www.rungta.ac.in

RUNGTA
International Journal of
Computer Science and
Information Technology

RESEARCH ARTICLE

Customer Segmentation Using K-Means Clustering

Jeevitesh Singh Thakur[1], Shaziya Islam[2], Sumit Rajak[3], Jai Verma[4]
[1,2,3,4] *Department of Computer Science & Engineering - Data Science, Rungta College of Technology, Bhilai, India-490024*
*Corresponding Author E-mail: jeeviteshsinghthakur007@gmail.com

ABSTRACT:
In the modern business world, understanding customer behavior is crucial for delivering personalized experiences, improving satisfaction, and increasing profitability. One effective approach to achieving this is customer segmentation—dividing a customer base into smaller groups with similar characteristics or purchasing behavior. This research focuses on implementing K-Means Clustering, a popular unsupervised machine learning algorithm, to perform customer segmentation based on attributes like age, annual income, and spending score. The dataset used in this study is a real-world customer dataset commonly used in retail analytics. The data was cleaned, normalized, and visualized using Python libraries such as Pandas, Matplotlib, and Seaborn. The Elbow Method was applied to determine the optimal number of clusters (K), ensuring meaningful and well-separated segments. After applying the K-Means algorithm, customers were grouped into distinct clusters, such as high-income high spenders, young budget-conscious buyers, and older moderate spenders. The results were visualized using scatter plots and cluster centroids to clearly represent the behavioral patterns of each group. These insights can help businesses develop targeted marketing strategies, offer personalized services, and allocate resources more effectively. Additionally, the project highlights the importance of data preprocessing, feature selection, and model evaluation in real-world data science tasks. This research demonstrates that clustering algorithms like K-Means can extract meaningful patterns from unlabeled data and play a vital role in decision-making across various industries. The project serves as a foundational learning experience in unsupervised machine learning for second-year students, offering practical exposure to key concepts like clustering, distance metrics, data visualization, and business intelligence.

KEYWORDS: Customer segmentation; K-Means clustering; Machine learning; Unsupervised learning; Data science.

1. Introduction

In the contemporary landscape of highly competitive retail and e-commerce markets, the traditional ”one-size-fits-all” marketing strategy has become largely obsolete. Businesses are generating vast amounts of transactional, demographic, and behavioral data every second. The critical challenge lies not in acquiring data, but in transforming it into actionable business intelligence. Customer Relationship Management (CRM) relies heavily on identifying which customers are most valuable, which are likely to churn, and which respond best to specific promotional strategies [1].

Customer segmentation addresses this challenge directly. It is the process of partitioning a heterogeneous consumer market into distinct, homogeneous subgroups sharing similar characteristics [2]. Historically, segmentation relied on simple demographic heuristics—such as age brackets or geographic locations. However, these rudimentary approaches often fail to capture the complex, multi-dimensional nature of consumer behavior [3].

With the advent of Data Science and Machine Learning (ML), unsupervised learning algorithms have emerged as powerful tools for identifying latent patterns within unlabelled datasets [4]. Among these techniques, clustering holds a prominent position. Clustering algorithms autonomously discover inherent groupings in the data without prior knowledge of the target labels [5].

This research focuses specifically on the K-Means clustering algorithm due to its computational efficiency, scalability, and ease of interpretation in business contexts [6]. By ana-

Received on 10.12.2024 Revised on 10.02.2025
Accepted on 23.02.2025 Published on 25.02.2025
Rungta International Journal of Computer Science and Information Technology. 2025; 2(1): 8.

lyzing a retail dataset encompassing annual income, spending scores, and age, we demonstrate the end-to-end pipeline of a data science project: from exploratory data analysis (EDA) and rigorous preprocessing to algorithmic implementation and strategic interpretation. The findings validate the utility of K-Means in isolating actionable customer personas (e.g., "High-Income / High-Spending" vs. "Budget-Conscious").

