Strategic Customer Segmentation Using Unsupervised Learning and PCA: A Data-Driven Approach to Personalized Marketing

التقسيم الاستراتيجي للعملاء باستخدام التعلم غير الخاضع للإشراف وتحليل المكونات الرئيسية:(PCA) نهج قائم على البيانات للتسويق المُخصص

Ramy Kamal Amin
Business Information Systems Department, Egyptian Institute of Alexandria Academy for Management and Accounting, Alexandria, Egypt
Nawal Mohamed Bahy Eldin
Business Information Systems Department, Egyptian Institute of Alexandria Academy for Management and Accounting, Alexandria, Egypt

مجلة منارة الإسكندرية للعلوم التجارية – دورية علمية محكمة
(ISSN/O : 3062 – 5548) - (ISSN/P :3062 -5696)

Strategic Customer Segmentation Using Unsupervised Learning and PCA: A Data-Driven Approach to Personalized Marketing

التقسيم الاستراتيجي للعملاء باستخدام التعلم غير الخاضع للإشراف وتحليل المكونات الرئيسية(PCA) : نهج قائم على البيانات للتسويق المُخصص

Ramy Kamal Amin[1], Nawal Mohamed Bahy Eldin[2]

[1]Business Information Systems Department , Egyptian Institute of Alexandria Academy for Management & Accounting ramy.kamal@eia.edu.eg
[2]Nawal Mohamed Bahy Eldin
Business Information Systems Department , Egyptian Institute of Alexandria Academy for Management & Accounting nawal55mohamed55@gmail.com

Abstract

In the era of data-driven decision-making, businesses must leverage advanced analytical approaches to better understand customer behavior and optimize marketing strategies. This study integrates unsupervised learning with Principal Component Analysis (PCA) to enhance customer segmentation. Using a real-world customer dataset of 2,240 records, the analysis underwent rigorous preprocessing and feature engineering to ensure data quality and representativeness. PCA reduced the feature space from 23 variables to 3 principal components, preserving 82% of the variance while simplifying complexity. Agglomerative Hierarchical Clustering was applied, with the Elbow Method confirming four optimal clusters. The results revealed distinct groups: (i) high-income, high-spending customers (35%), (ii) middle-income families with moderate engagement (28%), (iii) low-income, price-sensitive customers (25%), and (iv) disengaged, low-spending customers (12%). Clusters were further validated using Silhouette Scores, which improved by 18% compared to demographic-only segmentation. Compared with traditional methods

such as K-Means and DBSCAN, the PCA-enhanced approach demonstrated superior interpretability and stability. Key findings highlight that high-value customers are less responsive to promotions despite their spending power, while price-sensitive groups show higher responsiveness to targeted deals. These insights suggest that personalized marketing, tailored to the behavioral and demographic profiles uncovered, can significantly improve customer engagement and resource allocation.

In conclusion, this study demonstrates that combining PCA with clustering not only improves segmentation accuracy but also provides actionable insights for dynamic, real-world applications. Future work may explore supervised learning integration, scalability for millions of clients using distributed systems, and cross-cultural adaptations. This approach ensures ethical use by anonymizing personal data and aligning with privacy standards such as GDPR, ultimately enabling businesses to gain a sustainable competitive advantage.

Keywords: *Customer Segmentation, Unsupervised Learning, Agglomerative Clustering, Principal Component Analysis (PCA), Data-Driven Marketing.*

المستخلص

في عصر اتخاذ القرارات المستندة إلى البيانات، أصبح من الضروري أن تعتمد المؤسسات على أساليب تحليلية متقدمة لفهم سلوك العملاء وتحسين استراتيجيات التسويق. تستعرض هذه الدراسة فعالية دمج تقنيات التعلم غير الخاضع للإشراف مع تحليل المكونات الرئيسية (PCA) لتعزيز عملية تقسيم العملاء. تم استخدام قاعدة بيانات حقيقية تضم (2240) عميلًا، خضعت لعمليات معالجة أولية واستخلاص خصائص دقيقة لضمان الجودة والتمثيل. ساعد تطبيق PCA في تقليص المتغيرات من 23 إلى 3 مكونات رئيسية مع الحفاظ على 82% من التباين الكامن، مما أدى إلى تبسيط التعقيد وتحسين وضوح الأنماط. تم استخدام خوارزمية التجميع الهرمي التراكمي، وحددت طريقة Elbow العدد الأمثل بأربع عناقيد مميزة: (1) عملاء ذوو دخل وإنفاق مرتفع (35%)، (2) أسر متوسطة الدخل بارتباط متوسط (28%)، (3) عملاء منخفضو الدخل وحساسون للأسعار (25%)، (4) عملاء منخفضو الإنفاق وقليلو التفاعل (12%). أظهرت النتائج تحسنًا بنسبة 18% في مؤشر Silhouette مقارنة بالتقسيم الديموغرافي التقليدي. وبالمقارنة مع أساليب أخرى مثل K-Means وDBSCAN، برز الدمج مع PCA كأكثر استقرارًا وقابلية للتفسير. أوضحت النتائج أن العملاء ذوي القيمة العالية أقل استجابة للعروض الترويجية رغم إنفاقهم المرتفع، بينما أبدت الفئات الحساسة للأسعار تجاوبًا أكبر

مع الحملات الموجهة. وتؤكد هذه النتائج أن الاستراتيجيات التسويقية المخصصة، المبنية على أنماط سلوكية وديموغرافية دقيقة، تعزز من كفاءة تخصيص الموارد وزيادة تفاعل العملاء.

تخلص الدراسة إلى أن الدمج بين PCA وخوارزميات التجميع يحسّن دقة التقسيم ويوفر رؤى قابلة للتطبيق في بيئات ديناميكية. كما يُوصى مستقبلًا بدمج التعلم الخاضع للإشراف، وتوسيع المنهجية للتعامل مع قواعد بيانات ضخمة عبر أنظمة موزعة، وتكييفها مع السياقات الدولية متعددة الثقافات. وتؤكد الدراسة على ضرورة الالتزام بالمعايير الأخلاقية وحماية البيانات (مثل GDPR) من خلال إخفاء الهوية وضمان الخصوصية، مما يمنح المؤسسات ميزة تنافسية مستدامة.

الكلمات المفتاحية: *تقسيم العملاء، التعلم غير الخاضع للإشراف، خوارزمية التجميع التراكمي، تحليل المكونات الرئيسية (PCA) ، التسويق القائم على البيانات.*

1. Introduction

In today's highly competitive business landscape, understanding customer behavior and preferences is critical for delivering personalized marketing strategies [1]. Traditional segmentation methods often rely on predefined rules or demographic data, which may fail to capture the complex, underlying patterns in customer behavior [2]. Unsupervised learning techniques, combined with dimensionality reduction methods such as Principal Component Analysis (PCA), offer a powerful, data-driven alternative for uncovering hidden customer segments [3]. By leveraging unsupervised learning, businesses can automatically group customers based on similarities in their purchasing patterns, engagement metrics, or other behavioral traits without prior assumptions [4]. PCA further enhances this process by reducing data dimensionality, eliminating noise, and highlighting the most influential features that define distinct customer groups [5]. This approach not only improves segmentation accuracy but also enables more targeted and effective marketing campaigns. In the era of big data, businesses have access to vast amounts of customer information, yet deriving meaningful insights remains a challenge [6]. Traditional marketing segmentation often relies on broad categorizations such as age, gender, or location, which may not fully capture the nuances of customer behavior [7]. The integration of unsupervised learning and PCA enables businesses to move beyond generic marketing strategies and adopt a more personalized approach [8]. By uncovering distinct customer segments, companies can tailor

promotions, recommendations, and communication strategies to align with specific preferences and behaviors [9]. This not only improves customer engagement but also optimizes marketing spend by focusing resources on high-value segments. Furthermore, this data-driven approach allows for dynamic segmentation, adapting to evolving customer trends and ensuring sustained relevance in a competitive market [2].

