output and analysis", fully covering the entire process from data preparation to business application.

Firstly, the code enters the data completeness check phase, which is the basic guarantee for the normal operation of the recommendation system. The system first defines a list of core fields necessary for the recommendation function (including CustomerID for customer identification, StockCode for product code, Description for product description, and Quantity for purchase quantity). By comparing with the column names of the dataset, it identifies whether there are missing necessary fields; if there are missing fields, it immediately outputs a warning and prompts that the recommendation system may not work normally. On the premise that the field integrity meets the standard, it further counts the number of customers and products. If either of them is less than 10, it judges that the data volume is insufficient and prompts that the recommendation effect may be limited; otherwise, it confirms that the data volume meets the requirements for the operation of the recommendation system.

Next, the code enters the data preprocessing and filtering stage to prepare for subsequent recommendation generation. The system first checks whether there is an outlier dataset outliers_data in the current environment. If it exists, it extracts the abnormal customer IDs from it, filters out the transaction records of these abnormal customers from the original dataset, and obtains the cleaned dataset df_filtered; if there is no outlier dataset, it directly copies the original dataset as the data to be processed. At the same time, to avoid merging errors caused by inconsistent data types, the CustomerID in the customer segmentation result data customer_data_cleaned is uniformly converted to float type to ensure consistency with the customer identification type in the transaction data.

Subsequently, the code enters the core stage of personalized recommendation generation, which is designed with multi-layer robustness guarantee mechanisms. First, it checks whether the preprocessed transaction data and customer segmentation data are empty. If either is empty, it outputs a warning and initializes an empty recommendation result dataset; if the data are valid, it attempts to generate recommendations through steps such as data merging, popular product identification, and customer purchase history matching. The first step is to perform an inner join between the filtered transaction data and customer segmentation data by CustomerID to obtain the merged dataset merged_data containing customer clustering information; if the merged data is empty, the recommendation generation is also terminated and a prompt is given. The second step is to identify the best-selling products of each cluster. By grouping and counting the purchase quantity by "cluster-product code-product description", and sorting by cluster number in ascending order and purchase quantity in descending order, the list of popular products for each cluster is obtained; at the same time, the number of products in each cluster is counted. If there are clusters with less than 3 products, the global popular product supplement mechanism is activated. By counting the total purchase quantity of products in the full transaction data, the Top10 global popular products are selected for standby. The third step is to record the customer purchase history. By grouping and counting the purchase quantity by "customer ID-cluster-product code", the list of products purchased by each customer is clarified. The fourth step is to generate personalized recommendations. For each cluster, first obtain the Top10 popular products of the cluster and the list of customers belonging to it; for each customer, first filter out the cluster's popular products that have not been purchased as candidate recommendations. If there are less than 3 candidate recommendations, the global popular products are called to supplement to 3. Finally, the customer ID, cluster number, and the code and description of 3 recommended products are organized into a recommendation record. If there are less than 3 recommended products, null values are used for filling to ensure the uniform format of each recommendation record, and finally the recommendation

result dataset recommendations_df is constructed. The entire recommendation generation process is wrapped in an exception capture mechanism. If an error occurs during execution, an error message will be output and an empty recommendation dataset will be initialized.

After the recommendation results are generated, the code enters the result integration and output stage. If the recommendation result dataset is not empty, it is left-joined with the customer segmentation data by CustomerID and cluster (cluster number) to obtain the complete dataset customer_data_with_recommendations containing customer segmentation information and personalized recommendations, and 10 customers' recommendation results are randomly selected as examples for display; if the recommendation results are empty, the customer segmentation data is directly copied as the result dataset, and a prompt that recommendations cannot be generated is given.