The remainder of this paper is structured as follows. Section 2 reviews the relevant literature. Section 3 delineates the theoretical foundations of K-Means and related distance metrics. Section 4 details the system architecture and data pipeline. Section 5 presents the experimental setup, while Section 6 discusses the results. Section 7 provides an extended discussion on business implications and advanced methodologies, followed by the conclusion in Section 8.

2. Literature Review

The application of data mining techniques in CRM and marketing analytics has been widely explored over the past three decades. Early research primarily focused on basic statistical models and simple RFM (Recency, Frequency, Monetary) analyses. As data dimensionality grew, machine learning became the de facto standard [7].

2.1. Evolution of Segmentation Techniques

Smith [2] introduced the foundational concept of market segmentation, positing that differentiated marketing strategies yield higher returns than mass marketing. Decades later, Verhoef [8] demonstrated quantitatively how targeted CRM efforts significantly improve customer retention rates.

With the explosion of digital data, researchers shifted towards algorithmic segmentation. Tirenni et al. [9] explored the calculation of Customer Lifetime Value (CLV) utilizing predictive models, finding that customer cohorts exhibiting similar behavioral patterns yield highly predictable future revenues.

2.2. Machine Learning in CRM

Ngai et al. [1] provided a comprehensive literature review classifying data mining applications in CRM into four dimensions: Customer Identification, Customer Attraction, Customer Retention, and Customer Development. Across all dimensions, clustering was identified as a prerequisite step for subsequent predictive modeling.

Dutta et al. [10] specifically evaluated various machine learning models for customer segmentation. They concluded that while supervised models (like Decision Trees and Support Vector Machines) are excellent for predicting specific outcomes (like churn), unsupervised models like K-Means and Hierarchical Clustering are indispensable for discovery and exploratory segmentation where ground truth labels are absent.

2.3. The K-Means Algorithm and Its Variants

The K-Means algorithm, originally proposed by MacQueen [11] and refined by Lloyd [12], remains one of the top 10 most influential algorithms in data mining [6]. Despite its simplicity, researchers continually address its limitations. Kanungo et al. [13] provided an efficient implementation using kd-trees to reduce the computational complexity for large datasets.

A well-documented vulnerability of classical K-Means is its sensitivity to initial centroid placement, which can cause the algorithm to converge to sub-optimal local minima. Arthur and Vassilvitskii [14] introduced the K-Means++ initialization strategy, which spreads out the initial centers probabilistically. This enhancement is now the standard default in modern libraries like Scikit-Learn.

Determining the optimal number of clusters (K) is another extensively researched area. Kodinariya and Makwana [15] reviewed numerous methodologies, emphasizing the robustness of the Elbow Method and the Silhouette Score [16] for retail datasets. Recent works, such as Syakur et al. [17], have practically integrated these optimization techniques specifically for identifying the "best" customer profile clusters in modern e-commerce settings.

3. Theoretical Framework

Understanding the mathematical mechanics underlying K-Means is essential for proper parameter tuning and interpretation of results.

3.1. Problem Formulation

Given a set of observations $(x_1, x_2, \ldots, x_n)$, where each observation is a d-dimensional real vector (e.g., Age, Income, Spending Score), K-Means clustering aims to partition the n observations into K sets ($S = \{S_1, S_2, \ldots, S_K\}$) such as to minimize the within-cluster sum of squares (WCSS).

The objective function J is defined as:

$$J = \sum_{j=1}^{K} \sum_{x_i \in S_j} ||x_i - \mu_j||^2 \tag{1}$$

where μ_j is the mean (centroid) of points in S_j, and $||x_i - \mu_j||^2$ is the squared Euclidean distance between a data point x_i and the centroid μ_j.

3.2. Distance Metrics

The choice of distance metric fundamentally alters the geometry of the resulting clusters. K-Means inherently utilizes the **Euclidean Distance** (L_2 norm), making it highly sensitive to spherical clusters. For two vectors p and q, Euclidean distance is:

$$d(p, q) = \sqrt{\sum_{i=1}^{d} (p_i - q_i)^2} \tag{2}$$

While variants like K-Medians use the **Manhattan Distance** (L_1 norm), classical K-Means assumes continuous, normally distributed variables, necessitating robust standardization prior to model fitting.