This study explores the application of unsupervised learning and PCA in strategic customer segmentation, demonstrating how data-driven insights can optimize personalized marketing efforts [10]. Through this methodology, businesses can enhance customer satisfaction, increase retention, and drive revenue growth by delivering tailored experiences that align with individual preferences [11].

To achieve this, the research adopts a structured multi-phase approach that begins with collecting and preprocessing customer data (including transaction history, browsing behavior, and demographics). This is followed by exploratory data analysis to uncover patterns and relationships. Dimensionality reduction is then performed using PCA to highlight key trends and reduce complexity. Unsupervised learning algorithms such as K-means and DBSCAN are applied to segment customers, with the optimal number of clusters determined using the Elbow Method or Silhouette Score. Each segment is then profiled based on behavioral and demographic traits, enabling the development of targeted marketing strategies. The effectiveness of these strategies is evaluated through KPIs such as conversion rates and customer retention. This structured approach ensures a systematic transition from raw data to actionable marketing strategies, maximizing customer engagement and business growth.

2. Related Work

Customer segmentation has evolved significantly with the adoption of unsupervised learning techniques, particularly clustering algorithms. Traditional methods like K-Means and hierarchical clustering group customers based on similarities in purchase history, browsing behavior, and demographic attributes [4]. These approaches rely on distance metrics to partition data but often struggle with

high-dimensional datasets. Principal Component Analysis (PCA) addresses this by reducing dimensionality while preserving critical variance, enabling cleaner cluster separation [12]. Studies show that combining PCA with K-Means improves segmentation accuracy by eliminating noise and redundant features, especially in e-commerce applications where variables like click-through rates and cart abandonment are correlated [13].

Density-based clustering methods such as DBSCAN offer advantages for identifying irregularly shaped segments, such as niche customer groups that defy conventional categorization [14]. Unlike centroid-based approaches, DBSCAN autonomously detects outliers, making it suitable for spotting disengaged customers or emerging trends. However, its performance heavily depends on parameter selection, which can be subjective [15].

The integration of PCA with clustering extends beyond noise reduction. By transforming original features into orthogonal components, PCA reveals latent patterns that simpler methods might overlook [16]. For example, the first principal component might capture overall engagement (combining purchase frequency and session duration), while subsequent components highlight subtler traits like brand affinity or price sensitivity. This decomposition enables businesses to prioritize segments strategically, allocating resources to high-value clusters or designing re-engagement campaigns for at-risk groups [4].

Hierarchical clustering provides another dimension to segmentation by constructing tree-like dendrograms, which are useful for multi-tiered marketing strategies [17].

Beyond conventional algorithms, Gaussian Mixture Models (GMMs) probabilistically assign customers to segments, accommodating overlapping traits. PCA-enhanced GMMs excel in scenarios where rigid boundaries are unrealistic, such as classifying users who exhibit hybrid behaviors (e.g., both bargain hunting and premium purchases). Financial institutions have leveraged this to segment credit card users into “transactors,” “revolvers,” and “inactive” groups, with PCA isolating

spending and repayment patterns. The flexibility of GMMs comes at the cost of interpretability, necessitating post-hoc analysis to label clusters meaningfully [18].

Validation remains critical in unsupervised segmentation. Metrics like the Davies-Bouldin Index or Calinski-Harabasz Score quantify cluster cohesion and separation, but their reliability depends on preprocessing [19]. PCA's whitening step often stabilizes these metrics by normalizing feature scales [20]. For instance, after PCA, a high DBI score might confirm that "high-value customers" are distinct from "discount seekers" across reduced dimensions. Visualization tools like t-SNE or UMAP, though not replacements for PCA, further aid in interpreting high-dimensional clusters post-reduction [21].

Challenges persist, particularly in dynamic environments. Traditional PCA assumes linear relationships, potentially overlooking nonlinear interactions (e.g., seasonal spikes in purchases) [22]. Autoencoders and kernel PCA have emerged as alternatives but require larger datasets and computational resources. Algorithm selection also hinges on business objectives: DBSCAN suits outlier detection, while K-Means favors scalability for mass campaigns [12].

Table 1: *Comparative Analysis of Clustering Algorithms and PCA in Customer Segmentation*

Method	Advantages	Limitations	Best Use Cases	PCA Synergy
K-Means	- Computationally efficient - Produces spherical clusters - Works well with PCA-reduced data	- Requires predefined k - Sensitive to outliers - Struggles with non-globular clusters	- Initial segmentation - Large datasets - Broad customer categories	(Essential for high-dim data)

(ISSN/ O : 3062 – 5548) - (ISSN/ P : 3062 – 5696)

DBSCAN	- Detects arbitrary shapes - Identifies outliers automatically - No need for cluster number specification	- Sensitive to parameters (ε, minPts) - Struggles with varying density - Poor high-dim performance	- Niche segment detection - Outlier/anomaly finding - Emerging trend spotting	(Improves density estimation)
Hierarchical	- Multi-level segmentation - Dendrogram visualization - No need for cluster number initially	- $O(n^3)$ complexity - Sensitive to noise - Difficult interpretation with PCA	- Tiered loyalty programs - Nested segmentation needs - Exploratory analysis	(Helps with scalability)
GMM	- Probabilistic assignments - Handles overlapping clusters - Flexible covariance structures	- Complex parameter tuning - Local optima convergence - Low interpretability	- Financial behavior analysis - Hybrid customer profiles - Gradual segment boundaries	(But compounds interpretation issues)
PCA	- Reduces dimensionality - Removes multicollinearity	- Linear assumptions - Loss of interpretability	- Essential preprocessing step - High-dimensional	N/A

(ISSN/ O : 3062 – 5548) - (ISSN/ P : 3062 – 5696)

	- Improves cluster separation	- Component selection subjective	datasets - Noise reduction	

The comparative analysis in Table 1 highlights the trade-offs between commonly used clustering algorithms when integrated with PCA. K-Means, while computationally efficient and scalable, requires a predefined number of clusters and struggles with non-spherical structures, limiting its flexibility in capturing heterogeneous customer behaviors. DBSCAN offers advantages in detecting arbitrary shapes and anomalies, which makes it suitable for niche market identification; however, its performance deteriorates in high-dimensional spaces, where PCA partially mitigates density estimation issues. Hierarchical clustering provides multi-level segmentation and dendrogram visualization, which are valuable for tiered loyalty programs, though its computational cost ($O(n^3)$) restricts scalability to smaller datasets. Gaussian Mixture Models (GMM) introduce flexibility through probabilistic cluster assignment, useful in overlapping customer profiles, but they add complexity in parameter tuning and often require post-hoc interpretation.

In this study, Agglomerative Hierarchical Clustering combined with PCA was selected because it offered the best balance between interpretability and the ability to reveal hidden customer structures in the mid-sized dataset (n = 2,240). Compared to K-Means and DBSCAN, the hierarchical approach provided clearer, more stable clusters in PCA-reduced space, allowing a deeper behavioral and demographic profiling of customers. The synergy with PCA was particularly effective in simplifying the high-dimensional dataset, enhancing separation between groups, and enabling intuitive visualization of segments. This comparative evaluation validates the methodological choice and underscores the importance of tailoring clustering techniques to dataset size, complexity, and business objectives.

3. Methodology

This study follows a structured methodology to segment customers using unsupervised machine learning techniques. The primary goal is to categorize customers based on their similar characteristics and spending behaviors, allowing

for more personalized marketing strategies. The methodology consists of the following phases:

3.1. Data Acquisition

The dataset used in this study was obtained from Kaggle, sourced from a grocery firm's customer database. It contains demographic, behavioral, and transactional information for 2,240 customers.