Finally, the code enters the result saving and report generation stage. The system saves the customer segmentation result data and the customer data containing recommendation information as CSV files (customer_segmentation_results.csv and customer_data_with_recommendations.csv) respectively, which is convenient for subsequent analysis and application. At the same time, a customer segmentation report summary is automatically generated. For each cluster, core feature indicators such as the number of customers, average consumption amount, average number of transactions, average number of products purchased, average transaction value (if the field exists), average days since last purchase, and online shopping preference ratio (if the field exists) are counted and output. At the end of the report, targeted business recommendations are given based on customer value stratification, including providing VIP services and personalized recommendations for high-value customers to improve loyalty, increasing value for medium-value customers through promotions and cross-selling, and focusing on activation and retention strategies for low-value customers, forming a complete closed-loop from data processing to business decision-making. (Table 7) (Table 8)

Table 7. Recommendation System Results

CustomerID	Rec1_StockCode	Rec1_Description	Rec2_StockCode	Rec2_Description	Rec3_StockCode	Rec3_Description
79156	SKU_1162	WHITE MUG	SKU_1780	USB CABLE	SKU_1693	USB CABLE
15310	SKU_1803	WIRELESS MOUSE	SKU_1309	T-SHIRT	SKU_1341	WALL CLOCK
51472	SKU_1278	OFFICE CHAIR	SKU_1666	WIRELESS MOUSE	SKU_1427	OFFICE CHAIR
34367	SKU_1278	OFFICE CHAIR	SKU_1666	WIRELESS MOUSE	SKU_1427	OFFICE CHAIR
71648	SKU_1278	OFFICE CHAIR	SKU_1666	WIRELESS MOUSE	SKU_1427	OFFICE CHAIR
82502	SKU_1278	OFFICE CHAIR	SKU_1666	WIRELESS MOUSE	SKU_1427	OFFICE CHAIR
37653	SKU_1278	OFFICE CHAIR	SKU_1666	WIRELESS MOUSE	SKU_1427	OFFICE CHAIR
68269	SKU_1803	WIRELESS MOUSE	SKU_1309	T-SHIRT	SKU_1341	WALL CLOCK
17137	SKU_1278	OFFICE CHAIR	SKU_1666	WIRELESS MOUSE	SKU_1427	OFFICE CHAIR
67667	SKU_1162	WHITE MUG	SKU_1780	USB CABLE	SKU_1693	USB CABLE

Table 8. Recommendation System Validation

Item	Value
Precision	69.1%
Recall	88.85%

6. Theoretical and Managerial Implications

6.1 Theoretical Implications

This study adopts a generalized algorithmic framework, whereas existing research has mostly focused on the high-precision application of customized algorithms, with an emphasis on large enterprise scenarios. The algorithmic framework proposed in this study has the potential to be extended to individual online vendors. This not only fills the research gap from the perspective of technology inclusiveness, but also enriches the theoretical understanding of the hierarchical adaptation of AI algorithms in the e-commerce field, and validates the theoretical value of non-high-precision algorithmic frameworks in specific scenarios.

6.2 Managerial Implications

In terms of management practice, the concomitant value of the algorithm route in this study provides a feasible path for individual sellers. Such simple and generalized algorithms do not require professional technical reserves, enabling individual sellers to apply them to basic data analysis at low cost and optimize operational decisions, which to a certain extent makes up for their shortcomings in resources and technology. For e-commerce platforms, lightweight tools can be developed based on this, balancing usability and basic analysis needs, promoting the implementation of technological inclusion, and activating the micro vitality of the e-commerce ecosystem.

7. Result and Discussion

The presented cluster-aware popularity method offers a high-leverage, low-complexity approach to product recommendation. Its strengths lie in operational simplicity, interpretability, and robustness to sparse individual histories — properties that make it an excellent baseline and a reliable fallback in production stacks. The seven-step implementation neatly separates data hygiene (outliers, typing, returns) from model logic (within-cluster popularity and per-user exclusion) and serving output (top-3 per user with rationales).

While not a replacement for more sophisticated collaborative or sequential models, the approach integrates smoothly with them as a candidate generator or warm-start policy, and it supports a wide range of incremental improvements (time decay, margin weighting, diversity, and hybrid re-ranking). With disciplined evaluation and responsible governance, this method can deliver tangible commercial impact and provide a transparent bridge between descriptive segment analytics and personalized recommendations.

Funding

This research received no external funding.

Acknowledgements

The authors would like to thank the reviewers for their constructive feedback.