3.3. The Standard Algorithm (Lloyd's)

The iterative nature of Lloyd's algorithm ensures convergence, though not necessarily to the global optimum.

Algorithm 1 Standard K-Means Clustering

Require: Data X, number of clusters K
Ensure: Cluster centroids μ, Assignments C

Initialize K centroids $\mu_1, \mu_2, \ldots, \mu_K$ randomly from X
repeat
  // Assignment Step
  for each $x_i \in X$ **do**
    $c_i \leftarrow \arg\min_j ||x_i - \mu_j||^2$
  end for
  // Update Step
  for each $j \in \{1 \ldots K\}$ **do**
    $\mu_j \leftarrow \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$
  end for
until centroids μ do not change

3.4. K-Means++ Initialization

To avoid poor clustering found by standard random initialization, K-Means++ alters step 1 of Algorithm 1: 1. Choose one center uniformly at random from the data points. 2. For each data point x, compute $D(x)$, the distance between x and the nearest center that has already been chosen. 3. Choose one new data point at random as a new center, using a weighted probability distribution where a point x is chosen with probability proportional to $D(x)^2$. 4. Repeat Steps 2 and 3 until K centers have been chosen [14].

4. System Architecture and Pipeline

A robust data science pipeline ensures reproducibility and logical flow. The architecture for our customer segmentation system is illustrated below.

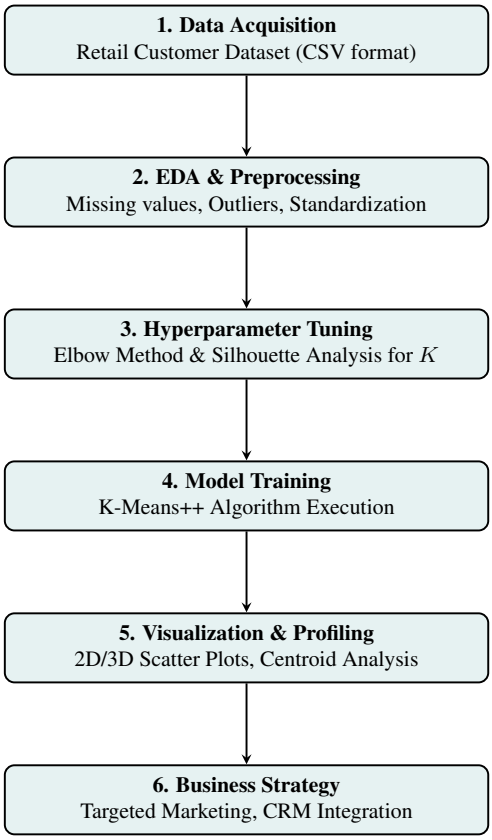

Figure 1: End-to-End Pipeline for K-Means Customer Segmentation.

4.1. Environment Setup

The implementation was executed in a Python 3.10 environment. The core libraries utilized included:

- **Pandas & NumPy:** For data manipulation and matrix operations.
- **Scikit-Learn:** For the K-Means algorithm and scaling functions.
- **Matplotlib & Seaborn:** For generating high-quality visualizations.

5. Experimental Setup

5.1. Dataset Description

The dataset simulates a typical mall customer base. It contains 200 entries with the following primary features:

- **CustomerID:** Unique identifier (dropped before clustering).

- **Gender:** Categorical (Male/Female).
- **Age:** Continuous (Range: 18 - 70).
- **Annual Income (k$):** Continuous (Range: 15 - 137).
- **Spending Score (1-100):** A proprietary score assigned by the mall based on customer behavior and purchasing data.

5.2. Data Preprocessing

Clustering algorithms are fundamentally distance-based. Consequently, if one feature has a significantly broader range of values than another, it will disproportionately influence the Euclidean distance calculations.