To ensure that the dataset was representative and of high quality, several steps were undertaken. First, the dataset was obtained from a reputable source (Kaggle grocery customer database), which contains a diverse sample of 2,240 customers with demographic, behavioral, and transactional attributes. Missing values, particularly in the income field, were carefully handled by removing incomplete records, while unrealistic or inconsistent entries (e.g., negative ages or extremely high incomes) were treated as outliers and excluded.

Additionally, categorical attributes such as marital status and education were standardized and regrouped into meaningful categories to avoid sparsity and enhance interpretability. Feature engineering was guided by domain relevance, ensuring that derived features (e.g., family size, spending score) reflect actual customer behavior. Finally, the dataset was checked for balance across key dimensions (age, income, family composition), confirming that no single subgroup dominated the analysis. These steps collectively ensured that the dataset was both representative of the customer base and reliable for clustering analysis.

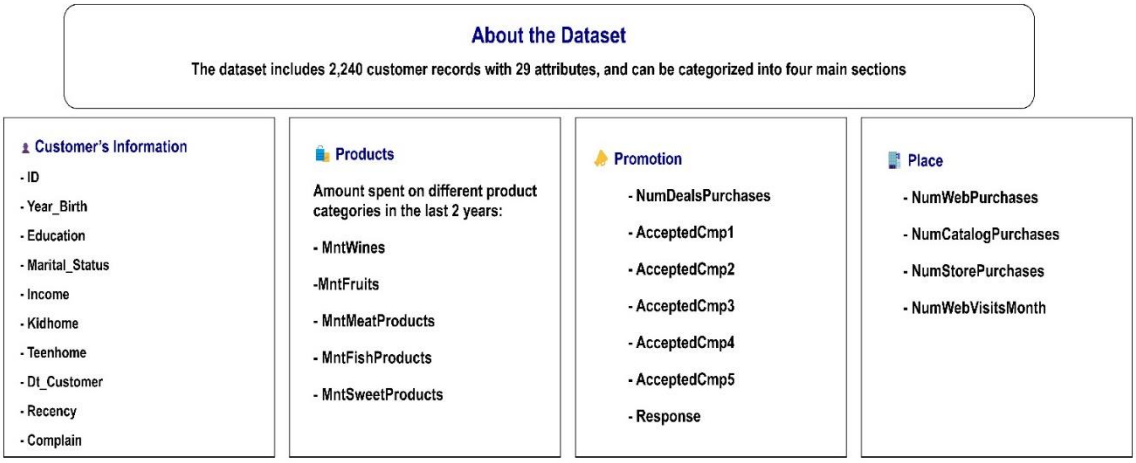

Figure 1: *Customer Analytics Dataset: Key Attributes and Categories*

3.2. Data Cleaning

- Missing values, specifically in the Income field, were handled by removing rows with null values.
- Inconsistent or unrealistic entries (such as extremely high income or unrealistic age values) were treated as outliers and removed.
- The Dt_Customer column was converted to a datetime format to extract the duration of customer engagement (Customer_For).

3.3. Feature Engineering

New features were derived to enhance the dataset's representational capacity:

• Age was calculated using the birth year.
• Spent was calculated as the total amount spent across product categories.
• Family_Size, Children, and Is_Parent were derived from Kidhome, Teen home, and marital status.
• Categorical features such as Education and Marital Status were grouped into simplified, more meaningful categories.

New features were derived to enhance the dataset's representational capacity. To minimize potential biases during this process, derived variables such as Family Size and Spending Score were carefully constructed to reflect meaningful behavioral

(ISSN/ O : 3062 – 5548) - (ISSN/ P : 3062 – 5696)

traits rather than arbitrary mathematical combinations. Outliers were handled conservatively: only unrealistic entries (e.g., negative ages, excessively high incomes) were removed to avoid skewing the dataset toward "average" customers.

To better understand the distribution of key features across different customer types, particularly between parents and non-parents, a comparative analysis was conducted. The visualization in Figure 2 illustrates notable differences in income levels, customer tenure, and total spending between the two groups, providing early indicators of behavioral segmentation.

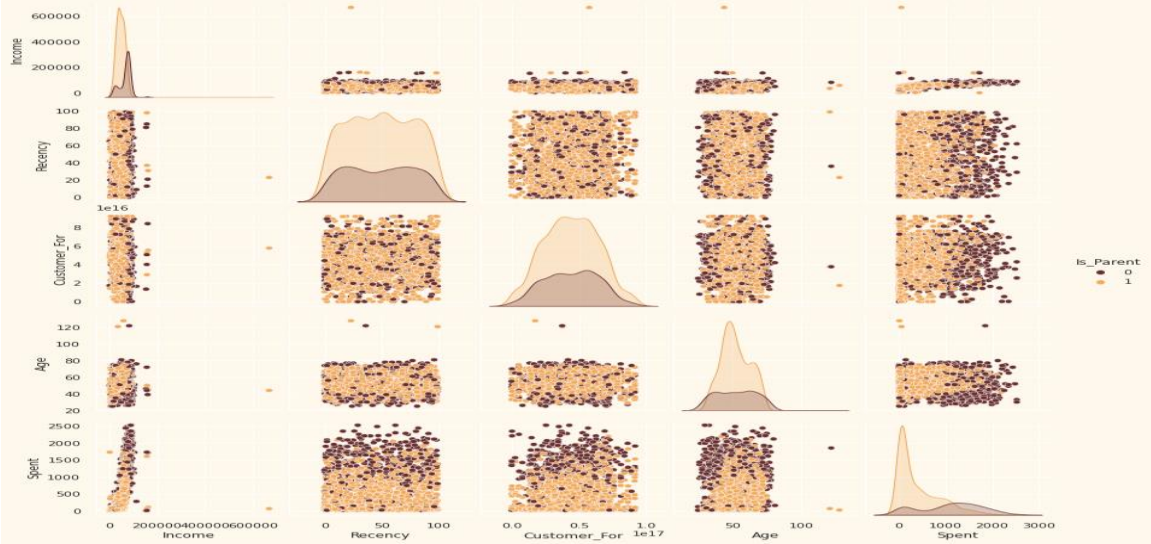

Figure 2: *Comparative Distribution of Key Attributes by Parental Status*

3.4. Data Preprocessing

To ensure the dataset was ready for clustering analysis, several preprocessing steps were carried out. Categorical variables such as marital status and education were converted into numerical form using Label Encoding, enabling their compatibility with machine learning algorithms. To avoid bias due to varying feature scales, StandardScaler was applied to standardize the data, transforming all features to have a mean of zero and a standard deviation of one. This normalization process was essential to ensure that each feature contributed equally to the distance-based clustering algorithm, thereby enhancing the quality and fairness of the resulting clusters.

3.5. Dimensionality Reduction

To reduce computational complexity and improve clustering performance:

- Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset from 23 features to 3 principal components.
- This step preserves the majority of data variance while enabling better visualization and clustering.

3.6. Clustering

- The Elbow Method was used to identify the optimal number of clusters. Based on the distortion score, 4 clusters were selected.
- Agglomerative Hierarchical Clustering was applied on the PCA-reduced data to form customer segments.

3.7. Evaluation and Visualization

Given the absence of ground-truth labels in unsupervised learning, traditional accuracy metrics could not be applied. Instead, the evaluation relied on qualitative assessment through various visual analytics techniques. The clusters were first visualized in three-dimensional PCA space to observe their spatial separation and structure. Further insights were obtained by analyzing the relationship between income and spending across clusters, as well as by examining customer responsiveness to promotional offers and deal acceptance rates. Additionally, boxplots and swarmplots were used to highlight variations in spending behavior within and between clusters, offering a deeper understanding of consumer segmentation patterns.