Data Availability Statement

This paper contain data can get from open network platform.

Ethics Statement

No human subjects are involved.

Conflicts of Interest

The authors declare no conflicts of interest.

Disclaimer of Artificial Intelligence (AI) Usage in Manuscript Preparation

The author(s) strictly adhered to academic norms in the process of using AI tools for manuscript preparation. No AI tools were employed to generate the content of the manuscript: AI was used solely for linguistic editing and proofreading. The author(s) have not engaged in any practices that violate academic ethics, including plagiarism, data forgery, or result falsification.

References

Bawack, R. E., Wamba, S. F., Guthrie, C., & Queiroz, M. M. (2022). Artificial intelligence in e-commerce: A bibliometric study and literature review. *Electronic Commerce Research, 22*, 885–917.https://doi.org/10.1007/s12525-022-00537-z

Davenport, T. H., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. Journal of the *Academy of Marketing Science, 48*, 24–42.https://doi.org/10.1007/s11747-019-00696-0

Elmaghraby, W., & Keskinocak, P. (2003). Dynamic pricing in the presence of inventory considerations: Research overview, current practices, and future directions. *Management Science/Manufacturing & Service Operations Management*. https://doi.org/10.1287/mnsc.49.10.1287.17315

Feng, J., Li, X., Sun, M., & Zhang, X. (2019). Online product reviews-triggered dynamic pricing. *Information Systems Research, 30(3)*, 1063–1081.https://doi.org/10.1287/isre.2019.0852

Grewal, D., Roggeveen, A. L., & Nordfält, J. (2017). *The future of retailing. Journal of Retailing, 93(1)*, 1–6.https://doi.org/10.1016/j.jretai.2016.12.008

Haleem, A., Javaid, M., Qadri, M. A., & Suman, R. (2022). Artificial intelligence (AI) applications for marketing: A literature-based study,*International Journal of Intelligent Networks,Volume 3*,2022,Pages 119-132,https://doi.org/10.1016/j.ijin.2022.08.005

Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research, 21(2)*, 155–172.https://doi.org/10.1177/1094670517752459

Huang, M.-H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science, 49*, 30–50.https://doi.org/10.1007/s11747-020-00749-9

Kanbach, D. K., & Stubner, S. (2024). The GenAI is out of the bottle: Generative artificial intelligence and business model innovation. *Review of Managerial Science, 18*, 291–317.https://doi.org/10.1007/s11846-023-00696-z

Kietzmann, J., Paschen, J., & Treen, E. (2018). Artificial intelligence in advertising: How marketers can leverage AI along the consumer journey. *Journal of Advertising Research, 58(3)*, 263–267.https://doi.org/10.2501/JAR-2018-035

Kshetri, N., & Dwivedi, Y. K. (2024). Generative artificial intelligence in marketing: Applications, opportunities, challenges, and research agenda. *International Journal of Information Management, 74*, 102716.https://doi.org/10.1016/j.ijinfomgt.2023.102716

Rahman, M. S., Al-Hakim, L., Bhuiyan, M. M. H., & Tarofder, A. K. (2023). Technology readiness of B2B firms and AI-based customer relationship management capability. *Industrial Marketing Management, 110*, 36–48.https://doi.org/10.1016/j.jbusres.2022.113525

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing, 90*, 106181.https://doi.org/10.1016/j.asoc.2020.106181

Shankar, V. (2018). How artificial intelligence (AI) is reshaping retailing. *Journal of Retailing, 94(4)*, v–xi.https://doi.org/10.1016/S0022-4359(18)30076-9

Verma, S., Sharma, R., Deb, S., & Maitra, D. (2021). Artificial intelligence in marketing: Systematic review and future research direction. *International Journal of Information Management Data Insights, 1(1)*, 100002.https://doi.org/10.1016/j.jjimei.2020.100002

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys, 52(1)*, 1–38.https://doi.org/10.1145/3285029

About Author(s)

Jiajun Li is a graduate student at Teesside University, majoring in Applied Artificial Intelligence. During their studies, the research focus is on the application of machine learning in data processing and prediction.