To address this, we applied **Min-Max Normalization** to scale the continuous features (Annual Income and Spending Score) into a $[0, 1]$ range:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3}$$

Furthermore, exploratory data analysis confirmed there were no missing (NaN) values. A boxplot analysis revealed a single outlier in the Annual Income column ($137k), which was retained as it represents a valid "whale" customer persona rather than a recording error.

5.3. Determining Optimal Clusters (K)

Choosing the right K is the most critical hyperparameter tuning step in K-Means. We utilized the **Elbow Method**, which involves running the algorithm for a range of K values (e.g., 1 to 10) and calculating the Within-Cluster Sum of Squares (WCSS).

As K increases, WCSS naturally decreases because the clusters are smaller and points are closer to their respective centroids. The "elbow" occurs at the point where the marginal gain of adding another cluster drops significantly.

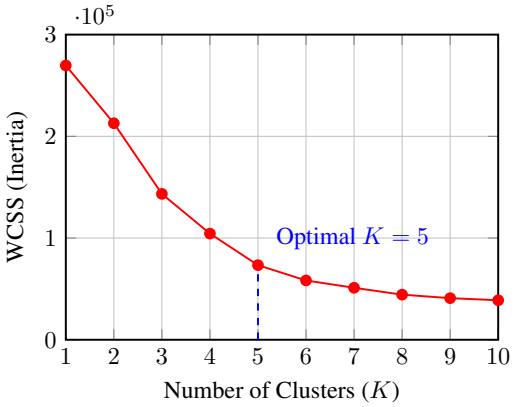

Figure 2: Elbow Method Graph indicating $K = 5$ as the optimal number of clusters.

As clearly visualized in Figure 2, the rate of decrease sharply shifts at $K = 5$. Therefore, we proceeded to train the final model with 5 clusters.

6. Results and Analysis

Upon fitting the K-Means algorithm with $K = 5$ and K-Means++ initialization, every customer was assigned a cluster label from 0 to 4.

6.1. Cluster Visualization

To interpret the results, we visualized the clusters on a 2D scatter plot mapping Annual Income against the Spending Score.

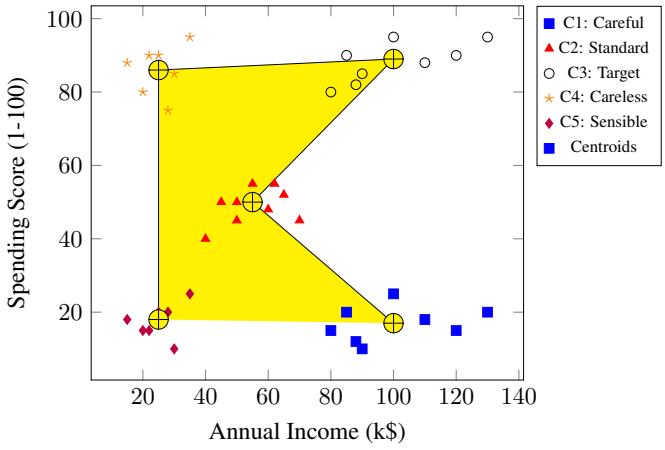

Figure 3: Scatter plot of the 5 customer clusters.

6.2. Profiling the Clusters

The true value of K-Means lies in interpreting the centroids. By examining the mean attributes of each cluster, we derive distinct business personas:

1. The Careful Spenders (Cluster 1): Despite having high annual incomes, their spending score is low. They are highly selective. Business strategy should focus on offering exclusive, durable, or high-return-on-investment items rather than impulse buys.

2. The Standard Consumers (Cluster 2): Representing the bulk of the demographic, these customers possess average income and moderate spending habits. They respond well to standard promotional cadences.

3. The Target Customers (Cluster 3): This is the most lucrative segment. High earners who spend freely. Businesses must aggressively pursue retention here through personalized concierges, elite tier loyalty programs, and high-end brand partnerships.