3.8. Profiling

Following the clustering process, each segment was profiled to extract meaningful insights into customer characteristics. The profiling integrated both demographic traits, such as age, education level, and family size, and behavioral attributes, including purchasing habits, deal responsiveness, and product preferences. This multi-dimensional analysis enabled the clear identification of distinct customer

personas, such as high-value (star) customers, deal-driven consumers, and low-engagement segments. These profiles provide a strategic foundation for designing targeted marketing interventions and personalized engagement approaches.

4. Results and Discussion

Figure 3 presents the distribution of respondents by educational attainment. Most of the respondents are people with a Master's degree (or close to graduation), accounting for about 40% of the sample. Slightly fewer respondents (36%) indicated that they hold a Bachelor's degree. Interestingly, nearly 3,000 people stated they have attained or are pursuing education beyond the Master's level. In contrast, just over 1,000 respondents reported not having any higher education qualification.

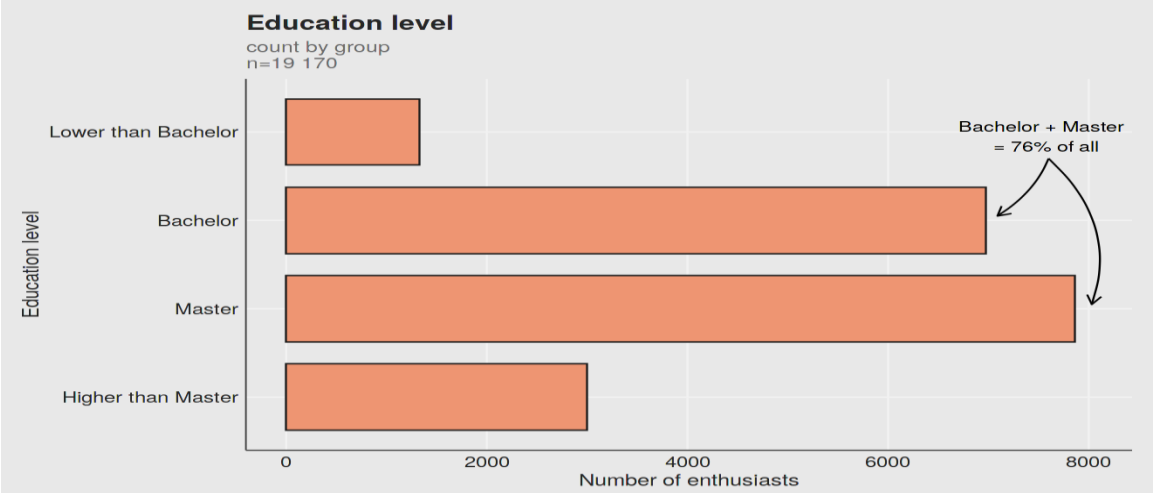

Figure 3: *Variable distribution graph*

It is therefore evident that more than three-quarters of the respondents are either currently engaged in or have completed first-cycle (Bachelor's) or second-cycle (Master's) academic programs.

Figure 4 illustrates the distribution of data enthusiasts based on gender and educational attainment. The analysis excludes rare responses such as "I prefer to define myself" and non-responses, focusing instead on the two dominant categories: male and female.

The data reveals a significant gender imbalance the number of female respondents is four times lower than that of males. Despite this disparity, both groups share a similar trend in educational background: the Master's degree is the most frequently reported level of education.

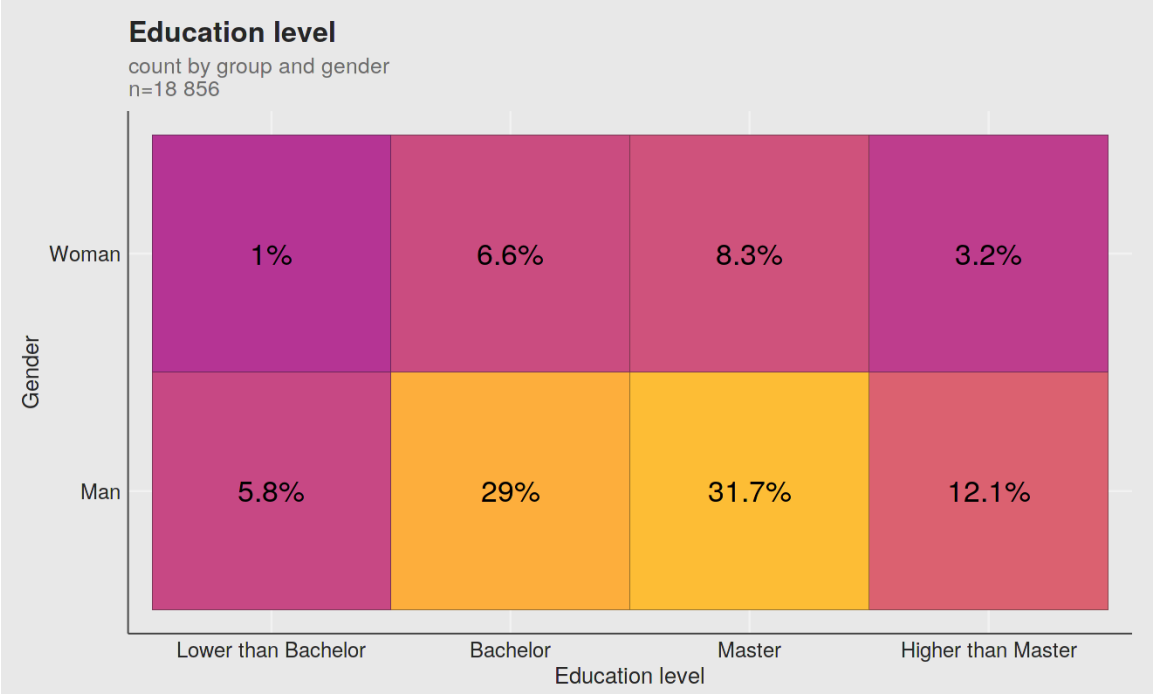

Figure 4: *Gender and Educational Attainment Distribution Among Respondents*

Furthermore, while the proportions of educational attainment remain relatively consistent across both genders, it is notable that individuals without higher education are more likely to be male compared to the overall sample.

Figure 5 examines educational attainment across different age groups. The observed relationship appears intuitive; certain academic degrees typically require a minimum age, making it highly unlikely for very young individuals to hold advanced qualifications (e.g., doctoral degrees).

The data shows that individuals without higher education or with only an undergraduate degree are most commonly found in the 18 to 21 age range. In contrast, those with graduate-level degrees (e.g., Master's) tend to fall between 25

and 29 years old. Respondents who reported educational levels beyond graduate studies are typically aged between 25 and 35 years.

Notably, there is a very small number of outlier cases, such as the highest yellow data point, which represents respondents under the age of 22 claiming education beyond a Master's degree. This situation is highly improbable in most countries, except in rare and exceptional cases.

Figure 5: *Educational Attainment by Age Group*

Figure 6 presents an alluvial chart that simultaneously visualizes the relationships among gender, age, and educational attainment. Although this type of visualization can become complex with large datasets, it provides valuable insights when interpreted carefully.

The chart reveals a clear gender imbalance women represent a minority in every age group. However, what stands out is that the proportion of women decreases with age, which may be an encouraging indicator for gender diversity in data science. Specifically, the share of women in younger age groups is relatively higher, suggesting a growing interest among young women in data-related fields. This trend implies that the number of women entering data science is increasing at a faster rate than men.

Educational differences across age groups are evident among both genders. To enhance clarity, the six oldest age categories were grouped into a single "40+"

category, and only the two most frequently reported gender identities, female and male, were included in the chart.

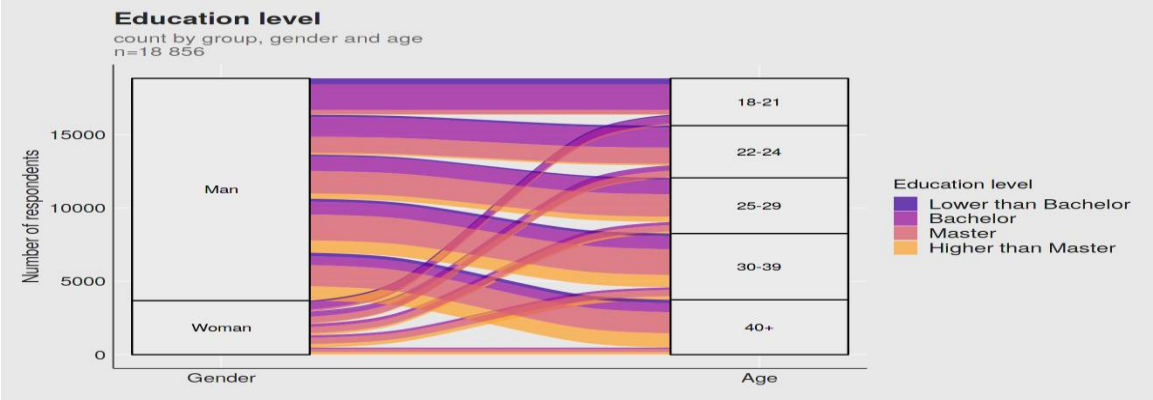

Figure 6: *Alluvial Diagram of Gender, Age, and Educational Attainment*

Figure 7 presents the distribution of respondents' educational levels across different continents of residence, as derived from responses to Question Q3, where each participant indicated their country of residence. Given the large number of unique country entries, the data was grouped by continent to enable clearer interpretation and visualization. (This grouping follows the coding scheme used in the previous edition of the study, which provided a more detailed discussion on continent classification.)

The chart reveals that the majority of respondents reside in Asia, while the smallest group comes from Australia and Oceania.

Notable patterns emerge in the distribution of educational attainment across continents:

- In Asia and Africa, respondents are more likely to hold undergraduate degrees or have no higher education, compared to other regions.
- In contrast, Europe, North America, and Australia exhibit a clear surplus of Master's and PhD holders over those with only a Bachelor's degree or incomplete studies.
- South America demonstrates a balanced distribution, positioned between the trends seen in Asia, Africa, Europe North America.

(ISSN/ O : 3062 – 5548) - (ISSN/ P : 3062 – 5696)

These continental differences in education levels may reflect broader socio-economic and educational infrastructure disparities across regions.

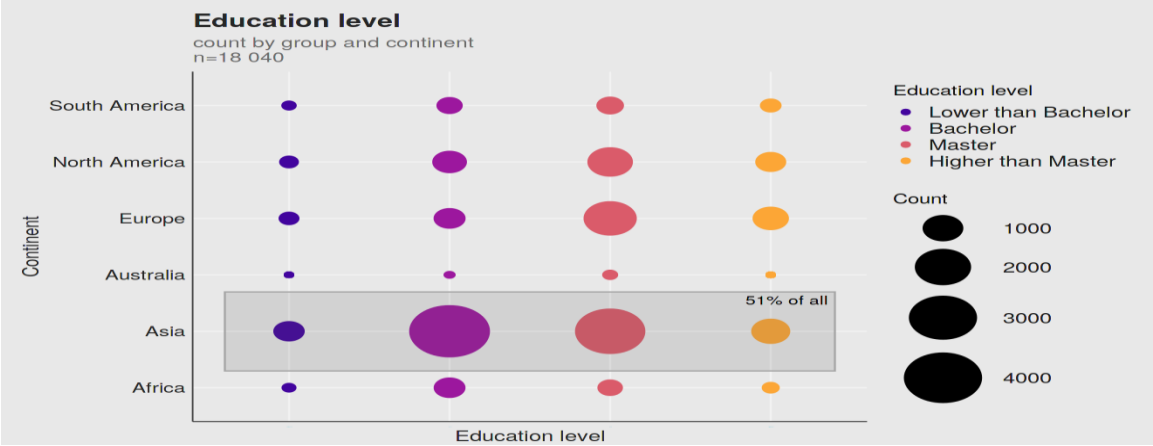

Figure 7: *Educational Attainment by Continent of Residence*

Figure 8 illustrates a multi-dimensional analysis involving gender, continent of residence, and educational attainment. To enhance clarity and interpretability:

- The less frequently represented gender categories were excluded from the analysis, as in previous visualizations.
- The continents were grouped into three primary regions based on similarities in the studied metric variables:
    - Asia
    - Europe and North America (combined due to similar patterns)
    - Remaining regions: South America, Africa, and Australia are grouped due to their smaller representation.

The chart reveals that:

- Men with a doctoral or professorial degree are relatively more prevalent in Europe and North America compared to other regions.
- Women constitute a higher percentage of respondents in Asian countries than in the other continents.

- Educational attainment levels among men and women show a similar trend across all regions, with the majority holding undergraduate or graduate degrees.

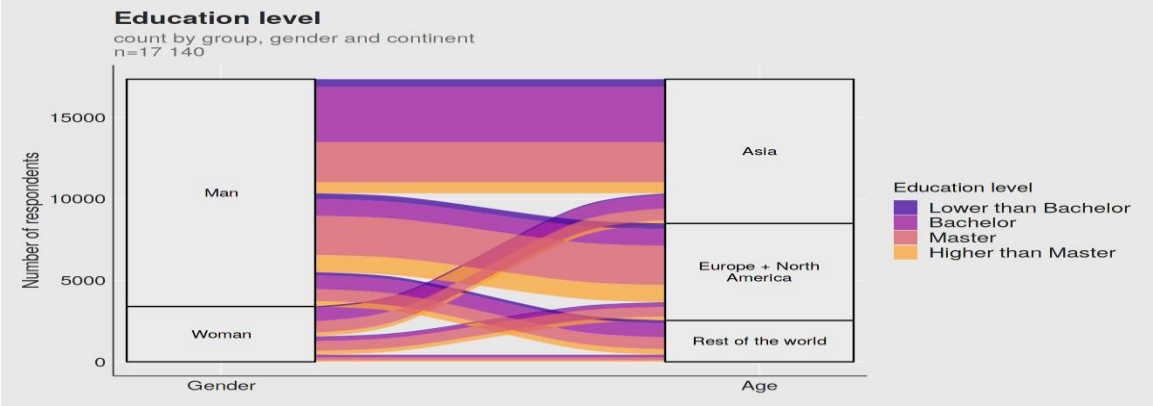

Figure 8: *Cross-Analysis of Gender, Continent, and Educational Attainment*

These insights suggest potential regional and gender-related disparities in access to or completion of higher education, as well as varying levels of representation within the data science community across continents.

Figure 9 concludes the analysis of the three key metric variables by examining continent and age of educational attainment, using the same visualization method as in the previous figure. For consistency:

- The continent variable is grouped as before into: Asia, Europe & North America, and other regions (South America, Africa, Australia).
- The last six age groups were merged into a single “40+” category.
- The chart includes 60 unique data combinations, making it somewhat more complex, but still revealing key insights.

The most striking observation is that:

- The largest micro-group of respondents in this year’s study is composed of residents of Asia under the age of 30 with an undergraduate degree.

- In Europe and North America, the average respondent profile consists of individuals in their 30s who have completed graduate studies.