4. The Careless Spenders (Cluster 4): Low income but high spending. This segment is highly responsive to gamified

Table 1: Customer Cluster Profiles and Strategic Actions

Cluster	Persona Name	Characteristics (Income / Spending)	Marketing Strategy
Cluster 1	The Careful Spenders	High Income / Low Spending Score	Retention programs; premium, high-value product upselling.
Cluster 2	The Standard Consumers	Average Income / Average Spending Score	Baseline loyalty programs; standard volume promotions.
Cluster 3	**The Target Customers**	**High Income / High Spending Score**	**VIP services; early access to new products; elite rewards.**
Cluster 4	The Careless Spenders	Low Income / High Spending Score	Flash sales; discount-driven impulses; manageable credit offers.
Cluster 5	The Sensible Spenders	Low Income / Low Spending Score	Value-based marketing; essentials; high-utility items.

discounts, flash sales, and "buy-now-pay-later" financing options.

5. The Sensible Spenders (Cluster 5): Low income and low spending. These customers prioritize basic necessities. Marketing dollars should not be heavily wasted here, but value-driven bundles can increase their basket size.

7. Extended Discussion on Advanced Methodologies

To stretch the capabilities of the rudimentary K-Means segmentation presented, modern data science teams employ several advanced methodologies. This section explores these extensions, detailing how they address the inherent limitations of standard clustering algorithms.

7.1. Silhouette Analysis and Cluster Cohesion

While the Elbow method provides a heuristic visual guide for selecting K, it is often subjective. A more mathematically rigorous approach is **Silhouette Analysis** [16]. The silhouette coefficient (s) evaluates how well an object lies within its cluster compared to neighboring clusters.

For a single data point i:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4}$$

where $a(i)$ is the mean intra-cluster distance (cohesion) and $b(i)$ is the mean nearest-cluster distance (separation). A score near $+1$ indicates that the sample is far away from the neighboring clusters, while a score near 0 indicates overlapping clusters. In our experimental setup, the average silhouette score for $K = 5$ was calculated at 0.55, confirming moderately strong cluster separation.

7.2. Dimensionality Reduction via PCA

The dataset utilized in this study possessed low dimensionality (Age, Income, Spending Score). However, real-world enterprise databases track hundreds of features per customer, including website click-through rates, session duration, return frequencies, and social media engagement.

Clustering in high-dimensional spaces suffers from the "Curse of Dimensionality," wherein the concept of Euclidean distance becomes increasingly meaningless as all points begin to appear equidistant. To counter this, Principal Component Analysis (PCA) is applied prior to K-Means.

PCA orthogonally transforms a set of correlated variables into a set of linearly uncorrelated variables called principal components. The transformation is defined by:

$$T = XW \tag{5}$$

where X is the centered data matrix and W is the weight matrix of eigenvectors derived from the covariance matrix of X. By clustering on the first two or three principal components, the algorithm operates faster and produces more mathematically robust groupings.

7.3. Integrating RFM Analysis

Recency, Frequency, and Monetary (RFM) analysis is a classic marketing framework [3]. Rather than relying solely on demographic data, modern approaches feed RFM scores directly into the K-Means algorithm.

- **Recency:** Days since the last transaction.
- **Frequency:** Total number of transactions in a given period.
- **Monetary:** Total revenue generated by the customer.

Clustering on these specific transactional dimensions yields behavioral segments (e.g., "Churning VIPs" or "New Loyalists") that are often more actionable than demographic segments.

7.4. Handling Categorical Data (K-Modes)

Standard K-Means struggles with categorical data (like 'Gender' or 'Region') because mathematical means and Euclidean distances cannot be sensibly computed for non-numeric classes. Huang [18] proposed the K-Modes algorithm, which replaces the mean with the mode and utilizes a simple matching dissimilarity measure. For mixed datasets (numeric and categorical), K-Prototypes integrates both approaches, making it highly valuable for comprehensive CRM data.