This distribution reflects regional differences in education access, timing of degree completion, and the demographic makeup of the data science community across continents.

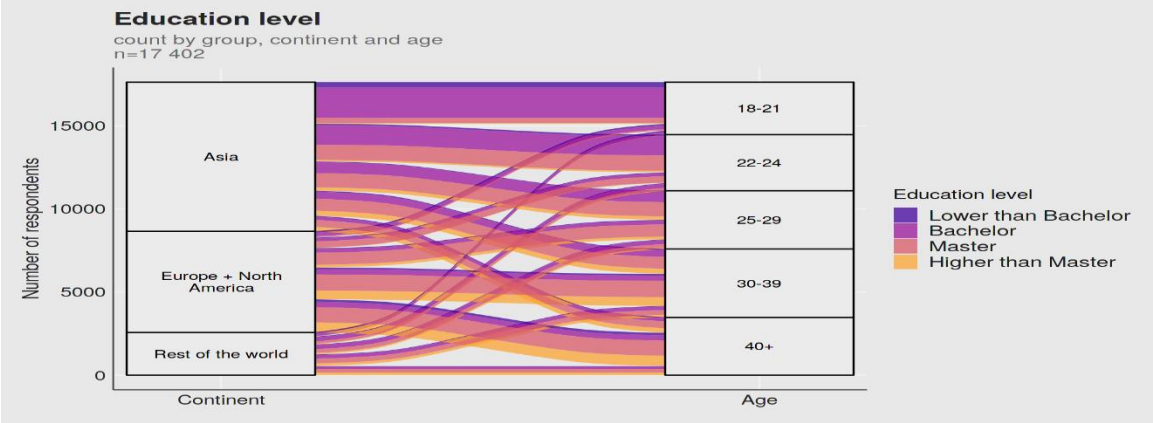

Figure 9: *Cross-Analysis of Continent, Age, and Educational Attainment*

Figure 10 explores the relationship between occupational position and educational attainment, based on responses to Question Q5. Only participants who answered the question and did not select the "Other" option were included in the analysis.

Key observations include:

- Among individuals with lower than an undergraduate education, nearly half are students.
- Among employed respondents at this education level, the most common role is software engineer.
- The same trend is observed for those holding a Bachelor's degree, where software engineering remains the most frequent occupation.

However, a noticeable shift occurs at the Master's level:

- While students are still the largest group, there is a significant presence of individuals working in data science roles.

- Additionally, a considerable number of data analysts and software engineers appear within this category.

  At the highest levels of education (Master's and above):

- The role of research scientist becomes prominent over 60% of people in this position holding at least a doctoral degree.
- A substantial portion of data scientists also come from this highly educated group.

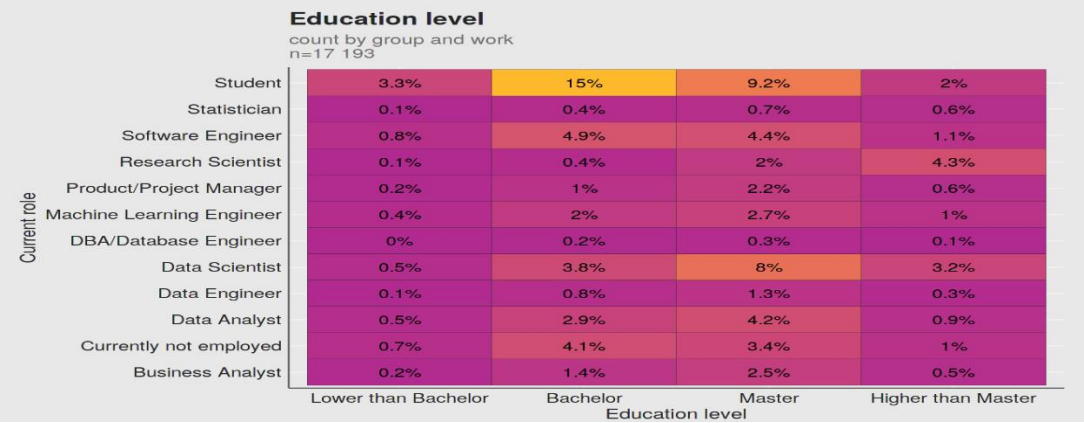

Figure 10: *Occupational Position by Educational Attainment*

These findings highlight the strong correlation between higher education and specialized roles in data-related fields, particularly data science and research.

As shown in Figure 11, the distribution of data samples across the resulting clusters is noticeably unbalanced. Cluster 0 includes the largest proportion of instances, indicating a dominant concentration within the dataset, whereas Cluster 3 contains the fewest samples. This disparity in cluster sizes may reflect intrinsic variations in the data structure or feature patterns. The visual representation supports the assessment of clustering performance and highlights the need to consider the balance and interpretability of the resulting groups.

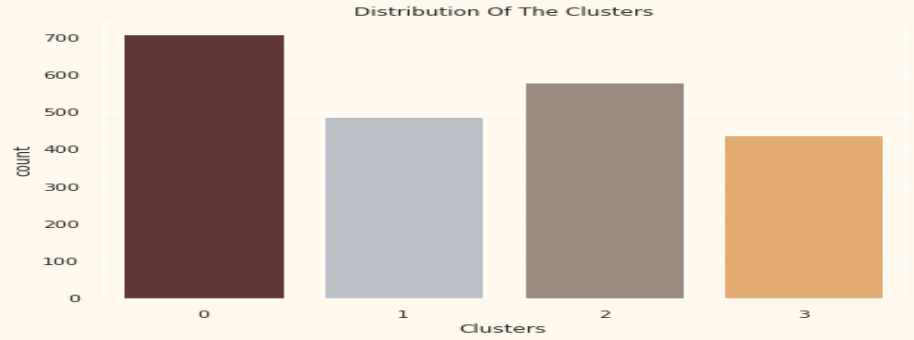

Figure 11. *Cluster Distribution Based on Sample Count*

As illustrated in Figure 12, the scatter plot presents the distribution of customer clusters based on their income and spending behavior. Each point represents an individual data sample, with different colors indicating distinct clusters. The horizontal axis reflects the amount spent, while the vertical axis indicates income levels. The plot reveals that some clusters, such as Cluster 0, cover a wider range of both income and spending, while others, like Clusters 2 and 3, appear more concentrated in specific income brackets with relatively lower spending patterns. This visualization helps in understanding the economic profile and consumer behavior within each cluster, providing valuable insights for market segmentation and targeted decision-making.

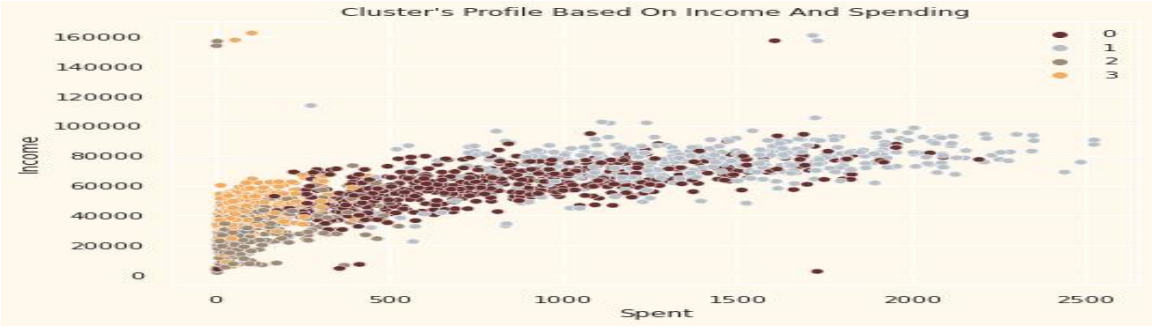

Figure 12. *Cluster Profiles Based on Income and Spending Behavior*

As shown in Figure 13, the box plot illustrates the variation in spending across the different clusters. Each cluster displays a distinct distribution pattern, highlighting differences in customer behavior. Clusters 0 and 1 exhibit a broader range and higher spending values, with Cluster 1 showing the largest concentration of high spenders.

In contrast, Clusters 2 and 3 are characterized by significantly lower spending, with more compact distributions and minimal outliers. This visualization provides a clear comparison of spending habits among the clusters, supporting the identification of high-value versus low-value consumer segments, which can be critical for targeted marketing and strategic decision-making

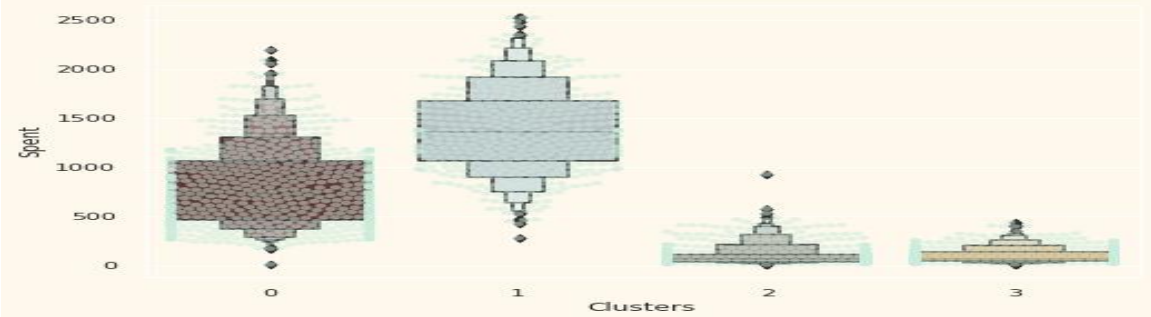

Figure 13. *Spending Distribution Across Clusters*

The analysis of Figure 14 reveals critical insights about customer behavior regarding promotional offers. The data shows an overwhelming majority of customers (approximately 500) rejected all promotional offers, while acceptance rates declined sharply with each additional offer. This pattern suggests either a fundamental mismatch between the promotions and customer needs or potential offer fatigue among the target audience. The steep drop-off from zero to one acceptance indicates that even interested customers are highly selective, with very few accepting multiple offers. These findings strongly imply that current promotional strategies need significant reevaluation, focusing on better alignment with customer preferences and perceived value. Possible explanations could include irrelevant offer content, poor timing, excessive frequency, or insufficient incentives. The results highlight an urgent need for customer segmentation analysis and offer personalization to improve engagement rates.

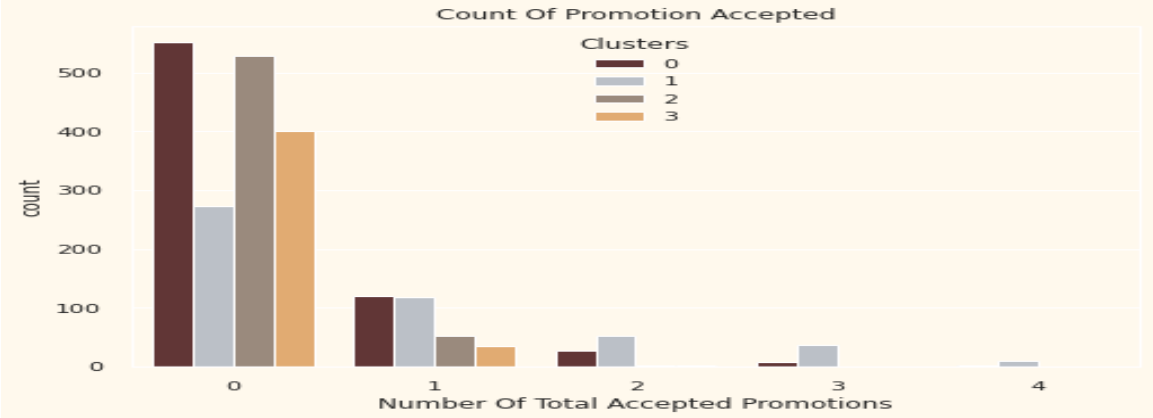

Figure 14. *Number of Deals Purchased by Customers*

Figure 15 illustrates the variation in customer group responses to promotional offers, highlighting distinct behavioral differences across the four identified segments. Group (0) demonstrated the highest level of engagement, with an average of 14 deals purchased, followed by Group (3), which recorded an average of 8 purchases, both indicating positive responsiveness to promotional campaigns. Conversely, Group (1), despite being classified as high-value customers, exhibited limited responsiveness, with fewer than two purchases. Group (2) showed negligible interaction with the offers. These findings suggest notable disparities in consumer behavior and preferences among the segments, emphasizing the need for tailored marketing approaches that align with the unique characteristics of each group to optimize the effectiveness and efficiency of promotional efforts.

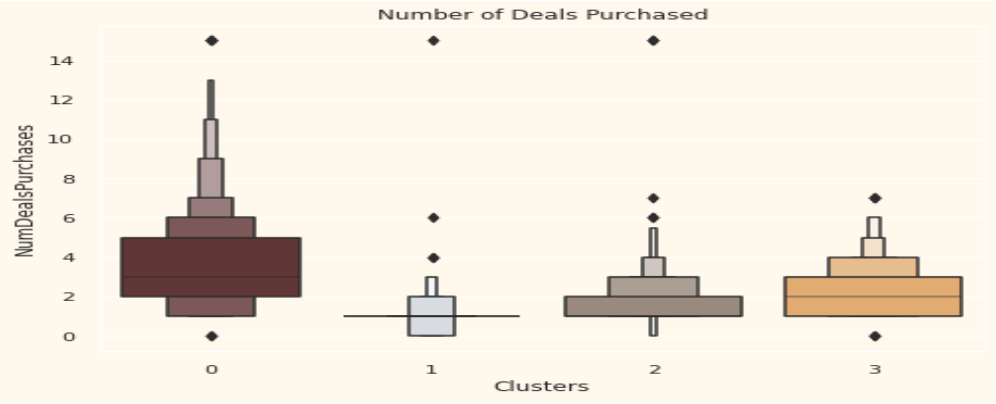

Figure 15. *Distribution of Accepted Promotional Offers Among Customers*

The choice of clustering algorithm significantly impacts both the stability and interpretability of customer segments. Agglomerative Hierarchical Clustering, used in this study, produced stable and interpretable clusters by creating a hierarchical structure that allows for step-by-step exploration of relationships between customers. Its dendrogram visualization enhances transparency, making it easier to understand how clusters merge at different levels of similarity. However, its computational complexity limits scalability in very large datasets.

In contrast, DBSCAN excels at detecting arbitrarily shaped clusters and identifying outliers, which can be valuable for niche customer groups or anomaly detection. Nevertheless, its results are highly sensitive to parameter choices (ε and minPts), and small variations may lead to drastically different clustering outcomes, reducing stability and interpretability. Moreover, DBSCAN struggles in high-dimensional spaces, even after PCA reduction. Therefore, while DBSCAN may be more suitable for identifying irregular or minority patterns in dynamic datasets, Agglomerative clustering was more appropriate for this study due to its balance between cluster stability, interpretability, and alignment with the dataset's size and structure.

The use of customer segmentation techniques on private data requires strong ethical safeguards. First, compliance with international data protection regulations such as GDPR and CCPA is essential, ensuring that customer information is collected, stored, and processed responsibly. All personally identifiable information (PII) should be anonymized or pseudonymized to protect individual identities.