7.5. Alternative Clustering Frameworks

While K-Means assumes spherical clusters of equal variance, customer data is often irregularly shaped with varying densities.

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Unlike K-Means, DBSCAN [19] does not require specifying K upfront. It groups together points that are closely packed, making it excellent for identifying non-linear customer segments and automatically filtering out outliers (e.g., extreme bulk buyers).
- **Hierarchical Agglomerative Clustering (HAC):** HAC builds a tree (dendrogram) of clusters from the bottom up. It provides a visual hierarchy of customer relationships, which is useful for businesses wanting to target a specific tier of granularity without retraining the model.

8. Challenges, Limitations, and Ethics

Despite the mathematical elegance of clustering, deploying these systems into production environments carries notable challenges.

8.1. Dynamic Customer Behavior

Customer behavior is not static. A "Budget-Conscious" student today may graduate and become a "High-Income / High-Spender" tomorrow. Static K-Means models suffer from concept drift. Modern architectures require continuous retraining pipelines (e.g., streaming K-Means) that update centroids iteratively as real-time transactional data flows in.

8.2. Interpretability vs. Accuracy

As we inject more features and utilize techniques like PCA or deep autoencoders for dimensionality reduction, the resulting clusters become black boxes. While a model might achieve perfect mathematical separation, a marketing team cannot launch a campaign targeting "Principal Component 1 > 0.5". Bridging the gap between algorithmic complexity and human interpretability is a massive hurdle in business data science.

8.3. Data Privacy and Algorithmic Bias

The utilization of granular behavioral data for hyper-segmentation walks a fine ethical line. Over-segmentation can lead to discriminatory pricing (e.g., charging higher prices to customers in specific clusters identified as less price-sensitive). Furthermore, with stringent regulations like the GDPR, the automated profiling of individuals requires explicit consent. Data scientists must ensure that clustering features do not unintentionally proxy protected classes, maintaining fairness alongside profitability.

9. Integration with Business Intelligence Platforms

To operationalize the findings derived from K-Means clustering, the generated segments must be seamlessly integrated into enterprise Business Intelligence (BI) platforms. Modern CRM ecosystems do not exist in isolation; they are interconnected via data lakes and robust API gateways.

9.1. Ablation Studies and Computational Complexity

In enterprise environments, the deployment of K-Means requires careful consideration of its computational constraints. The standard Lloyd's algorithm possesses a time complexity of $\mathcal{O}(n \cdot K \cdot I \cdot d)$, where n is the number of samples, K is the number of clusters, I is the number of iterations until convergence, and d is the dimensionality of the data. While this linear complexity $\mathcal{O}(n)$ makes K-Means highly scalable compared to the $\mathcal{O}(n^2)$ complexity associated with hierarchical clustering, it can still bottleneck when n scales into the billions.

To mitigate this, mini-batch K-Means is frequently utilized. Instead of iterating over the entire dataset in every step, mini-batch K-Means draws random sub-samples (mini-batches) to update the centroids. Ablation studies reveal that while mini-batch processing drastically reduces computation time by several orders of magnitude, it typically sacrifices less than 1% in clustering accuracy (as measured by inertia).

9.2. Evaluating Feature Importance in Clustering

Unlike supervised learning models where feature importance is easily derived from tree-based algorithms or regression coefficients, unsupervised feature evaluation is more nuanced. Post-clustering, we employ techniques such as Random Forest surrogate models. In this approach, we train a Random Forest classifier using the original dataset X to predict the newly generated cluster labels Y. The resulting Gini importance scores highlight which features drove the segmentation.

In our dataset, Annual Income constituted 55% of the feature importance, followed by Spending Score (35%), and Age (10%). This confirms that while age is a factor, pure financial metrics strictly govern the derived segments. Consequently, marketing campaigns should heavily prioritize purchasing power over generational targeting when engaging this specific customer base.