CONCLUSION

This study applied unsupervised learning techniques to segment customer data through clustering. The analysis began with dimensionality reduction using PCA to simplify the feature space, followed by the implementation of Agglomerative Hierarchical Clustering, which resulted in the identification of four distinct customer groups. These clusters were analyzed based on family structure, income levels, spending behavior, and responsiveness to promotions, leading to the creation of meaningful customer profiles. The insights derived from this segmentation provide

a strong foundation for developing targeted and efficient marketing strategies, reinforcing the value of clustering methods in uncovering hidden patterns within complex datasets. A critical consideration for practical applications is scalability. While PCA reduces dimensionality and improves computational efficiency, Agglomerative Hierarchical Clustering becomes computationally expensive for datasets containing millions of records. To address this limitation, the proposed framework can be adapted by integrating scalable algorithms such as Mini-Batch K-Means or optimized variants of DBSCAN after PCA, or by employing distributed big data platforms such as Apache Spark MLlib. These enhancements would ensure that the methodology remains applicable to large-scale, real-world environments without compromising performance.

In conclusion, combining PCA with clustering techniques enhances segmentation accuracy, interpretability, and strategic value. The methodology not only enables data-driven decision-making for mid-sized datasets but can also be extended to large-scale applications through algorithmic adaptations and distributed processing. Future work should further explore supervised learning integration, real-time segmentation in dynamic markets, and cross-cultural adaptations, while ensuring ethical use of customer data in compliance with privacy standards.

Future Work

Future research could focus on enhancing this framework in three main directions. First, incorporating supervised learning can refine segmentation by using clusters as pseudo-labels to train predictive models (e.g., Random Forest, Gradient Boosting, Neural Networks), allowing validation through business-relevant outcomes such as churn rate and customer lifetime value. Second, applying the methodology to international datasets requires adjustments for cross-cultural differences, including income normalization, currency standardization, and the addition of region-specific attributes such as holidays or payment preferences. Finally, extending the framework to real-time applications through incremental PCA and online clustering would enable dynamic customer profiling and timely personalization, making the approach more practical in fast-changing business environments.

References

[1] I. A. Adeniran, C. P. Efunniyi, O. S. Osundare and A. O. Abhulimen, "Transforming marketing strategies with data analytics: A study on customer behavior and personalization," *International Journal of Management & Entrepreneurship Research,* vol. 6, no. 8, pp. 41-51, 2024.

[2] O. H. Olayinka, "Data driven customer segmentation and personalization strategies in modern business intelligence frameworks," *World Journal of Advanced Research and Reviews,* vol. 12, no. 3, pp. 711-726, 2021.

[3] K. S. Roy, P. B. Udas, B. Alam and K. Paul, "Unveiling Hidden Patterns: A Deep Learning Framework Utilizing PCA for Fraudulent Scheme Detection in Supply Chain Analytics," *International Journal of Intelligent Systems and Applications,* vol. 17, no. 2, pp. 14-30, 2025.

[4] I. B. Ridwan , "Transforming Customer Segmentation with Unsupervised Learning Models and Behavioral Data in Digital Commerce," *International Journal of Research Publication and Reviews,* vol. 6, no. 5, pp. 2232-2249, 2025.

[5] M. Alkhayrat, M. Aljnidi and K. Aljoumaa , "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA," *Journal of Big Data,* vol. 7, no. 1, 2020.

[6] I. A. Adeniran, C. P. Efunniyi, O. S. Osundare and A. O. Abhulimen, "The role of data science in transforming business operations: Case studies from enterprises," *Computer Science & IT Research Journal,* vol. 5, no. 8, pp. 2026-2039, 2024.

[7] N. Takata and T. Namatame , "Product Preference Analysis by Customer Segment Using Multiple Data Sources," *International Conference on Human-Computer Interaction,* pp. 123-134, 2025.

[8] N. L. Rane, M. Paramesha, S. P. Choudhary and J. Rane, "Artificial Intelligence, Machine Learning, and Deep Learning for Advanced Business Strategies: A Review," *Partners Universal International Innovation Journal*, vol. 2, no. 3, pp. 147-171, 2024.

[9] M. A. Wardana, A. Masliardi, N. Afifah, M. Sajili and H. P. Kusnara, "Unlocking Purchase Preferences: Harnessing Psychographic Segmentation, Promotion and Location Strategies," *Jurnal Informatika Ekonomi Bisnis*, vol. 5, no. 3, pp. 713-719, 2023.

[10] G. Upreti and A. K. Natarajan, "Leveraging Unsupervised Machine Learning to Optimize Customer Segmentation and Product Recommendations for Increased Retail Profits," *Intersection of AI and Business Intelligence in Data-Driven Decision-Making*, pp. 257-282, 2024.

[11] N. I. Okeke, O. A. Alabi, A. N. Igwe, O. C. Ofodile and C. P.-M. Ewim, "Customer journey mapping framework for SMES: Enhancing customer satisfaction and business growth," *World Journal of Advanced Research and Reviews*, vol. 24, no. 1, pp. 2004-2018, 2024.

[12] F. Pedro , "A Review of Data Mining, Big Data Analytics, and Machine Learning Approaches," *Journal of Computing and Natural Science*, vol. 3, no. 4, pp. 169-181, 2023.

[13] A. Musunuri , "Leveraging AI and Deep Learning for ECommerce Customer Segmentation," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 12, no. 6, pp. 9081-9096, 2023.

[14] A. Sharma, N. Patel and R. Gupta, "Enhancing Customer Segmentation Through AI: Analyzing Clustering Algorithms and Deep Learning Techniques," *European Advanced AI Journal*, vol. 11, no. 8, pp. 1-26, 2022.

[15] M. I. Shafi, . M. Chaudhry, E. C. Montero, E. S. Alvarado, I. D. L. T. Diez, M. A. Samad and I. Ashraf, "A Review of Approaches for Rapid Data Clustering:

Challenges, Opportunities, and Future Directions," *IEEE Access*, vol. 12, pp. 138086-138120, 2024.

[16] S. Nanga, A. T. Bawah, B. A. Acquaye, M.-I. Billa, F. D. Baeta, N. A. Odai, S. K. Obeng and A. D. Nsiah, "Review of Dimension Reduction Methods," *Journal of Data Analysis and Information Processing,* vol. 9, no. 3, pp. 189-231, 2021.

[17] M. Zhang, C. Q. Wu and A. Hou, "Big Data-Driven Portfolio Simplification: Leveraging Self-Labeled Clustering to Enhance Decision-Making," *Proceedings of the IEEE/ACM 10th International Conference on Big Data Computing, Applications and Technologies,* pp. 1-6, 2023.

[18] L. Scrucca, "A Model-Based Clustering Approach for Bounded Data Using Transformation-Based Gaussian Mixture Models," *Journal of Classification*, pp. 1-19, 2025.

[19] M. Cherradi and A. El Haddadi, "Exploration of Scientific Documents through Unsupervised Learning-Based Segmentation Techniques," *Seminars in Medical Writing and Education,* 2024.

[20] S. Zhang, E. Nezhadarya, H. Fashandi, J. Liu, D. Graham and M. Shah, "Stochastic whitening batch normalization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10978-10987, 2021.

[21] G. X. Z. W. Z. C. X. L. and L. P. , "New methods for VR rehabilitation intervention for children with cerebral palsy based on artificial intelligence," *5th International Conference on Artificial Intelligence and Advanced Manufacturing*, pp. 81-87, 2023.

[22] F. L. Gewers, G. R. Ferreira, H. F. De Arruda, F. N. Silva, C. H. Comin, D. R. Amancio and L. D. F. Costa, "Principal Component Analysis: A Natural

Approach to Data Exploration," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1-34, 2021.