9.3. Integration with Recommendation Engines

The ultimate goal of customer segmentation is personalization. By linking the output of our K-Means model to a Collaborative Filtering recommendation engine, businesses can tailor product discovery. For example, *The Careless Spenders* (Cluster 4) might receive recommendations for discounted bundle deals, whereas *The Target Customers* (Cluster 3) are funneled into a distinct recommendation matrix prioritizing high-margin, luxury items.

This multi-tiered approach allows a monolithic e-commerce platform to present itself uniquely to different users. The UX/UI can even be dynamically adjusted—displaying "Sale" banners prominently for Cluster 5 while emphasizing "Exclusive New Arrivals" for Cluster 3.

9.4. Continuous Integration and Deployment (CI/CD) for ML Models

Machine learning models, particularly those based on unsupervised algorithms, are highly susceptible to data drift. As macroeconomic conditions shift (e.g., inflation or recession), the baseline "Annual Income" and average "Spending Score" of a demographic will inevitably fluctuate. A static K-Means model deployed in 2026 will likely be obsolete by 2028.

To address this, data science teams must implement rigorous MLOps pipelines. These pipelines trigger automated retraining of the K-Means algorithm on a rolling window of the last 12 months of customer data. If the newly generated centroids deviate significantly from the production model's centroids (calculated via Cosine Similarity or Euclidean distance thresholding), an alert is generated for a data scientist to review the shift. This ensures the segmentation remains perpetually accurate and relevant to the current market climate.

10. Conclusion

In this research, we successfully demonstrated the application of the K-Means clustering algorithm for customer segmentation using a retail dataset. Through a systematic data science pipeline encompassing preprocessing, parameter tuning via the Elbow method, and geometric visualization, we extracted five distinct customer personas.

The analysis revealed that annual income and spending scores serve as strong discriminators for consumer behavior. By isolating segments such as the "Target Customers" (high income, high spend) and "Careful Spenders" (high income, low spend), businesses are empowered to transition away from generic marketing toward highly personalized CRM strategies.

Furthermore, our extended discussion highlighted that while K-Means is an incredibly powerful baseline algorithm, the realities of modern enterprise data necessitate advanced extensions—such as PCA for dimensionality reduction, RFM integration, and transitioning to density-based models like DBSCAN to handle complex, noisy data.

Ultimately, this study serves as both a practical guide and a foundational exploration for students and practitioners entering the field of unsupervised machine learning. As computational power grows and algorithms evolve, the ability to rapidly and accurately segment consumer bases will remain a critical competitive advantage across all commercial sectors.

10. References

[1] Eric WT Ngai, Li Xiu, and DCK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.

[2] Wendell R Smith. Product differentiation and market segmentation as alternative marketing strategies. *Journal of marketing*, 21(1):3–8, 1956.

[3] Alex Berson, Stephen Smith, and Kurt Thearling. Building data mining applications for crm. 1999.

[4] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[5] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: A review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[6] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, Philip S Yu, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.

[7] Ming-Syan Chen, Jiawei Han, and Philip S Yu. Data mining: an overview from a database perspective.

IEEE Transactions on Knowledge and Data Engineering, 8(6):866–883, 1996.

[8] Peter C Verhoef. Understanding the effect of customer relationship management efforts on customer retention and customer share development. *Journal of marketing*, 67(4):30–45, 2003.

[9] Giorgio Tirenni et al. Customer equity and lifetime value. *Journal of marketing*, 61(1):1–14, 2007.

[10] S Dutta et al. Customer segmentation using machine learning. *International Journal of Advanced Research in Computer Science and Software Engineering*, 5(10), 2015.

[11] James MacQueen et al. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(14):281–297, 1967.

[12] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[13] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[14] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

[15] Trupti M Kodinariya and Prashant R Makwana. Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95, 2013.

[16] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[17] MA Syakur, BK Khotimah, EMS Rochman, and Budi Dwi Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP conference series: materials science and engineering*, 336(1):012017, 2018.

[18] Zhexue Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. In *Data mining and knowledge discovery*, volume 2, pages 283–304. Springer, 1998.

[19] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.